WILEY ECOLOGY

Hybrid datasets: integrating observations with experiments in the era of macroecology and big-data

Journal:	*Ecology*
Manuscript ID	ECY18-0495.R1
Wiley - Manuscript type:	Concepts & Synthesis
Date Submitted by the Author:	n/a
Complete List of Authors:	Benedetti-Cecchi, Lisandro; Universita di Pisa, Biology Bulleri, Fabio; Università di Pisa, Dipartimento di Biologia Dal Bello, Martina ; Massachusetts Institute of Technology, Department of Physics Maggi, Elena; Università di Pisa, Biology Ravaglioli, Chiara; Universita di Pisa, Biology Rindi, Luca; Universita degli Studi di Pisa Dipartimento di Biologia, Biologia
Substantive Area:	Population Ecology < Substantive Area, Community Analysis/Structure/Stability < Community Ecology < Substantive Area, Disturbance < Community Ecology < Substantive Area, Species Interactions < Community Ecology < Substantive Area
Organism:	Bacterial < Viruses, Algae (specify type in field below)
Habitat:	Intertidal/Tidal/Coastal < Marine < Aquatic Habitat < Habitat
Geographic Area:	Europe < Geographic Area
Additional Keywords:	Causal inference, convergent cross-mapping, distributed experiments, empirical dynamic modelling, macroecology, species distribution models, time series, hybrid dataset approach

Abstract:	Understanding how increasing human domination of the biosphere affects life on earth is a critical research challenge. This task is facilitated by the increasing availability of open-source data repositories, which allow ecologists to address scientific questions at unprecedented spatial and temporal scales. Large datasets are mostly observational, so they may have limited ability to uncover causal relations among variables. Experiments are better suited at attributing causation, but they are often limited in scope. We propose hybrid datasets, resulting from the integration of observational with experimental data, as an approach to leverage the scope and ability to attribute causality in ecological studies. We show how the analysis of hybrid datasets with emerging techniques in time series analysis (Convergent Cross Mapping) and macroecology (Joint Species Distribution Models) can generate novel insights into causal effects of abiotic and biotic processes that would be difficult to achieve otherwise. We illustrate these principles with two case-studies in marine ecosystems

	and discuss the potential to generalize across environments, species and ecological processes. If used wisely, the analysis of hybrid datasets may become the standard approach for research goals that seek causal explanation for large-scale ecological phenomena.

Running title: hybrid datasets in ecology

Hybrid datasets: integrating observations with experiments in the era of macroecology and big-data

Lisandro Benedetti-Cecchi[1*], Fabio Bulleri[1], Martina Dal Bello[2], Elena Maggi[1], Chiara Ravaglioli[1], Luca Rindi[1]

[1]Department of Biology, University of Pisa, CoNISMa, Via Derna 1, 56126, Pisa, Italy
[2]Physics of Living Systems Group, Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139.

*Correspondence to:
Lisandro Benedetti-Cecchi
e-mail lbenedetti@biologia.unipi.it
Voice: +39 050 2211413
Fax: +39 050 2211410

Paper type: Concept and Synthesis

ABSTRACT

Understanding how increasing human domination of the biosphere affects life on earth is a critical research challenge. This task is facilitated by the increasing availability of open-source data repositories, which allow ecologists to address scientific questions at unprecedented spatial and temporal scales. Large datasets are mostly observational, so they may have limited ability to uncover causal relations among variables. Experiments are better suited at attributing causation, but they are often limited in scope. We propose hybrid datasets, resulting from the integration of observational with experimental data, as an approach to leverage the scope and ability to attribute causality in ecological studies. We show how the analysis of hybrid datasets with emerging techniques in time series analysis (Convergent Cross Mapping) and macroecology (Joint Species Distribution Models) can generate novel insights into causal effects of abiotic and biotic processes that would be difficult to achieve otherwise. We illustrate these principles with two case-studies in marine ecosystems and discuss the potential to generalize across environments, species and ecological processes. If used wisely, the analysis of hybrid datasets may become the standard approach for research goals that seek causal explanations for large-scale ecological phenomena.

Keywords: causality, convergent cross-mapping, distributed experiments, empirical dynamic modelling, hybrid dataset approach, macroecology, species distribution models, time series

INTRODUCTION

Ecology for the Anthropocene aspires to become a global-scale science, addressing changes in biodiversity and ecosystems in an increasingly human-dominated world (Sutherland et al. 2014, Corlett 2015). This aspiration builds on recent theoretical and empirical advances, but faces

important challenges. Large repositories of ecological and environmental data are becoming increasingly accessible as the result of technological developments in remote sensing and rapid advances in genomic techniques and computational capabilities (Hampton et al. 2013). This boost of 'big-data' allows ecologists to test hypotheses at unprecedented scales and promotes the integration of disciplines that typically focus on different levels of biological organization, such as population genomics, community ecology and macroecology (McGaughran 2015).

Two common approaches to address global change problems are time series analysis and species distribution models (SDM) (Elith and Leathwick 2009, Magurran et al. 2015). Time series allow modeling population fluctuations in relation to environmental change, but they rarely extrapolate beyond current or past conditions. SDM build on contemporary species distributions to make projections under various future climate scenarios. Both time series and SDM have contributed considerably to our understanding of species responses to global change. However, these approaches, and the analysis of 'big data' in general, are largely correlative and cannot identify cause-effect relations with the same level of confidence offered by manipulative experiments. Observational studies, in contrast, can identify relationships among variables at spatial and temporal scales that are difficult or impossible to address experimentally.

Identifying procedures that capitalize on the strengths of both observations and experiments is a great research challenge that can contribute major breakthroughs in the way we investigate, understand and forecast ecological responses to global change. The logical connections between observations and experiments have been widely recognized (Underwood 1997), but how to integrate observational and experimental data remains an open question (Fig. 1a). The 'comparative experimental approach', where identically designed manipulations are distributed along environmental or latitudinal gradients, addresses this question by embedding manipulative experiments in a correlative framework that allows evaluating interactions between local and

regional processes (Menge et al. 2002, Hewitt et al. 2007) (Fig. 1b). This distributed experiment approach can be traced back to pioneering studies on kelp forests in California, where experimental perturbations designed to probe the resilience and resistance of kelps and associated organisms were repeated at multiple sites in contrasting environments (Dayton et al. 1984). Although challenging, distributed experiments can have great power to reveal geographic trends in ecological processes (Coleman et al. 2006) and incorporating environmental covariates in the analysis can improve the ability to determine causal relations by controlling for potentially confounding effects (Hewitt et al. 2007). However, the attribution of causality in distributed experiments is limited to manipulated factors and does not extend to covariates because these involve only observational data.

Here, we propose novel strategies to leverage the scope and attribution of causality in ecological studies based on the hybridization of observational and experimental data and using emerging analytical techniques that go beyond the use of observations as covariates: convergent cross mapping (CCM) and joint SDM (jSDM) (Sugihara et al. 2012, Iknayan et al. 2014, Warton 2015). Central to our idea is the synthesis of hybrid datasets that combine the ability to establish causality typical of experiments (causal inferential strength) with the large scope allowed by observational studies (Fig. 1c). Hybridization is possible whenever observational data (spatial and temporal series) are available for variables whose causal relation has also been examined experimentally in the same environmental setting (e.g., same locations). Indeed, distributed experiments are a great resource to generate the kind of data needed to implement the hybrid datasets approach. Many climate change experiments are now embedded in long-term monitoring programs, offering an extraordinary opportunity to generate hybrid datasets from many different ecosystems and levels of biological organization. Below, we describe the key procedures that we believe can make the best use of hybrid datasets, provide real-world examples of their application

and discuss potential advantages and limitations of the approach. The code and datasets needed to reproduce all the results presented in this paper have been made available through figshare (see Appendix S1 for details).

HYBRIDIZATION OF TIME SERIES

Empirical dynamic modeling

Recent developments in time series analysis have focused on non-parametric (equation-free) methods to model nonlinear systems (Sugihara 1994, Sugihara et al. 2012). This empirical dynamic modeling framework (EDM, BOX 1) accommodates both the nonequilibrium dynamics and nonlinearities typical of complex ecological systems. EDM builds on the concept that time series originate from what in dynamical system theory is called an attractor manifold, a high-dimensional space where axes define the possible states of the system (e.g., the environmental and biological variables in an interacting ecosystem) and trajectories along axis coordinates describe deterministic temporal changes (Fig. 2a). Time series that share the same attractor are causally linked, implying that each variable can identify the state of the other. For example, time series of two species that compete for a limited resource are dynamically linked and past values of abundance of one species can be recovered from time series of the other species. If one variable is a stochastic driver external to the system – e.g., a climate variable such as temperature regulating the abundance of a population – information about the state of the driver can be recovered from time series of population abundance, but the opposite is not possible. This counterintuitive asymmetry originates because only population abundance contains information about temperature, whereas the forcing variable does not depend on population abundance and so it contains no information about it (Sugihara et al. 2012).

Convergent Cross Mapping (CCM) is a technique that allows examining whether the time series of a variable contains the signatures of another variable (Deyle and Sugihara 2011,

Sugihara et al. 2012). This is achieved by using the points on the reconstructed manifold of, say, variable *X* (M_X) to predict concurrent points (at the same time) on the manifold reconstructed from variable *Y* (M_Y) (Fig. 2b). Cross-validation methods can be used to assess the forecasting skill of CCM (Fig. 2c,d). If causation is asymmetrical (e.g., temperature forcing population abundance, but not the reverse) forecasting skill will improve with time series length when cross-mapping from the response to the predictor, but not in the opposite direction. This allows distinguishing between causal and response variables (Fig. 2c,d).

Box 1. Empirical Dynamic Modeling and Convergent Cross-mapping

Empirical dynamic modeling (EDM) is a non-parametric framework that allows reconstructing the underlying attractor of a time series (Sugihara 1994, Sugihara et al. 2012). Attractor manifolds are complex geometric structures describing the possible states and trajectories of a system. Sequential projections of the motion on the manifold to an axis coordinate generates a time series of the corresponding variable (Fig. 2a). A fundamental mathematical theorem proves that a time series of an individual variable contains all the information about the entire system and therefore can be used to reconstruct a 'shadow' version of the original attractor (Takens 1981). This is achieved by using lags of the observed time series as surrogates for the unknown or unobserved variables (Sugihara et al. 2012). That is, a time series of a variable *X* is sufficient to reproduce the fundamental geometry of the system using lagged-coordinate embedding. For each data point *X(t)*, one generates a lagged vector of data points $\overrightarrow{X(t)} = X(t), X(t-\tau), X(t-2\tau), \ldots, X(t-(E-1)\tau)$, where *E* is the embedding dimension – i.e. the number of time steps used for prediction – and τ is the time lag (usually set to one). Thus, each *E*-dimensional point $\overrightarrow{X(t)}$ consists of the present value *X(t)* and the *E*-1 previous values each separated by lag time τ.

Time series originating from the same attractor are causally linked and this relation can be deciphered from the reconstructed manifolds through Convergent Cross-Mapping (CCM) (Fig. 2). This is done by using the points on the reconstructed manifold of variable *X* (M_X) to predict concurrent points on the manifold reconstructed from variable *Y* (M_Y) (Fig. 2b). Cross-mapping

means that the relationship between points is evaluated at the same time, so the procedure does not involve prediction of the future in the classical sense. In practice, for causally related time series,

the time indices of nearby points on the manifold of one variable will correspond to nearby points on the manifold of the other. Thus, when predicting from a given point *X(t)* at time *t*, a set of nearby points on M_X will be projected onto M_Y using the corresponding time indices, and the centroid of these points on M_Y gives the target prediction $\hat{Y}(t)|M_X$ (Fig. 2b). The set of points projected for each *X(t)* is formed by the closest E+1 neighbors, the smallest number of points for a bounding simplex in E-dimensional space (Sugihara and May 1990). The distance among projected values provides a measure of the uncertainty of prediction. If there is symmetrical causality between variables (*X* causes *Y* and *vice versa*: $X \Leftrightarrow Y$), prediction will be possible in both directions with low uncertainty. If causation is asymmetrical (e.g., *X* is an external forcing variable: $X \Rightarrow Y$), only the time indices of

nearby points on the manifold of *Y* will correspond to nearby points on the manifold of *X* and will allow predictions with low uncertainty, but the reverse will not be true.

The forecasting skill of CCM is assessed through cross-validation. The typical approach uses Pearson-product moment correlation to compare the target predictions $\hat{Y}(t)|M_X$ to the actual values *Y(t)*. This is repeated with increasing time series length, i.e., the number of points used to reconstruct the manifolds M_Y and M_X. Estimation skill is expected to increase with time series length because more trajectories will fill the reconstructed attractor with longer series, so that points on both M_Y and M_X will be closer, resulting in more precise and accurate estimates. Cross-validation is also used to select the optimal embedding dimension *E*. This requires assessing the forecasting skill for a set of *E* values (e.g., from 1 to 10) and selecting the embedding dimension on the basis of optimal prediction (Sugihara and May 1990).

In principle, forecasting skill should converge to one when cross-mapping causally related variables. In practice, convergence will be limited with short time series and with large background noise. Nevertheless, convergence is a key property of CCM to distinguish between causation and correlation (Sugihara et al. 2012). In asymmetrical causation, forecasting skill will improve with time series length when cross-mapping from the response to the predictor, but not in the opposite

direction. This is a key diagnostic to identifying causal environmental forcing variables. The simulated data in Fig. 2c illustrate the case of two competing species where species A causally affects species B more than B affects species A (see Appendix S1 for details of the simulation). Cross map skill converges faster with increasing time series length when predicting from the manifold of species B to that of species A than in the opposite direction. This is the expected outcome that correctly identifies the stronger causal effect of species A on species B (Fig. 2d). Data in Fig. 2c were generated using the system of coupled difference equations described in Sugihara *et al*. (2012) and implemented with the make_ccm_data function in the R library multispatialCCM (Clark et al. 2015). We used functions simplex and ccm in the rEDM package (Sugihara et al. 2012, Ye and Sugihara 2016) to perform CCM on scaled data, with best embedding dimension E set to 4 and 2 for species A and B, respectively, and the other arguments left to default values (further details including links to R code are provided in Appendix S1).

Recent applications of CCM in biology and ecology include studies uncovering predator-prey interactions, the influence of environmental variables such as sea surface temperature (SST) on the recruitment and dynamics of fish stocks and the spread of influenza (Sugihara et al. 2012, Deyle et al. 2016a, Deyle et al. 2016b). The possibility of extending CCM to short, spatially replicated time series has opened the opportunity for validation using data from field experiments (Clark et al. 2015). Although these examples show that CCM can identify causation in some circumstances, CCM has greatest power to distinguish causation from correlation when dealing with weakly coupled processes and when observational and process error are limited (Sugihara et al. 2012). How many empirical time series will meet these criteria is unclear, but limited noise may be more an exception than the rule in real-world datasets, due to the non-additivity of multiple processes, feedbacks and nonlinearities.

Empirical dynamic modeling with hybrid time series

A key requirement for data hybridization is that both the observational and experimental data originate from the same attractor manifold – i.e. they belong to the same dynamical system. The ideal setting is a field experiment where estimates of both manipulated (e.g., temperature) and response (e.g., population abundance) variables are available at the same location as part of observational time series and over periods encompassing the duration of the experiment. The hybridization of observational and experimental data collected at distant locations or in different periods (e.g., years) would not be appropriate, since these data may not originate from the same dynamical system. Similarly, assimilating data from laboratory experiments into time series obtained in natural conditions is questionable, since the laboratory data do not contain information about other relevant, but unobserved processes that operate in real-world conditions. Caution is also needed in using perturbation experiments that may drive the system to an alternative state, so that experimental and observational data are no longer on the same attractor.

Under reasonable assumptions, hybrid time series can be obtained by replacing observational with experimental values for both the driver and the response variable, for the period encompassing the duration of the experiment. Averages can be taken across experimental units to have a unique time series for the predictor and response variables (Fig. 3a). Alternatively, hybridization is needed (or is possible) only for the driver. This will be the case with pulse experiments, where experimental plots are treated in some periods and left untouched in others (Fig. 3b). The hybrid dataset is generated by substituting natural with experimental values of the driver at the time of the manipulation, whereas the time series of the response variable are already hybridized, since they incorporate the effects of experimental and naturally fluctuating environmental conditions (Fig. 3b).

Hybrid time series should improve the ability of CCM to correctly identify causation for at least two reasons. First, in experiments where the causal process is held constant (or

approximately so), such as with press perturbations, hybrid time series of the driver should be less variable than purely observational data and this should increase the forecasting skill of CCM (Fig. 3a). Also the response variable may display dampened fluctuations in press experiments, so hybridization of the response variable may further contribute to reduce noise in hybrid datasets. Second, experimental data may improve the ability of CCM to correctly cross-map the state of the driver from the response variable by reinforcing the causal signal in the appropriate direction.

Comparing the outcome of CCM between observational and hybrid time series may improve the causal inferential strength of the analysis. A significant causal relation resulting from hybrid time series will increase confidence in the attribution of causality compared to the same outcome based only on observational data (Fig. 1c). In contrast, lack of evidence of a causal relation from hybrid time series may cast doubt on a positive effect that may result when analyzing only observational data. Nonetheless, care is needed to avoid confirmatory bias when deciding to analyze hybrid datasets. The analysis should be motivated by well-defined scientific questions and should not be driven by previous knowledge of the outcome of an experiment. Thus, negative or counterintuitive experimental results should also be included in hybrid datasets, if the analysis is motivated by a clear hypothesis.

Case study 1: Effect of experimental warming on biofilms

Rocky intertidal biofilms (or epilithic microphytobenthos, EMPB) consist primarily of photosynthetic organisms such as cyanobateria, diatoms and algal sporelings that occur on many rocky shores globally (Murphy et al. 2006, Maggi et al. 2017). These primary producers undergo large fluctuations in biomass in response to environmental change and to variation in grazing pressure (Sanz-Lazaro et al. 2015). The rocky intertidal is a highly variable environment where extreme events such as prolonged hot and dry periods, storms and sediment accretion can cause severe impacts to EMPB biomass (Dal Bello et al. 2017). Grazers can also eradicate biofilms

from rocky shores, but EMPB biomass usually recovers from even the most extreme disturbances (Underwood 1984, Sanz-Lazaro et al. 2015, Dal Bello et al. 2017).

The rapid response of biofilms to changing environmental conditions makes EMPB biomass an ideal variable to examine the consequences of data hybridization on causation. Time series of this variable were available to us as part of a research program on the ecology of EMPB that included both observations and experiments on rocky shores in the north-west Mediterranean (Dal Bello et al. 2015, Dal Bello et al. 2017, Maggi et al. 2017). We took advantage of the spatial extension of CCM that uses short, but spatially replicated time series. The experiment was conducted between April and August 2013 and examined the effect of extreme warming on chlorophyll *a* concentration (*chl*), an indirect measure of EMPB biomass. Extreme temperature conditions were imposed to plots of 35 x 55 cm marked on the rock and were repeated twice, either 15 or 60 days apart, to reflect separate and clustered events in time. Each warming event consisted of heating experimental plots between 11 a.m. and 3 p.m. in one day, by means of aluminium chambers equipped with stoves. Warming simulated an extreme temperature value with a return periods of 100yrs for the corresponding month (with a mean temperature of 31°C over the study period). There were nine plots for each of the clustered and non-clustered treatments sampled at approximately 15 day intervals four and seven times, respectively, for a total of 99 experimental data points.

Observational data were obtained from three plots located on the same shore used in the experiment and consisted of monthly measurements of *chl* concentration collected over a period of 24 months, between 2012 and 2014. We augmented these observations with data from the three control plots available from the experiment, which provided nine data points each. The vector of observational data had 99 points as the vector of experimental data. Data on aerial temperature for the study location were obtained from a local meteorological station, whereas

temperature in heated plots was measured with digital thermometers (Dal Bello et al. 2017). All plots were sampled with an IR-sensitive camera (Agricultural Digital Camera ©) that allowed us to quantify the Ratio Vegetational Index (RVI, the ratio of reflectance between near-infrared and red bands), which was subsequently converted into *chl* concentration using a previously calibrated relation (Dal Bello et al. 2015). This spatial version of CCM and associated probabilistic tests were performed using the library multispatialCCM in the R computational environment (Clark et al. 2015).

CCM did not identify any causal forcing of temperature on *chl* when using only observational data (Fig. 3c). Indeed, cross map skill increased significantly in the wrong causal direction, suggesting that *chl* causes variation in temperature (bootstrap test: $P>0.05$ for *chl* cross-mapping temperature and $P<0.01$ for temperature cross-mapping *chl*) (Fig. 3d; see Appendix S1 for details on bootstrap and probability tests in CCM). Hybrid time series were assembled only for temperature, by replacing natural with experimentally imposed values in the periods in which warming was applied. C*hl* values from experimental plots did not need any further hybridization, since they already incorporated natural fluctuations in aerial temperature and the effects of pulse warming perturbations (Fig. 3e). With this approach, observational and hybrid time series had equal sample size (99 data points each), allowing a fair comparison between the two groups. Cross map skill increased with time series length when cross-mapping from *chl* to temperature, but not in the opposite direction, indicating that temperature information was encoded in EMPB biomass (Fig. 3f). This asymmetrical causal relation with temperature forcing *chl* was supported statistically (bootstrap test: $P<0.001$ for *chl* cross-mapping temperature and $P>0.05$ for temperature cross-mapping *chl*). In addition to clarifying the signal of causality, reconstructing the attractor from hybrid time series also improved the ability to forecast *chl* data compared to the analysis based on observations, as indicated by cross-validation (Pearson

correlation coefficient: 0.789 for experimental data and 0.675 for observational data; see Appendix S1 for details).

These results show how hybrid time series can increase the ability of CCM to identify a causal relation between a driver and a biological response variable, compared to purely observational data. Our example focuses on pulse warming treatments, but it is not difficult to envision similar applications in other contexts. Many studies manipulate temperature and other ocean drivers, such as nutrients and acidification, in regions where observational data are also available. The analysis of hybrid datasets would increase the scope and causal inferential strength of these local, often short-term experiments. Many terrestrial studies also manipulate environmental drivers such as temperature, nutrients and rainfall in areas where these variables are also measured regularly as part of ongoing monitoring programs, offering great opportunities for integration.

HYBRIDIZATION IN SPACE

Joint species distribution models (jSDM)

jSDM use a hierarchical Bayesian framework to model species ensembles rather than one species at a time (Warton et al. 2015, Ovaskainen et al. 2017). This emerging analytical technique expands the traditional approach of modeling individual species in relation to one or more environmental drivers, recognizing the multivariate nature of species assemblages. Using appropriate frequency distributions to model species occurrences (e.g. binomial, probit) or abundances (normal, Poisson) and regression parameters (e.g., multivariate normal), jSDM allow a community-wide analysis of species distributions in relation to environmental covariates. Residual correlation matrices quantify networks of species co-occurrence after accounting for environmental filters and these matrices can be further modeled as a function of species traits. jSDM allow the estimation of random effects to account for variability at multiple hierarchical

scales in space and time and variance partition methods can be used to determine the percentage of variation explained by fixed and random effects (Ovaskainen et al. 2017).

How to incorporate species interactions in jSDM has been an intense area of research in recent years (Araujo et al. 2011, Pellissier et al. 2013, Pollock et al. 2014, Mod et al. 2015, Morueta-Holme et al. 2016). One approach uses the abundance of competitors or consumers as covariates in the analysis. Alternatively, positive or negative species interactions can be inferred from the residual species correlation matrix, after accounting for environmental filters. If a representative set of environmental covariates has been considered, residual species associations are most parsimoniously explained in terms of species interactions. However, it is difficult to ascertain whether all relevant covariates have been included in any particular study, and both approaches are based on observations and remain largely correlative.

Hybrid spatial datasets and jSDM

There are two possibilities of generating hybrid spatial datasets. One approach (Path 1 in Fig. 4) simply consists of concatenating experimental (treatments and controls) and observational data in a single dataset. This requires matching the observational and experimental data at the level of both the predictor and the response variables. That is, the covariates measured in the observational study should also be recorded in the experimental plots and manipulated factors should have corresponding observational values. The advantage of a spatial hybrid dataset is greater power due to increased degrees of freedom and, possibly, a stronger signal of the causal relation between predictor and response variables compared to individual datasets. However, spatial sampling programs often include a massive number of observations, whereas experiments typically involve few spatial replicates. Thus, concatenating observational and experimental data will probably be most valuable when the number of observations is limited and the experiment identifies a strong causal signal.

A second and perhaps more powerful approach to generate spatial hybrid datasets consists of translating experimentally determined effect sizes into covariates to be incorporated into jSDM (Path 2 in Fig. 4). This framework allows examining biotic interactions with jSDM in an unprecedented way. As an example, consider the case of a monitoring program that includes observations on the abundance and distribution of habitat-forming species (e.g. forest trees, corals, macroalgal forests) and associated biodiversity. Loss of habitat-forming species usually has strong effects on associated assemblages and these effects can be mediated by environmental filters and other spatially variable drivers, such as disturbance and species dispersal (Bulleri et al. 2012, Krumhansl et al. 2016). jSDM can assess these relations and would typically include the abundance of the habitat-forming species as a covariate to model biotic interactions. However, if the interspecific effect of the habitat-former is quantified directly (e.g., from a removal experiment), hybrid datasets that incorporate causal relations can be developed and analyzed with jSDM as follows. First, the effect of the habitat-former on species *i* in the assemblage can be expressed using one of the several interaction strength indices available to quantify biotic interactions (Berlow et al. 1999). The relative interaction intensity index (RII) is appropriate to measure interactions that range from competition to facilitation, as in the case of habitat-forming species (Armas et al. 2004):

$$RII_i = \frac{\bar{x}_{iC} - \bar{x}_{iT}}{\bar{x}_{iC} + \bar{x}_{iT}} \quad (1)$$

where $\bar{x}_{iT}$ and $\bar{x}_{iC}$ are the mean abundance of species *i* in treatments and controls, respectively. This index varies between -1 and 1, depending on the direction of the interaction of the habitat former on species *i*, with negative (positive) values reflecting competition (facilitation). A mean relative interaction index can then be obtained for each observed plot by averaging RII_i values across the species present in each plot. The vector of average values describes the relative

intensity of positive and negative interactions across observational plots, so we can expect positive values in plots where the habitat-former is abundant and negative values where it is absent. This will be the case only if the habitat-former is a strong interactor and causally affects other species in the assemblage. This is a crucial point that makes hybrid datasets a superior approach than using species abundances or co-occurrences as surrogates for biotic interactions in a purely observational context. This last approach will be misleading if shared causal processes, such as environmental drivers, induce positive or negative covariance between the hypothesized strong interactor and other species. In this case 'mirage' correlations (Sugihara et al. 2012) may be erroneously interpreted as evidence of biotic interactions. This will not happen with hybrid datasets, because if the experiment reveals no causal signal, the resulting average *RII* values will be close to zero and will explain no variation in a jSDM, regardless of the patterns of 'mirage' correlations that may be induced by external drivers. Below, we provide a real-world example on the effect of loss of habitat-forming species in the marine benthos.

Case study 2: Effect of loss of macroalgal forests on subtidal rocky reefs.

Macroalgal forests (kelps and fucoids) are amongst the most diverse and productive coastal marine ecosystems, yet they are declining dramatically worldwide in response to global warming and increasing environmental degradation (Benedetti-Cecchi et al. 2001, Strain et al. 2014, Krumhansl et al. 2016, Vergés et al. 2016). A direct consequence of loss of canopy-forming species is a shift from a macroalgal forest into less diverse and productive assemblages dominated by encrusting coralline algae (barren habitat) or intricate mats of low-lying algae (algal turfs) (Benedetti-Cecchi et al. 2015, Rindi et al. 2017).

Although these relations are well established, understanding the relative contribution of species interactions, environmental filters and spatiotemporal context in driving habitat shifts remains a critical gap. We use data from a long-term sampling program and repeated canopy-

removal experiments in the north-west Mediterranean, to show how hybrid datasets can incorporate empirically determined species interactions and how these effects can be partitioned along with other drivers of species distribution using jSDM. Observational data were collected at six locations, four islands in the Tuscan Archipelago (Capraia, Pianosa, Giannutri and Montecristo, 42°46'N, 10°11'E) and two locations along the main coast of Tuscany (Livorno and Rosignano, 43°28'N, 10°19'E) between 2005 and 2013 (Bulleri et al. 2018). We used a hierarchical sampling design, including two to four sampling years within each location, 6-10 sites in each year and ten replicate plots at each site (years were reasonably well interspersed among locations). Assemblages were sampled non-destructively using photo-quadrats and images were processed in the laboratory to extract for each plot the percentage cover of all identifiable species. When species could not be identified unambiguously, they were lumped into higher taxonomic or morphological categories (e.g., filamentous, coarsely branched and sheet-like algae). The observational datasets consisted of 1327 plots and 55 species (higher taxa).

Canopy-removal experiments involving full canopy-removal and control plots were performed during the observational study at Capraia and Pianosa islands and lasted two and three years, respectively. Overall, the experimental dataset consisted of 168 canopy-removal and 80 control plots. Several controls were discarded due to the disappearance of the canopy during the course of the experiment at Pianosa. We used Eq. 1 to quantify the average intensity and direction of canopy-removal effects across species in each observational plot. Environmental covariates included sea surface temperature (SST), nutrients (nitrates and phosphates) and a wave exposure index. SST and nutrients were derived from the Bio-Oracle database (Tyberghein et al. 2012), which provides data layers at the 5 arcmin resolution (*c.* 9.2 Km). These covariates reflected average environmental conditions at the location level. In contrast, the wave exposure index was obtained from a high resolution hydrodynamic model and varied at the site level

(Bulleri et al. 2018). We used the HMSC package in the R computational environment to fit a hierarchical Bayesian jSDM with spatiotemporal random effects (Ovaskainen et al. 2017). To increase the performance of the model we focused on the presence/absence of species that occurred in at least 10% of the quadrats. This resulted in 30 species (higher taxa) for the analysis. Each of these species had a RII_i value obtained as the mean between the two experiments.

An assumption implicit to SDMs is that species are near equilibrium with their environment (Guisan and Thuiller 2005). Short-term snapshots of presence-absence or abundance data may violate this assumption, reflecting transient effects rather than long-term average conditions. When using hybrid datasets, it is also important that experiments are maintained long enough to reflect steady-state effects. In our analysis, both observational and experimental data were obtained at comparable, multi-year time scales and involved a large number of spatial replicates. We are confident that these well-replicated, relatively long-term observations minimized the influence of transient effects. Furthermore, we know from previous studies that canopy removals may trigger a shift to a turf-dominated assemblage in less than one year (Benedetti-Cecchi et al. 2001, Rindi et al. 2017, Bulleri et al. 2018). Thus, experiments at Capraia and Pianosa run for long enough (two and three years, respectively) to ensure that biotic interactions estimated through the *RII* index reflected steady conditions.

When averaged over species, biotic interactions explained 19% of variation (measured by Tjur's R^2, specific for binary data) (Ovaskainen et al. 2017), with most of the variability occurring among years (24%) and localities (13%) (Fig. 5a). Environmental variables collectively explained 38% and 56% of variation in models with and without biotic interactions, respectively (Fig. 5a,b). Thus, biotic interactions accounted for 19% of variability that would have otherwise been ascribed to environmental filtering. We used 10-fold cross-validation to quantify the predictive power of the model. Biotic interactions improved predictive power for several species

as evidenced by three common validation statistics: R^2, area under the curve (AUC) and the true skill statistic (TTS) (Fig. 5c-e).

Biotic interactions also affected the network of species associations in space and time. As an example we illustrate species associations at the location scale (Fig. 5f,g). Three groups of species were distinguishable at this scale when species interactions were included in the model. The largest group (delimited by species 15 and 26) included positively associated species that were prevalent in the understory of macroalgal forests on islands; a second group (delimited by species 1 and 17) included species that typically co-occurred on the mainland or in gaps of macroalgal forests and that were negatively associated with the first group; finally, the third group included the remaining species that occurred mostly independently of the other species. The network of species associations became much less structured when biotic interactions were not included in jSDM, now erroneously indicating a positive association of species that typically occur in the understory of macroalgal forests with those dominating in the absence of a canopy (Fig. 5g).

These results show how biotic interactions can increase the predictive ability of jSDM, explaining up to 58% of variation in species occurrences. Environmental filters appeared less important after accounting for biotic interactions, suggesting that the effect of biological forcing can be mistakenly ascribed to abiotic variables if not addressed explicitly. Overestimating environmental filters may exacerbate the problem of 'mirage' correlations, where spurious species associations may emerge due to shared species response to environmental change. Indeed, we have shown that positive species associations were more frequent when the influence of biotic interactions was erroneously ascribed to environmental covariates. Hybrid datasets can mitigate this problem, because experimentally derived measures of biotic interactions are not subjected to 'mirage' correlations.

The data hybridization approach shown here can be extended to other species, types of interactions and environments. For example, consumer-resource interactions can be implemented in jSDM in a similar way as with habitat-forming species and causal effects can be quantified using any appropriate measure of interaction strength other than the *RII* index (Berlow et al. 1999). Furthermore, in addition to biotic interactions, experiments may probe abiotic drivers (warming, rainfall, pH, among others) and appropriate effect sizes can be derived for integration in hybrid datasets. Finally, experiments do not need to examine one factor at the time. Multifactorial experiments addressing several factors and their interactions can also be implemented in jSDM by translating each effect size in a covariate, as we have done for the average *RII* index.

CONCLUSIONS

Increasing causal inferential strength of ecological studies is becoming overwhelmingly important in an era where observational data and ecological models play a prominent role to address large-scale, long-term environmental problems (Connolly et al. 2017, McGill and Potochnik 2018). Here, we argue that in addition to articulating better ecological models to make sense of observations, ecological experiments still have a key role to play to reduce uncertainty in attributing causality in large-scale ecological problems. We propose hybrid datasets and their implementation through emerging techniques in time series analysis and macroecology, as a strategy to leverage the scope and causal inferential strength of ecological studies. This approach builds on the increasing availability of observational and experimental datasets, owing to more effective sharing practices among scientists and technological innovation for the acquisition, storage and dissemination of digital information. Large datasets are mostly observational, but distributed experiments are becoming more common and the data they generate are made increasingly accessible. Thus, the time is ripe to develop formal approaches to data hybridization

that go beyond the use of observations as covariates in the analysis of distributed experiments (Fig. 1).

CCM has been designed to identify causality from observational time series, but the power of this technique may be compromised with noisy data and when the driving and response variables are strongly coupled. We have shown how observational time series may fail to identify the causal effect of temperature on EMPB. Hybrid datasets that included data from a warming experiment, with realistic treatments, correctly identified this asymmetrical causal relation.

jSDM allow macroecologists to unravel the response of assemblages to global change and to project species distributions under future climate scenarios. We have discussed two ways to generate hybrid datasets in a spatial context and provided an example focusing on biotic interactions. Empirically derived measures of species interactions can be readily incorporated in jSDM as a covariate, providing a simple and intuitive way to model biotic effects. Our approach differs from that of other studies where biotic interactions have been inferred by including the abundance of potential competitors or consumers as covariates in jSDM, which is essentially a correlative approach (Pellissier et al. 2013, Pollock et al. 2014, Mod et al. 2015, Bueno de Mesquita et al. 2016). Our method is also more direct than a recently proposed two-stage analysis that incorporates species interactions after fitting a species distribution model (Staniczenko et al. 2017).

Hybrid datasets can also be developed and analyzed in the context of 'big-data'. Opportunities arise with studies examining the molecular mechanisms regulating the response of organism to stress. For example, recent studies on microbes and corals have integrated observational and experimental approaches to evaluate the effects of global warming on gene expression (Barshis et al. 2013, Mock et al. 2016). Hence, the hybridization strategy proposed here can also be implemented to increase the scope and causal inferential strength of studies that

make extensive use of DNA-sequencing techniques, which are amongst the most important contemporary sources of 'big-data' in ecology.

There are potential caveats in the analysis of hybrid datasets that require attention. Most importantly, scientific questions should have priority over other considerations in deciding whether to embark in such analysis. For example, one may be tempted to proceed with the analysis of hybrid datasets only when experimental outcomes go in the expected direction or corroborate a trend already present in the observational data. To avoid confirmatory bias, previous knowledge of the outcome of separate analyses on observational and experimental data should have no bearing on the final decision to analyze hybrid datasets. Another important aspect to keep in mind is that a significant relation between variables in a hybrid dataset does not necessarily underscore causality over the entire temporal or spatial domain of the analysis. This will depend on how many experiments are available for integration and their degree of interspersion with the observational data. Well-distributed experiments in space and time will increase our confidence in the analysis of hybrid datasets and will have utmost ability to distinguish causal relations from mirage correlations.

The potential for developing and analyzing hybrid datasets in ecology and biology is enormous. We have discussed several applications and provided real-world examples of the use of hybrid datasets, with the hope of motivating further research in this direction. The analysis of hybrid datasets should become the standard for research goals that seek causal explanation for large-scale phenomena, beyond the limits to causal inference inherent in observations and beyond the scales encompassed by individual manipulative experiments.

Acknowledgments. This work was partially supported by the University of Pisa through the PRA_2017_19 project to LBC and the Italian Ministry for Education, University and Research

through project HI-BEF (protocol RBFR12RXWL). We also thank the numerous students who have assisted with field work in the last decade.

References

Araujo, M. B., A. Rozenfeld, C. Rahbek, and P. A. Marquet. 2011. Using species co-occurrence networks to assess the impacts of climate change. Ecography **34**:897-908.

Armas, C., R. Ordiales, and F. I. Pugnaire. 2004. Measuring plant interactions: A new comparative index. Ecology **85**:2682-2686.

Barshis, D. J., J. T. Ladner, T. A. Oliver, F. O. Seneca, N. Traylor-Knowles, and S. R. Palumbi. 2013. Genomic basis for coral resilience to climate change. Proceedings of the National Academy of Sciences of the United States of America **110**:1387-1392.

Benedetti-Cecchi, L., F. Pannacciulli, F. Bulleri, P. S. Moschella, L. Airoldi, G. Relini, and F. Cinelli. 2001. Predicting the consequences of anthropogenic disturbance: large-scale effects of loss of canopy algae on rocky shores. Marine Ecology Progress Series **214**:137-150.

Benedetti-Cecchi, L., L. Tamburello, E. Maggi, and F. Bulleri. 2015. Experimental Perturbations Modify the Performance of Early Warning Indicators of Regime Shift. Current Biology **25**:1867-1872.

Berlow, E. L., S. A. Navarrete, C. J. Briggs, M. E. Power, and B. A. Menge. 1999. Quantifying variation in the strengths of species interactions. Ecology **80**:2206-2224.

Bueno de Mesquita, C. P., A. J. King, S. K. Schmidt, E. C. Farrer, and K. N. Suding. 2016. Incorporating biotic factors in species distribution modeling: are interactions with soil microbes important? Ecography **39**:970-980.

Bulleri, F., L. Benedetti-Cecchi, M. Cusson, E. Maggi, F. Arenas, R. Aspden, I. Bertocci, T. P. Crowe, D. Davoult, B. K. Eriksson, S. Fraschetti, C. Gollety, J. N. Griffin, S. R. Jenkins, J. Kotta, P. Kraufvelin, M. Molis, I. S. Pinto, A. Terlizzi, N. Valdivia, and D. M. Paterson. 2012. Temporal stability of European rocky shore assemblages: variation across a latitudinal gradient and the role of habitat-formers. Oikos **121**:1801-1809.

Bulleri, F., A. Cucco, M. Dal Bello, E. Maggi, C. Ravaglioli, and L. Benedetti-Cecchi. 2018. The role of wave-exposure and human impacts in regulating the distribution of alternative habitats on NW Mediterranean rocky reefs. Estuarine, Coastal and Shelf Science **201**:114-122.

Clark, A. T., H. Ye, F. Isbell, E. R. Deyle, J. Cowles, G. D. Tilman, and G. Sugihara. 2015. Spatial convergent cross mapping to detect causal relationships from short time series. Ecology **96**:1174-1181.

Coleman, R. A., A. J. Underwood, L. Benedetti-Cecchi, P. Aberg, F. Arenas, J. Arrontes, J. Castro, R. G. Hartnoll, S. R. Jenkins, J. Paula, P. Della Santina, and S. J. Hawkins. 2006. A continental scale evaluation of the role of limpet grazing on rocky shores. Oecologia **147**:556-564.

Connolly, S. R., S. A. Keith, R. K. Colwell, and C. Rahbek. 2017. Process, Mechanism, and Modeling in Macroecology. Trends in Ecology & Evolution **32**:835-844.

Corlett, R. T. 2015. The Anthropocene concept in ecology and conservation. Trends in Ecology & Evolution **30**:36-41.

Dal Bello, M., E. Maggi, L. Rindi, A. Capocchi, D. Fontanini, C. Sanz-Lazaro, and L. Benedetti-Cecchi. 2015. Multifractal spatial distribution of epilithic microphytobenthos on a Mediterranean rocky shore. Oikos **124**:477-485.

Dal Bello, M., L. Rindi, and L. Benedetti-Cecchi. 2017. Legacy effects and memory loss: how contingencies moderate the response of rocky intertidal biofilms to present and past extreme events. Global Change Biology **23**:3259-3268.

Dayton, P. K., V. Currie, T. Gerrodette, B. D. Keller, R. Rosenthal, and D. Ventresca. 1984. Patch Dynamics and Stability of Some California Kelp Communities. Ecological Monographs **54**:253-289.

Deyle, E. R., M. C. Maher, R. D. Hernandez, S. Basu, and G. Sugihara. 2016a. Global environmental drivers of influenza. Proceedings of the National Academy of Sciences of the United States of America **113**:13081-13086.

Deyle, E. R., R. M. May, S. B. Munch, and G. Sugihara. 2016b. Tracking and forecasting ecosystem interactions in real time. Proceedings of the Royal Society B-Biological Sciences **283**:10.1098/rspb.2015.2258.

Deyle, E. R., and G. Sugihara. 2011. Generalized Theorems for Nonlinear State Space Reconstruction. Plos One **6**:10.1371/journal.pone.0018295.

Elith, J., and J. R. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology Evolution and Systematics **40**:677-697.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters **8**:993-1009.

Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter. 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment **11**:156-162.

Hewitt, J. E., S. F. Thrush, P. K. Dayton, and E. Bonsdorff. 2007. The effect of spatial and temporal heterogeneity on the design and analysis of empirical studies of scale-dependent systems. American Naturalist **169**:398-408.

Iknayan, K. J., M. W. Tingley, B. J. Furnas, and S. R. Beissinger. 2014. Detecting diversity: emerging methods to estimate species diversity. Trends in Ecology & Evolution **29**:97-106.

Krumhansl, K. A., D. K. Okamoto, A. Rassweiler, M. Novak, J. J. Bolton, K. C. Cavanaugh, S. D. Connell, C. R. Johnson, B. Konar, S. D. Ling, F. Micheli, K. M. Norderhaug, A. Pérez-Matus, I. Sousa-Pinto, D. C. Reed, A. K. Salomon, N. T. Shears, T. Wernberg, R. J. Anderson, N. S. Barrett, A. H. Buschmann, M. H. Carr, J. E. Caselle, S. Derrien-Courtel, G. J. Edgar, M. Edwards, J. A. Estes, C. Goodwin, M. C. Kenner, D. J. Kushner, F. E. Moy, J. Nunn, R. S. Steneck, J. Vásquez, J. Watson, J. D. Witman, and J. E. K. Byrnes. 2016. Global patterns of kelp forest change over the past half-century. Proceedings of the National Academy of Sciences **113**:13785-13790.

Maggi, E., L. Rindi, M. Dal Bello, D. Fontatnini, A. Capocchi, L. Bongiorni, and L. Benedetti-Cecchi. 2017. Spatio-temporal variability in Mediterranean rocky shore microphytobenthos. Marine Ecology Progres Series **575**:17-29.

Magurran, A. E., M. Dornelas, F. Moyes, N. J. Gotelli, and B. McGill. 2015. Rapid biotic homogenization of marine fish assemblages. Nature Communications **6**:10.1038/ncomms9405.

McGaughran, A. 2015. Integrating a population genomics focus into biogeographic and macroecological research. Frontiers in Ecology And Evolution **3**:doi.org/10.3389/fevo.2015.00132.

McGill, B. J., and A. Potochnik. 2018. Mechanisms Are Causes, Not Components: A Response to Connolly et al. Trends Ecol Evol **33**:304-305.

Menge, B. A., E. Sanford, B. A. Daley, T. L. Freidenburg, G. Hudson, and J. Lubchenco. 2002. Inter-hemispheric comparison of bottom-up effects on community structure: Insights revealed using the comparative-experimental approach. Ecological Research **17**:1-16.

Mock, T., S. J. Daines, R. Geider, S. Collins, M. Metodiev, A. J. Millar, V. Moulton, and T. M. Lenton. 2016. Bridging the gap between omics and earth system science to better understand how environmental change impacts marine microbes. Global Change Biology **22**:61-75.

Mod, H. K., P. C. le Roux, A. Guisan, and M. Luoto. 2015. Biotic interactions boost spatial models of species richness. Ecography **38**:913-921.

Morueta-Holme, N., B. Blonder, B. Sandel, B. J. McGill, R. K. Peet, J. E. Ott, C. Violle, B. J. Enquist, P. M. Jørgensen, and J.-C. Svenning. 2016. A network approach for inferring species associations from co-occurrence data. Ecography **39**:1139-1150.

Murphy, R. J., A. J. Underwood, and M. H. Pinkerton. 2006. Quantitative imaging to measure photosynthetic biomass on an intertidal rock-platform. Marine Ecology Progress Series **312**:45-55.

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

Pellissier, L., R. P. Rohr, C. Ndiribe, J. N. Pradervand, N. Salamin, A. Guisan, and M. Wisz. 2013. Combining food web and species distribution models for improved community projections. Ecol Evol **3**:4572-4583.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, M. A. McCarthy, and J. McPherson. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution **5**:397-406.

Rindi, L., M. D. Bello, L. Dai, J. Gore, and L. Benedetti-Cecchi. 2017. Direct observation of increasing recovery length before collapse of a marine benthic ecosystem. Nature Ecology and Evolution **1**:10.1038/s41559-41017-40153.

Sanz-Lazaro, C., L. Rindi, E. Maggi, M. Dal Bello, and L. Benedetti-Cecchi. 2015. Effects of grazer diversity on marine microphytobenthic biofilm: a 'tug of war' between complementarity and competition. Marine Ecology Progress Series **540**:145-155.

Staniczenko, P. P. A., P. Sivasubramaniam, K. B. Suttle, and R. G. Pearson. 2017. Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. Ecol Lett **20**:693-707.

Strain, E. M., R. J. Thomson, F. Micheli, F. P. Mancuso, and L. Airoldi. 2014. Identifying the interacting roles of stressors in driving the global loss of canopy-forming to mat-forming algae in marine ecosystems. Glob Chang Biol **20**:3300-3312.

Sugihara, G. 1994. Nonlinear Forecasting for the Classification of Natural Time-Series. Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences **348**:477-495.

Sugihara, G., R. May, H. Ye, C. H. Hsieh, E. Deyle, M. Fogarty, and S. Munch. 2012. Detecting causality in complex ecosystems. Science **338**:496-500.

Sugihara, G., and R. M. May. 1990. Nonlinear Forecasting as a Way of Distinguishing Chaos from Measurement Error in Time-Series. Nature **344**:734-741.

Sutherland, W. J., R. Aveling, T. M. Brooks, M. Clout, L. V. Dicks, L. Fellman, E. Fleishman, D. W. Gibbons, B. Keim, F. Lickorish, K. A. Monk, D. Mortimer, L. S. Peck, J. Pretty, J. Rockstrom, J. P. Rodriguez, R. K. Smith, M. D. Spalding, F. H. Tonneijck, and A. R. Watkinson. 2014. A horizon scan of global conservation issues for 2014. Trends in Ecology & Evolution **29**:15-22.

Takens, F. 1981. Detecting strange attractors in turbulence. Pages 366-381 Dynamical Systems and Turbulence, Warwick 1980, Lecture Notes in Mathematics. Springer, Berlin Heidelberg.

Tyberghein, L., H. Verbruggen, K. Pauly, C. Troupin, F. Mineur, and O. De Clerck. 2012. Bio-ORACLE: a global environmental dataset for marine species distribution modelling. Global Ecology and Biogeography **21**:272-281.

Underwood, A. J. 1984. Vertical and Seasonal Patterns in Competition for Microalgae between Intertidal Gastropods. Oecologia **64**:211-222.

Underwood, A. J. 1997. Experiments in ecology: their logical design and interpretation using analysis of variance. Cambridge University Press.

Vergés, A., C. Doropoulos, H. A. Malcolm, M. Skye, M. Garcia-Pizá, E. M. Marzinelli, A. H. Campbell, E. Ballesteros, A. S. Hoey, A. Vila-Concejo, Y.-M. Bozec, and P. D. Steinberg. 2016. Long-term empirical evidence of ocean warming leading to tropicalization of fish communities, increased herbivory, and loss of kelp. Proceedings of the National Academy of Sciences **113**:13791-13796.

Warton, D. I. 2015. New opportunities at the interface between ecology and statistics. Methods in Ecology and Evolution **6**:363-365.

Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. 2015. So Many Variables: Joint Modeling in Community Ecology. Trends in Ecology & Evolution **30**:766-779.

Ye, H., and G. Sugihara. 2016. Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. Science **353**:922-925.

680 **Legend to figures**

Figure 1. Conceptual approaches integrating observations with experiments. In each panel (a-c), arrows illustrate the cyclical process where observations (red) motivate experiments (cyan) and experiments generate new observations. The interspersion of arrows provides a qualitative indication of the level of integration between observational and experimental data. Bars on the 685 right side of each panel indicate the relative importance (thickness) and degree of integration (distance between bars) between scope (SC) and causal inferential strength (CI). (a) The logical dependence of observations on experiments is at the core of the scientific method and it is the typical level of integration between observational and experimental areas of scientific inquiry (Underwood 1997). Observational and experimental data are logically linked when 690 experiments seek to explain the processes underlying observed patterns and trends in ecological variables. In this framework, observations do not contribute to increase the causal inferential strength of experiments. (b) A further level of integration is possible by treating observations as covariates in the comparative experimental approach (Menge et al. 2002, Hewitt et al. 2007). A statistical relationship between observational and experimental data can 695 be established when identically designed experiments are distributed along relevant axes of environmental variation (e.g., latitudinal or elevational gradients) and ancillary data are collected to characterize the environment. Factoring out potential confounding effects through covariates strengthens causal inferential strength in the analysis of distributed experiments. (c) Hybrid datasets: short experimental time series of predictor and response variables can be 700 joined with longer observational time series of the same variables and probed for causality using emerging techniques such as convergent cross-mapping (CCM; see Case study 1). Similarly, hybrid species abundance matrices combining experimental data with large-scale observations can be used to increase the causal inferential strength and statistical power of

joint species distribution models (jSDM). Experimentally-derived effect sizes (e.g., species interaction strengths) can also be hybridized with species occurrence data to obtain a plot-level covariate to use in jSDM (see Case study 2).

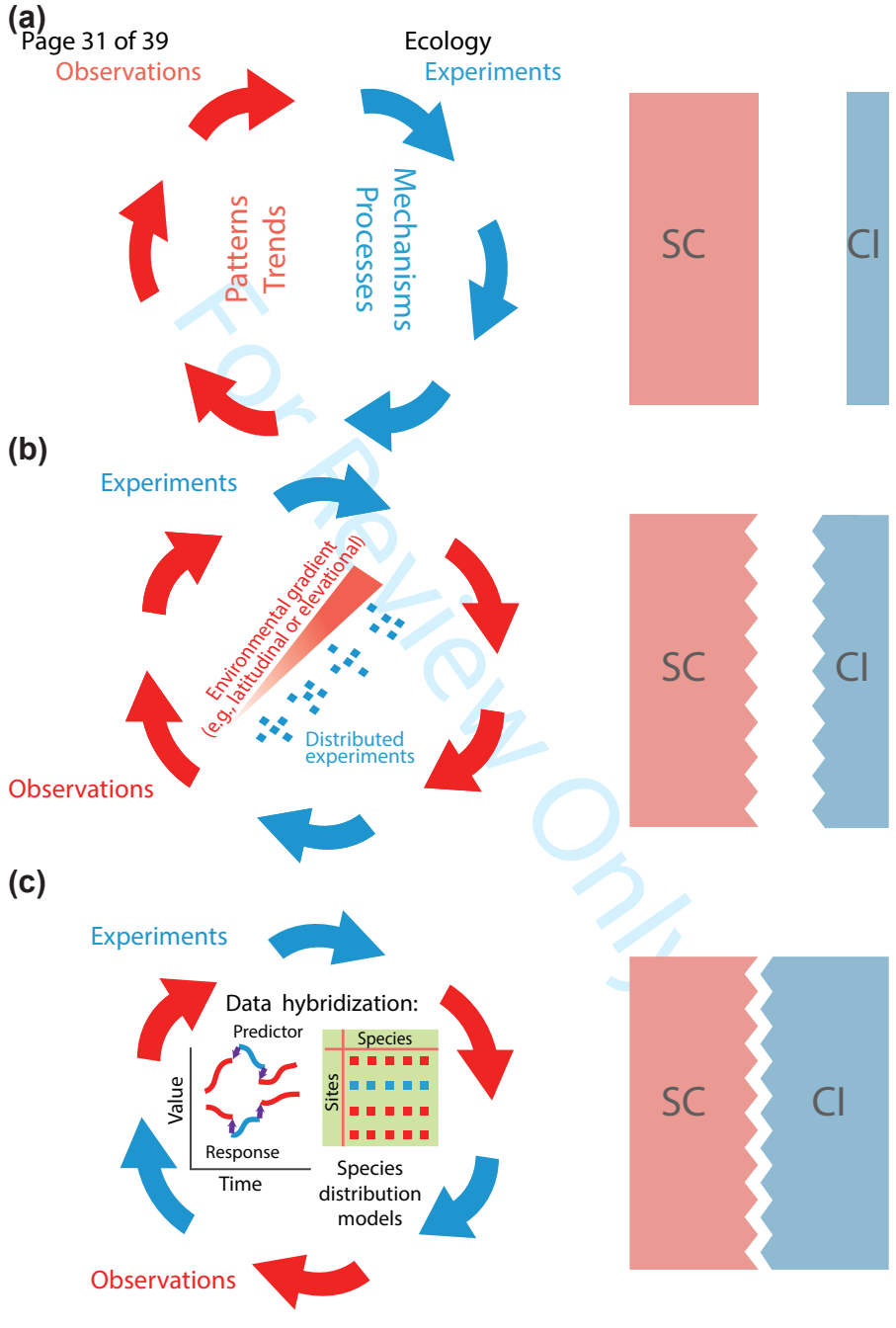
Figure 2. Detecting causality from time series. (a) The canonical Lorenz attractor (a classical example of an attractor manifold originating from a system of ordinary differential equations originally developed to model atmospheric circulation). Time series of individual variables originate as projections of the system state onto one of the coordinate axes as time unfolds (shown here by the red line for variable Z); (b) a generic example of CCM: the nearest neighbors to a focal point on predictor manifold M_X (white and black dots, respectively) are mapped to M_Y and used to predict the target value $\hat{Y}(t)|M_X$; (c) simulated time series of two competing species and (d) CCM for the simulated time series of the two competing species.

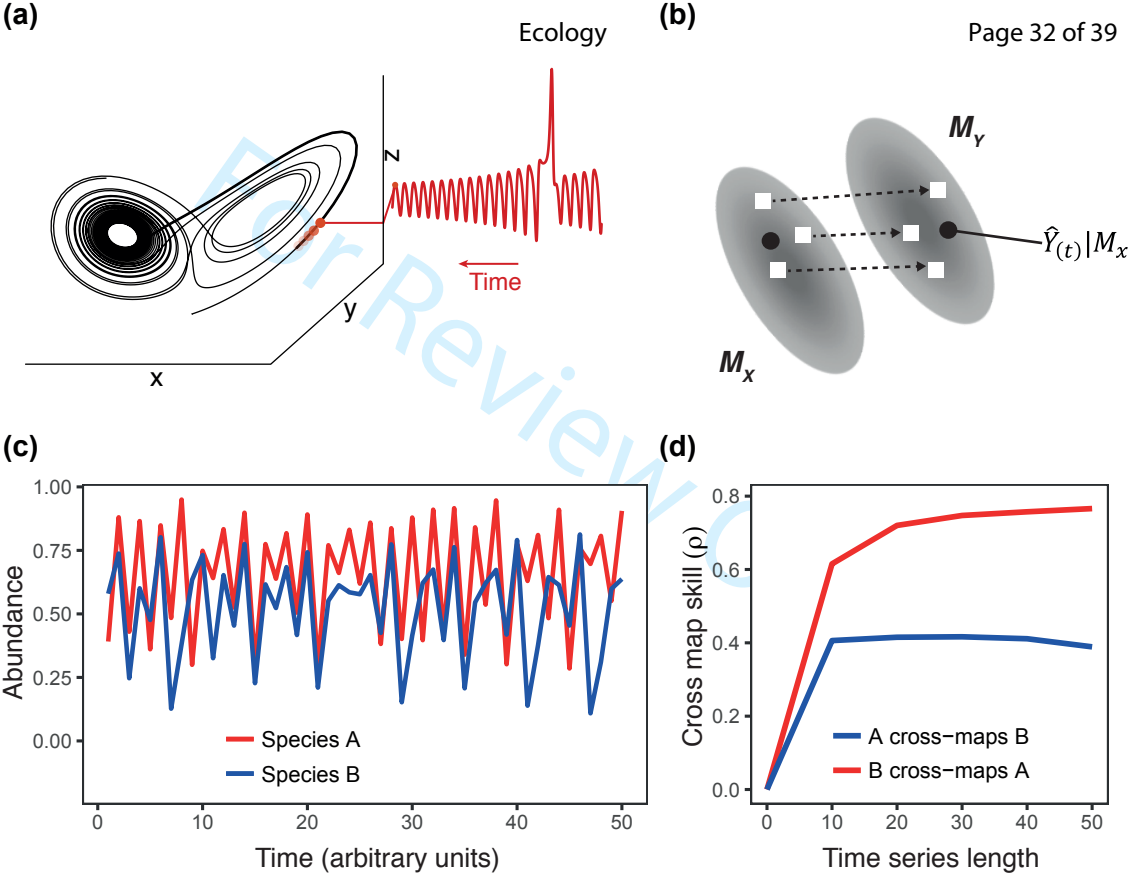
Figure 3. Application of CCM to hybrid time series. (a) hypothetical hybrid time series integrating observations of a driver (orange) and a response (blue) variable with data from a press experiment (red; see text and Appendix S1 for further details); (b) example of a pulse experiment where a hybrid dataset is generated for the driver (experimental data shown as a continuous red line, observations in orange). The response variable is shown as a dotted blue line; (c) real-world observational time series of temperature (orange) and chlorophyll *a* (*chl*) concentration on a rocky intertidal shore (blue lines, with line type showing different replicates). The plot shows daily temperature values, although only monthly maxima were used in the analysis to match *chl* data; (d) CCM fails to identify the correct causal relation between temperature and *chl* in the observational dataset (bootstrap test: $P>0.05$ for *chl* cross-mapping temperature and $P<0.01$ for temperature cross-mapping *chl*; analysis on scaled data; embedding dimension $E=2$ for temperature and 3 for *chl*; shaded regions are 95% bootstrapped confidence intervals); (e) hybrid time series integrating observational (orange)

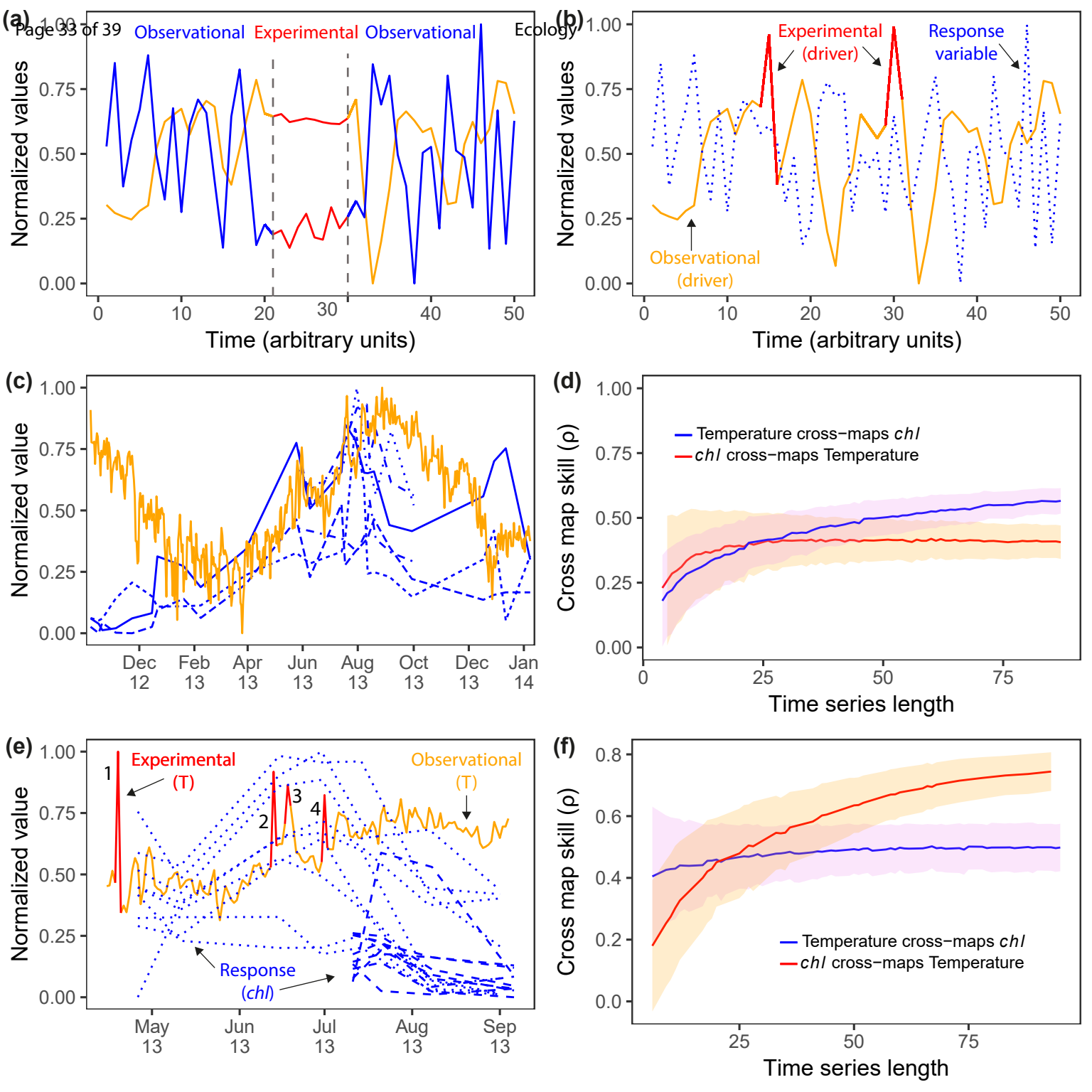
and experimental (red) values of temperature and *chl* concentration (blue) from replicated experimental plots (dotted lines indicate plots exposed to warming events 1 and 2, occurring 60 days apart; dashed lines indicate plots exposed to warming events 3 and 4, 15 days apart); (f) CCM correctly identifies the asymmetrical causal relation between temperature and *chl*. Cross map skill increased significantly for *chl* cross-mapping temperature (red line, $P<0.001$, bootstrap test), but not when using temperature to cross map *chl* (blue line, $P>0.05$) (analysis on scaled data; embedding dimension $E=4$ for temperature and 2 for *chl*; shaded regions are bootstrapped standard errors).

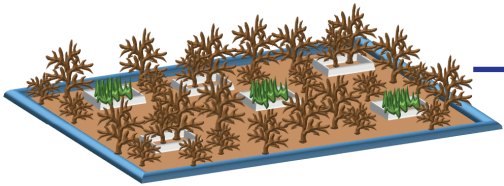
Figure 4. Two paths to generate hybrid spatial datasets. These include direct concatenation of experimental and observational data to increase sample size (Path 1) or the derivation of an effect size for each species from the experiment that is then translated into a covariate to be included in jSDM (Path 2). Depending on the nature of the experiment (a hypothetical canopy-removal experiment in this example), effect sizes can reflect the importance of species interactions such as the *RII* index discussed in the main text or other measures of interaction strength. Experimentally estimated species-specific effects are indicated here for the generic species *i* as a_i (with *i* varying from 1 to *s*). The average effect size over the species present in the generic observational plot *p* are indicated as $\bar{a}_p$ (plots from 1 to *n*).

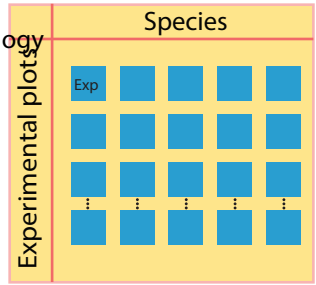
Figure 5. Application of jSDM to hybrid datasets. (a) variance partitioning for jSDM with biotic interactions and (b) without biotic interactions. (c) comparison of R^2 (d) area under the curve (AUC) and (e) true skill statistic (TSS) values between models with and without biotic interactions (identity lines in red); (f) species association networks (Pearson correlation) at the location scale for jSDM with biotic interactions and (g) without biotic interactions.






(a)
Observational
Experimental
Observational
Normalized values
Time (arbitrary units)
(b)
Experimental
(driver)
Response
variable
Observational
(driver)
Normalized values
Time (arbitrary units)
(c)
Normalized value
Dec 12
Feb 13
Apr 13
Jun 13
Aug 13
Oct 13
Dec 13
Jan 14
(d)
Temperature cross−maps chl
chl cross−maps Temperature
Cross map skill (ρ)
Time series length
(e)
Experimental
(T)
Observational
(T)
1
2
3
4
Response
(chl)
Normalized value
May 13
Jun 13
Jul 13
Aug 13
Sep 13
(f)
Temperature cross−maps chl
chl cross−maps Temperature
Cross map skill (ρ)
Time series length



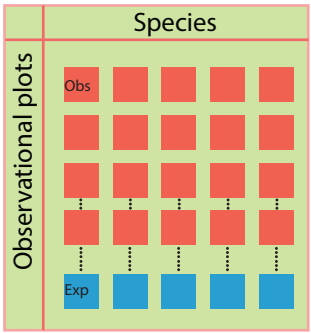


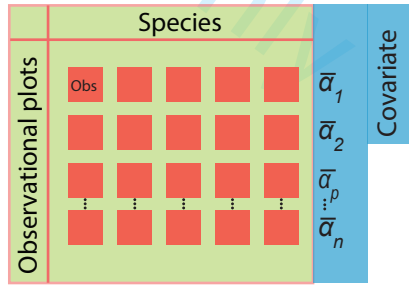
Path 1: direct integration of observational and experimental plots

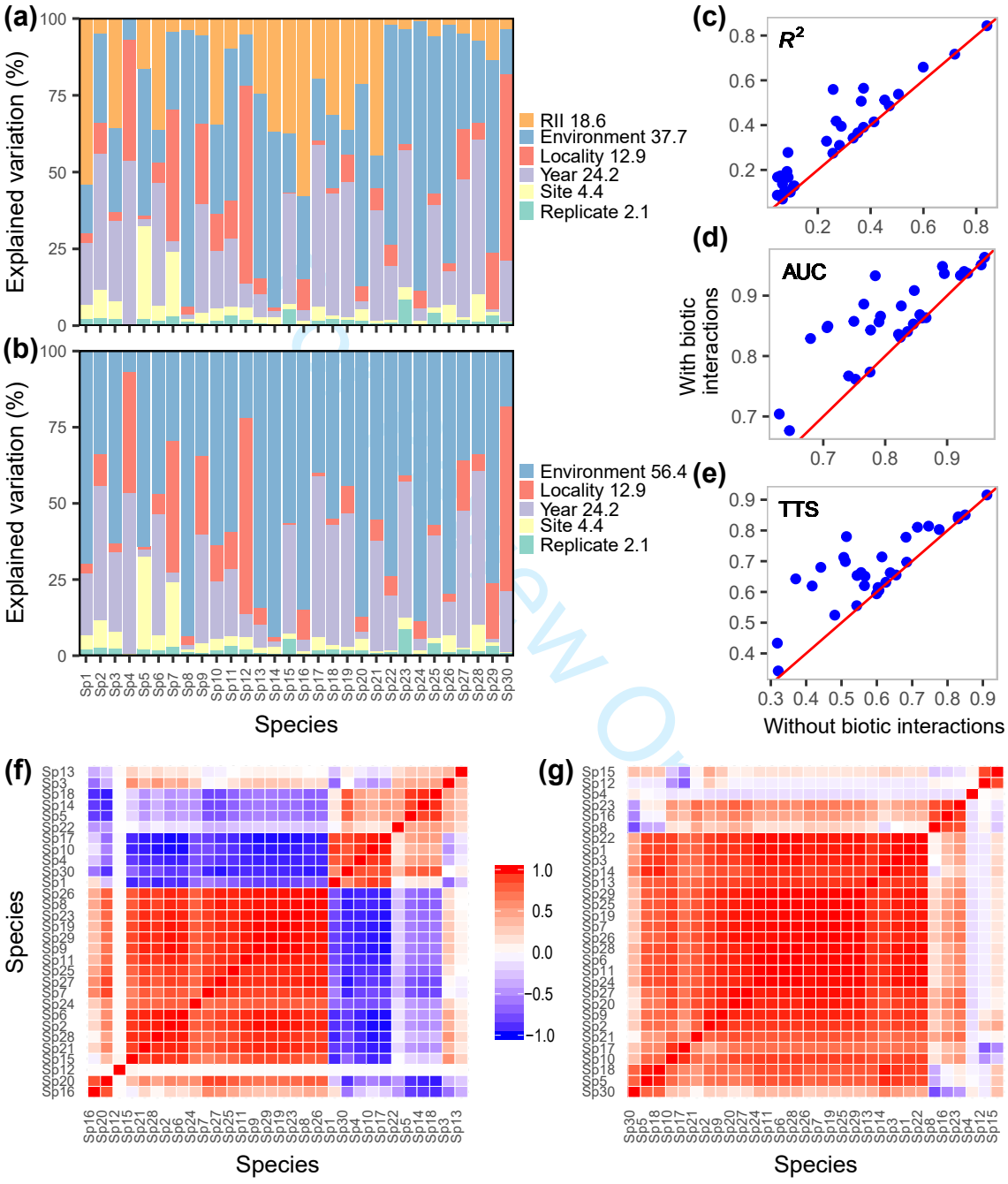
Path 2: an effect size is estimated for each species from the experiment and an average effect size is obtained for each observational plot on the basis of species presence/absence

Species' effect sizes

α_1 α_2 α_3 α_i ··· α_s






(a)
Explained variation (%)
RII 18.6
Environment 37.7
Locality 12.9
Year 24.2
Site 4.4
Replicate 2.1
(b)
Explained variation (%)
Environment 56.4
Locality 12.9
Year 24.2
Site 4.4
Replicate 2.1
Species
(c)
R^2
(d)
AUC
With biotic interactions
(e)
TTS
Without biotic interactions
(f)
Species
Species
(g)
Species

APPENDIX S1

Hybrid datasets: integrating observations with experiments in the era of macroecology and big-data

Lisandro Benedetti-Cecchi[1*], Fabio Bulleri[1], Martina Dal Bello[2], Elena Maggi[1], Chiara Ravaglioli[1], Luca Rindi[1]

[1]Department of Biology, University of Pisa, CoNISMa, Via Derna 1, 56126, Pisa, Italy
[2]Physics of Living Systems Group, Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139

*Correspondence to:
Lisandro Benedetti-Cecchi
e-mail lbenedetti@biologia.unipi.it
Voice: +39 050 2211413
Fax: +39 050 2211410

CONTENT

Convergent Cross Mapping (CCM), bootstrap and probability test

CCM uses data points on the reconstructed manifold of one variable (e.g., variable X on manifold M_X) to predict concurrent points (i.e. at the same time) on the manifold reconstructed from another variable (e.g., variable Y on manifold M_Y). If variables X and Y are causally related, the time indices of nearby points on the manifold of one variable will correspond to nearby points on the manifold of the other variable allowing for accurate predictions using cross-mapped points form one manifold to the other (Fig. 2b in main text). If causation is asymmetrical – e.g., variable X causes variable Y – only the time indices of nearby points on the manifold of Y will correspond to nearby points on the manifold of X, but the reverse is not true because the forcing variable X contains no information about the response variable Y. Due to this asymmetry, forecasting skill (ρ) will increase with increasing time series length (L) only when predicting from the response to the forcing variable (Figs. 2d, 3d and 3f in main text).

Two statistical approaches are used to differentiate between causal and non-causal relations in CCM: bootstrapping and probability testing. Bootstrapping, as implemented by function CCM_boot in package multispatialCCM, involves resampling with replacement n spatial replicates n times and repeating CCM on these resampled replicates for the best embedding dimension E (Clark et al. 2015). This procedure is repeated 1000 times to build bootstrapped standard errors (the sample standard deviation of the bootstrapped distribution) around mean ρ (Figs. 3d and 3f in main text). The probability test assesses whether ρ is significantly larger than zero and increases significantly with L. Probability is computed as 1 minus the percentage of times ρ at maximum L is simultaneously greater than zero and larger than ρ at the shortest L in the bootstrapping iterations (Clark et al. 2015).

Diagnostics and forecasting ability of CCM

Embedding dimension
The reconstruction of an attractor manifold requires the identification of the best embedding dimension E – i.e., the number of time steps used for prediction (see Box 1 in main text) (Sugihara et al. 2012). We used function SSR_pred_boot in package multispatialCCM (Clark et al. 2015), which employs cross-validation to assess the predictive ability of reconstructed attractors for a range of E values. We run the analysis for chlorophyll and temperature time series separately. The best E is the one providing the highest predictive power from one time-step to the next, based on Pearson correlation. Results are illustrated in Fig. S1a,c, with best values of E identified by dashed lines.

Nonlinearity and forecasting
An underlying assumption of CCM is that the system under investigation undergoes nonlinear dynamics (Sugihara et al. 2012, Clark et al. 2015). A diagnostic of this behavior is the decay of predictive ability with increasing prediction step. Our expectation is that the analysis has reasonable predictive ability for short time steps and that predictive ability decreases with increasing time horizon.

We used function SSR_check_signal for this analysis, with the best *E* selected before. Results are illustrated in Fig. S1b,d, separately for chlorophyll and temperature time series. We note that forecasting ability at step one is larger for hybrid (Fig. S1d) than purely observational (Fig. S1b) time series (0.789 vs. 0.675). One caveat of this analysis is that a linear system dominated by autocorrelated ('red') noise could still show the same decay of prediction skill with time, but in this event the CCM analysis should highlight no causal link in either direction, since increasing information about the system will not increase predictive power. Our CCM identified a significant increase of cross map skill with time series length for *chl* cross-mapping temperature, but not in opposite direction (Fig. 3c-e), indicating that spurious effects due to autocorrelation were unlikely and suggesting a true causal effect of temperature on *chl*.

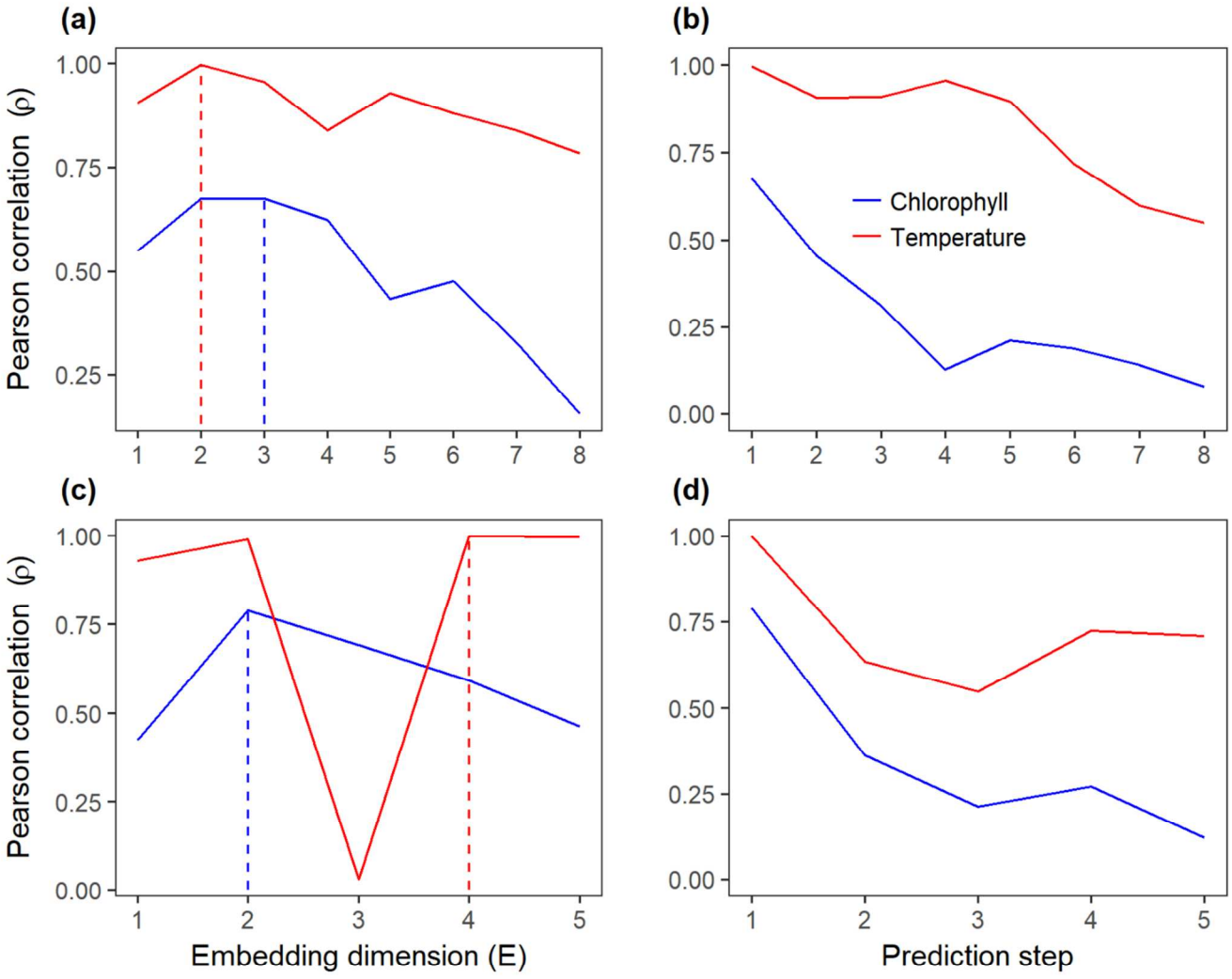


Figure S1. Best embedding dimension *E* and test of nonlinearity for purely observational (a,b) and experimental (c,d) time series of chlorophyll and temperature data. Hybrid time series were assembled only for temperature, by replacing observed daily maxima with the values imparted experimentally. *Chl* data from pulse experimental plots were already hybrid time series, incorporating natural and experimental effects of temperature variation (See main text for details). Dashed lines highlight the best *E* used in CCM.

Simulation of a two species competition system

This simulation follows the procedure described in Sugihara *et al.* (2012) and it is based on a two-species, discrete-time competition model as implemented in function make_ccm_data of package multispatial CCM (Clark et al. 2015). Time series of abundances of species *X* and *Y* were generated as:

$$X(t+1) = X(t)(r_x - r_x X(t) - \beta Y(t))$$

$$Y(t+1) = Y(t)(r_y - r_y Y(t) - \alpha X(t))$$

where t is time, r_x and r_y are species' intrinsic growth rates and α and β are interspecific interaction coefficients. Fig. 2c illustrates the results of one realization of the simulation with parameters chosen to reflect asymmetrical competition. The full set of parameters and code for this simulation are specified in R function CompetingSpecies.R (see *Description of code* below to access the function).

Simulation of hybrid time series (Fig. 3a in main text)

We started this toy example by simulating a time series of a hypothetical predictor variable using an autoregressive integrated moving average (ARIMA) model (orange line in Fig. 3a of main text). The value of the response variable at each time step was generated as a 30% reduction of the value of the predictor at the previous time step, plus random noise (blue line in Fig. 3a of main text). These series were then scaled to vary between zero and one. In this way we simulated observations that were causally related and where the predictor caused a lagged response one step ahead. Time series from a hypothetical press experiment were generated by maintaining the value of the predictor 20% above the mean of the observational time series and imposing an effect size on the response variable consisting of a 50% reduction of the value of the predictor at the previous time step. A small amount of random variation was added to both the predictor and the response variables. The hybrid dataset was obtained by replacing observational with experimental values of the hypothetical predictor and response time series between time steps 21 and 30 (red lines in Fig. 3a of main text). For the pulse experiment example (Fig. 3b in main text) we simply introduced two spikes in the response variable at times 15 and 30. Parameters and code are specified in R function Fig3.R (see *Description of code* below to access the function).

Description of datasets

Datasets are provided in R (extension: .RData) and are available on figshare at https://figshare.com/s/5dcf6dae011c15d71c39

Chlorophyll data

biofilm_cont – dataset containing observational chlorophyll and temperature data for CCM analysis

biofilm_treat – dataset containing experimental chlorophyll and hybrid temperature data for CCM analysis

biof_control_plot – includes sampling date, the original chlorophyll values as in biofilm_cont, normalized chlorophyll values and paired temperature maxima. Used for plotting

temp_C – daily observed temperature data for plotting

biof_treat_plot – includes sampling date, the original chlorophyll values as in biofilm_treat, normalized chlorophyll values and paired temperature maxima. Used for plotting

temp_T – daily hybrid temperature data for plotting
plot_ccm_cont – results of CCM on observational data. Used for plotting
plot_ccm_treat – results of CCM on observational data. Used for plotting

Spatial community data
hmsc_dat – species presence-absence data
pred_vars – environmental covariates and mean interaction strength index (mean.rii)
cross_test_rii_30 – results of cross-validation analysis using function hmsc.crossval.parallel, designed for parallel computing. Used for plotting the results.

Description of code

R scripts are available on figshare at https://figshare.com/s/ff3d1f4f01095e36977e

Chlorophyll data
CCM_biofilm.R – function to perform CCM analysis on chlorophyll data
CompetingSpecies.R – code for competing species simulation; used to generate Fig. 2c,d
Fig.3.R – code to generate Fig. 3 in the main text

Spatial community data
hmsc.crossval.parallel.R – function to perform cross-validation on hmsc models
hmsc.R – function to reproduce the results of the jSDM analysis in the main text

References

Clark, A. T., H. Ye, F. Isbell, E. R. Deyle, J. Cowles, G. D. Tilman, and G. Sugihara. 2015. Spatial convergent cross mapping to detect causal relationships from short time series. Ecology **96**:1174-1181.

Sugihara, G., R. May, H. Ye, C. H. Hsieh, E. Deyle, M. Fogarty, and S. Munch. 2012. Detecting causality in complex ecosystems. Science **338**:496-500.