Reconstructing coupled time series in climate systems using three kinds of machine learning methods

Yu Huang[1], Lichao Yang[1], Zuntao Fu[1*]

[1]Lab for Climate and Ocean-Atmosphere Studies, Dept. of Atmospheric and Oceanic Sciences, School of Physics, Peking University, Beijing, 100871, China

Correspondence to: Zuntao Fu (fuzt@pku.edu.cn)

Abstract.

Despite the great success of machine learning, its applications in climate dynamics have not been well developed. One concern might be how well the trained neural networks could learn a dynamical system and what can be the potential applications of this kind of learning. In this paper, three machine learning methods are used: reservoir computer (RC), back propagation based artificial neural network (BP), and long short-term memory neural network (LSTM). It is shown that the coupling relations or dynamics among variables in linear or nonlinear systems can be well learnt by RC and LSTM, which can be further applied to reconstruct one time series from the other dominated by the common coupling dynamic. Specifically, we analyze the climatic toy models to address two questions: (i) what factors significantly influence machine learning reconstruction; and (ii) how to select suitable explanatory variables for machine learning reconstruction. The results reveal that both linear and nonlinear coupling relations between variables do influence the reconstruction quality of machine learning. If there is a strong linear coupling between two variables, the reconstruction can be bi-directional, where any one of these two variables is able to be an explanatory variable for reconstructing the other variable. When the linear coupling among variables is absent, but with the significant nonlinear coupling, the machine learning reconstruction between two variables is direction-dependent and it may be only uni-directional. Then we propose using the convergent cross mapping (CCM) causality index to determine which variable can be taken as the reconstructed one and which can be taken as the explanatory variable. In a real-world example, the Pearson correlation between the average Tropical Surface Air Temperature (TSAT) and the average Northern Hemispheric SAT (NHSAT) is as weak as 0.08, but the CCM index of NHSAT cross maps TSAT is 0.70, it means that NHSAT could be taken as the explanatory variable. Then we find that TSAT can be well reconstructed from NHSAT by the machine learning method. However, the reconstruction quality in the opposite direction is poor, where the CCM index of TSAT cross maps NHSAT is only 0.24. These results also provide insights on machine learning approaches for paleoclimate reconstruction, parameterization scheme, and prediction in related climate research.

Key words: Reconstruction, Climate time series, Machine learning, Causality, Surface air temperature

Highlights:

i) Learnt coupling dynamics between series by machine learning can be used to reconstruct series.

ii) Reconstruction quality is direction- and variable-dependent for nonlinear systems.

iii) The CCM index is a potential indicator to choose reconstructed and explanatory variables.

iv) The tropical average SAT can be well reconstructed from the average Northern Hemispheric SAT.

1 Introduction

Neural network-based machine learning provides effective tools for studying climatic data (Reichstein et al., 2019), which attracts great attention recently. The machine learning approach is widely applied to downscaling and data mining analyses (Mattingly et al., 2016; Racah et al., 2017), and it can be also used to predict the time series of climate variables, such as temperature, humidity, runoff and air pollution (Zaytar and Amrani, 2016; Biancofiore et al., 2017; Kratzert et al., 2019; Feng et al., 2019). Recently, it is demonstrated that a large potential application of machine learning is to reconstruct the temporal dynamics of complex systems (Pathak et al., 2017; Du et al., 2017; Watson, 2019). Studies (Pathak et al., 2017; Lu et al., 2018; Carroll, 2018) have shown that the chaotic attractors in Lorenz system and Rossler system can be described by machine learning. Since chaos is the key property of the underlying climate system giving rise to climatic time series (Lorenz, 1963; Patil et al., 2001), these studies provide a theoretical explanation why the machine learning can be well applied in reconstructing climate temporal dynamics.

Though applying machine learning to climatic series attracts much attention, it is still open questions what can be learnt by machine learning during the training process, and what is the key factor determining the performance of machine learning approach to climatic time series. This is crucial for investigating why machine learning cannot perform well with some datasets, and how to improve the performance for them. One possible key factor is the coupling between different variables. Because different climate variables are coupled with one another (Donner and Large, 2008), and the coupled variables will share their information content with one another through the information transfer (Takens, 1981; Schreiber, 2000; Sugihara et al., 2012). Furthermore, a coupling often results in that the observational time series are statistically correlated (Brown, 1994).

Correlation is a crucial property for the climate system, and often influences the climatic time series analysis. "Pearson Coefficient" is often used to detect the correlation, which only detects the linear correlation. It is known that when the Pearson correlation coefficient is weak, most of traditional regression methods will fail in dealing with the climatic data, such as fitting, reconstruction and prediction (Brown, 1994; Sugihara et al., 2012; Emile-Geay and Tingley, 2016). However, a weak linear correlation does not mean that there is no coupling relation between the variables. Previous studies (Sugihara et al., 2012; Emile-Geay and Tingley, 2016) have suggested that, although the linear correlation of two variables is potentially absent, they might be nonlinearly coupled and can be exploited by analysis. For instance, the linear cross-correlations of sea surface temperature series observed in different tropical areas are unstable and vary with time, which leads to an overall weak linear correlation, but this non-linear correlation is conductive to the better El Niño predictions (Ludescher et al., 2014; Conti et al., 2017). The linear correlations between ENSO/PDO index and some proxy variables are weak but their nonlinear coupling relations can be detected, which contributes greatly to reconstructing longer paleoclimate time series (Mukhin et al., 2018). These studies indicate that nonlinear coupling relations would contribute to the better analysis, reconstruction, and prediction (Hsieh et al., 2006; Donner, 2012; Schurer et al., 2013; Badin et al., 2014; Drótos et al., 2015; Van Nes et al., 2015; Comeau et al., 2017; Vannitsem and Ekelmans, 2018). Accordingly, when applying machine learning to climatic series, is it necessary to give attention to the linear or nonlinear relationships induced by the physical couplings? This is worth to be addressed.

In a recent study (Lu et al., 2017), a machine learning method called reservoir computer was used to reconstruct the unmeasured time series in the Lorenz 63 model (Lorenz, 1963). It is found that the Z variable can be well reconstructed from the X variable by reservoir computer, but it failed to reconstruct X with Z. Lu et al. (Lu et al., 2017) demonstrated that the nonlinear coupling dynamic between X and Z was responsible for this asymmetry in the reconstruction. This was explained by the nonlinear observability in control theory (Hermann and Krener, 1977; Lu et al., 2017): for the Lorenz 63 equation, both ($X(t)$, $Y(t)$, $Z(t)$) and ($-X(t)$, $-Y(t)$, $Z(t)$) could be its solutions. Therefore, when $Z(t)$ was acting as an observer, it cannot distinguish $X(t)$ from $-X(t)$, and the information content of X was incomplete for $Z(t)$, which determined that X cannot be reconstructed by machine learning. The nonlinear observability for a nonlinear system with known equation can be easily analyzed (Hermann and Krener, 1977; Schumann-Bischoff et al., 2016; Lu et al., 2017). But for the observational data from a complex system without explicit equation, the nonlinear observability is hard to analyze and few studies ever investigated that. Furthermore, does such asymmetric nonlinear observability in the reconstruction also exist in other climatic time series which are nonlinearly coupled? This is still an open question.

In this paper, we apply machine learning approaches to learn the coupling relation, and then reconstruct the coupled climatic time series. Specifically we aim to make progress on how machine learning approach is influenced by the physical couplings of climatic series, and the abovementioned questions can be addressed. There are several variants of machine learning methods (Reichstein et al., 2019), and recent studies (Lu et al., 2017; Reichstein et al., 2019; Chattopadhyay et al., 2019) suggest that three of them are more applicable to sequential data like time series: reservoir computer (RC), back propagation based artificial neural network (BP), and long short-term memory (LSTM) neural network. Here we adopt these three methods to carry out our study, and provide a performance comparison among them. We first investigate their performance dependence on different coupling dynamics by analyzing a hierarchy of climatic conceptual models. Then we use a novel method to select explanatory variables for machine learning, which can further detect the nonlinear observability (Hermann and Krener, 1977; Lu et al., 2017) for a complex system without any known explicit equations.

Finally, we will discuss a real-world example from climate system. It is known that there exist atmospheric energy transportations between the tropics and the Northern Hemisphere, which results in the coupling between the climate systems in these two regions (Farneti and Vallis, 2013). Due to the underlying complicated processes, it is difficult to use a formula to cover this coupling between the tropical average surface air temperature (TSAT) series and the Northern Hemispheric surface air temperature (NHSAT) series. We employ machine learning methods to investigate whether the NHSAT time series can be reconstructed from the TSAT time series, and whether the TSAT time series can be also reconstructed from the NHSAT time series. Accordingly, the conclusions from our model simulations can be further tested and generalized.

Our paper is organized as follows. In section 2, the methods for reconstructing time series and detecting coupling relation are introduced. The used data and climatic conceptual models are introduced in section 3. In section 4, the association between the coupling relation and reconstruction quality by machine learning is investigated, and an application to real-world climate series is presented. Summary is made in section 5.

2 Methods

2.1 Learning coupling relations and reconstructing coupled time series

Firstly, we introduce our workflow for learning couplings of dynamical systems by machine learning, and reconstructing the coupled time series. The total time series can be divided into two parts: the training series (time lasting denoted as *t*) and the testing series (time lasting denoted as *t*'). For the systems of toy models, the coupling relation or dynamics is stable and unchanged with time, i.e., there is the stable coupling or dynamic relation $b(t)=F[a_1(t),a_2(t),...,a_n(t)]$ among inputs $a_1(t),a_2(t),...,a_n(t)$ and output $b(t)$. If this inherent coupling relation can be reconstructed by machine learning in the training series, the reconstructed coupling relation should be reflected by machine learning in the testing series. Therefore, the workflow of our study can be summarized as follows (see Fig. 1):

(i) During the training period, $a_1(t),a_2(t),...,a_n(t)$ and $b(t)$ are input into the machine learning frameworks to learn the coupling or dynamic relation $b(t)=F[a_1(t),a_2(t),...,a_n(t)]$. The inferred coupling relation is denoted as $b(t)=\hat{F}[a_1(t),a_2(t),...,a_n(t)]$. Then it is tested whether this coupling relation can be reconstructed by machine learning.

(ii) The second step is accomplished with the testing series to apply the reconstructed coupling relation $\hat{F}$ together with only $a_1(t'),a_2(t'),...,a_n(t')$ to derive $b(t')$, denoted as $\hat{b}(t')$. $\hat{b}(t')$ is called "the reconstructed $b(t')$" since only $a_1(t'),a_2(t'),...,a_n(t')$ and the reconstructed coupling relation $\hat{F}$ have been taken into account.

(iii) The first objective of this study is to answer whether the coupling relation $b(t)=F[a_1(t),a_2(t),...,a_n(t)]$ can be reconstructed by machine learning, i.e., whether the reconstructed coupling relation $\hat{F}$ can well approximate the real coupling relation F. Since we do not intend to reach an explicit formula of the reconstructed coupling relation $\hat{F}$, we will answer this question indirectly by comparing the reconstructed series $\hat{b}(t')$ with the original series $b(t')$. If $\hat{b}(t')\approx b(t')$, then it can be regarded as $\hat{F}\approx F$, and the machine learning can indeed learn the intrinsic coupling relation among $a_1(t), a_2(t), \ldots, a_n(t)$ and $b(t)$.

(iv) If the machine learning can infer the intrinsic coupling relation between $a_1(t), a_2(t), \ldots, a_n(t)$ and $b(t)$, the inferred coupling relation $\hat{F}$ can be applied to reconstruct output $b(t')$ even if only $a_1(t'), a_2(t'), \ldots, a_n(t')$ are available.

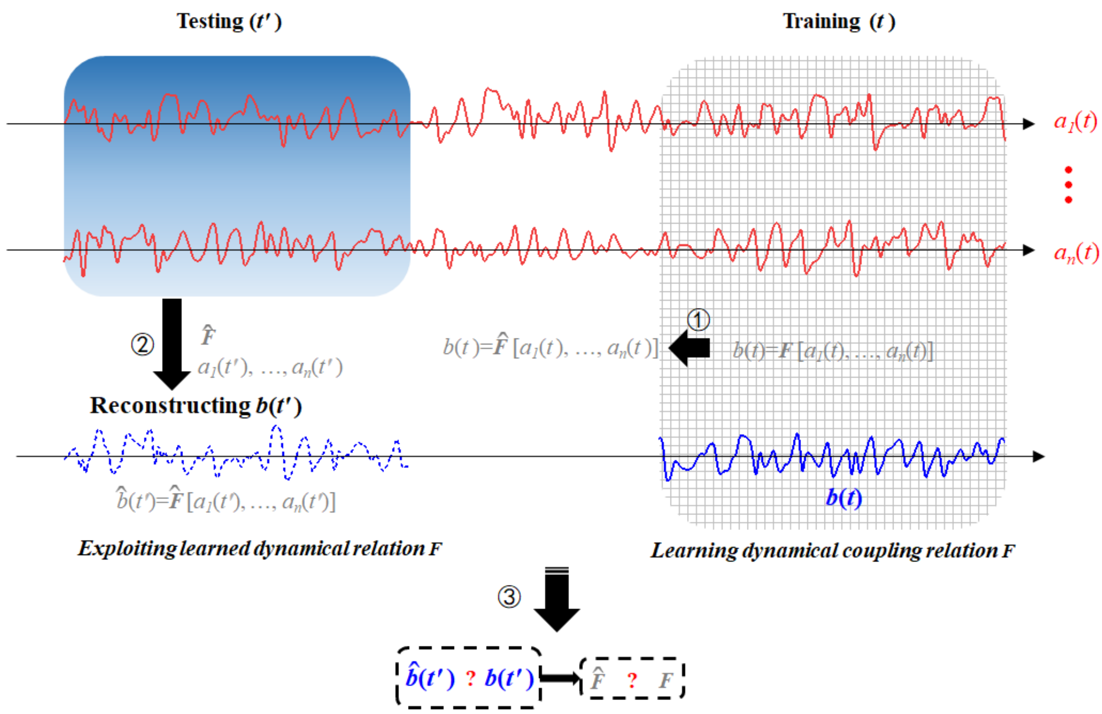

Figure 1 Diagram illustration for reconstructing time series by machine learning. (1) The available part of the dataset $\{a_1(t), \ldots, a_n(t), b(t)\}$ is used to train the neural network ($a_1(t), \ldots, a_n(t)$ and $b(t)$ are the time series of the variables $a_1, \ldots, a_n, b$). So that the inherent coupling relation F among these variables can be learnt by the neural network, and the learnt coupling relation is noted as $\hat{F}$. (2) $b(t')$ is unknown, but the dataset $\{a_1(t'), a_2(t'), \ldots, a_n(t')\}$ is available which is input into the trained neural network, and the unknown series $b(t')$ can be reconstructed, denoted as $\hat{b}(t')$. (3) If $\hat{b}(t') \approx b(t')$, then $\hat{F} \approx F$ can be derived, and it indicates that the machine learning framework have learnt the intrinsic coupling relation.

2.2 Machine learning methods

2.2.1 Reservoir computer

A newly developed neural network called RC (Du et al., 2017; Lu et al., 2017; Pathak et al., 2018) has three layers: the input layer, the reservoir layer and the output layer (see Fig. 2). If $a(t)$ and $b(t)$ denote two time series from a system, and then the following steps can estimate $b(t)$ from $a(t)$:

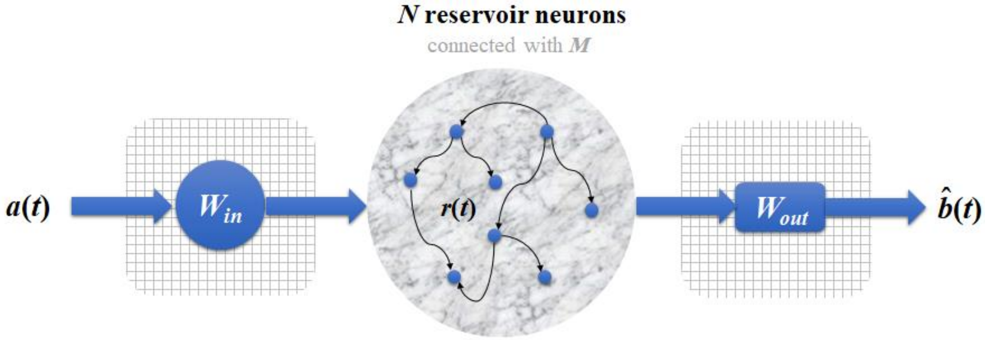

Figure 2 Schematic of the RC neural network: the three layers are the input layer, the reservoir layer, and the output layer. The input layer consists of a matrix "W_{in}" (whose elements are randomly chosen from the interval [-1, 1]). The reservoir layer consists of N reservoir neurons whose connectivity is through the adjacent matrix "M", and $r(t)$ represents the activations of the N neurons. The output layer consists of a matrix "W_{out}", whose elements are trainable in the training process. A time series $a(t)$ is input into the RC neural network. After the training process, the time series of b variable can be reconstructed by machine learning, denoted as $\hat{b}(t)$.

(i) $a(t)$ (a vector with length L) is input into the input layer and reservoir layer. There are four components in this process: the initial reservoir state $r(t)$ (a vector with dimension N, representing the N neurons), the adjacent matrix "M" (size $N\times N$) representing connectivity of the N neurons, the input-to-reservoir weight matrix "W_{in}" (size $N\times L$), and the unit matrix "E" (size $N\times N$) which is crucial for modulating the bias in the training process (Lu et al., 2018). The elements of "M" and "W_{in}" are randomly chosen from a uniform distribution in [–1, 1], and we set N = 1000 here (we have tested that this yields the good performance). These components are employed by Eq. (1), and then an updated reservoir state $r^*(t)$ is output.

$$r^*(t) = \tanh\,[M \cdot r(t) + W_{in} \cdot a(t) + E]\,, \tag{1}$$

(ii) $r^*(t)$ then gets into the output layer that consists of the reservoir-to-output matrix "W_{out}". As Eq. (2) shows, $r^*(t)$ will be trained as the estimated value $\hat{b}(t)$. The mathematical form of "W_{out}" is shown by Eq. (3), which is a trainable matrix that fits the relation between $r^*(t)$ and $b(t)$ in the training process. "$\|\cdot\|$" denotes the L_2-norm of a vector (L_2 represents the least square method) and α is the ridge regression coefficient, whose values are determined after the training.

$$\hat{b}(t) = W_{out} \cdot r^*(t)\,, \tag{2}$$

$$W_{out} = \arg\min_{W_{out}} \left\| W_{out} \cdot r^*(t) - Y(t+\tau) \right\| + \alpha \| W_{out} \|\,, \tag{3}$$

After this reservoir neural network has been trained, we can use it to estimate $b(t)$, where the estimated value is noted as $\hat{b}(t)$.

2.2.2 Back propagation based artificial neural network

Here, the used BP artificial neural network is a traditional neural computing framework which has been widely used in climate research (Chattopadhyay et al., 2019; Watson, 2019; Reichstein et al., 2019). There are six layers in the BP neural network: the input layer with 8 neurons; 4 hidden layers with 100 neurons each; the output layer with 8 neurons. In each layer, the connectivity weights of the neurons need to be computed during training process, where the back propagation optimization with the complicated gradient decent algorithm is used (Dueben and Bauer, 2018). A crucial difference between the BP and the RC neural networks is as follows: unlike RC, all neuron states of the BP neural network are independent on the temporal variation of time series (Chattopadhyay et al., 2019; Reichstein et al., 2019), while the neurons of RC can track temporal evolution (such as the neuron state $r(t)$ in Fig. 2) (Chattopadhyay et al., 2019). If $a(t)$ and $b(t)$ are two time series of a system, through the BP neural network, we can also reconstruct $b(t)$ from $a(t)$.

2.2.3 Long short-term memory neural network

The LSTM neural network is an improved recurrent neural network to deal with time series (Reichstein et al., 2019; Chattopadhyay et al., 2019). As Fig. 3 shows, LSTM has a series of components: a memory cell, input gate, output gate, and a forget gate in addition to the hidden state in traditional recurrent neural network. When a time series $a(t)$ is input to train this neural network, the information of $a(t)$ will flow through all these components, and then the parameters at different components will be computed for fitting the relation between $a(t)$ and $b(t)$. The govern equations for the LSTM architecture are shown in the Appendix. After the training is accomplished, $a(t)$ can be used to reconstruct $b(t)$ by this neural network.

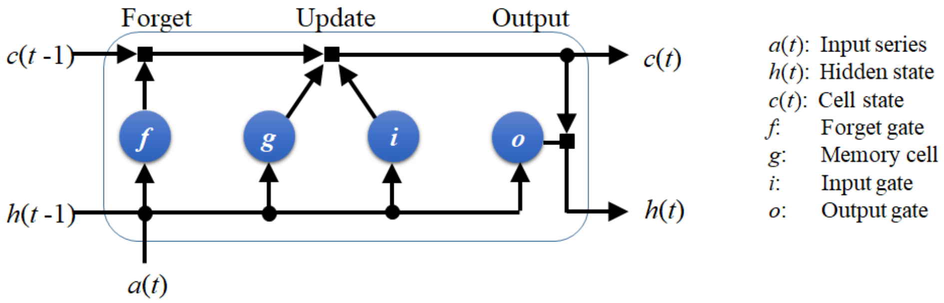

Figure 3 Schematic of the LSTM architecture. LSTM has a memory cell, input gate, output gate, and a forget gate to control the information of the previous time to flow into the neural network.

The crucial improvement of LSTM on the traditional recurrent neural network (Reichstein et al., 2019) is, that LSTM has the forget gate which controls the information of the previous time to flow into the neural network. This will make the neuron states of LSTM have ability to track the temporal evolution of time series (Chattopadhyay et al., 2019; Kratzert et al., 2019; Reichstein et al., 2019), which is also the crucial difference between the LSTM and the BP neural networks.

Here, we also test the LSTM neural network without the forget gate, and call it LSTM*. This means that the information of the previous time cannot flow into the LSTM* neural network, which does not have the memory for the past information. We will compare the performance of LSTM with that of LSTM*, so that the role of the neural network memory for the previous information can be presented.

2.3 Evaluation of reconstruction quality

To evaluate the quality of reconstruction by machine learning, the root mean squared error (RMSE) of residual series (Hyndman and Koehler, 2006) is adopted (Eq. (4)), which represents the difference between the real series $b(t')$ and the reconstructed series $\hat{b}(t')$. In order to fairly compare the errors of reconstructing different processes with different variability and units (Hyndman and Koehler, 2006; Pennekamp et al., 2018; Huang and Fu, 2019), we normalize the RMSE as Eq. (5) shows.

$$RMSE = \sqrt{\frac{1}{k}\sum_{t}[b(t') - \hat{b}(t')]^2}, \tag{4}$$

$$nRMSE = \frac{RMSE}{\max[b(t')] - \min[b(t')]}. \tag{5}$$

2.4 Coupling detection

2.4.1 Linear correlation

As the introduction mentioned, the linear Pearson correlation is a commonly-used method to quantify the linear relationship between two observational variables. The Pearson correlation between two series $a(t)$ and $b(t)$, is defined as

$$corr. = \frac{mean[(a-\bar{a})\cdot(b-\bar{b})]}{std(a)\cdot std(b)}. \tag{6}$$

The symbols "*mean*" and "*std*" denote the average and standard deviation for series $a(t)$ and $b(t)$, respectively.

2.4.2 Convergent cross mapping

To measure the nonlinear coupling relation between two observational variables, we choose the convergent cross mapping method that has been demonstrated to be useful for many complex nonlinear systems (i.e. Sugihara et al., 2012; Tsonis et al., 2018; Zhang et al. 2019). Considering $a(t)$ and $b(t)$ as two observational time series, we begin with the cross mapping (Sugihara et al., 2012) from $a(t)$ to $b(t)$ through the following steps:

i) Embedding $a(t)$ (with length L) into the phase space with a vector $M_a(t_i)=\{a_{t_i}, a_{t_i-\tau 0}, \ldots, a_{t_i-(m-1)\tau}\}$ ("t_i" represents a historical moment in the observations), where embedding dimension (m) and time delay (τ) can be determined through the false nearest neighbor algorithm (Hegger and Kantz, 1999).

ii) Estimating the weight parameter w_i which denotes the associated weight between two vectors "$M_a(t)$" and "$M_a(t_i)$" ("t" denotes the excepted time in this cross mapping), defined as:

$$w_i = \frac{u_i}{\sum_{i=1}^{m+1} u_i}, \tag{7}$$

$$u_i = exp\{-\frac{d\,[M_a(t), M_a(t_i)]}{d\,[M_a(t), M_a(t_1)]}\}, \tag{8}$$

where $d\,[M_a(t), M_a(t_i)]$ denotes the Euler distance between vectors "$M_a(t)$" and "$M_a(t_i)$". The nearest neighbor to "$M_a(t)$" generally corresponds to the largest weight.

iii) Cross mapping the value of $b(t)$ by

$$\hat{b}(t) = \sum_{i=1}^{m+1} w_i b(t_i). \tag{9}$$

$\hat{b}(t)$ denotes the estimated value of $b(t)$ with this phase-space cross mapping. Then, we will evaluate the cross mapping skill (Sugihara et al., 2012; Tsonis et al., 2018) as the follows:

$$\rho_{a \to b} = corr.\ [b(t),\ \hat{b}(t)] \tag{10}$$

The cross mapping skill from b to a is also measured according to the above steps, marked as $\rho_{b \to a}$. Sugihara et al. and Tsonis et al. ever defined the causal inference according to $\rho_{a \to b}$ and $\rho_{b \to a}$ like that: (i) if $\rho_{a \to b}$ is convergent when L is increased, and $\rho_{a \to b}$ is of high magnitude, then b is suggested to be a causation of a. (ii) Besides, if $\rho_{b \to a}$ is also convergent when L is increased, and is of high magnitude, then the causal relationship between a and b is bidirectional (a and b cause each other). In our study, all values of the CCM indices are measured when they are convergent with the data length (Tsonis et al. 2018).

According to literature (Sugihara et al., 2012; Ye et al., 2015), the CCM index is related to the ability of using one variable to reconstruct another variable: if b influence a but a does not influence b, the information content of b can be encoded in a (through the information transfer from b to a), but the information content of a is not encoded in b (there exists no information transfer from a to b). Therefore, the time series of b can be reconstructed from the records of a. For the CCM index ($\rho_{a \to b}$), its magnitude represents how much information content of b is encoded in the records of a. Therefore, the high magnitude of $\rho_{a \to b}$ means that b causes a, and we can get good results of reconstruction from a to b. In this paper, we will test the association between the CCM index and the reconstruction performance of machine learning.

3 Data

3.1 Time series from conceptual climate models

A linearly coupled model: The autoregressive fractionally integrated moving average (ARFIMA) model (Granger and Joyeux, 1980) maps a Gaussian white noise $\varepsilon(t)$ into a correlated sequence $x(t)$ (Eq. (11)), which could simulate the linear dynamics of oceanic-atmospheric coupled system (Hasselmann, 1976; Franzke, 2012; Massah and Kantz, 2016; Cox et al., 2018).

$$\varepsilon(t) \xrightarrow{ARFIMA(p,d,q)} x(t) \tag{11}$$

In this model, d is a fractional differencing parameter, and p and q are the orders of the autoregressive and moving average components. Here, the parameters are set as: p = 3, d = 0.2 and q = 3. Hence $x(t)$ is a time series composited with three components: the third-order autoregressive process whose coefficients are 0.6, 0.2 and 0.1, the fractional differencing process whose Hurst exponent is 0.7, and the third-order moving average process whose coefficients are 0.3, 0.2 and 0.1 (Granger and Joyeux, 1980). These two time series $\varepsilon(t)$ and $x(t)$ are used for the reconstruction analysis.

A nonlinearly coupled model: The Lorenz 63 chaotic system (Lorenz, 1963) depicts the nonlinear coupling relation in a low-dimensional chaotic system. The system reads

$$\begin{aligned} \frac{dx}{dt} &= -\sigma(x-y) \\ \frac{dy}{dt} &= \mu x - xz - y \\ \frac{dz}{dt} &= xy - Bz \end{aligned} \tag{12}$$

When the parameters are fixed at (σ, μ, B) = (10, 28, 8/3), the state in the system is chaotic. We employed the fourth-order Runge-Kutta integrator to acquire the series output from this Lorenz 63 system. The time steps were 0.01. The time series $X(t)$ and $Z(t)$ are used for the reconstruction analysis.

A high-dimensional model: The two-layer Lorenz 96 model (Lorenz, 1996) is a high-dimensional chaotic system, which is commonly used to mimic mid-latitude atmospheric dynamics (Chorin and Lu, 2015; Hu and Franzke, 2017; Vissio and Lucarini, 2018; Chen and Kalnay, 2019; Watson, 2019). It reads

$$\begin{aligned}\frac{dX_k}{dt} &= X_{k-1}(X_{k+1}-X_{k-2})-X_k+F-\frac{h_1}{\mathrm{J}}\sum_{j=1}^{J}Y_{j,k}\\ \frac{dY_{k,j}}{dt} &= \frac{1}{\theta}[Y_{k,j+1}(Y_{k,j-1}-Y_{k,j+2})-Y_{k,j}+h_2X_k].\end{aligned} \tag{13}$$

In the first layer of the Lorenz 96 system there are 18 variables marked as X_k (k is a integer ranging from 1 to 18), and each X_k is coupled with $Y_{k,j}$ ($Y_{k,j}$ is from the second layer). The parameters are set as fellows: J = 20, h_1 = 1, h_2 = 1, and F=10. The parameter θ can alter the coupling strength: when θ is decreased, the coupling strength between X_k and $Y_{k,j}$ will be enhanced. The fourth-order Runge-Kutta integrator and periodic boundary condition are adopted (that is: $X_0 = X_K$ and $X_{K+1} = X_1$; $Y_{k,0} = Y_{k-1,J}$ and $Y_{k,J+1} = Y_{k+1,1}$), and the integral time unit was taken as 0.05. The time series $X_1(t)$ and $Y_{1,1}(t)$ are used for the reconstruction analysis.

3.2 Real-world climatic time series

TSAT, NHSAT and the Nino3.4 index are chosen to represent real-world climatic time series, which are used for reconstruction analysis. The original data is obtained from National Centers for Environmental Prediction (https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis2.html) and KNMI Climate Explorer (http://climexp.knmi.nl). The series of TSAT and NHSAT are obtained from the regional average of gridded daily data in NCEP Reanalysis 2. The selected spatial range is 20^0N–20^0S for the tropics and 20^0N–90^0N for the Northern Hemisphere. The selected temporal range is from 1981/09/01 to 2018/12/31.

Training and testing datasets: Before analysis, all the used time series are standardized to take zero mean and unit variance so that any possible impact of mean and variance on the statistical analysis is avoided (Brown, 1994; Hyndman and Koehler, 2006; Chattopadhyay et al., 2019). We divide the total series into two parts: 60% of the time series training the neural network and 40% being the testing series. Specific data lengths of the training series and testing series will be also listed in the results section.

4 Results

4.1 Coupling relation learning

4.1.1 Linear coupling relation and machine learning

We first consider the simplest case: the linear coupling relation between two variables. Here, two time series $x(t)$ and $\varepsilon(t)$ in ARFIMA (3, 0.2, 3) model, are analyzed. Obviously, there are different temporal structures in $x(t)$ and $\varepsilon(t)$, especially for their large-scale trends (Fig. 4a) and power spectra (Fig. 4b). The marked difference between $x(t)$ and $\varepsilon(t)$ is in their low-frequency variations, and there are more low-frequency and larger-scale structures in $x(t)$ than in $\varepsilon(t)$. We employ neural networks (RC, LSTM, LSTM*, and BP) to learn the dynamics of this model (Eq. (11)) by the procedure shown in Fig. 1. The training parts of $\varepsilon(t)$ are selected from the gray shadow in Fig. 4a. RC, LSTM, LSTM*, and BP are trained to learn the coupling relation between $x(t)$ and $\varepsilon(t)$. Then, the trained neural networks together with $\varepsilon(t')$ is used to reconstruct $x(t')$. The reconstruction results and the performance of different neural networks are presented in Table 1. It shows that there is a strong linear correlation (0.88) between $x(t')$ and $\varepsilon(t')$. This reconstruction result suggests that the strong linear coupling can be well captured by these three neural networks since all values of nRMSE are low.

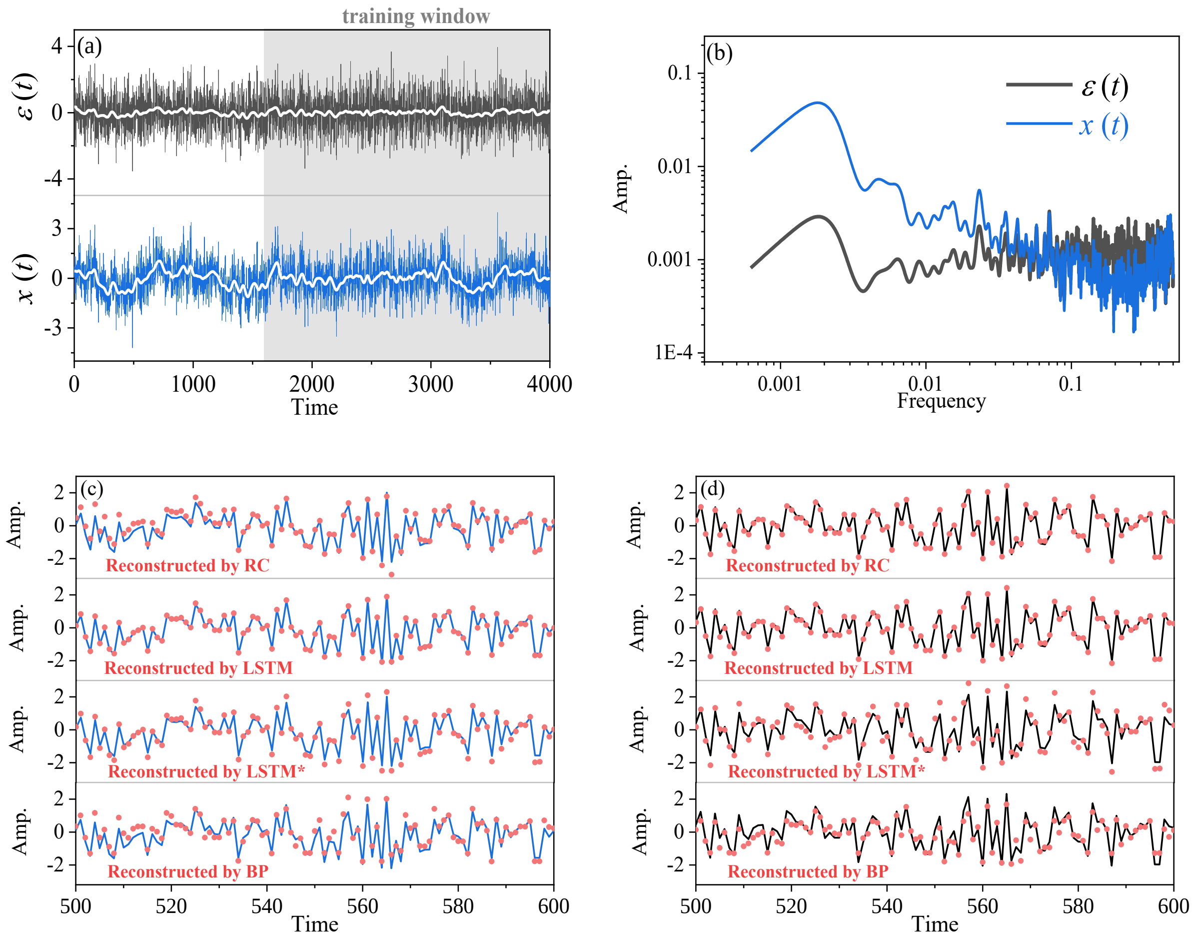

Figure 4 (a) The $x(t)$ time series (blue) and the $\varepsilon(t)$ time series (black) of the ARFIMA(3,0.2,3) model. White lines depict the large-scale trends of these time series acquired by 50-step smoothing average. (b) Comparison of the power spectrum of $x(t)$ (blue) with the power spectrum of $\varepsilon(t)$ (black). (c) Comparison of the reconstructed time series of $x(t)$ by RC, LSTM, LSTM* and BP respectively (red dots), and the original $x(t)$ time series are presented by the blue lines. (d) Comparison of the reconstructed time series of $\varepsilon(t)$ by RC, LSTM, LSTM* and BP respectively (red dots), and the original $\varepsilon(t)$ time series are presented by the black lines. Only partial segments of

the reconstructed series are shown.

Detailed comparisons between the real and reconstructed series are shown in Fig. 4c. When inputting $\varepsilon(t')$, the trained RC and LSTM neural networks can be applied to accurately reconstruct the original $x(t')$. When $x(t')$ is reconstructed from $\varepsilon(t')$ by LSTM, the minimum of nRMSE (0.01) is reached; all reconstructed data are nearly overlapped with the real ones and cannot be visually differentiated (see Fig. 4c). When reconstructing $x(t')$ from $\varepsilon(t')$ by the RC, the reconstruction quality is also well. The best performance of LSTM among the three neural networks benefits from its memory function for the past information (Reichstein et al., 2019; Chattopadhyay et al., 2019). When the memory function of LSTM is stopped, then the reconstruction of LSTM* is no longer better than that of the RC (see Table 1). The reconstruction by BP is successful in this linear system (Fig. 4), but its performance is not as good as LSTM and RC (Table 1). This performance difference might be due to that, unlike LSTM and RC, the neuron states of BP cannot track the temporal evolution of a time series (Chattopadhyay et al., 2019).

Table 1 Details of reconstructing ARFIMA (3, 0.2, 3)

Input (a)	Output (b)	*corr.*	Data length (training/testing)	Neural network	RMSE	nRMSE
$\boldsymbol{\varepsilon(t')}$	$\boldsymbol{x(t')}$	**0.88**	2400/1600	RC	0.31	0.04
				LSTM	0.07	0.01
				LSTM*	0.46	0.06
				BP	0.52	0.07
$\boldsymbol{x(t')}$	$\boldsymbol{\varepsilon(t')}$	**0.88**	2400/1600	RC	0.09	0.01
				LSTM	0.08	0.01
				LSTM*	0.45	0.06
				BP	0.50	0.07

4.1.2 Nonlinear coupling relation and machine learning

It is known that a strong linear correlation is useful for training neural networks and reconstructing time series. When the linear correlation between variables is very weak, could these machine learning methods still be applied to learn the underlying coupling dynamics? To address this question, two nonlinearly coupled time series X and Z in a Lorenz 63 system (Lorenz, 1963) are analyzed.

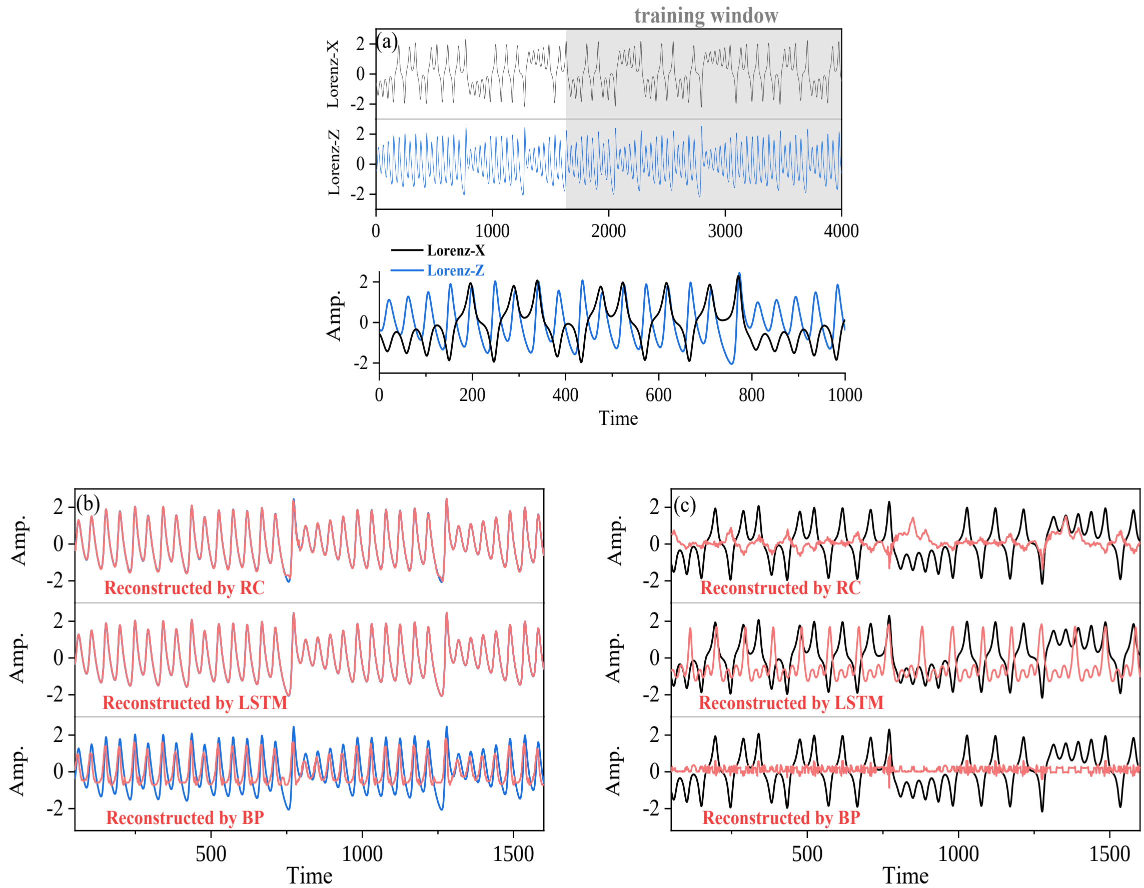

Figure 5 (a) The X time series (black) and the Z time series (blue) of the Lorenz 63 model. (b) Comparison of the reconstructed time series of Z by RC, LSTM and BP respectively (red lines), and the original Z time series are presented by the blue lines. (c) Comparison of the reconstructed time series of X by RC, LSTM and BP respectively (red lines), and the original X time series are presented by the black lines.

There is a very weak linear correlation between variables X and Z (with a Pearson correlation of

0.002) in the Lorenz63 model (Table 2), and such a weak linear correlation is resulted from the time-varying local correlation between variables X and Z (see Fig. 5a): For example, X and Z are negatively correlated in the time interval of 0-200, but positively correlated in 200-400. This alternation of negative and positive correlation appears over the whole temporal evolutions of X and Z, which leads to an overall weak linear correlation. In this case, we cannot use a feasible linear regression model between X and Z to reconstruct one from the other, since there is no such good linear dependency as found in the ARFIMA (p, d, q) system (see Figs. 6a and 6b).

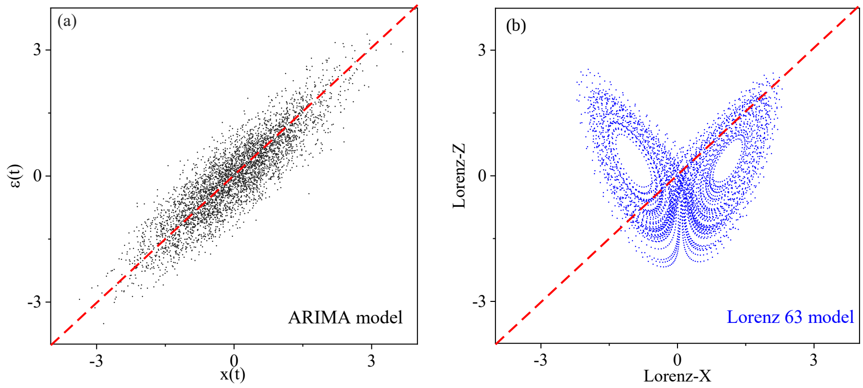

Figure 6 (a) Scatter plot of $x(t)$ versus $\varepsilon(t)$ from ARFIMA(3,0.2,3) model (black dots). (b) Scatter plot of X time series and Z time series of the Lorenz 63 model (blue dots).

Table 2 Details of Lorenz63 system reconstruction

Input (a)	Output (b)	*corr.*	$\rho_{a\to b}$	Data length (training/testing)	Neural network	RMSE	nRMSE
Lorenz -X	Lorenz-Z	**0.002**	0.91	2400/1600	**RC**	**0.04**	**0.008**
					LSTM	0.02	0.004
					LSTM*	1.02	0.24
					BP	0.77	0.17
Lorenz -Z	Lorenz-X	**0.002**	**0.03**	2400/1600	**RC**	**1.13**	**0.34**
					LSTM	1.03	0.31
					LSTM*	1.08	0.33
					BP	1.01	0.31

In a nonlinear coupled system, it is known that the coupling strength between two variables cannot be estimated by the linear Pearson correlation (Brown, 1994; Sugihara et al., 2012). Here, we use CCM to estimate the coupling strength between X and Z, and then it shows a high magnitude of the CCM index: $\rho_{X \to Z} = 0.91$. According to the CCM theory (see Method), such a high magnitude of the CCM index indicates that the information content of Z is encoded in the time series of X. Therefore, we conjecture that: when inputting X to the neural network, not only the information content of X, but also the information content of Z can be learned by the neural network. And then it is possible to reconstruct Z from the trained neural network. We will test it in the following.

Figure 5b shows the results of RC, LSTM and BP applied to reconstructing Z from X. Different from the case of linear system, the successful reconstruction for the time series of the Lorenz63 system depends on the used machine learning methods. The series reconstructed by LSTM nearly overlaps with the real series (Fig. 5b), and has the minimum nRMSE (0.004, see Table 2); moreover, the RC performs quite well, with only a little difference found at some peaks and dips (Fig. 5b). These reconstruction results suggest that, even though the linear correlation is very weak, a strong nonlinear correlation will allow RC and LSTM to fully capture the underlying coupling dynamics. However, BP and LSTM* perform poorly, and their reconstruction results have large errors (nRMSE = 0.17 for BP, and nRMSE = 0.24 for LSTM*). The reconstructed series heavily depart from the real series, especially for all peaks and dips, and the reconstructed values for each extreme point are underestimated (Fig. 5b). This means that both of BP and LSTM* cannot learn the nonlinear coupling.

As mentioned in section 2.2, a BP neural network does not track the temporal evolution, since its neuron states are independent to the temporal variation of time series. For LSTM*, it does not include the information of previous time. Previous studies have revealed that the temporal evolution and memory are very important properties for a nonlinear time series (Kantz and Schreiber, 2003; Franzke et al. 2015), which could not be neglected when modeling nonlinear dynamics. These might be responsible for that BP and LSTM* fail in dealing with this nonlinear Lorenz 63 system. Investigations for the application of BP in other different nonlinear relationships needs to be further addressed in the future.

4.2 Reconstruction quality and impact factors

From the above results, it is revealed that RC and LSTM are able to learn both linear and nonlinear coupling relations, and then the coupled time series can be well reconstructed. In this section, we further investigate what factors could influence the reconstruction quality.

4.2.1 Direction dependence and variable dependence

When reconstructing time series of the linear model of Eq. (11), it can be found that the reconstruction is invertible (see Fig. 4d and Table 1): one variable can be taken as explanatory variable to reconstruct another variable well; oppositely, it can be also well reconstructed by another variable. In fact, when there is a strong linear correlation between variables, the invertible (or bi-directional) reconstruction can also be accomplished by using a traditional regression approach (Brown, 1994). Further, when the linear correlation is weak but the nonlinear coupling is strong, will the bi-directional reconstruction still be allowed? The answer is usually no. For example, when comparing the reconstruction quality of reconstructing Z from X (Fig. 5b) with that of reconstructing X from Z (Fig. 5c), it is obvious that all the used machine learning methods fail (large values of nRMSE are all close to 0.3) in reconstructing X from Z. This result is consistent with the nonlinear observability mentioned by Lu et al. (Lu et al., 2017). The reconstruction direction is no longer invertible in this nonlinear system, where the reconstruction quality is direction-dependent and variable-dependent.

Therefore, we further discuss how to select the suitable explanatory variable or reconstruction direction. Tables 1 and 2 show that the reconstruction quality in a linear coupled system highly depends on the Pearson correlation, however it is different for a nonlinear system. For the Lorenz 63 system, the two-direction CCM coefficients between the variables X and Z are asymmetric (with a stronger $\rho_{X \to Z} = 0.91$ and weaker $\rho_{Z \to X} = 0.03$), and then Z can be well reconstructed from X by machine learning but variable X cannot be reconstructed from variable Z (Fig. 5b and 5c). The CCM index can be taken as a potential indicator to determine the explanatory variable and reconstructed variable for this nonlinear system. Here the asymmetric reconstruction quality is resulted from the asymmetric information transfer between the two nonlinearly coupled variables (Hermann and Krener, 1977; Sugihara et al., 2012; Lu et al., 2017). In this coupling between X and Z, much more information content of Z is encoded in X, so that it performs well for reconstructing Z from X (Lu et al., 2017), which can be detected by the CCM index (Sugihara et al., 2012; Tsonis et al., 2018).

4.2.2 Generalization to a high-dimensional chaotic system

The selection for direction and variable is important for the application of neural networks to reconstructing nonlinear time series, but this is derived from the low-dimensional Lorenz 63 system. In this subsection, we present the results from a high-dimensional chaotic system of Lorenz 96 model. Furthermore, we will investigate the association between the CCM index and reconstruction quality in the machine learning frameworks.

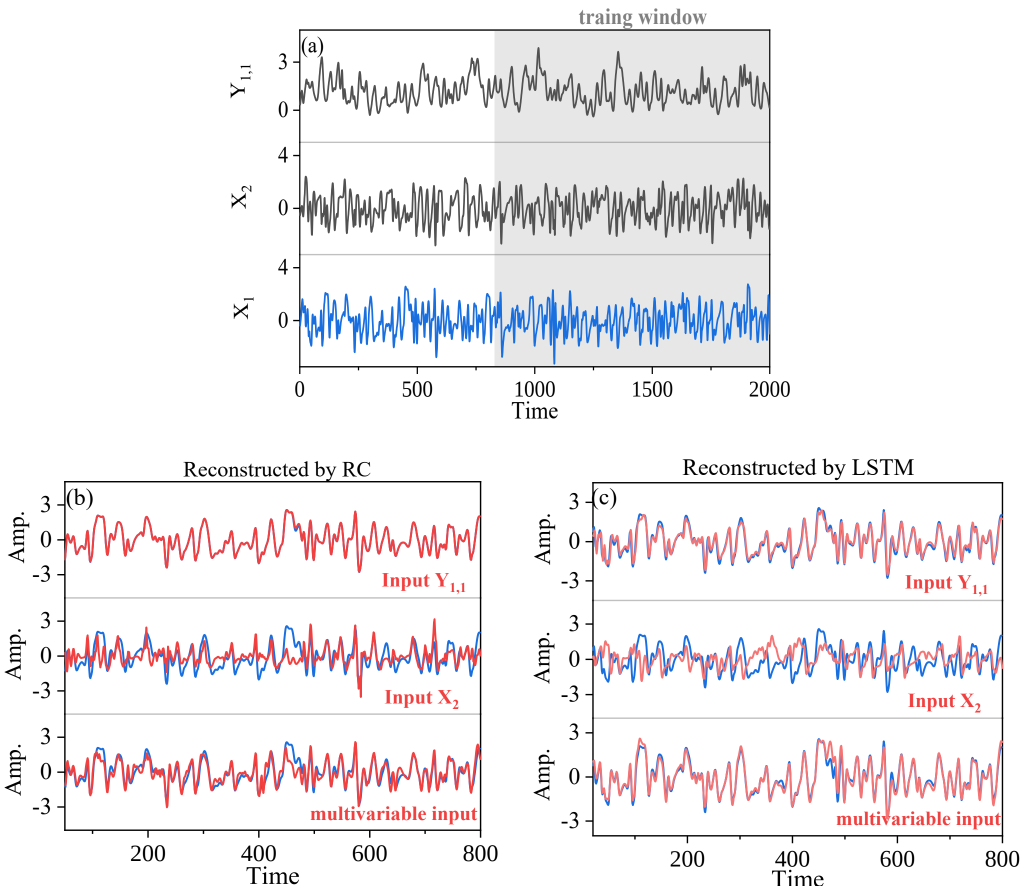

Figure 7 (a) The $Y_{1,1}$ time series(black), X_2 time series (black) and X_1 time series(blue)of the Lorenz 96 model. (b) By means of the RC machine learning, when using $Y_{1,1}$, X_2 and multivariate to be the explanatory variable respectively, the corresponding reconstructed X_1 time series are showed respectively from the top panel to the bottom panel (red lines), and the original X time series are presented by the blue lines. (c) By means of the LSTM machine learning, when using $Y_{1,1}$, X_2 and multivariate to be the explanatory variable respectively, the corresponding reconstructed X_1 time series are showed respectively from the top panel to the bottom panel (red lines), and the original X time series are presented by the blue lines.

Firstly, we use variables X_1 and $Y_{1,1}$ in Eq. (13) to illustrate the direction dependence in the high-dimensional system. Details of X_1 and $Y_{1,1}$ are shown in Fig. 7a, and the Pearson correlation between X_1 and $Y_{1,1}$ is weak (only -0.11, see Table 3). In Eq. (13), the forcing from X_1 to $Y_{1,1}$, is much stronger than the forcing from $Y_{1,1}$ to X_1. The CCM index shows: $\rho_{Y_{1,1}\to X_1}=0.98$ and $\rho_{X_1\to Y_{1,1}}=0.61$. It indicates that reconstructing X_1 from $Y_{1,1}$ may obtain a better quality than the opposite direction. As expected, by means of RC, the error of reconstructing X_1 from $Y_{1,1}$ is nRMSE = 0.01, and in the opposite direction it is nRMSE = 0.06 (Table 3). The result of LSTM is similar to that of RC in this case. Thus, direction dependence does exist in reconstructing this high-dimensional system, and the result is consistent with the indication of the CCM index. In this case, the reconstruction results of BP and LSTM* are not good (not shown here), and we will analyze them in the latter.

Table 3 Details of reconstructing the Lorenz 96 model

Input (a)	Target (b)	***corr.***	$\rho_{a\to b}$	Data length (training/testing)	Neural network	RMSE	nRMSE
$Y_{1,1}$	X_1	-0.11	0.98	1200/800	RC	0.03	0.01
					LSTM	0.34	0.05
$\boldsymbol{X_1}$	$\boldsymbol{Y_{1,1}}$	**-0.11**	**0.61**	**1200/800**	**RC**	**0.35**	**0.06**
					LSTM	**0.42**	**0.08**
X_2	X_1	-0.06	0.37	1200/800	RC	0.69	0.13
					LSTM	1.09	0.20
$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	**-0.06**	**0.25**	**1200/800**	**RC**	**0.95**	**0.17**
					LSTM	**0.84**	**0.16**
X_2, X_{17}, X_{18}	X_1	-0.06, -0.24, 0.06	0.37, 0.29, 0.41	1200/800	RC	0.41	0.08
					LSTM	0.32	0.06

The reconstruction between X_1 and X_2 in the same layer of Lorenz 96 system is also shown. There is an asymmetric causal relation ($\rho_{X_2\to X_1}=0.37$ and $\rho_{X_1\to X_2}=0.25$) between X_1 and X_2, and their linear correlation is very weak (see Table 3). The RC gives better result of reconstructing X_1 from X_2 (nRMSE=0.13) than reconstructing X_2 from X_1 (nRMSE=0.17). LSTM also has different results for X_1 and X_2 (Table 3), where the quality of reconstructing from X_1 to X_2 (nRMSE=0.16) is better than reconstructing from X_2 to X_1 (nRMSE=0.20). In this case, the reconstruction quality of LSTM is worse than the RC, and the reconstruction results by LSTM are not consistent with the indication of the CCM index. A previous study (Chattopadhyay et al., 2019) also suggests that LSTM performs worse than RC in some cases, and this might be related to that only a simple variant of the LSTM architecture used. So in this high-dimensional system, the reconstruction quality is also influenced by the chosen explanatory variables: The quality of reconstructing X_1 from $Y_{1,1}$ is better than the quality of reconstructing X_1 from X_2 by RC and LSTM (see Fig. 7b and 7c).

Besides, the number of the chosen explanatory variables can also influence the reconstruction quality. If more than one explanatory variable in the same layer is considered, the reconstruction of X_1 from X_2 can be greatly improved (see Figs. 7b and 7c). For example, when all of X_2, X_{17} and X_{18} are acting as the explanatory variables, the nRMSE of reconstructed X_1 is reduced from 0.13 to 0.08 (Table 3). For both of RC and LSTM, the multivariable reconstruction reaches lower error than those from unit-variable reconstruction.

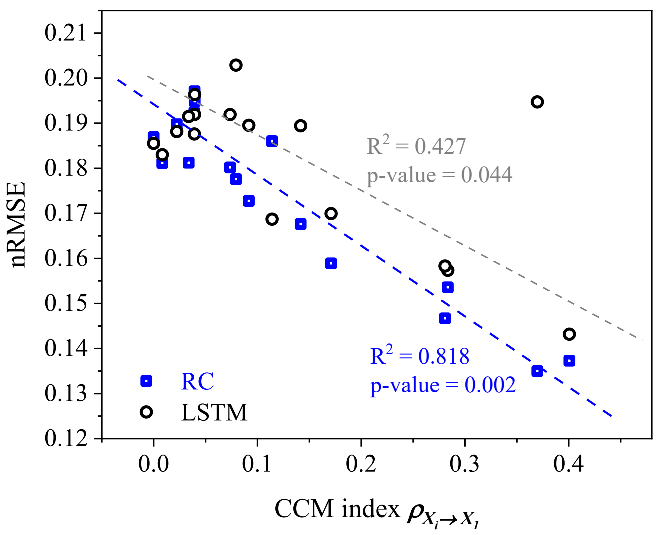

Figure 8 Scatter plot of nRMSE values and CCM index values. The blue boxes are results of the RC machine learning, and the black cycles are results of the LSTM machine learning. The blue and grey dashed lines are the fitted linear trends of the blue boxes and black cycles respectively, and these two dependency trends are both significant because their p-values are both smaller than 0.05.

In the above results, the CCM index is used to select explanatory variable for RC and LSTM. Now we employ more variables to test the association between the CCM index of the data and the performances of RC and LSTM. The values of CCM index are calculated between X_1 and X_2, X_3 …, X_{18}; meanwhile, X_1 is reconstructed from X_2, X_3 …, X_{18}, respectively. We find a significant correspondence exists between the nRMSE and the CCM index (Fig. 8), for both results of RC and LSTM. Here we only use a simple LSTM architecture, and there are many other variants of this architecture where the abnormal point of LSTM in Fig. 8 might be reduced. The result of Fig. 8 reveals the robust association between the CCM index and reconstruction quality in the machine learning frameworks of RC and LSTM. For other machine learning methods, such association deserves further investigation.

4.2.3 Performance of BP and LSTM* in Lorenz 96 system

Since that BP and LSTM* cannot track the temporal evolutions of a nonlinear time series, in the above cases of nonlinear system, we did not obtain similar result to RC and LSTM (not shown here). Here we present a simple experiment, to illustrate what might influence the performances of BP and LSTM* in a nonlinear system.

The experiment is set as follows: in Eq. (13), the value of h_1 is set as 0, and the value of θ is decreased from 0.7 to 0.3. When θ is equal to 0.7, the forcing from X_1 to $Y_{1,1}$ is weak. At that time, the Pearson correlation between X_1 and $Y_{1,1}$ is only 0.48, and the performances of BP and LSTM* are not good. When θ is equal to 0.3, the forcing is dramatically magnified. As the second panel of Fig. 9a shows, this strong forcing makes $Y_{j,i}$ synchronized to X_i, and the Pearson correlation between X_1 and $Y_{1,1}$ is greatly increased to 0.8. When the forcing strength is magnified, the performance of machine learning is also enhanced (Fig. 9b): the reconstructed series by BP and the reconstructed series by LSTM* are much closer to the real target series. This means that the reconstruction quality of BP and LSTM* is greatly improved when the linear correlation is increased. This experiment reveals that, the coupling strength in a nonlinear system can alter the Pearson correlation of two time series, which further influences the performance of BP and LSTM* in a nonlinear system.

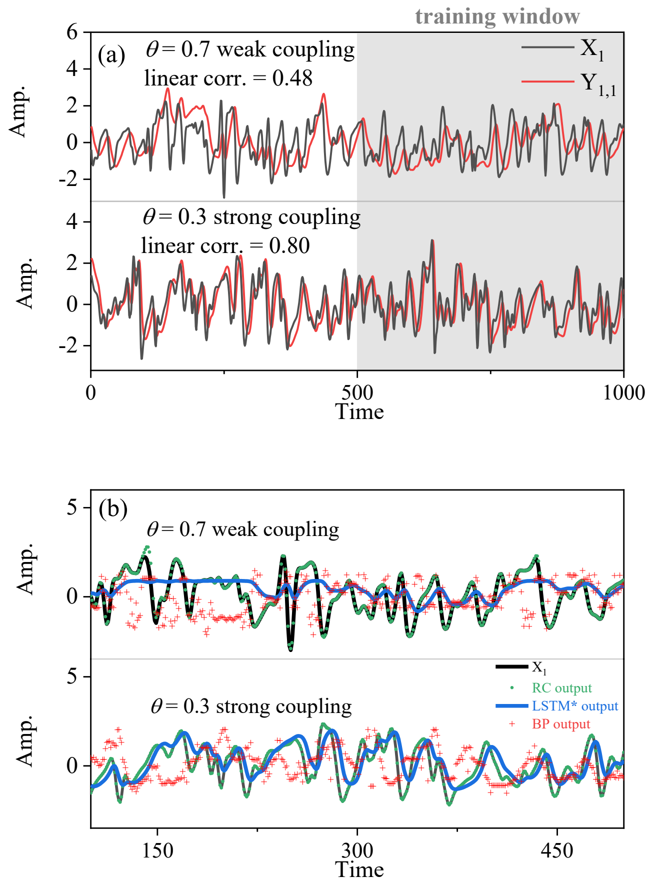

Figure 9 Influence of strong nonlinear coupling on linear Pearson correlation and machine learning performances. (a) Comparison of the linear correlation when the coupling strength is different. The top panel corresponds to the weak coupling strength, and the bottom panel corresponds to the strong coupling. The red lines present the input explanatory variable and the black lines present the target series of machine learning. (b) Comparison of the machine learning performances when the coupling strength is different. The top panel corresponds to the weak coupling strength, and the bottom panel corresponds to the strong coupling. The black lines are the original series;

the reconstructed series by RC (green lines), LSTM*(blue lines) and BP (red dots) are shown respectively. In this case, the results of LSTM are overlapped with that of RC.

4.3 Application to real-world climate series: reconstructing SAT

The natural climate series are usually nonstationary, and are encoded with the information of many physical processes in the earth system. In the following, we illustrate the utility of the above methods and conclusions by investigating a real-world example mentioned in the introduction.

The daily NHSAT and TSAT time series are shown in Fig. 10a. It is quite different for the oscillation shapes of the NHSAT and TSAT series, and there is a weak linear correlation (0.08, see Table 4) between them. In the scatter plot for the NHSAT and TSAT (Fig. 10b), the marked nonlinear structure is observed between NHSAT and TSAT. Such a weak linear correlation will make the linear regression model fail to reconstruct one series from the other. Likewise, there is no explicit physical expression that can transform TSAT and NHSAT to each other. Now we try to use machine learning to reconstruct these climate series. The CCM index of that NHSAT cross maps TSAT is 0.70, and the CCM index of that TSAT cross maps NHSAT is 0.24 (Table 4). The CCM index means that the information content of TSAT is well encoded in the records of NHSAT, and the information transfer might be mainly from TSAT to NHSAT, which is consistent with previous studies (Farneti and Vallis, 2013). Further, the CCM analysis indicates that the reconstruction from NHSAT to TSAT might obtain a better quality than the opposite direction.

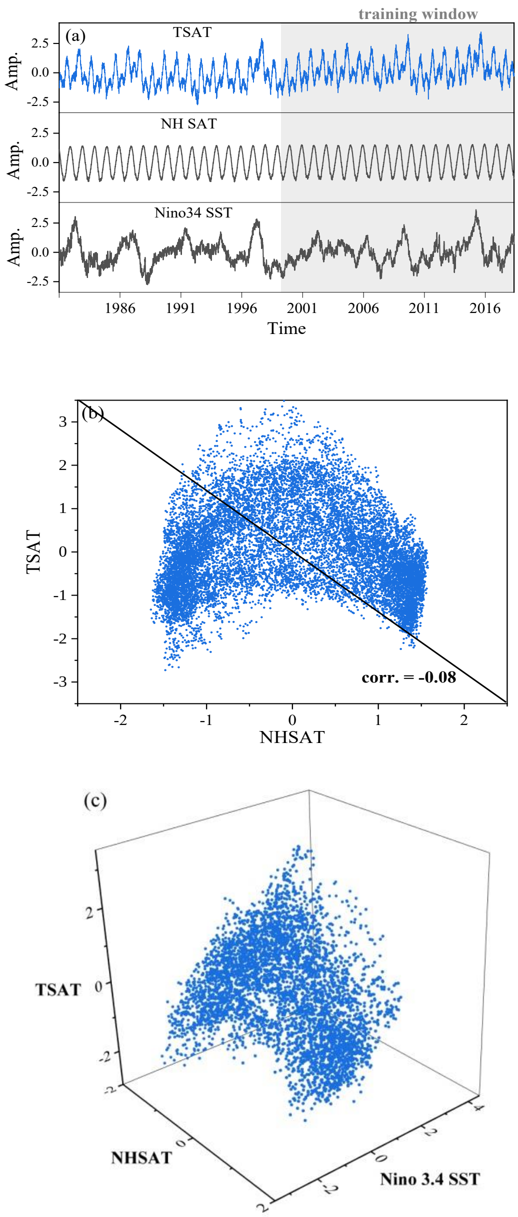

Figure 10 (a) Daily time series of TSAT, NHSAT and Nino 3.4 index. (b) Scatter plot of normalized NHSAT and normalized TSAT. (c) Three-dimensional scatter plot of normalized NHSAT, normalized TSAT and normalized Nino 3.4 SST.

The results are consistent with our conjecture that the nRMSE of reconstruction from NHSAT to TSAT is lower than that from TSAT to NHSAT (Table 4). By using RC, the TSAT time series can be relatively well described by the reconstructed ones (Fig. 11a), with nRMSE equal to 0.13. It is a bit high because some extremes of the TSAT time series have not been well described (Fig. 11b).

When using TSAT to reconstruct the time series of NHSAT, the reconstructed time series cannot describe the original time series of NHSAT (Fig. 11c), and the corresponding nRMSE is equal to 0.21. Besides, we also use LSTM and BP to reconstruct these natural climate series, the performances of these two neural networks are worse than RC (Table 4). For BP, this might be due to its inability to deal with nonlinear coupling (As mentioned in method, the BP neurons cannot track the temporal evolution of a time series). LSTM performs worse than RC in this real-world case might be induced by the used simple variant of LSTM architecture.

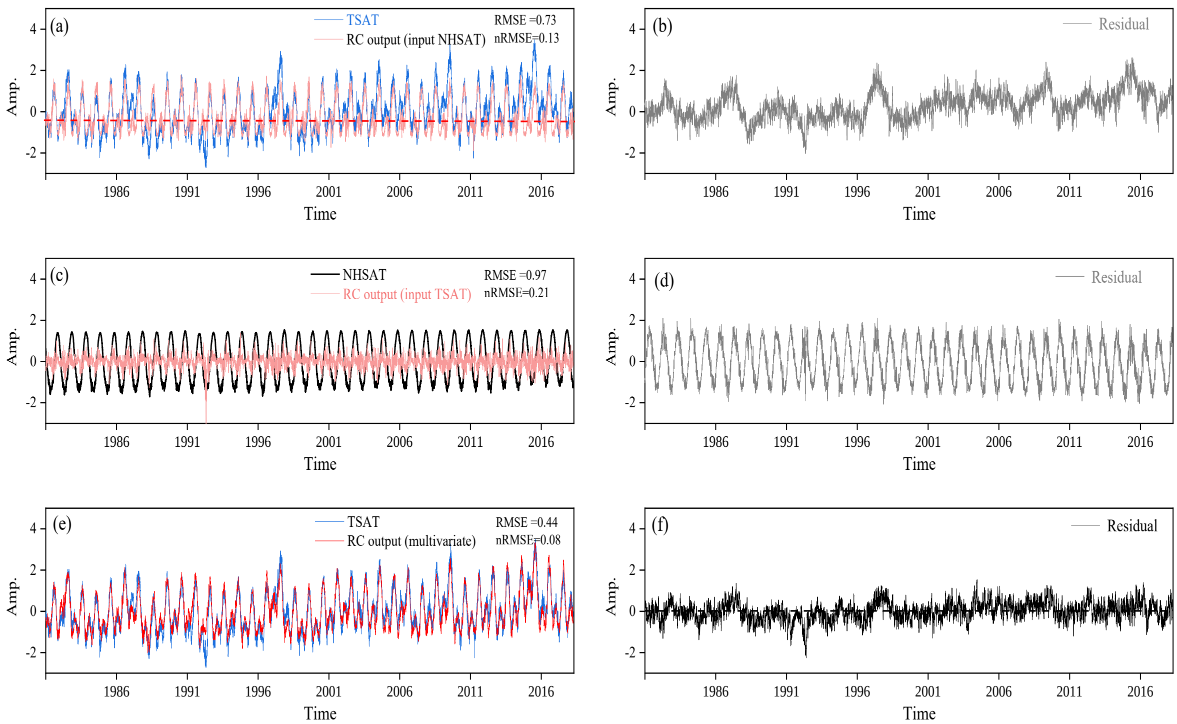

Figure 11 (a) Reconstructed TSAT time series (red) when NHSAT is the explanatory variable; (b) Residual series given by the original TSAT series and the reconstructed TSAT series of (a). (c) Reconstructed NHSAT time series (red) when TSAT is the explanatory variable. (d) Residual series given by the original NHAST series and the reconstructed NHSAT series of (c). (e) Reconstructed TSTA time series (red) when NHSAT and Nino3.4 index are the explanatory variables. (f) Residual series given by the original TSAT series and the reconstructed TSAT series of (e).

Table 4 Details of temperature records' reconstruction

Input (*a*)	Output (*b*)	***corr.***	$\rho_{a \to b}$	Data length (training/testing)	Neural network	RMSE	nRMSE
NHSAT	TSAT	**0.08**	0.70	8182/5454	**RC**	**0.73**	**0.13**
					LSTM	1.14	0.20
					BP	1.45	0.26
TSAT	NHTSAT	**0.08**	**0.24**	8182/5454	**RC**	**0.97**	**0.21**
					LSTM	1.04	0.23
					BP	1.23	0.37

We can further improve the reconstruction quality of TSAT. Considering that the tropics climate system do not only interact with the Northern Hemisphere climate system, we can use the information of other subsystems to improve the reconstruction. Looking at the time series of Nino 3.4 index (Fig. 10), some of its extremes occur at the same time regions as the extremes of TSAT. Moreover, when Nino 3.4 index is included into the scatter plot (Fig. 11c), a nonlinear attractor structure is revealed. We combine NHSAT with Nino 3.4 index to reconstruct the time series of TSAT by means of RC. The reconstructed TSAT (Fig. 11e) is much closer to the original TSAT series, and the corresponding nRMSE has been improved to 0.08.

Finally, we make a further comparison between the real TSAT and the reconstructed TSAT: (i) the annual variations of TSAT and the reconstructed TSAT are close to each other (Fig. 12a). (ii) The power spectrum of TSAT and the reconstructed TSAT are compared in Fig. 12b, and it can be seen that the main deviation is in the frequency bands corresponding to around 0-15 days. The reason might be that the local weather processes are not input into this RC reconstruction. This conjecture can be further confirmed by red-noise test with response time 15 days for the residual series (red-noise test is the same as the method used in Roe, 2009). All data points of the residual series lie within the confidence intervals (Fig. 12c), and this means, the residual is possibly induced by local weather processes that is not input into RC.

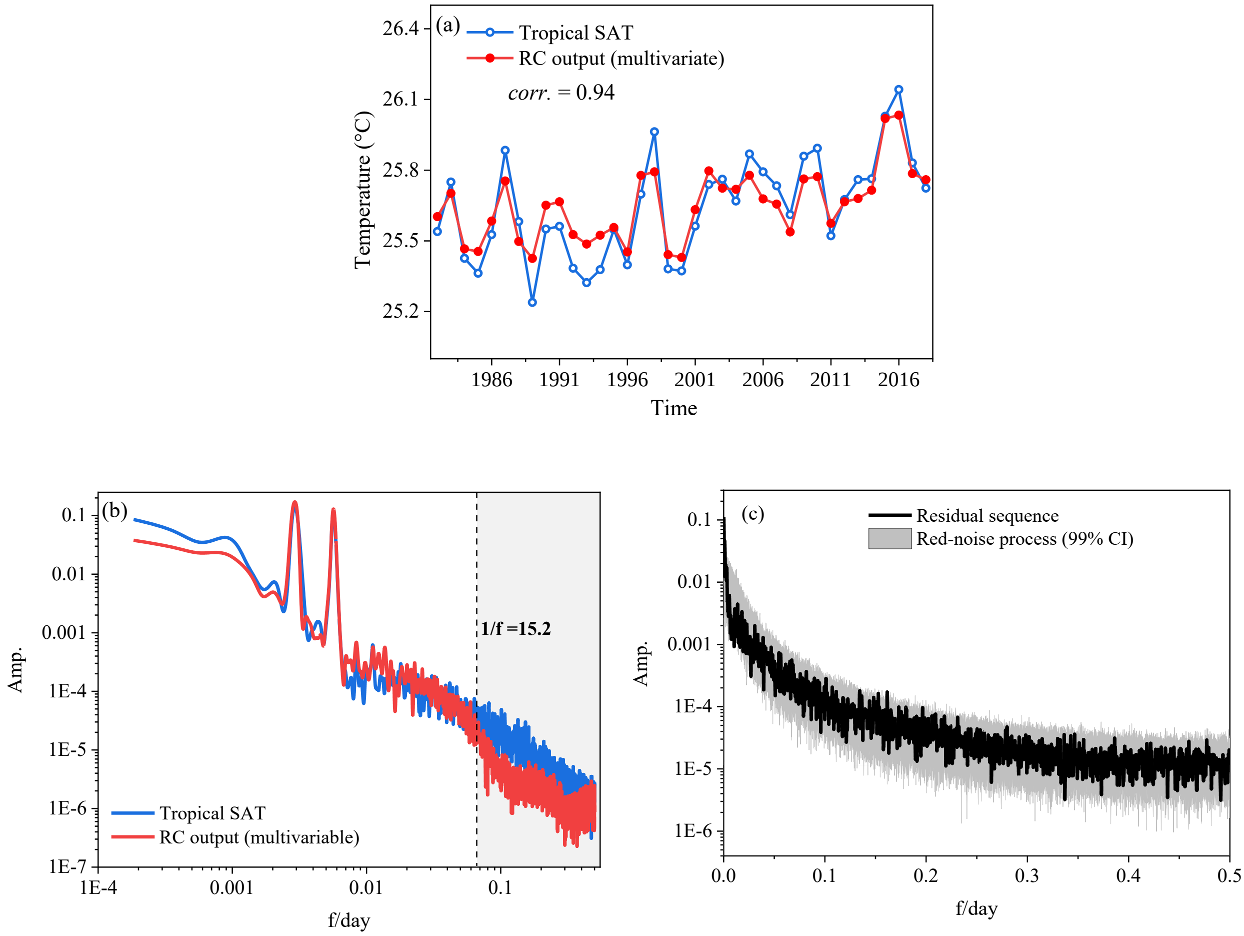

Figure 12 (a) Comparison between the annual mean values of reconstructed TSAT (red) and the annual mean values of original TSAT (blue). (b) Comparison between the power spectrum of reconstructed TSAT (red) and the power spectrum of original TSAT (blue). (c) Red-noise test for residual series, the gray shaded area is the 99% CI of red-noise process.

5 Conclusions and discussions

In this study, three kinds of machine learning methods are used to reconstruct the time series of toy models and real-world climate systems. One series can be reconstructed from the other series by machine learning when they are governed by the common coupling relation. For the linear system, variables are coupled by the linear mechanism, and a strong Pearson coefficient benefits to machine learning with bi-directional reconstruction. For a nonlinear system, the time series often have a weak Pearson coefficient, but the machine learning can still well reconstruct the time series when the CCM index is strong; moreover, the reconstruction quality is direction-dependent and variable-dependent, which is determined by the coupling strength and causality between the dynamical variables.

Considering the reconstruction quality dependency, selecting the suitable explanatory variables is crucial for obtaining a good reconstruction quality. But the results show that machine learning performance cannot be only explained by linear correlation. Hence, we propose using the CCM index to select explanatory variables. Especially for the time series of nonlinear systems, when the CCM index is strong enough, the corresponding variable can be selected as an explanatory variable. When the CCM index is higher than 0.5 in this study, the nRMSE is often smaller 0.1, where the reconstructed series is very close to the real series in the presented results. Therefore, the CCM index that is higher than 0.5 may be considered for selecting explanatory variables. It is well known that atmospheric or oceanic motions are nonlinearly coupled over most of time scales, and therefore, in the natural climate series, there would be similar nonlinear coupling relation to the Lorenz 63 and the Lorenz 96 systems (the Pearson correlation is weak but the CCM indices are of high magnitudes). If only Pearson coefficient is used to select the explanatory variable, then some useful nonlinearly correlated variables might be left out.

Finally, it is worth noting the potential applications for machine learning in the climate studies. For instance, a series $b(t)$ is unmeasured during some periods for the measuring instrument failure, but there are other kinds of variables without missing observations. Moreover, CCM can be applied to select the suitable variables coupled with $b(t)$, and then RC or LSTM can be employed to reconstruct the unmeasured part of $b(t)$ (following Fig. 1). This is useful for some climate studies, such as paleoclimate reconstruction (Brown, 1994; Donner 2012; Emile-Geay and Tingley, 2016), interpolation for the missing points in measurements (Hofstra et al., 2008), and the parameterization schemes (Wilks, 2005; Vissio and Lucarini, 2018). Our study in this article is only a beginning for reconstructing climate series by machine learning, and more detailed investigations will be reported soon.

Appendix

Govern equations for the LSTM neural network

The If $a(t)$ and $b(t)$ denote two time series from a system, and $a(t)$ is input into LSTM to estimate $b(t)$, then the govern equations for the LSTM architecture (Fig. 3) are as follows:

$$f(t) = \sigma_f\left(W_f\left[h\left(t-1\right), a(t)\right] + s_f\right), \tag{14}$$

$$i(t) = \sigma_f\left(W_i\left[h\left(t-1\right), a(t)\right] + s_i\right), \tag{15}$$

$$\tilde{c}(t) = \tanh\left(W_c\left[h\left(t-1\right), a(t)\right] + s_h\right), \tag{16}$$

$$c(t) = f(t)c\left(t-1\right) + i(t)\tilde{c}(t), \tag{17}$$

$$o(t) = \sigma_h(W_h\left[h\left(t-1\right), a(t)\right] + s_h), \tag{18}$$

$$h(t) = o(t)\tanh(\ c(t)\ ), \tag{19}$$

$$b(t) = W_{oh}\ h(t), \tag{20}$$

$f(t)$, $i(t)$, $o(t)$ are the forget gate, input gate, and output gate respectively. $h(t)$ and $c(t)$ represent the hidden state and the cell state, the dimension of the hidden layers are set as 200 which could yield the good performance in our experiment. All these components can be found in Fig. 3, and the information flow among these components are realized by the Eqs. (14)-(20). There are many parameters in the LSTM architecture: σ_f is the softmax activation function; s_f, s_i, and s_h are the biases in the forget gate, the input gate, and the hidden layers; the weight matrixes "W_f", "W_i", "W_c" and "W_{oh}" denote the neuron connectivity in each layers. These parameters need to be computed during training (Chattopadhyay et al., 2019). $a(t)$ and $b(t)$ represent the input and output time series.

Code and data availability. All code and data used in this paper are available on request from authors once the manuscript is accepted.

Author contribution. Yu Huang and Zuntao Fu designed this study. All of the authors contributed to the preparation and writing of the manuscript.

Competing interests. The authors declare no competing interest.

Acknowledgement. The authors thank the constructive suggestions from the editor, the two anonymous reviewers and Dr. Zhixin Lu. We also thank the in-depth and helpful discussion with Dr. Christian L.E. Franzke and Dr. Naiming Yuan. We acknowledge the supports from National Natural Science Foundation of China through Grants (No. 41675049 and No. 41475048).

References

Badin, G., Domeisen, D. I.: A search for chaotic behavior in stratospheric variability: comparison between the Northern and Southern Hemispheres. J. Atm. Sci., 71(12), 4611-4620, 2014.

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., ... Di Carlo, P.: Recursive neural network model for analysis and forecast of PM10 and PM2. 5. Atmos. Pollut. Res., 8(4), 652-659, 2017.

Brown, P. J.: Measurement, Regression, and Calibration, vol. 12 of Oxford Statistical Science Series, Oxford University Press, USA, 216 pp, 1994.

Carroll, T. L.: Using reservoir computers to distinguish chaotic series. Phys. Rev. E. 98(5), 052209, 2018.

Chattopadhyay A., Hassanzadeh P., Palem K., Subramanian D.: Data-driven prediction of a multi-scale Lorenz 96 chaotic system using a hierarchy of deep learning methods: reservoir computing, ANN, and RNN-LSTM. arXiv preprint arXiv:1906.08829, 2019.

Chen, T. C., Kalnay, E.: Proactive quality control: observing system simulation experiments with the Lorenz'96 Model. Mon. Wea. Rev., 147(1), 53-67, 2019.

Chorin, A. J., Lu, F.: Discrete approach to stochastic parameterization and dimension reduction in nonlinear dynamics. P. Natl. Acad. Sci., 112(32), 9804-9809, 2015.

Comeau, D., Zhao, Z., Giannakis, D., Majda, A. J.: Data-driven prediction strategies for low-frequency patterns of North Pacific climate variability. Clim. Dyn., 48(5-6), 1855-1872, 2017.

Conti, C., Navarra, A., Tribbia, J.: The ENSO Transition Probabilities. J. Clim., 30 (13), 4951-4964, 2017.

Cox, P. M., Huntingford, C., Williamson, M. S.: Emergent constraint on equilibrium climate sensitivity from global temperature variability. Nature, 553(7688), 319, 2018.

Donner, L. J., Large, W. G.: Climate modeling. Annual Review of Environment and Resources, 33, 2008.

Donner, R. V.: Complexity concepts and non-integer dimensions in climate and paleoclimate research. Fractal Analysis and Chaos in Geosciences, Nov 14:1, 2012.

Drótos, G., Bódai, T., Tél, T.: Probabilistic concepts in a changing climate: A snapshot attractor picture. J. Clim., 28(8), 3275-3288, 2015.

Du, C., Cai, F., Zidan, M. A., Ma, W., Lee, S. H., Lu, W. D.: Reservoir computing using dynamic memristors for temporal information processing. Nat. Commun., 8(1), 2204, 2017.

Dueben, P.D., Bauer, P.. Challenges and design choices for global weather and climate models based on machine learning. Geoscientific Model Development, 11(10), 3999-4009, 2018.

Emile-Geay, J., Tingley, M.: Inferring climate variability from nonlinear proxies: application to paleo-ENSO studies. Clim. Past., 12(1), 31-50, 2016.

Farneti, R., Vallis, G. K.: Meridional energy transport in the coupled atmosphere–ocean system: Compensation and partitioning. J. Clim., 26(18), 7151-7166, 2013.

Feng, X., Fu, T. M., Cao, H., Tian, H., Fan, Q., Chen, X.: Neural network predictions of pollutant emissions from open burning of crop residues: Application to air quality forecasts in southern China. Atmos. Environ., 204, 22-31, 2019.

Franzke, C. L.: Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. J. Clim., 25(12), 4172-4183, 2012.

Franzke C. L., Osprey, S. M., Davini, P., Watkins, N. W.: A dynamical systems explanation of the Hurst effect and atmospheric low-frequency variability. Sci. Rep., 5, 9068, 2015.

Granger, C. W., Joyeux, R.: An introduction to long-memory time series models and fractional differencing. J. Time. Ser. Anal., 1(1), 15-29, 1980.

Hasselmann, K.: Stochastic climate models part I. Theory. Tellus, 28(6), 473-485, 1976.

Hegger, R, Kantz, H.: Improved false nearest neighbor method to detect determinism in time series data. Phys. Rev. E, 60(4), 4970, 1999.

Hermann R, Krener A. Nonlinear controllability and observability. IEEE Transactions on automatic control, 22(5), 728-740, 1977.

Hofstra, N., Haylock, M., New, M., Jones, P., Frei, C.: Comparison of six methods for the interpolation of daily European climate data. J. Geophys. Res., 113(D21), 2008.

Hsieh, W. W., Wu, A., Shabbar, A.: Nonlinear atmospheric teleconnections. Geophys. Res. Lett., 33(7): L07714, 2006.

Hu, G., Franzke, C. L.: Data assimilation in a multi-scale model. Mathematics of Climate and Weather Forecasting, 3(1), 118-139, 2017.

Huang, Y., Fu, Z.: Enhanced time series predictability with well-defined structures. Theor. Appl. Climatol., 138, 373–385, 2019.

Hyndman, R. J., Koehler, A. B.: Another look at measures of forecast accuracy. Int. J. Forecasting., 22(4), 679-688, 2006.

Kantz, H., Schreiber, T.: Nonlinear time series analysis (Vol. 7). Cambridge university press, 2004.

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., Klambauer, G. Neural Hydrology-Interpreting LSTMs in Hydrology. arXiv:1903.07903, 2019.

Lorenz, E. N.: Deterministic nonperiodic flow. J. Atmos. Sci., 20(2), 130-141, 1963.

Lorenz, E. N.: Predictability: a problem partly solved. Proc. ECMWF Seminar on Predictability, vol I, Reading, United Kingdom, ECMWF, pp 40–58, 1996.

Lu, Z., Pathak. J., Hunt, B., Girvan, M., Brockett, R., Ott, E.: Reservoir observers: Model-free inference of unmeasured variables in chaotic systems. Chaos, 27(4), 041102, 2017.

Lu, Z., Hunt, B. R., Ott, E.: Attractor reconstruction by machine learning. Chaos, 28(6): 061104, 2018.

Ludescher, J., Gozolchiani, A., Bogachev, M. I., Bunde, A., Havlin, S., Schellnhuber, H. J.: Very early warning of next El Niño. P. Natl. Acad. Sci., 111(6), 2064-2066, 2014.

Massah, M., Kantz, H.: Confidence intervals for time averages in the presence of long-range correlations, a case study on Earth surface temperature anomalies. Geophys. Res. Lett., 43(17), 9243-9249, 2016.

Mattingly, K. S., Ramseyer, C. A., Rosen, J. J., Mote, T. L., Muthyala, R.: Increasing water vapor transport to the Greenland Ice Sheet revealed using self-organizing maps. Geophys. Res. Lett., 43(17), 9250-9258, 2016.

Mukhin, D., Gavrilov, A., Loskutov, E., Feigin, A., Kurths, J.: Nonlinear reconstruction of global climate leading modes on decadal scales. Clim. Dyn., 51(5-6), 2301-2310, 2018.

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., Ott, E.: Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. Chaos, 27(12), 121102, 2017.

Patil, D. J., Hunt, B. R., Kalnay, E., Yorke, J. A., Ott, E.: Local low dimensionality of atmospheric dynamics. Phys Rev Lett 86(26): 5878, 2001.

Pennekamp, F., Iles, A. C., Garland, J., Brennan, G., Brose, U., Gaedke, U., Novak, M.: The intrinsic predictability of ecological time series and its potential to guide forecasting. Ecol, Monogr., e01359, 2019.

Racah, E., Beckham, C., Maharaj, T., Kahou, S. E., Prabhat, M., Pal, C.: ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In Advances in Neural Information Processing Systems (pp. 3402-3413), 2017.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N.: Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195, 2019.

Roe, G.: Feedbacks, timescales and seeing red. Ann. Rev. Earth. Plan. Sci., 37: 93-115, 2009.

Schreiber T.: Measuring information transfer. Phys. Rev. Lett., 85(2), 461, 2000.

Schurer, A. P., Hegerl, G. C., Mann, M. E., Tett, S. F., Phipps, S. J.: Separating forced from chaotic climate variability over the past millennium. J. Clim., 26(18), 6954-6973, 2013.

Schumann-Bischoff J, Luther S, Parlitz U. Estimability and dependency analysis of model parameters based on delay coordinates. Phys. Rev. E, 94(3), 032221, 2016.

Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., Munch, S.: Detecting causality in complex ecosystems. Science, 338(6106), 496-500, 2012.

Takens, F.: Detecting strange attractors in turbulence. Dynamical Systems and Turbulence, Lecture Notes in Mathematics, 898, 366–381 (Springer Berlin Heidelberg), 1981.

Tsonis, A. A., Deyle, E. R., Ye, H., Sugihara, G.: Convergent cross mapping: theory and an example. In Advances in Nonlinear Geosciences (pp. 587-600), Springer, Cham., 2018.

Vallis, G. K., Farneti, R.: Meridional energy transport in the coupled atmosphere–ocean system: Scaling and numerical experiments. Q. J. Roy. Meteor. Soc., 135(644), 1643-1660, 2009.

Van, Nes, E. H., Scheffer, M., Brovkin, V., Lenton, T. M., Ye, H., Deyle, E., Sugihara, G.: Causal feedbacks in climate change. Nat. Clim. Change, 5(5): 445, 2015.

Vannitsem, S., Ekelmans, P. Causal dependences between the coupled ocean–atmosphere dynamics over the tropical Pacific, the North Pacific and the North Atlantic. Earth Syst. Dyn., 9(3), 1063-1083, 2018.

Vissio, G., Lucarini, V.: A proof of concept for scale-adaptive parameterizations: the case of the Lorenz 96 model.

Q. J. Roy. Meteor. Soc., 144(710), 63-75, 2018.

Watson, P. A.: Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. J. Adv. Model Earth. Sys., doi.org/10.1029/2018MS001597, 2019.

Wilks, D. S.: Effects of stochastic parametrizations in the Lorenz'96 system. Q. J. Roy. Meteor. Soc., 131(606), 389-407, 2005.

Ye H., Deyle E. R., Gilarranz L. J., Sugihara G.: Distinguishing time-delayed causal interactions using convergent cross mapping, Sci. Rep., 5, 14750, 2015.

Zaytar, M. A., El, Amrani, C.: Sequence to sequence weather forecasting with long short-term memory recurrent neural networks. Int. J. Comput. Appl., 143(11), 7-11, 2016.

Zhang, N. N., Wang, G. L., Tsonis, A. A.: Dynamical evidence for causality between Northern Hemisphere annular mode and winter surface air temperature over Northeast Asia. Clim. Dyn., 52, 3175-3182, 2019.