Decomposing predictability to identify dominant causal drivers in complex ecosystems.

Kenta Suzuki[1]*. Shin-ichiro S. Matsuzaki[2] and Hiroshi Masuya[1]

1. BioResource Research Center, RIKEN, Tsukuba, Ibaraki 305-0074, Japan.

2. Biodiversity Division, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan.

* Corresponding author: Kenta Suzuki

Email: kenta.suzuki.zk@riken.jp

Author Contributions: K.S. conceptualized the study, developed methodology, performed the formal analysis, and wrote the first draft. S.S.M. contributed to data curation and analysis, and wrote the paper with K.S.. All authors critically reviewed and edited the paper.

Competing Interest Statement: We declare that we have no conflict of interest.

Classification: Biological sciences/Ecology, Physical sciences/Environmental Sciences

Keywords: Ecosystem monitoring, Lake ecosystem, Causal network, Echo state network, Granger causality

Abstract

Ecosystems are complex systems of various physical, biological, and chemical processes. Since ecosystem dynamics are composed of a mixture of different levels of stochasticity and nonlinearity, handling these data is a challenge for existing methods of time-series based causal inferences. Here we show that, by harnessing contemporary machine learning approaches, the concept of Granger causality can be effectively extended to the analysis of complex ecosystem time series and bridge the gap between dynamical and statistical approaches. The central idea is to use an ensemble of fast and highly predictive artificial neural networks to select a minimal set of variables that maximizes the prediction of a given variable. It enables decomposition of the relationship among variables through quantifying the contribution of an individual variable to the overall predictive performance. We show how our approach, EcohNet, can improve interaction network inference for a mesocosm experiment and simulated ecosystems. The application of the method to a long-term lake monitoring dataset yielded new but interpretable results on the drivers causing cyanobacteria blooms, which is a serious threat to ecological integrity and ecosystem services. Since performance of EcohNet is enhanced by its predictive capabilities, it also provides an optimized forecasting of overall components in ecosystems. EcohNet could be used to analyze complex and hybrid multivariate time series in many scientific areas not limited to ecosystems.

Significance Statement

Effective use of ecosystem monitoring data to resolve global environmental issues is a major challenge of the 21st century ecology. A promising solution to address this challenge is a time-series-based causal analysis which can provide insight on the mechanical links between ecosystem components. In this work, a model-free framework named EcohNet is proposed. EcohNet utilizes ensemble predictions of echo state networks, which are known to be fast, accurate, and highly relevant for a variety of dynamical systems, and can robustly predict causal networks of ecosystem components. It also can provide an optimized forecasting of overall ecosystem components, and could be used to analyze complex and hybrid multivariate time series in many scientific areas, not limited to ecosystems.

Main Text

Introduction

Various systems in the world, from cells, organisms, ecosystems, and our own societies, are complex and driven by many interacting components. An attempt to infer relationships among the components as a causal network is an important step in understanding their mechanistic basis. Great efforts have been made to infer such networks from time series [1, 2]. In particular, methods that can overcome the limitations of classical methods [3, 4], such as CCM [5], which is suitable for systems under the influence of dynamic processes, and PCMCI [6], which is suitable for stochastic systems, have been proposed and are becoming widely used.

However, there is a problem when attempting to apply the existing time-series-based causal analyses to ecosystems. Ecosystems are complex systems of various physical, biological, and chemical processes [7]. Meteorological variables such as temperature and precipitation are under the influence of atmospheric and oceanic fluid dynamics [8]. Recent studies showed that these earth system dynamics are well captured as stochastic processes, and approaches based on statistical causal analysis are effective in elucidating their relationships [2, 6, 9]. In contrast, the density and abundance of organisms often show strong nonlinear dynamics, driven by interactions between organisms. This motivates the application and success of approaches based on nonlinear dynamical systems [5, 10-12]. Some chemicals control ecological dynamics as essential resources and are produced by organismal activities, as well they are under the control of global geochemical cycles [13]. These processes influence each other, but it is through rather weak causal couplings, and the dynamics of any one process may not become dominant [5, 14-17]. Methods for revealing causal relationships among ecosystem components from time series data need to be robust to dynamical complexity, i.e., the different levels of stochasticity and nonlinearity. However, the assumptions of most methods do not satisfy this requirement [1, 18].

In this paper, we introduce a new method called EcohNet. It is based on the ensemble prediction of neural networks [19-21] that can seamlessly handle stochastic/deterministic and linear/nonlinear dynamics [22]. Therefore, it is expected to be robust to dynamical complexity. The ensemble prediction is used to decompose relationships among variables in terms of predictability [23]. Here, the contribution of one variable to predictive performance can be evaluated apart from that of the other variables, as in the concepts of partial correlation and conditional independence. It is expected that even if some variables are driven by a strong driver, weak relationships among variables can be detected separately from the effect of the driver without special treatments [24-27].

Although various methods have been proposed and used for descriptive purposes, the relevance of their application to actual ecosystem monitoring data has not been seriously examined. How well can interactions between ecosystem components be captured as causal relationships in EcohNet? What advantages does it have over conventional methods used to assess causal analysis? To answer these questions, we performed benchmarking with data from long-term observation of an aquatic mesocosm, as well as a simulated dataset to test robustness to different dynamical complexity (equilibrium, equilibrium forced by an external oscillator, and intrinsic oscillatory dynamics, as well as different magnitudes of noise) under two interaction types (food web and random interaction) and three observational conditions (different data size, sampling interval and the presence of unobserved species). In addition to the performance criteria for network inference, we focused on susceptibility to the often-problematic interaction topologies such as chain relationships (e.g., in a three-species food chain $X \leftarrow Y \leftarrow Z$, a causal relationship may be identified between X and Z) and fan-out relationships (e.g., in a one-predator-two-prey-relationships $X \leftarrow Z \rightarrow Y$, a causal relationship may be identified between X and Y) [5, 24, 28]. We then applied our method to a long-term lake monitoring dataset that includes heterogeneous components such as meteorological, chemical, and biological variables, and interpreted the results. Here, we mainly focused on how EcohNet provides insight on the drivers

of cyanobacteria blooms, which is often a serious threat to ecological integrity and ecosystem services [29], but also we show how well the detection of unrealistic causal relationships such as those from biological to meteorological variables were avoided.

EcohNet

EcohNet combines a type of recurrent neural network, called echo state network (ESN) [19-20, 30], with a progressive selection of variables [31] (Figure 1; see Materials and Methods for detail). Initially, a target variable is selected (Fig. 1A). Then, N ESNs are generated and the ensemble predictive performance (prediction skill) of one-time-step ahead is evaluated when the past state of the target itself is given as an input of ESNs (Figure 1B,C). As the result, we obtain a distribution of prediction skills (illustrated by a gray mountain shape in Fig.1B). In the next step, we choose a second variable other than the target variable one by one and adopt the one that improves the prediction skill the most. Here, N ESNs are generated for each pair of variables. If the prediction skill does not improve, no further new variables are adopted. This process is repeated for variables that have not yet been selected, as long as new variables are adopted. As a result of the above process, a set of variables $\overline{\Omega}$ that maximizes the prediction skill for the target variable is obtained. Then, $\overline{\Omega}$ is used to evaluate the *unique prediction skill*, which represents the unique contribution of each variable in $\overline{\Omega}$ to the overall prediction skill (Figure 1D). If $\overline{\Omega} = \{X, Y\}$ and the target variable is X, for each variable (here, X and Y), we evaluate how much the prediction changes when one variable is excluded from it (e.g., $\overline{\Omega}^{(-Y)} = \{X\}$). For example, when denoting the prediction skill for $\overline{\Omega}$ and $\overline{\Omega}^{(-Y)}$ as, $\rho_{\overline{\Omega}}$ and $\rho_{\overline{\Omega}^{(-Y)}}$ respectively, the unique contribution of Y on X is $\rho_{Y\to X} = \rho_{\overline{\Omega}} - \rho_{\overline{\Omega}^{(-Y)}}$. In this stage, a variable is removed from $\overline{\Omega}$ if $\rho_{\overline{\Omega}} - \rho_{\overline{\Omega}^{(-Y)}} \leq 0$. By performing the above procedure with all variables as targets in turn, a prediction-skill-based causal network is obtained for the entire system.

To show how EcohNet works, we explain what would be $\overline{\Omega}$ for X in some representative relationships of the three variables X, Y and Z. First, we assume a transitive relationship (Z→Y→X and Z→X; SI Appendix Fig. S1A). Here, in addition to $\rho_{X\to X}$, both $\rho_{Y\to X}$ and $\rho_{Z\to X}$ are non-zero corresponding to the direct relationship Y→X and Z→X. In a chain relationship (Z→Y→X; SI Appendix Fig. S1B) the direct relationship Z→X is removed. In this case, the contribution of Z on X is always mediated by Y. Thus, if all else is equal, only $\rho_{Y\to X}$ and $\rho_{X\to X}$ are expected to have non-zero values. In a fan-out relationship (X←Z→Y; SI Appendix Fig. S1C) the direct relationship Y→X is removed. In this case, both X and Y have a direct relationship with Z but remain disconnected from each other. X and Y share dynamic imprint of Z that may contribute to an ability to predict X from Y but it is simply involved in the direct influence of X on Z. Thus, if all else is equal, only $\rho_{X\to X}$ and $\rho_{Z\to X}$ are expected to have non-zero value. The importance of evaluating the unique prediction skill is highlighted by another representative example (SI Appendix Fig. S2). In this case, both Z_1 and Z_2 have a fan-out relationship with X and Y ($X \leftarrow Z_1 \rightarrow Y$, and $X \leftarrow Z_2 \rightarrow Y$). As explained in Runge et al. [1], a simple forward stepwise algorithm would select Y first followed by Z_1 and Z_2, because Y can have a larger individual contribution than either Z_1 or Z_2. However, evaluation of the unique prediction skill can remove Y from $\overline{\Omega}$ since contribution of Y is included in the union of the contribution of Z_1 and Z_2, and thus $\rho_{Y\to X} = \rho_{\overline{\Omega}} - \rho_{\overline{\Omega}^{(-Y)}} \leq 0$.

Conventional approaches

We selected a correlation-based method (Spearman rank correlation), two equation-free methods from nonlinear time series analysis (convergent cross mapping (CCM) [5] and partial cross mapping (PCM) [28]) and an equation-based method (LIMITS) [32] as the conventional approaches for comparison with EcohNet (please see SI Appendix A for the details of the implementations), and compared them using the evaluation criteria of a binary classification task

(SI Appendix Fig. S3). CCM and PCM are intended to detect causality in dynamical system and are therefore comparative to EcohNet. Since there are several different implementations for CCM, we tested three representative approaches (SI Appendix A2). PCM is an extension of CCM that accounts for direct causality between two variables conditional on indirect causation through a third variable. Thus, PCM implements an idea similar to the unique prediction skill of EcohNet. LIMITS infers the strength of ecological interactions directly rather than the causal relationships among variables. In the benchmark using simulation data, the results of LIMITS should be carefully interpreted because there is the advantage that the basic process is consistent, i.e., it assumes Lotka-Volterra (LV) equation used to generate the data. In other words, it contains more prior knowledge than the other methods. Applying it to relationships of all ecosystem components such as chemical and meteorological variables exceeds its scope of application. Although it is frequently pointed out that correlation does not signal the presence of interaction, we considered Spearman rank correlation as a baseline.

Results

Benchmarking

When applied to the data from a long-term mesocosm experiment [33], EcohNet identified 11 out of 13 interacting pairs of components, with 3 false positives, and was superior to conventional methods (Fig. 2a-d). It outperformed other methods in all evaluation criteria (Fig. 2e). The results were not sensitive to different parameter values that might affect the performance of echo state networks, except when the forgetting factor λ was very close to 1 (see SI Appendix B). A benchmark with datasets generated by food web models (SI Appendix Fig. S4) also supported the superiority of our method. ROC-AUC and F1-score of EcohNet outperformed CCM, PCM and Spearman rank correlation except for ROC-AUC of CCM in $O_{S,L}$ in which the signal of interaction was considered to be strong (Fig. 2g,h). Performance of CCM was largely reduced in $E_{S,L}$ and $EO_{S,L}$. PCM performed better than CCM in $E_{S,L}$, but it was reduced in $EO_{S,L}$. CCM was better than PCM in $EO_{S,L}$ potentially because of the use of seasonal surrogate method. Performance of Spearman rank correlation was comparable to other methods only in $E_{S,L}$. LIMITS outperformed EcohNet in $O_{S,L}$, especially in ROC-AUC. Its high applicability to cases of intrinsic oscillations has been reported in previous studies [34]. It is also worth noting that EcohNet's performance was comparable to LIMITS in most cases, despite not explicitly assuming underlying LV processes.

Although accuracy of EcohNet, CCM and PCM was consistent among $O_{S,L}$, $E_{S,L}$ and $EO_{S,L}$ except for PCM for $EO_{S,L}$, there were large variations in sensitivity and specificity for CCM and PCM (Fig. 3a-c). The number of detected links varied depending on the simulation conditions (Fig. 3d). It suggests that the fundamental sensitivity of the methods used to detect presence of interactions is affected by dynamic complexity. In comparison, performance of EcohNet was relatively stable, and its sensitivity, specificity and the number of links detected were not greatly affected by the simulation conditions. Our benchmarking also showed that false positives due to chain and fan-out relationships were best suppressed in EcohNet (Fig. 2i,j).

The above conclusions did not change significantly based on different dataset sizes (SI Appendix Fig. S5,6), sampling intervals (SI Appendix Fig. S7-10), and the presence of unobserved species (SI Appendix Fig. S11) that we tested. While these results are based on food web models where the interactions are bidirectional, the superiority of EcohNet was shown also in a random interaction model where the interactions were unidirectional and the detectability of directional interactions were evaluated (SI Appendix Fig. S12; please see SI Appendix C for the evaluation of directional interactions in the food web models).

Phytoplankton dynamics in a real ecosystem

We applied EcohNet to a time-series (Lake Kasumigaura Long-term Monitoring Dataset) to examine the top-down and bottom-up causal factors of phytoplankton community composition (Fig.4a). The causal network was naturally organized in a top-down structure with temperature at the top and temperature had the most numbers of interaction links across multiple trophic levels. EcohNet showed that each seven dominant phytoplankton groups were determined by different factors and those networks were complex. Four out of seven phytoplankton groups (all three cyanobacteria and Thalassiosiraceae) were forced by NO_3-N. We also detected the top-down control of rotifers and calanoids on not only diatoms but also cyanobacteria (Rotifers→Nitzschia, Rotifers→Oscillatoriales, Calanoida→Microcystis, Clanoida→Thalassiosiraceae, Cyclopoida→Aulacoseira), while large- and small-cladocerans did not influence dominant phytoplankton groups. In contrast, three bottom-up links from phytoplankton to zooplankton (Nitzschia→Rotifers, Oscillatoriales→Cyclopoida, Fragilaria→Small-cladocera) were also detected. Importantly, in addition to the effects of environmental variables and zooplankton, we identified three interactions among dominant phytoplankton groups (Nostocales →Oscillatoriales, Oscillatoriales→Aulacoseira, and Fragilaria→Microcystis).

An additional analysis on the direction of causation between meteorological and other components showed that the total number of unrealistic causalities was lowest in PCM, followed by EcohNet and CCM in this order (Fig. 4b). For the causal links with the meteorological variables as factors, EcohNet detected an equal number of links as PCM for temperature and precipitation and found more links for wind speed. Moreover, unrealistic causal links were likely to be detected as strong links in CCM and PCM, but appeared only as weak links in EcohNet (Fig. 4c).

Discussion

We developed and evaluated the effectiveness of a method, EcohNet, that utilizes ensemble predictions of online echo state networks. In short, our method encompasses the scope of Granger causality [3] and nonlinear time series analysis [35], and implements a framework for decomposing relationships among variables in terms of predictability [23]. Although ESNs have been applied to causal analysis [36-38], we developed for the first time a reliable method of causal network inference by integrating adaptive online ESNs [20] into an ensemble machine learning framework. In addition to its applicability to nonlinear dynamics, our approach is not affected by the reliability of nonlinear prediction methods to the dynamics of a given system [39], which differs from previous approaches [5, 26, 28, 34, 40, 41]. As we have confirmed (Fig. 2e-g), performance of CCM and PCM, which relies on a specific nonlinear forecasting method (state space reconstruction and simplex-projection) was sensitive to the differences in dynamic complexity [42, 43], although it performed better than the Spearman rank correlation coefficient. Another advantage of our method is its capability to decompose relationships among variables; we confirmed that false positives due to chain and fan-out relationships were best suppressed in EcohNet followed by PCM (Fig. 2e,f). This result is plausible because PCM also address both relationships [28].

EcohNet has two prominent features. First, it was robust in ability to detect links and has relatively stable sensitivity and specificity (Fig. 3). This robustness is important because it means that the error rate of an inferred network is consistently controlled by the regularization scheme irrespective of the dynamics. For example, comparison of the connectivity of ecological networks from several sites would be difficult if the method applied was sensitive to the dynamical complexity [16]. Second, we confirmed that EcohNet can identify causation between meteorological and other components while filtering out the unrealistic direction of causalities (Fig. 4b,c) [18]. The unrealistic causal links were likely to be detected only as weak links in EcohNet. Because of observational noise and other factors, it is hard to completely avoid detecting unrealistic causalities. Therefore, it is a desirable feature that the importance of unrealistic causality is underestimated. Among the rest of the methods, PCM had stable sensitivity and specificity and suppressed the detection of unrealistic links. On the other hand, its overall performance (Fig. 2g,h) was no better than CCM. One reason would be the absence of

steps for evaluating the convergence of a prediction skill in PCM whereas in CCM this helps to discount the predictability of the target variable itself. In EcohNet, it is considered as $\rho_{X \to X}$.

EcohNet revealed that temperature determined nutrient dynamics, phytoplankton, and zooplankton communities in Lake Kasumigaura. This result is similar to Tanentzap et al. [44] that demonstrates strong direct and indirect impacts of temperature on phytoplankton and zooplankton. Quantifying the causal effects of temperature is an important advantage of EcohNet. CCM failed to distinguish causal relationships from seasonality-driven synchronization, leading to misidentification of causality [45]. Likewise, previous work on Lake Kasumigaura did not detect the causal effects of temperature on phytoplankton and zooplankton [25]. Our results suggest that temperature may determine the whole food web structure of Lake Kasumigaura.

Clearly, each seven dominant phytoplankton groups were determined by different factors (Fig. 4). Some studies reported that phytoplankton species or genera have unique physiological and ecological features and thus respond differently to environmental factors and grazing [46, 47]. These differences can result in dynamic changes in phytoplankton community composition, since it is reported that the dominant phytoplankton group changed temporally in Lake Kasumigaura [48-50]. NO_3-N determined four out of seven phytoplankton groups. This is consistent with previous work on Lake Kasumigaura showing nitrogen limits phytoplankton primary production [25]. Not only diatoms but also cyanobacteria were influenced by rotifers, calanoids, and cyclopods. Although our analysis did not identify whether these are direct predation or indirect effects, it is known that some copepods can ingest and shorten the filament size of cyanobacteria and rotifers can graze or utilize decomposed cyanobacteria [51, 52]. Our results suggest that these zooplankton groups might have an important role in the food web of shallow hypereutrophic lakes.

We identified significant interactions among dominant phytoplankton groups. It has been reported that phytoplankton groups may have compensatory responses to environmental factors [53]. One possible mechanism for the interaction between Anabaena and Oscillatoriales could be nitrogen availability, because N_2-fixing cyanobacteria, including Nostocales, compete with non-N_2-fixing bacteria, including Oscillatoriales [54]. Another mechanism could be light availability. Mixing regimes determined by climatic factors like heat exchange and wind action affect competition for light between phytoplankton species (e.g., buoyant cyanobacteria and sinking diatoms), and shading by cyanobacteria blooms also influences other phytoplankton [55]. As pointed out in Freeman et al. [47], previous studies have not incorporated these interactions in models for identifying species-specific factors. Explicitly including not only environmental factors and grazing impacts but also community interactions in complex systems is another advantage of EcohNet.

As well as predicting the flow of causation, EcohNet enables cyanobacteria bloom forecasts for a lake (SI Appendix Fig. S13, 14). The application of EcohNet to long-term monitoring data can be a useful tool for lake management. Harmful cyanobacteria blooms are a serious threat to ecological integrity and ecosystem services, despite management efforts, and climate change is predicted to promote the occurrence and severity of cyanobacterial blooms [56]. As show here, the system is complex and therefore, identifying the drivers causing cyanobacteria blooms and forecasting these blooms using EcohNet are critically important for water quality management [29].

There are two developmental directions that we did not fully address in this paper. First, in this paper we did not consider the possibility of synergistic effects of two or more variables that may contribute more to the prediction than simple addition of their effect. Of course, we know this can occur [57] but we believe that a further extension of our methodology, e.g., evaluating the combinatorial effect of removing multiple variables, could address this issue. Second, we did not consider data-dependent parameter fitting of ESN-RLS. It may be true that optimal settings of parameters would maximize the performance of EcohNet's network inference. Guiding principles for such a setup would be to use autocorrelation to determine the forgetting coefficient λ [58], and to determine parameters to maximize the predictability of the time series [59]. However, we have demonstrated the benefits of EcohNet well enough without these further optimizations.

The development of frameworks to facilitate time-series-based causal analysis would advance the field of ecology in the coming decades, where advanced techniques for ecosystem monitoring [60-64] will lead to unprecedented increases in dataset size [7, 65, 66]. Such analysis could promote studies to estimate the effects of global environmental change on biological communities [27], prevent regime shifts that lead to catastrophic impacts on biodiversity [67], and support efficient management of biological resources [68]. Further development by applying EcohNet to time-series of high-frequency data, such as sensor data, could lead to real-time, near-term forecasting to prepare for or preempt future impairment of ecological functions and services. It is expected that with integration to active hypothesis testing, these studies will aid the development of predictive and manipulative ecology in the 21st century.

Materials and Methods

Echo state network with a recursive least squares algorithm (ESN-RLS)

An echo state network (ESN) [19, 21] implements a type of reservoir computing that uses a recurrent neural network (RNN) as a dynamical reservoir (an internal structure which is made up of individual, non-linear units, and can store information). ESNs can skillfully reconstruct and predict time series from different nonlinear dynamical systems [69-75]. It has been demonstrated that ESNs can track the temporal evolution of time series better than backpropagation-based artificial neural networks [71].

In an ESN, an input signal induces a nonlinear response to the dynamical reservoir RNN, and the reservoir states are converted to an output signal by linear weights (Fig. 1C). Here, for a multivariate time series $X = \{x_1, x_2, \dots, x_n\}$, we assume that $X^{\Omega}(t) = \{x_{\lambda t}\}_{\lambda \in \Omega}$ is the input at time t specified by a set of indices $\Omega \subset \{1,2, \dots, \mathrm{n}\}$, and $x_i = (x_{i1}, x_{i2}, \dots) \in X$ is a target variable to which an ESN is trained to output its one-step-ahead prediction. Following the typical implementation, we defined an ESN by three matrices and one function. The matrices are: 1) the input weight matrix, $W_{in} \in R^{n_{res} \times n_{in}}$; 2) the reservoir weight matrix, $W_{res} \in R^{n_{res} \times n_{res}}$ and 3) the output vector, $W_{out} \in R^{n_{res}}$, where n_{in}, n_{res} is the number of nodes in the input layer (number of elements in Ω), dynamical reservoir, respectively. The state update equations for an ESN are as follows:

$z(t) = f(W_{in} X^{\Omega}(t) + W_{res} z(t-1))$, (1)

$y(t) = W_{out}(t) z(t)$. (2)

Here, $z(t)$ is a vector of neural states in the dynamical reservoir, $y(t)$ is the output of an ESN as the prediction of $x_i(t+1)$, and $f(z) = tanh\, z$ is a neural activation function. The neural connections of an ESN are randomly generated, except for the output weights W_{out}. Here, W_{out} is sequentially updated by the RLS (recursive least squares) method. This defines an online implementation of ESN, namely, ESN-RLS, which is robust to the non-stationary dynamics and easily applicable to predictive purposes [20].

RLS is widely used in linear signal processing and has a desirable feature of fast convergence [76]. It incorporates the error history of a system into the calculation of the present error compensation. The recursive updating rule of RLS for W_{out} is as follows:

$v(t) = x_i(t) - y(t-1)$, (4)

$g(t) = \frac{P(t-1)z(t)}{\lambda + z(t)^T P(t-1) z(t)}$, (5)

$P(t) = \lambda^{-1}[P(t-1) - g(t) z^T(t) P(t-1)]$, (6)

$W_{out}(t) = W_{out}(t-1) + v(t) g^T(t)$. (7)

Here, λ is the forgetting factor, which determines the effect of past errors on the update, $v(t) \in R$ is the output error at time t when applying the weight matrix before the update, $g(t) \in R^{n_{res}}$ is the gain vector, and $P(t) \in R^{n_{res} \times n_{res}}$ is the inverse of the self-correlation matrix of $z(k)$ $(k = 1,2, \dots, t)$ weighted by the forgetting factor. This update can be repeated multiple times at each time step, and we define the number of iterations as τ. We set the initial conditions of the update rule as,

$P(0) = \delta^{-1} I$, (8)

$W_{out}(0) = O$, (9)

where I is a $n_{res} \times n_{res}$ identity matrix, O is a zero-vector of length n_{res} and δ is the regularization factor. We also set the initial reservoir state as $z(0) = O$.

This algorithm minimizes the following cost function, which is a weighted sum of the output errors at time $k = (1,2,\ldots,t)$ when applying the output weight matrix W_{out} at time t with a regularization term:

$E_{RLS}(t) = \frac{1}{2}\sum_{k=1}^{n} \lambda^{n-k} \|x_i(t) - W_{out}(t)z(k)\|^2 + \frac{\delta}{2}\lambda\|W_{out}\|^2$. (10)

Here, $\|\cdot\|$ represents the L_2 norm.

Parameterization of ESN-RLS

The parameter values were set according to the standard recommendation for parameterization of the ESN and RLS [20, 21 76-78] (Table 1, SI Appendix B). The elements of W_{in} are randomly drawn from a uniform distribution between -0.1 and 0.1. We scaled the spectral radius of W_{res} by ρ_0 after it is generated by randomly drawing its elements from a uniform distribution between -1 and 1 (with probability 0.1, and otherwise zero). The number of nodes in the dynamic reservoir n_{res} was 32, which is smaller than that of a typical ESN. This is intended to reduce computation time. Also, this number is related to the memory capacity of the reservoir [20], and the number of nodes must be scaled to the time series length. In this sense, a large number of nodes was not needed in this study. In our benchmark using aquatic microcosm data, we examined the stability of the results for four representative parameters ($\rho_0, \lambda, \tau, \delta$) that may affect the performance of ESNs (SI Appendix Fig. S16).

Progressive selection of input variables

For a target variable x_i (corresponds to X in Figure 1A), we used a progressive selection to obtain the set of variables Ω_{x_i} that minimizes the prediction skill for x_i ({X, Y} in Fig. 1B). For this purpose, we first define the prediction skill in this paper as:

$\rho_{\Omega_{x_i}} = \rho\left(x_i, y\left(X^{\Omega_{x_i}}|ESN_j\right)\right) = e^{-\left\|x_i - y\left(X^{\Omega_{x_i}}|ESN_j\right)\right\|}$. (11)

Here, x_i is the time series of target variable (blue lines in Fig.1C) and $y\left(X^{\Omega_{x_i}}|ESN_j\right)$ represents the prediction of x_i given a set of variables $X^{\Omega_{x_i}}$ and an echo state network indexed by j (yellow lines in Fig.1C). Practically, we calculated the prediction skill after truncating the first 20% of the time series. It is intended to remove the effect of initial condition ($z(0)$, $P(0)$ and $W_{out}(0)$).

To prevent the results from being dependent on a particular network, the evaluation of prediction skill is done by an ensemble of N ESNs. This allows us to account for the estimation error in prediction skill and ensures more reliable variable selection. Moreover, to avoid overfitting, the input is replaced by a zero vector with a fixed probability at each time. This means that we eventually use the following equation to update the reservoir state instead of eq. (1):

$z(t) = f\left(W_{in}\eta X^{\Omega}(t) + W_{res}z(t-1)\right)$, (12)

where η is a random number that is 1 with probability 0.5 and 0 otherwise.

The progressive selection of variables proceeds as follows. First, the set of indices of input variables to predict variable x_i is initialized to $I_{active} = \{i\}$ to account for the predictability inherent in the target variable itself (first step in Fig.1B). Correspondingly, the set of indices of the rest of the variables is initialized to $I_{inactive} = \{j\}_{j \neq i}$. Then, N ESNs are generated to perform predictions of x_i by setting $\Omega_{x_i} = I_{active}$ (Fig.1C). This returns a distribution of prediction skills $F_{T=0}$ (T indicates the step of variable selection). To select the next variable to be added, for each index $j^{(h)}$ in $I_{inactive}$, $I_{test}^{(h)} = I_{active} \cap \{j^{(h)}\}$ are generated, where the suffix h indicates that $j^{(h)}$ is the hth element of $I_{inactive}$, and then, for each $I_{test}^{(h)}$, N ESNs are generated to obtain predictions of x_i by setting $\Omega_{x_i} = I_{test}^{(h)}$ (2nd step in Fig.1B). Among the distribution of the prediction skills for hs, one

that has the highest median value ($h = h^*$) is set as F_{T+1} (second step in Fig.1B). The criterion used to accept F_{T+1} as $\Omega_{x_i} = I_{test}^{(h^*)}$ is that the median of F_{T+1} is larger than $q \times 100\%$ of values in F_T, formally written as, $\#\{\rho \in F_T | \rho > median(F_{T+1})\}/N < q$, where $\#\{\rho \in F | criteria\}$ is the number of ρ in F that satisfies the *criteria*. If the criterion is satisfied, we replace F_T by F_{T+1}, I_{active} by $I_{test}^{(h^*)}$, $I_{inactive}$ by $I_{inactive} \backslash \{j^{(h^*)}\}$ (removing $j^{(h^*)}$ from $I_{inactive}$), increment T by 1, and proceed to the next step only if $I_{inactive} \neq \{\}$, otherwise the procedure is stopped (3rd step in Fig.1B) with returning $\bar{\Omega}_{x_i} = I_{active}$ as the set of the indices of optimal variables that maximizes the prediction of x_i.

The number of ESNs generated for each evaluation, N, is related to the stability of the results. If N is small, the variability in results from trial to trial may be large, especially for weak causality. In this paper, we set $N = 10000$ for real-world data to obtain stable results, and $N = 5000$ for simulation data sets to reduce the time required for evaluation. The threshold value for accepting a new variable, q, was set to 0.48 throughout this paper. The closer the number is to 0.5, the more likely it is that weak links will remain. Therefore, instead of increasing the sensitivity (true positive rate), there is a possibility of decreasing the specificity (increasing the false positive rate). In this paper, the value is set close to 0.5, taking into account the evaluation by ROC-AUC.

Calculation of unique prediction skill

Unique prediction skill quantifies the unique contribution of a single variable to the overall prediction skill (Fig. 1D). For each $j \in \bar{\Omega}_{x_i}$, the unique prediction skill is defined as,

$\rho_{x_j \to x_i} = median\left(F_{\bar{\Omega}_{x_i}}\right) - median(F_{\bar{\Omega}_{x_i}^{(-j)}})$. (12)

Here, F_Ω is the distribution of prediction skill for Ω and,

$\bar{\Omega}_{x_i}^{(-j)} = \bar{\Omega}_{x_i} \backslash \{j\}$, (13)

i.e., $\bar{\Omega}_{x_i}^{(-j)}$ is obtained by removing j from $\bar{\Omega}_{x_i}$. $F_{\bar{\Omega}_{x_i}}$ and $F_{\bar{\Omega}_{x_i}^{(-j)}}$ is again calculated by N ESNs. $\rho_{x_j \to x_i}$ represents how x_j uniquely contributes to the prediction of x_j (illustrated as the difference of the positions of two distributions in Fig.1D). We identify $\rho_{x_j \to x_i}$ as the indicator of causal influence from x_j to x_i. In this step, $\bar{\Omega}_{x_i}^{(-j)}$ such that $\rho_{x_j \to x_i} \leq 0$ is removed from $\bar{\Omega}_{x_i}$, and we set $\rho_{x_j \to x_i} = 0$ as well as other variables that does not included in $\bar{\Omega}_{x_i}$.

Dataset

We benchmarked our method using both experimental and simulation data (SI Appendix D) and then applied it to a long-term monitoring dataset from a lake ecosystem. One of the benchmarking datasets was from a real ecological system, which was obtained by long-term observation of a mesocosm [33]. The other datasets were obtained by simulations of two ecological models, namely food-web and random interaction model (all simulation conditions are listed in SI Appendix Table S1), and allowed us to examine how dynamic complexity of a time series affect identification of causal relationships, and how the effects of two types of interaction topologies (chain and fan-out relationships) can cause false positives.

Lake Kasumigaura Long-term Monitoring Dataset

We applied the EcohNet to the long-term monitoring data from Lake Kasumigaura, a hypereutrophic lake. Lake Kasumigaura is the second largest lake in Japan (167.7 km^2) and is

shallow (mean depth: approximately 4 m, maximum depth: 7.4 m). The National Institute for Environmental Studies has been conducting monthly monitoring in Lake Kasumigaura since 1976 and publishing the long-term data on the Lake Kasumigaura Database. In this lake, cyanobacteria blooms occur and disappear repeatedly, and the dominant cyanobacteria group changes. Phytoplankton communities can interact with many components, such as temperature, nutrients, and zooplankton, and are complex and dynamic. Using EcohNet, we quantified the causal relationships among environmental variables, seven dominant phytoplankton groups (three cyanobacteria (Microcystis, Nostocales, Oscillatoriales) four diatoms (Thalassiosiraceae, Aulacoseira, Fragilaria, Nitzschia)), and zooplankton, and examined the interactions among these phytoplankton groups.

We analyzed the monitoring data at the center of Takahamairi Bay (Station 3), which is shallow (ca. 3.2 m) and the most eutrophic site with cyanobacteria blooms. Phytoplankton biovolume data was obtained from Takamura and Nakagawa [79]. For zooplankton, we used the abundance data of five functional groups: large cladocerans (>1.0 mm), small cladocerans (<1.0 mm), rotifers, adult calanoids, and adult cyclopoids [25]. Zooplankton data were obtained from Takamura et al. [80]. Our key environmental variables were surface water temperature, soluble reactive phosphorus (PO_4-P), and nitrate nitrogen (NO_3-N) from the Lake Kasumigaura Database (https://db.cger.nies.go.jp/gem/moni-e/inter/GEMS/database/kasumi/index.html). We also included 30-day moving average of wind speed and precipitation, which were collected from the Tsukuba-Tateno Meteorological Station of the Japan Meteorological Agency (https://www.jma.go.jp/jma/menu/menureport.html). Since CCM, PCM and LIMITS requires taking consecutive time lags of observed variables, we analyzed data from April 1996 to March 2019 (276 months). There were no missing data for any variables during this time interval. All time series were square-root-transformed and normalized to have a mean 0 and variance of 1 to adjust for rapid increases in some phytoplankton species.

Evaluation

We evaluated the performance of EcohNet and conventional methods (SI Appendix A) in detecting interactions using the evaluation criteria of a binary classification task, namely, ROC-AUC (the area under the curve of a receiver-operator characteristic curve), F1 score, accuracy, sensitivity and specificity (SI Appendix Fig. S3). Since accuracy is affected by the degree of connctance, we considered ROC-AUC and F1-score as criteria for overall performance. We specifically considered prediction skill (EcohNet, CCM and PCM), correlation coefficient (Spearman rank correlation) and interaction coefficient (LIMITS) as a classifier to identify interactions. We adjusted the sparsity of the Spearman rank correlation matrix based on the p-value, i.e., we made the correlation matrix sparse by replacing elements with $p > 0.01$ with 0. In the same manner, we set the threshold p-value of CCM as 0.05 (SI Appendix A2). For PCM, following the author implementation [28], we used the value of prediction skill instead of p-value and set the threshold as 0.2. For LIMITS, as in EcohNet, the sparsity of an interaction matrix depends on the forward stepwise algorithm. Here, the threshold value for the improvement of prediction error (SI Appendix A4) was set as 0, i.e., new variable is added if it at least improves the prediction.

In predator-prey relationships species affect each other directly, whereas in a causal relationship, a relationship between species may only be detected in one direction due to a difference in the time-dependence of each effect [5, 24, 25]. Therefore, in this paper, when evaluating EcohNet, CCM and PCM for the long-term mesocosm experiment and the food web models using the evaluation criteria of the binary classification task, we consider the ability to detect interacting pairs (pairs where at least one has a direct influence on the other). Specifically, we symmetrized the matrices of the prediction skill of EcohNet, CCM and PCM so that $a_{ij}, a_{ji} \leftarrow max(|a_{ij}|, |a_{ij}|)$ before evaluation, and evaluated the performance of prediction skill as a classifier in detecting elements for which $i < j$ and whose values are non-zero in the actual interaction matrix. The matrix of Spearman rank correlation, which is originally symmetric, was evaluated in the same manner. For the random interaction model, we considered the ability of EcohNet, CCM

and PCM to detect direct influences. Specifically, we tested the presence of influence from one species to another (corresponding to a non-zero a_{ij} in the interaction matrix) when there are causal links from the former to the later. For LIMITS, we evaluated the performance of detecting non-zero components of the interaction matrix directly in all cases according to its original definition.

Data availability

We used Mathematica 12.3 and 13.1 for our analysis. The computer codes used for the analysis can be downloaded from: https://github.com/kecosz/EcohNet (DOI: 10.5281/zenodo.5492161)

Acknowledgments

We especially thank Shinji Nakaoka for valuable discussions, and Megumi Nakagawa for counting phytoplankton and providing helpful comments. We also thank all members of the Lake Kasumigaura long-term monitoring group of the National Institute for Environmental Studies. This work was supported by funding from the Management Expenses Grant for RIKEN BioResource Research Center, MEXT, and the Japan Society for the Promotion of Science (JSPS) KAKENHI JP20K06820 and JP20H03010 (to K.S.).

References

1. J. Runge, Causal network reconstruction from time series: From theoretical assumptions to practical estimation. Chaos 28, 075310 (2018).
2. J. Runge et al., Inferring causation from time series in Earth system sciences. Nat. Commun. 10, 1-13 (2019).
3. C. W. Granger, Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424-438 (1969).
4. H. Akoglu, User's guide to correlation coefficients. Turkish journal of emergency medicine 18, 91-93 (2018).
5. G. Sugihara, et al. Detecting causality in complex ecosystems. Science 338, 496-500 (2012).
6. J. Runge et al., Detecting and quantifying causal associations in large nonlinear time series datasets. Sci. Adv. 5, eaau4996 (2019).
7. J. M. Durden, Integrating "big data" into aquatic ecology: Challenges and opportunities. Limnol. Oceanogr. Bull. 26, 101-108 (2017).
8. G. K. Vallis, Atmospheric and oceanic fluid dynamics. (Cambridge University Press, 2017).
9. M. Kretschmer, D. Coumou, J. F. Donges, J. Runge, Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. J. Clim. 29, 4069–4081 (2016).
10. C. T. Perretti, S. B. Munch, G. Sugihara, Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. Proc. Natl. Acad. Sci. 110, 5253-5257 (2013).
11. E. R. Deyle, R. M. May, S. B. Munch, G. Sugihara, Tracking and forecasting ecosystem interactions in real time. Proc. R. Soc. London, Ser. B, 283, 20152258 (2016).

12. J. A. McGowan et al. Predicting coastal algal blooms in southern California. Ecology, 98, 1419-1433 (2017).
13. W. H. Schlesinger, J. J. Cole, A. C. Finzi, E. A. Holland, Introduction to coupled biogeochemical cycles. Front. Ecol. Environ. 9, 5-8 (2011).
14. A. A. Tsonis et al., Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. Proc. Natl. Acad. Sci., 112, 3253-3256 (2015).
15. E. R. Deyle, Global environmental drivers of influenza. Proc. Natl. Acad. Sci. 113, 13081-13086 (2016).
16. T. L. Rogers, S. B. Munch, Hidden similarities in the dynamics of a weakly synchronous marine metapopulation. Proc. Natl. Acad. Sci. 117, 479-485 (2020).
17. C. W. Chang et al. Causal networks of phytoplankton diversity and biomass are modulated by environmental context. Nat. Commun. 13, 1-11 (2022).
18. E. B. Baskerville, S. Cobey, Does influenza drive absolute humidity? Proc. Natl. Acad. Sci. 114, E2270-E2271 (2017).
19. H. Jaeger, The “echo state” approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report 148, 13 (2001).
20. H. Jaeger, Adaptive nonlinear system identification with echo state networks. Adv. Neural. Inf. Process. Syst. 15, 609-616 (2002).
21. H. Jaeger, H. Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Science 304, 78-80 (2004).
22. E. Bollt, On explaining the surprising success of reservoir computing forecaster of chaos? The universal machine learning dynamical system with contrast to VAR and DMD. Chaos 31, 013108 (2021).
23. F. Pennekamp et al. The intrinsic predictability of ecological time series and its potential to guide forecasting. Ecol. Monogr. 89, e01359 (2019).
24. H. Ye, E. R. Deyle, L. J. Gilarranz, G. Sugihara, Distinguishing time-delayed causal interactions using convergent cross mapping. Sci. Rep. 5, 1-9 (2015).
25. S. I. S. Matsuzaki et al. Bottom-up linkages between primary production, zooplankton, and fish in a shallow, hypereutrophic lake. Ecology 99, 2025-2036 (2018).
26. M. Ushio et al., Fluctuating interaction network and time-varying stability of a natural fish community. Nature, 554, 360-363 (2018).
27. C. W. Chang et al., Long-term warming destabilizes aquatic ecosystems through weakening biodiversity‐mediated causal networks. Global Change Biol. 26, 6413-6423 (2020).
28. S. Leng et al., Partial cross mapping eliminates indirect causal influences. Nat. Commun. 11, 1-9 (2020).
29. C. C. Carey et al., Advancing lake and reservoir water quality management with near-term, iterative ecological forecasting. Inland Waters 12, 1-14 (2021).
30. K. Nakajima, I. Fischer, Reservoir Computing. (Springer, 2021).
31. J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning (Vol. 1, No. 10). (Springer New York, 2001).
32. C. K. Fisher, P. Mehta, Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PloS One 9, e102451 (2014).
33. E. Benincà et al., Chaos in a long-term experiment with a plankton community. Nature 451, 822-825 (2008).
34. K. Suzuki, K. Yoshida, Y. Nakanishi, S. Fukuda, An equation-free method reveals the ecological interaction networks within complex microbial ecosystems. Methods Ecol. Evol. 8, 1774-1785 (2017).
35. H. Kantz, T. Schreiber, Nonlinear time series analysis (Vol. 7). (Cambridge university press, 2004).
36. Y. Huang, Z. Fu, C. L. Franzke, Detecting causality from time series in a machine learning framework. Chaos 30, 063116 (2020).

37. A. Duggento, M. Guerrisi, N. Toschi, Echo state network models for nonlinear granger causality. Philos. Trans. R. Soc. London, Ser. A, 379, 20200256 (2021).
38. M. Wang, Z. Fu, A new method of nonlinear causality detection: Reservoir computing Granger causality. Chaos, Solitons Fractals 154, 111675 (2022).
39. K. Judd, A. Mees, On selecting models for nonlinear time series. Physica D 82, 426-444 (1995).
40. S. Cenci, G. Sugihara, S. Saavedra, Regularized S-map for inference and forecasting with noisy ecological time series. Methods Ecol. Evol. 10, 650-660 (2019).
41. C. W. Chang, et al. Reconstructing large interaction networks from empirical time series data. Ecol. Lett. 24, 2763-2774 (2021).
42. S. Cobey, E. B. Baskerville, Limits to causal inference with state-space reconstruction for infectious disease. PloS One 11, e0169050 (2016).
43. D. Mønster et al., Causal inference from noisy time-series data—testing the convergent cross-mapping algorithm in the presence of noise and external influence. Future Gener. Comput. Syst. 73, 52-62 (2017).
44. A. J. Tanentzap, Climate warming restructures an aquatic food web over 28 years. Global Change Biol. 26, 6852-6866 (2020).
45. K. Kitayama, M. Ushio, S. I. Aiba, Temperature is a dominant driver of distinct annual seasonality of leaf litter production of equatorial tropical rain forests. J. Ecol. 109, 727-736 (2021).
46. V. Huber, C. Wagner, D. Gerten, R. Adrian, To bloom or not to bloom: contrasting responses of cyanobacteria to recent heat waves explained by critical thresholds of abiotic drivers. Oecologia 169, 245-256 (2012).
47. E. C. Freeman, I. F. Creed, B. Jones, A. K. Bergström, Global changes may be promoting a rise in select cyanobacteria in nutrient-poor northern lakes. Global Change Biol. 26, 4966-4987 (2020).
48. N. Takamura, A. Otsuki, M. Aizaki, Y. Nojiri, Phytoplankton species shift accompanied by transition from nitrogen dependence to phosphorus dependence of primary production in Lake Kasumigaura, Japan. Archiv für Hydrobiologie 124, 129-148 (1992).
49. N. Tomioka, A. Imai, K. Komatsu, Effect of light availability on Microcystis aeruginosa blooms in shallow hypereutrophic Lake Kasumigaura. J. Plankton Res. 33, 1263-1273 (2011).
50. T. Fukushima, H. Arai, Regime shifts observed in Lake Kasumigaura, a large shallow lake in Japan: Analysis of a 40-year limnological record. Lakes Reservoirs Res. Manage. 20, 54-68 (2015).
51. S. Kâ, et al., Can tropical freshwater zooplankton graze efficiently on cyanobacteria?. Hydrobiologia, 679, 119-138 (2012).
52. K. A. Ger, et al. The interaction between cyanobacteria and zooplankton in a more eutrophic world. Harmful Algae 54, 128-144 (2016).
53. M. C. Jochimsen, R. Kümmerlin, D. Straile, Compensatory dynamics and the stability of phytoplankton biomass during four decades of eutrophication and oligotrophication. Ecol. Lett. 16, 81-89 (2013).
54. P. H. Moisander, H. W. Paerl, J. P. Zehr, Effects of inorganic nitrogen on taxa-specific cyanobacterial growth and nifH expression in a subtropical estuary. Limnol. Oceanogr. 53, 2519-2532 (2008).
55. J. Huisman et al. Changes in turbulent mixing shift competition for light between phytoplankton species. Ecology 85, 2960-2970 (2004).
56. J. Huisman et al. Cyanobacterial blooms. Nat. Rev. Microbiol. 16, 471-483 (2018).
57. P. L. Williams, R. D. Beer, Nonnegative decomposition of multivariate information. [Preprint] (2010). https://arxiv.org/abs/1004.2515 (accessed 6 August 2022).
58. S. H. Leung, C. F. So. Gradient-based variable forgetting factor RLS algorithm in time-varying environments. IEEE Transactions on Signal Processing 53, 3141-3150 (2005).

59. L. Wang, Z. Wang, S. Liu, An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm. Expert Systems with Applications 43, 237-249 (2016).
60. J. G. Caporaso et al. QIIME allows analysis of high-throughput community sequencing data. Nat. methods 7, 335-336 (2010).
61. D. A. Bohan et al., Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. Trends Ecol. Evol. 32, 477-487 (2017).
62. M. Balint et al., Environmental DNA time series in ecology. Trends Ecol. Evol. 33, 945-957 (2018).
63. P. Haase et al. The next generation of site-based long-term ecological monitoring: Linking essential biodiversity variables and ecosystem integrity. Sci. Total Environ. 613, 1376-1384 (2018).
64. P. Taberlet, A. Bonin, L. Zinger, E. Coissac, Environmental DNA: For biodiversity research and monitoring. (Oxford University Press, 2018).
65. S. E. Hampton et al., Big data and the future of ecology. Front. Ecol. Environ. 11, 156-162 (2013).
66. N. Mouquet et al., Predictive ecology in a changing world. J. Appl. Ecol. 52, 1293-1310 (2015).
67. M. Scheffer, S. R. Carpenter, Catastrophic regime shifts in ecosystems: linking theory to observation. Trends Ecol. Evol. 18, 648-656 (2003).
68. E. Chivian, A. Bernstein, Eds., Sustaining life: how human health depends on biodiversity. (Oxford University Press, 2008).
69. J. Pathak et al., Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. Chaos 27, 121102 (2017).
70. C. Du et al., Reservoir computing using dynamic memristors for temporal information processing, Nat. Commun. 8, 2204 (2017).
71. Z. Lu, et al., Reservoir observers: Model-free inference of unmeasured variables in chaotic systems, Chaos 27, 041102 (2017).
72. Z. Lu, B. R. Hunt, E. Ott, Attractor reconstruction by machine learning, Chaos 28, 061104 (2018) .
73. J. Pathak et al., Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach, Phys. Rev. Lett. 120, 024102 (2018).
74. A. Chattopadhyay, P. Hassanzadeh, K. Palem, D. Subramanian, Data-driven prediction of a multi-scale Lorenz 96 chaotic system using a hierarchy of deep learning methods reservoir computing, ANN, and RNN-LSTM. [Preprint] (2019). https://arxiv.org/abs/1906.08829 (accessed 6 August 2022)
75. M. Reichstein et al., Deep learning and process understanding for data-driven earth system science, Nature 566, 195 (2019).
76. B. Farhang-Boroujeny, Adaptive filters: theory and applications. (John Wiley & Sons, 2013).
77. E. C. Ifeachor, B. W. Jervis, Digital signal processing: a practical approach. (Pearson Education, 2002).
78. M. Lukoševičius, A practical guide to applying echo state networks. In Neural networks: Tricks of the trade. (Springer Berlin, 2012), pp. 659-686.
79. N. Takamura, M. Nakagawa, Phytoplankton species abundance in Lake Kasumigaura (Japan) monitored monthly or biweekly since 1978. Ecol. Res. 27, 837-837 (2012).
80. N. Takamura, M. Nakagawa, T. Hanazato, Zooplankton abundance in the pelagic region of Lake Kasumigaura (Japan): monthly data since 1980. Ecol. Res. 32, 1-1 (2017).

Figures and Tables

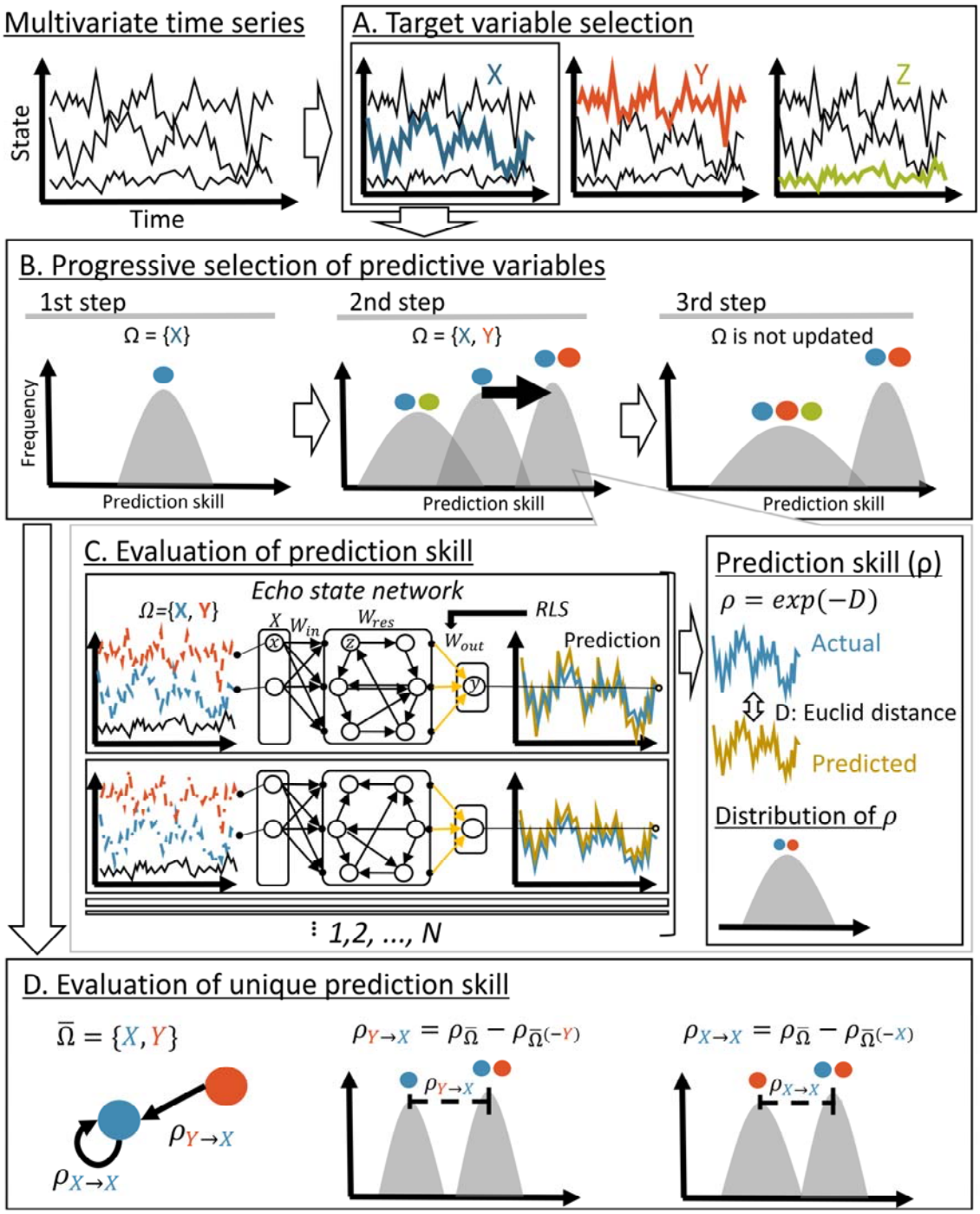

Figure 1. Graphical explanation of EcohNet. See main body and Materials and Methods for the explanation.

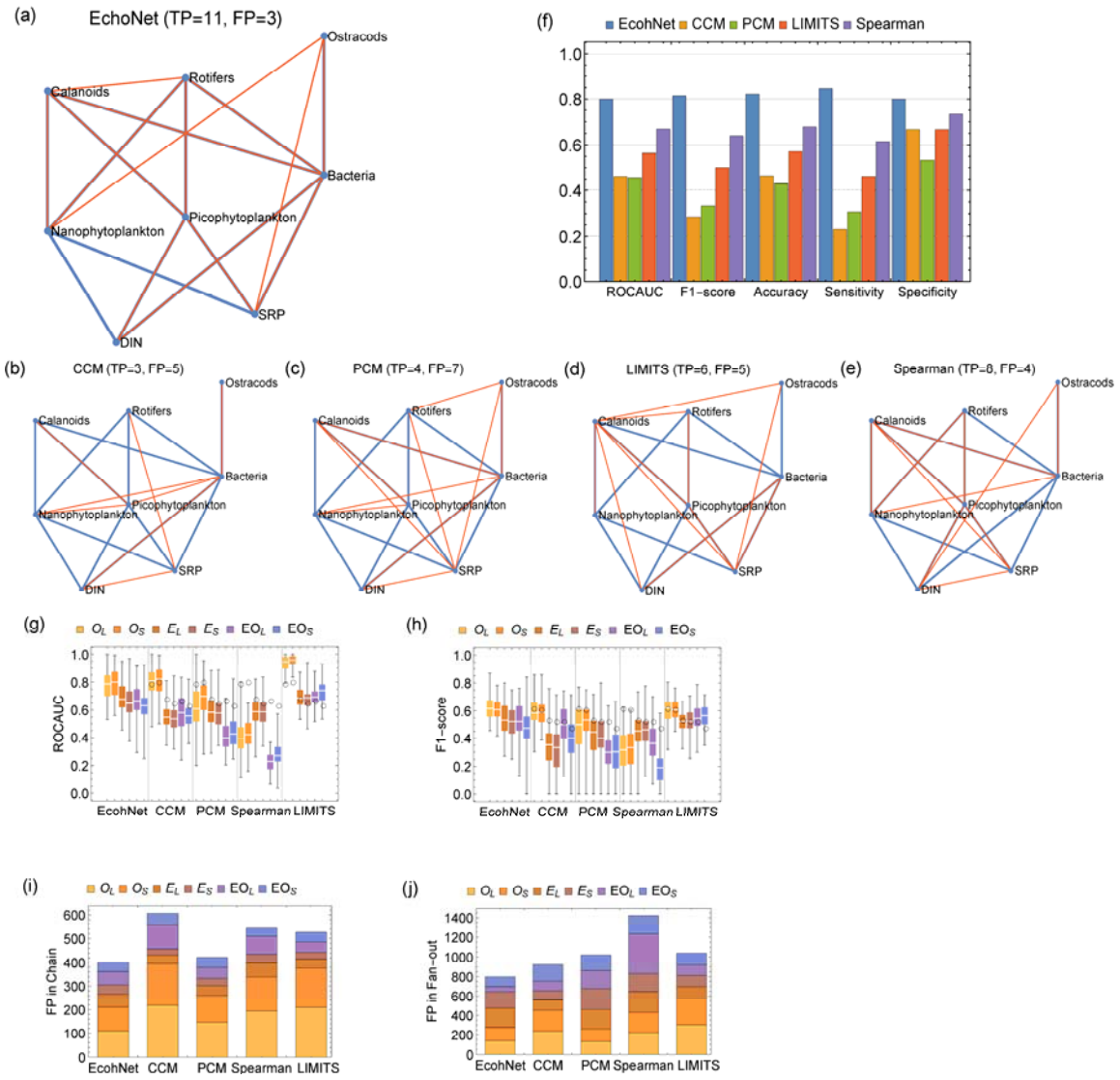

Figure 2. Benchmarking results. (a-e) Actual (blue) and inferred (red) network of EcohNet and conventional approaches (CCM, PCM, Spearman rank correlation and LIMITS) for mesocosm data (DIN and SRP stands for total dissolved inorganic nitrogen and soluble reactive phosphorus, respectively). The number of true positives (TP) and false positives (FP) are shown above the panel. (f)The value of evaluation criteria for mesocosm data. (g, h) Two performance criteria (ROC-AUC and F1-Score) for the food web models. Here, O is intrinsic oscillation, E is equilibrium and EO is equilibrium forced by an external oscillator and the suffixes S and L indicate small and large noise magnitude, respectively (see SI Appendix D2 for simulation parameter values). In the box plot, white lines indicate the median, box edges indicate the first and third quartile value, and whiskers indicate maximum and minimum values. Black circles indicate the median value of EcohNet. In (g) and (h), the result of LIMITS should be carefully interpreted because it assumes Lotka-Volterra (LV) equation used to generate the benchmark data. (i) Cumulative number of incorrectly identified links in chain relationships (for 6×100=600 time series in total). (j) Cumulative number of incorrectly identified links in fan-out relationships (for 6×100=600 time series in total).

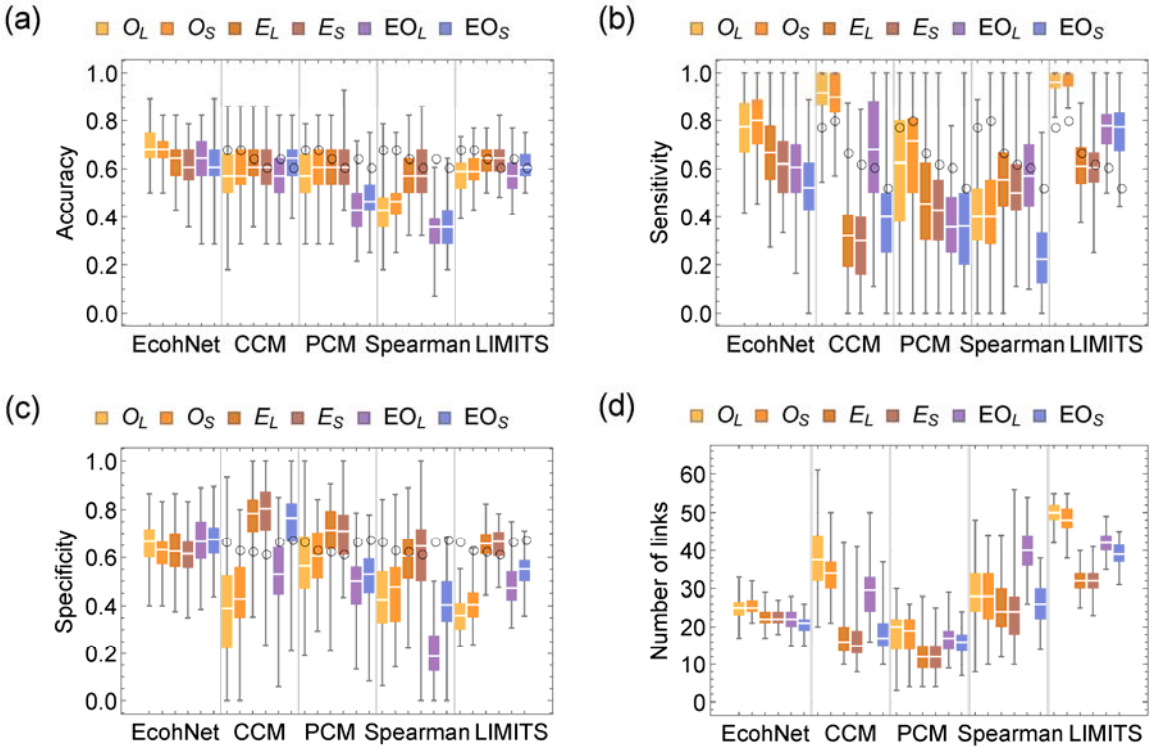

Figure 3. Values of accuracy (a), sensitivity (b), specificity (c) and number of inferred links (d) for the food web models. Since connectance was fixed at 0.33 when generating an interaction matrix, the number of links detected should ideally be a constant value (since the number of species was fixed at 8, average number of links was (8^2-8)*0.33=18). In the box plot, white lines indicate the median, box edges indicate the first and third quartile value, and whiskers indicate maximum and minimum values. Black circles indicate the median value of EcohNet.

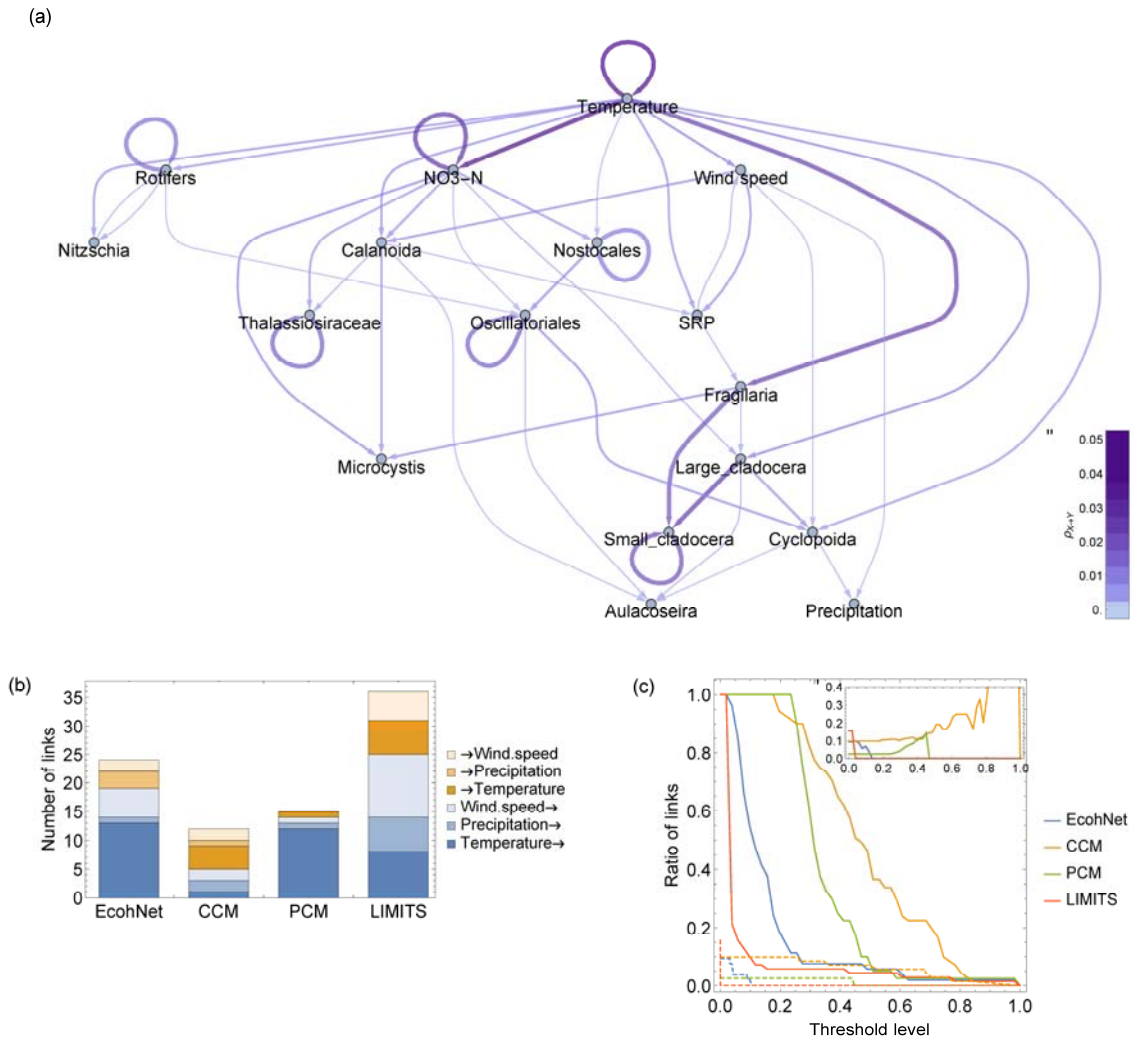

Figure 4. Analysis of Lake Kasumigaura Long-term Monitoring Dataset. (a) Causal network obtained by EcohNet. We removed causal links with a prediction skill of less than 0.002 (23 out of 70 links were removed). The magnitude of the prediction skill is indicated both by the thickness and color of arrows. Seven dominant phytoplankton groups were underlined. (b) We identified causal relationships including meteorological variables (Wind speed, Precipitation, Temperature) and examined how well unrealistic causal links (i.e., organisms/chemicals to meteorological variables) were avoided while the effect of meteorological variables were detected. (c) Ratio of causal links larger than the threshold level (x-axis) is shown for all links (solid lines) and unrealistic links (dashed lines). The inserted panel shows the ratio of the unrealistic links to all links. The faster the unrealistic/all ratio converge to zero, the smaller the unique prediction skill of unrealistic causal links, and the easier it is to remove them while retaining more reliable links.

Table 1. Parameters of ESN-RLS. For the parameters marked by *, we tested the impact on EcohNet's performance with different values. The tested values are $\tau \in \{2,4,6,8,10\}$, $\lambda \in \{0.8,0.9,0.95,0.98,0.995\}$, $\delta \in \{0.0001,0.001,0.01,0.1\}$, $\rho_0 \in \{0.7,0.8,0.9,0.95,0.98\}$.

Description	Value
Connectance of W_{in}	1
Range of values in W_{in}	[-0.1,0.1]
Number of nodes in W_{res} (n_{res})	32
Connectance of W_{res}	0.1
Range of values in W_{res}	[-1,1]
Spectral radius of W_{res} (ρ_0)	0.95*
Forgetting factor (λ)	0.95*
Number of iterations of RLS updates (τ)	8*
Regularization factor (δ)	0.001*