Response Letter

Dear Prof. Dhanya,

We sincerely thank you and all reviewers for concerning our manuscript entitled “Reconstructing coupled time series in climate systems using three kinds of machine learning methods” (ID: esd-2019-63). Your comments are very helpful for revising and improving our paper. We have made revision and correction by taking these comments into account carefully, and we hope all of these revisions meet with approval. Revised changes are marked in yellow in the paper. Below you will find the main revisions and corrections in the paper and the point-to-point responses to the reviewers’ comments:

Reply to the comments of Editor:

Editor's comments to the Author:

Reviewers have expressed their satisfaction over the technical suitability of the article to be published in ESD. However, they raise serious concerns over the presentation style/ grammatical errors in the manuscript. The manuscript is recommended for minor revision; but I suggest authors to please consider these comments seriously and rework on the presentation of the manuscript. Reviewer comments are enclosed.

Response: Many thanks for your comments and suggestions. We carefully addressed all issues proposed by the reviewers, and we made point-to-point responses to these comments in the following. In this revised manuscript, we thoroughly inspected the language errors, figures and formulation mistakes, and we have made correction which we hope meet with approval.

Reply to the comments of Reviewer 2:

Comments of anonymous Reviewer 2:

The authors have answered all the technical queries and modified the manuscript based on the suggestions. However, I feel the authors need to check for English language corrections. I am mentioning some specific ones below:

Response: Many thanks for your comments and suggestions. In this revised manuscript, we thoroughly inspected the language errors, figures and formulation mistakes, and all revisions and corrections are marked in yellow in the paper.

Specific Points:

1. Line No. 474-478: The following lines should be written as

"Chattopadhyay et al. (2019) also suggests that LSTM performs worse than RC in some cases, and this might be related to the use of a simple variant of the LSTM architecture. This variant of LSTM was tasted and it was found that the time-varying local mean in time series would sometimes influence its performance. However further investigation is required for a deeper understanding of the same. "

Response: Thank you! We rewrote these lines in the revised manuscript, as the following screenshot shows:

474 (nRMSE=0.16) is better than reconstructing from X_2 to X_1 (nRMSE=0.20). In this case, the
475 reconstruction quality of LSTM is worse than that of RC, and the reconstruction results by LSTM
476 are consistent with the indication of the CCM index. Chattopadhyay et al. (2020) also suggests that
477 LSTM performs worse than RC in some cases, and this might be related to the use of a simple
478 variant of the LSTM architecture. This variant of LSTM was tested and it was found that the
479 time-varying local mean in time series would sometimes influence its performance. However further
480 investigation is required for a deeper understanding of the real reason. In this high-dimensional

2. Line No. 523 - 528: The following lines can be rewritten as:

"However, RC and LSTM are not restrained by the Pearson correlation in this nonlinear system. When θ is altered from 0.7 to 0.3, although the Pearson correlation changed a lot, the values of CCM index stayed

consistently above 0.9. Throughout the alterations in θ, RC is able to produce a good quality reconstruction of X1. Fig. 9b shows that the reconstructed series by RC and LSTM always overlap with the real time series. Thus it can be inferred that the performance of both RC and LSTM is sensitive to the value of CCM index. This has been analyzed in section 4.2.2."

Response: Many thanks for your suggestion! We rewrote these lines in the revised manuscript, as the following screenshot shows:

525 However, RC and LSTM are not restricted to the Pearson correlation in this nonlinear system.
526 When θ is altered from 0.7 to 0.3, although the Pearson correlation is changed a lot, the values of
527 the CCM index are kept consistently above 0.9. For all values of θ, RC is able to equally produce a
528 good quality reconstruction of X_1. Fig. 9b shows that the reconstructed series through RC and LSTM
529 always overlap with the real time series. These results indicate that the performance of both RC and
530 LSTM is sensitive to the value of CCM index, which is in line with the results given in section 4.2.2.

3. Line No. 47-48: I suggest rewriting this line as:

"For example, chaos is a crucial property of climatic time series (Lorenz, 1963; Patil et al., 2001)."

Response: Many thanks for your suggestion! We rewrote these lines in the revised manuscript, as the following screenshot shows:

47 series. For example, chaos is a crucial property of climatic time series (Lorenz, 1963; Patil et al.,
48 2001). Thus, there is significant concern regarding the ability of machine learning algorithms to
49 reconstruct the temporal dynamics of the underlying complex systems (Pathak et al., 2017; Du et al.,

4. Line no. 48-50: The following lines can be rewritten as:

"Thus, there is significant concern regarding the ability of machine learning algorithms to reconstruct the temporal dynamics of the underlying complex systems (Pathak et al., 2017; Du et al., 2017; Lu et al., 2018; Carroll, 2018; Watson, 2019)."

Response: Thank you! We have rewritten these sentences, as the following screenshot shows:

47 series. For example, chaos is a crucial property of climatic time series (Lorenz, 1963; Patil et al.,
48 2001). Thus, there is significant concern regarding the ability of machine learning algorithms to
49 reconstruct the temporal dynamics of the underlying complex systems (Pathak et al., 2017; Du et al.,

Other main corrections:

1. In the lines 72-74, for the better presentation, we rewrote the sentences, as the following screenshot shows:

72 might be nonlinearly coupled. For instance, the linear cross-correlations of sea air temperature series
73 observed in different tropical areas are overall weak, but they can be strong locally and vary with
74 time (Ludescher et al., 2014), and such time-varying correlation is an indicator of non-linear
75 correlation (Sugihara et al., 2012). These non-linear correlations of the sea air temperature series

2. Line 663 and line 713, the two cited papers were from the arXiv preprint. Now their corresponding published articles are available , and we update their information in the reference, as the following screenshot shows:

663 Chattopadhyay, A., Hassanzadeh, P., and Subramanian, D.: Data-driven predictions of a multiscale Lorenz 96
664 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long
665 short-term memory network, Nonlin. Processes Geophys., 27, 373–389, 2020.

713 Kratzert F., Herrnegger M., Klotz D., Hochreiter S., Klambauer G.: NeuralHydrology – Interpreting LSTMs in
714 Hydrology. In: Samek W., Montavon G., Vedaldi A., Hansen L., Müller KR. (eds) Explainable AI:
715 Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol 11700.
716 Springer, Cham, 2019.

3. We also carefully inspected the sentences, figures, figure captions, tables and formulas for avoiding any possible typos and errors. Here we did not list the changes but marked in yellow in revised paper.

We have improved the manuscript and made some changes in the manuscript. These changes will not alter the content and framework of the paper.

We appreciate the Editor's and Reviewers' work earnestly, and hope that the correction will meet with approval.

Once again, thank you very much for your comments and suggestions.

Best regards.

Yours sincerely,

Yu Huang, Lichao Yang, Zuntao Fu

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 application in climate dynamics has not been well developed. One concern might be how well the trained neural networks could learn a dynamical system and what will be the potential application 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 shows that the coupling relations or dynamics among variables in linear or nonlinear systems can be inferred by RC and LSTM, which can be further applied to reconstruct one time series from the other. Specifically, we analyzed 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, then the reconstruction can be bi-directional, and both of these two variables can be an explanatory variable for reconstructing the other. 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 the convergent cross mapping (CCM) causality index is proposed to determine which variable can be taken as the reconstructed one and which 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 weak (0.08), but the CCM index of NHSAT cross maps TSAT is large (0.70). And this indicates that TSAT can be well reconstructed from NHSAT through machine learning.

All results shown in this study could provide insights on machine learning approaches for paleoclimate reconstruction, parameterization scheme, and prediction in related climate research.

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

Highlights:

i) The coupling dynamics learnt by machine learning can be used to reconstruct time 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

Applying neural network-based machine learning in climate fields has attracted great attention (Reichstein et al., 2019). Machine learning approach can be applied to downscaling and data mining analyses (Mattingly et al., 2016; Racah et al., 2017), and is 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). Besides, previous studies found that some temporal dynamics of the underlying complex systems can be encoded in these climatic time series. For example, chaos is a crucial property of climatic time series (Lorenz, 1963; Patil et al., 2001). Thus, there is significant concern regarding the ability of machine learning algorithms to reconstruct the temporal dynamics of the underlying complex systems (Pathak et al., 2017; Du et al., 2017; Lu et al., 2018; Carroll, 2018; Watson, 2019). The chaotic attractors in Lorenz system and Rossler system can be reconstructed by machine learning (Pathak et al., 2017; Lu et al., 2018; Carroll, 2018), and the Poincare return map and Lyapunov exponent of the attractor can be recovered as well (Pathak et al., 2017; Lu et al., 2017). These results are important to deeply understand the applicability of machine learning in climate fields.

Though applying machine learning to climate fields has been attracting much attention, there are 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 in different ways (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 it often influences the analysis of climatic time series. “Pearson Coefficient” is often used to detect the correlation, but it can only detect the linear correlation. It is known that when the Pearson correlation coefficient is weak, most of tasks based on 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. For instance, the linear cross-correlations of sea air temperature series observed in different tropical areas are overall weak, but they can be strong locally and vary with time (Ludescher et al., 2014), and such time-varying correlation is an indicator of non-linear correlation (Sugihara et al., 2012). These non-linear correlations of the sea air temperature series have been found to be 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 also overall weak but nonlinear coupling relations between them can be detected, and they contribute 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 pay attention to the linear or

84 nonlinear relationships induced by the physical couplings? This is what we want to address in this
85 study.

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

101 In this paper, we apply machine learning approaches to learn the coupling relation between
102 climatic time series (training period), and then reconstruct the series (testing period). Specifically we
103 aim to make progress on how machine learning approach is influenced by the physical couplings of
104 climatic series, and the abovementioned questions can be addressed. There are several variants of
105 machine learning methods (Reichstein et al., 2019), and recent studies (Lu et al., 2017; Reichstein et

al., 2019; Chattopadhyay et al., 2020) 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 neural network (LSTM). 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, and this 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, and this can result 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 set of formulas to cover the coupling relation 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. By this way, 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 analyzed data and climate conceptual models are described in section 3. In section 4, we will investigate the association between the coupling relation and reconstruction quality by machine learning, and present an application to real-world climate series. Finally 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 inferred by machine learning in the training series, the inferred coupling relation should be reflected by machine learning in the testing series. Therefore, the workflow of our study can be summarized as follows (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 inferred 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 inferred 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 inferred 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 $a_1(t), a_2(t), ..., a_n(t)$ and $b(t)$.

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

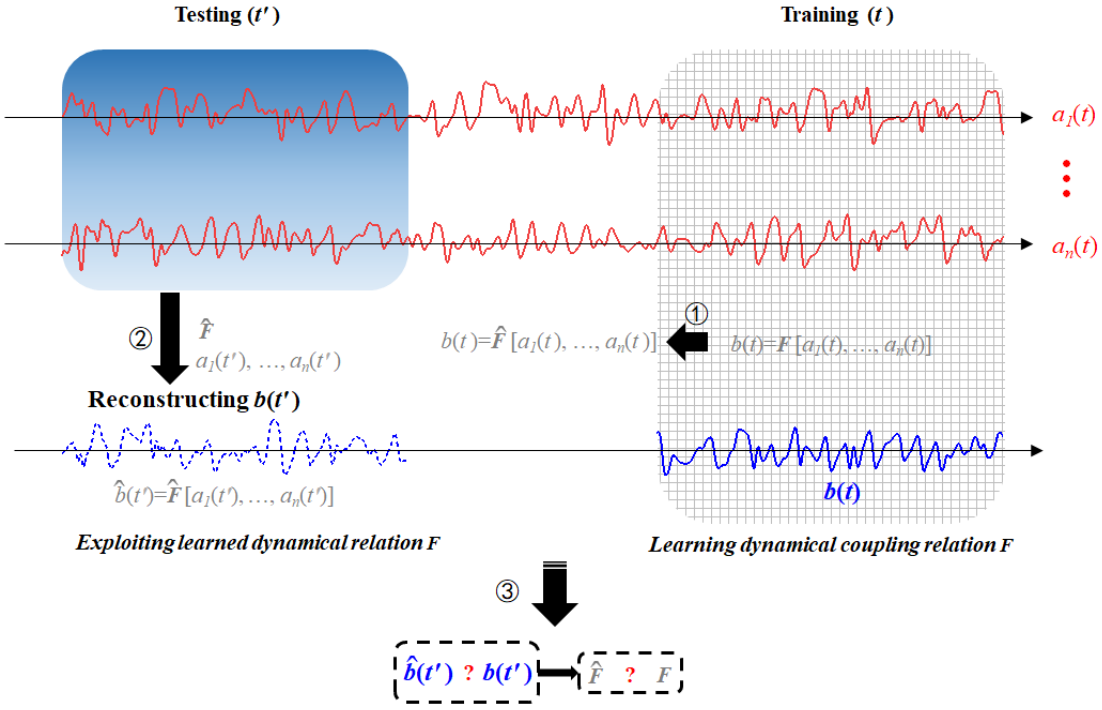

Figure 1 Diagram illustration for reconstructing time series by machine learning. (1) The available part of the dataset $\{a_1(t), \ldots, a_n(t), b(t)\}$ is used to train the neural network. F denotes the inherent coupling relation function among the variables $a_1, \ldots, a_n$, and b, and $\hat{F}$ denotes the inferred coupling relation by neural network. (2) The dataset $\{a_1(t'), a_2(t'), \ldots, a_n(t')\}$ 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, which indicates that the coupling relation is well reconstructed.

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 be taken to estimate $b(t)$ from $a(t)$:

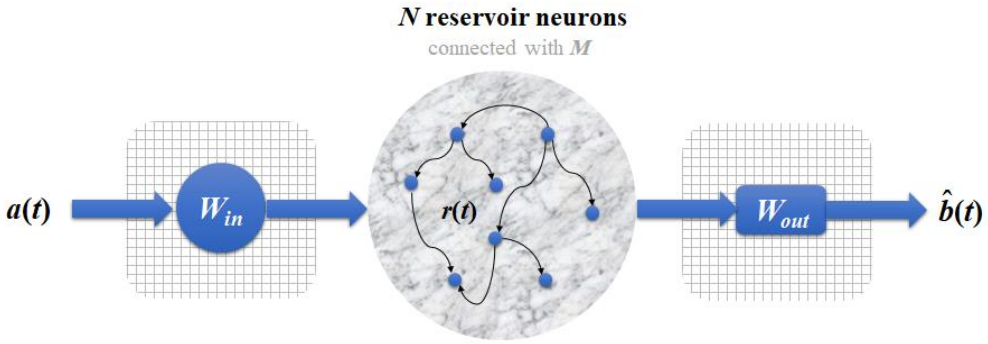

Figure 2 Schematic of the RC neural network: the three layers are input layer, reservoir layer, and output layer. Input layer consists of a matrix "W_{in}". Reservoir layer consists of N reservoir neurons whose connectivity is through the adjacent matrix "M", and $r(t)$ denotes the activation of neurons. Output layer consists of a matrix "W_{out}". $a(t)$ and $\hat{b}(t)$ denote the input and output time series.

(i) $a(t)$ (a vector with length L) is input into the input layer and reservoir layer. There are four components in these layers: 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 choice yields the good performance). These components are employed to derive

output as an updated reservoir state $r^*(t)$:

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

(ii) $r^*(t)$ then gets into the output layer that consists of the reservoir-to-output matrix "W_{out}". And $r^*(t)$ will be used to estimate the value of $\hat{b}(t)$ (see Eq. (2)).

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

The mathematical form of "W_{out}" is defined as

$$W_{out} = \arg\min_{W_{out}} \| W_{out} \cdot r^*(t) - b(t) \| + \alpha \| W_{out} \|, \quad (3)$$

and it 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.

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

2.2.2 Back propagation based artificial neural network

Here, the used BP neural network is a traditional neural computing framework, and it has been widely used in climate research (Watson, 2019; Reichstein et al., 2019; Chattopadhyay et al., 2020). 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 of the temporal variation of time series (Reichstein et al., 2019; Chattopadhyay et al., 2020), while the neurons of RC can track temporal evolution (such as the neuron state $r(t)$ in Fig. 2) (Chattopadhyay et al., 2020). If $a(t)$ and $b(t)$ are two time series of a system, through the BP neural network, we can 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., 2020). As Fig. 3 shows, LSTM has a series of components: a memory cell, an input gate, an output gate, and a forget gate. 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 governing 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.

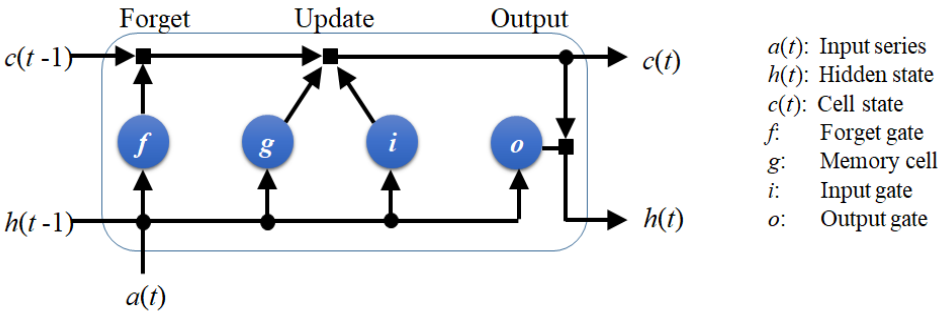

Figure 3 Schematic of the LSTM architecture. LSTM has a memory cell, an input gate, an output gate, and a forget gate to control the information of the previous time to flow into the neural network.

The crucial improvement of LSTM on the traditional recurrent neural network (Reichstein et al., 2019) is, that LSTM has a 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 (Kratzert et al., 2019; Reichstein et al., 2019; Chattopadhyay et al., 2020), and this is 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

The Root Mean Square Error (RMSE) of residuals is used here to evaluate the quality of reconstruction (Hyndman and Koehler, 2006). The residual represents the difference between the real series $b(t')$ and the reconstructed series $\hat{b}(t')$, and it is defined as

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

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

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

2.4 Coupling detection

2.4.1 Linear correlation

As mentioned in the introduction, 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

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

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 (CCM) method that has been demonstrated to be useful for many complex nonlinear systems (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 $M_a(t_i) = \{a_{t_i},\ a_{t_i\ -\tau}, \ldots, a_{t_i\ -\ (m-1)\tau}\}$ ("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), and it is defined as:

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

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

$d\,[M_a(t), M_a(t_i)]$ 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 $b(t)$ through

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

$\hat{b}(t)$ 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:

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

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. (Sugihara et al. 2012) and Tsonis et al. (Tsonis et al. 2018) defined the causal inference according to $\rho_{a \to b}$ and $\rho_{b \to a}$ as: (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 previous studies (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 influences 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). By this way, 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 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 can 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 with Hurst exponent 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 step was taken as 0.01. The time series $X(t)$ and $Z(t)$ are used for the reconstruction

299 analysis.

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

304
$$\begin{aligned} \frac{dX_k}{dt} &= X_{k-1}(X_{k+1} - X_{k-2}) - X_k + F - \frac{h_1}{\mathrm{J}}\sum_{j=1}^{J} Y_{j,k} \\ \frac{dY_{k,j}}{dt} &= \frac{1}{\theta}[Y_{k,j+1}(Y_{k,j-1} - Y_{k,j+2}) - Y_{k,j} + h_2 X_k]. \end{aligned} \tag{13}$$

305 In the first layer of the Lorenz 96 system, there are 18 variables marked as X_k (k is an integer ranging
306 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
307 as fellows: J = 20, h_1 = 1, h_2 = 1, and F=10. The parameter θ can alter the coupling strength: when
308 θ is decreased, the coupling strength between X_k and $Y_{k,j}$ will be enhanced. The fourth-order
309 Runge-Kutta integrator and periodic boundary condition are adopted (that is: $X_0 = X_K$ and $X_{K+1} = X_1$;
310 $Y_{k,0} = Y_{k-1,J}$ and $Y_{k,J+1} = Y_{k+1,1}$), and the integral time step is taken as 0.05. The time series $X_1(t)$ and
311 $Y_{1,1}(t)$ are used for the reconstruction analysis.

312 ## 3.2 Real-world climatic time series

313 TSAT, NHSAT and the Nino3.4 index are chosen as the example from real-world climatic time
314 series used for reconstruction analysis. The original data was obtained from National Centers for
315 Environmental Prediction (https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis2.html)
316 and KNMI Climate Explorer (http://climexp.knmi.nl). The series of TSAT and NHSAT were
317 obtained from the regional average of gridded daily data in NCEP Reanalysis 2. The selected spatial
318 range is 20^0N–20^0S for the tropics and 20^0N–90^0N for the Northern Hemisphere. The selected

319 temporal range is from 1981/09/01 to 2018/12/31.

320 **Training and testing datasets:** Before analysis, all the used time series are standardized to
321 take zero mean and unit variance so that any possible impact of mean and variance on the statistical
322 analysis is avoided (Brown, 1994; Hyndman and Koehler, 2006; Chattopadhyay et al., 2020). The
323 total series were divided into two parts: 60% of the time series training the neural network and 40%
324 being the testing series. Specific data lengths of the training series and testing series will be also
325 listed in the results section.

326 # 4 Results

327 ## 4.1 Coupling relation learning

328 ### 4.1.1 Linear coupling relation and machine learning

329 We first consider the simplest case: the linear coupling relation between two variables. Here,
330 two time series $x(t)$ and $\varepsilon(t)$ in ARFIMA (3, 0.2, 3) model are analyzed. Obviously, there are
331 different temporal structures in $x(t)$ and $\varepsilon(t)$, especially for their large-scale trends (Fig. 4a) and
332 power spectra (Fig. 4b). The marked difference between $x(t)$ and $\varepsilon(t)$ lies in their low-frequency
333 variations, and there are more low-frequency and larger-scale structures in $x(t)$ than in $\varepsilon(t)$. We
334 employ neural networks (RC, LSTM, LSTM*, and BP) to learn the dynamics of this model (Eq. (11))
335 through the procedure shown in Fig. 1. The training parts of $\varepsilon(t)$ are selected from the gray shadow
336 in Fig. 4a. RC, LSTM, LSTM*, and BP are trained to learn the coupling relation between $x(t)$ and
337 $\varepsilon(t)$. Then, the trained neural networks together with $\varepsilon(t)$ are used to reconstruct $x(t)$. The
338 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.

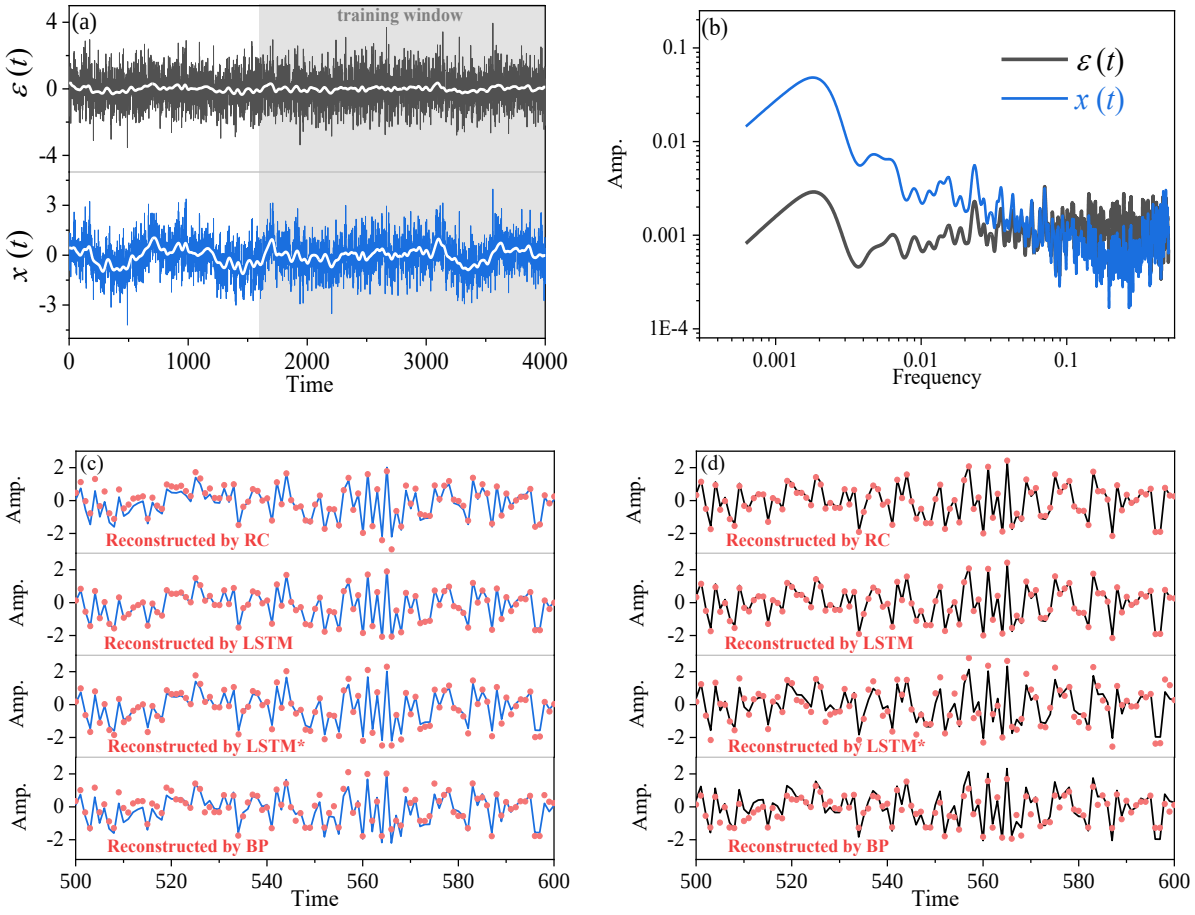

Figure 4 (a) The $x(t)$ time series (blue) and the $\varepsilon(t)$ time series (black) of the ARFIMA(3,0.2,3) model. White lines denote the results of 50-step running average. (b) Comparison of power spectra between $x(t)$ and $\varepsilon(t)$. (c) Comparison of the reconstructed time series of $x(t)$ by RC, LSTM, LSTM* and BP respectively (red dots), and blue lines denote the real $x(t)$. (d) Comparison of the reconstructed time series of $\varepsilon(t)$ through RC, LSTM, LSTM* and BP respectively (red dots), and black lines denote the real $\varepsilon(t)$.

Detailed comparisons between the real and reconstructed series are shown in Figs. 4c and 4d. When $\varepsilon(t)$ is input, the trained RC and LSTM neural networks can be applied to accurately

reconstruct $x(t)$. When $x(t)$ is reconstructed from $\varepsilon(t)$ through LSTM, the minimum of nRMSE (0.01) is reached; all reconstructed series are nearly overlapped with the real ones and cannot be visually differentiated (see Fig. 4c). For RC, the reconstruction quality is also good. The good performance of LSTM benefits from its memory function for the past information (Reichstein et al., 2019; Chattopadhyay et al., 2020). When the memory function of LSTM is stopped, the reconstruction of LSTM* is no longer better than that of 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 may 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., 2020).

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 be applied to learn the underlying coupling dynamics? To address this question, two nonlinearly coupled time series $X(t)$ and $Z(t)$ in the Lorenz 63 system (Lorenz, 1963) are analyzed.

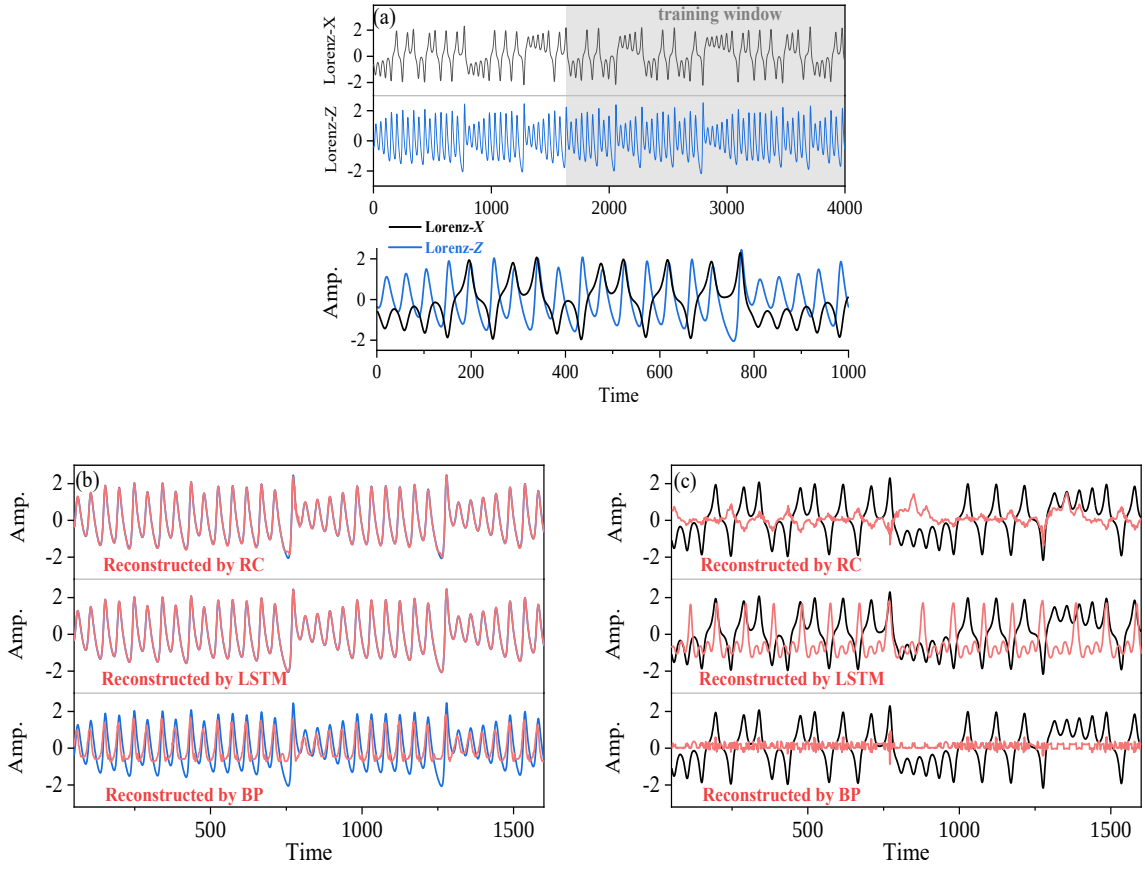

Figure 5 (a) The X time series (black) and the Z time series (blue) of the Lorenz 63 model. (b) Comparison of the reconstructed time series of Z (red) through RC, LSTM and BP respectively. Blue lines denote the real $Z(t)$. (c) Comparison of the reconstructed time series of X (red) through RC, LSTM and BP respectively. Black lines are the real $X(t)$.

There is a very weak overall linear correlation between variables X and Z (Pearson correlation: 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 $X(t)$ and $Z(t)$, which leads to an overall weak linear correlation. In this case, we cannot use a feasible linear regression model between $X(t)$ and $Z(t)$ 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).

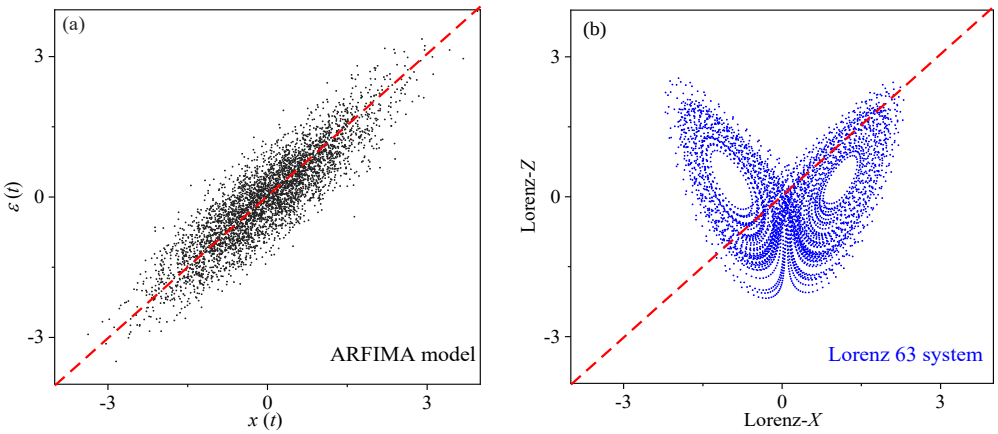

Figure 6 (a) Scatter plot of $x(t)$ and $\varepsilon(t)$ of the ARFIMA(3,0.2,3) model. (b) Scatter plot of X time series and Z time series of the Lorenz 63 model.

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 nonlinearly 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 $X(t)$ to the neural network, not only the information

content of $X(t)$, but also the information content of $Z(t)$ can be learned by the neural network. And then it is possible to reconstruct $Z(t)$ from the trained neural network. We will test it in the following.

Figure 5b shows the results of reconstructing Z time series through RC, LSTM and BP respectively. 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), and this 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 need 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 can 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 bi-directional (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. Furthermore, when the linear correlation is weak but the nonlinear coupling is strong, will the bi-directional reconstruction be still allowed? The answer is usually **NO**. For example, when comparing the reconstruction quality of reconstructing $Z(t)$ from $X(t)$ (Fig. 5b) with that of reconstructing $X(t)$ from $Z(t)$ (Fig. 5c), all of the used machine learning methods fail in reconstructing $X(t)$ from $Z(t)$ (large values of nRMSE are all close to 0.3). This result is consistent with the nonlinear observability mentioned by Lu et al. (Lu et al., 2017). The reconstruction direction is no longer bi-directional in this nonlinear system, but the reconstruction quality is direction-dependent and variable-dependent.

Therefore, we further discuss how to select the suitable explanatory variable or the 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 bi-directional CCM coefficients between the variables X and Z are asymmetric (with a stronger $\rho_{X\to Z}=0.91$ but weaker $\rho_{Z\to X}=0.03$), and then variable Z can be well reconstructed from variable X by machine learning but X cannot be reconstructed from 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 the coupling relation 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

Choosing direction and variable is important for the application of neural networks in 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 the Lorenz 96 model. Furthermore, we will investigate the association between the CCM index and reconstruction quality in the machine learning frameworks.

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 $\rho_{X_1\to Y_{1,1}}=0.61$. This indicates that reconstructing X_1 from $Y_{1,1}$ may obtain a better quality than from

X_1 to $Y_{1,1}$. As expected, by means of RC, the error of reconstructing X_1 from $Y_{1,1}$ is: nRMSE = 0.01, however it is nRMSE = 0.06 in the opposite direction (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, performance of the reconstruction through BP and LSTM* is not good and just as analyzed in section 4.2.3.

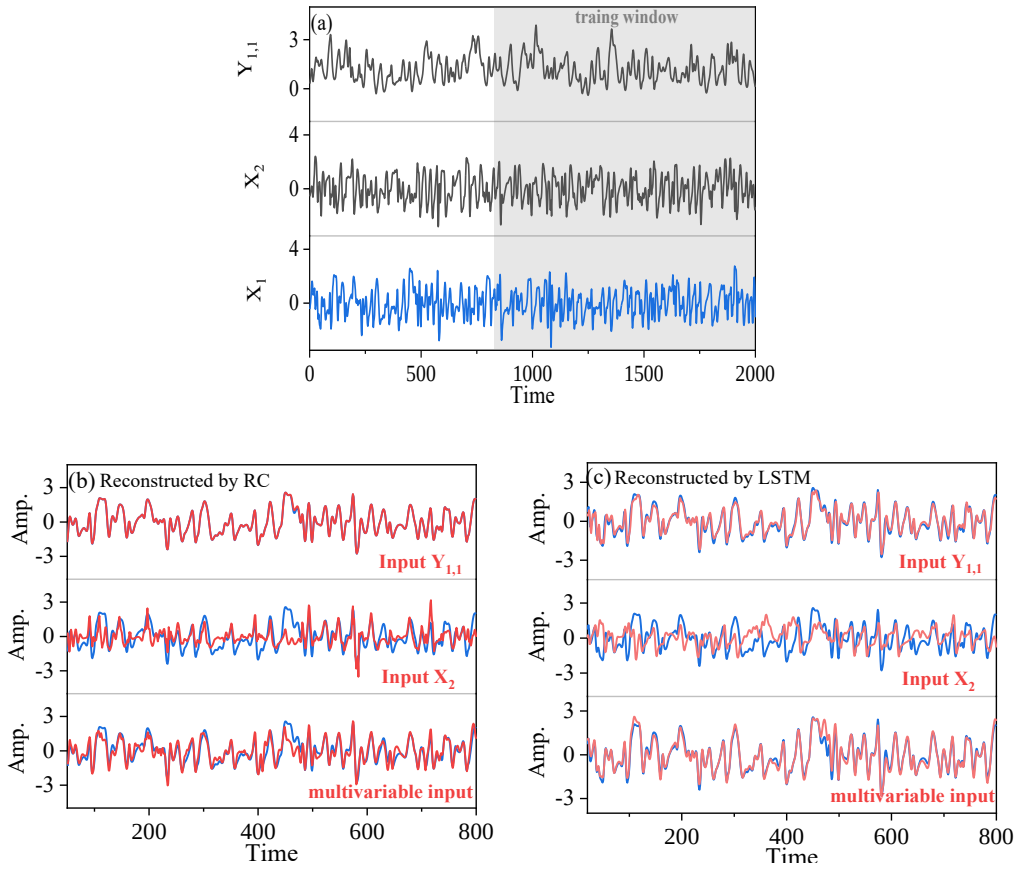

Figure 7 (a) The $Y_{1,1}$ time series (black), X_2 time series (black) and X_1 time series (blue) of the Lorenz 96 system. (b) Through RC, when $Y_{1,1}$, X_2 and multivariate are acting as the explanatory variable respectively, the corresponding reconstructed X_1 time series (red) are shown, and blue lines denote the real X_1 time series. (c) Through LSTM, when $Y_{1,1}$, X_2 and multivariate (multiple variables: X_2, X_{17} and X_{18}) are acting as the explanatory variable respectively, the corresponding reconstructed X_1 time series (red) are shown, and blue lines denote the real

X_1 time series.

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
X_1	***$Y_{1,1}$***	**-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
X_1	***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 the reconstructed 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 that of RC, and the reconstruction results by LSTM are consistent with the indication of the CCM index. Chattopadhyay et al. (2020) also suggests that LSTM performs worse than RC in some cases, and this might be related to the use of a simple variant of the LSTM architecture. This variant of LSTM was tested and it was found that the time-varying local mean in time series would sometimes influence its performance. However further investigation is required for a deeper understanding of the real reason. 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 through RC and LSTM (see Fig. 7b and 7c).

Besides, the number of the chosen explanatory variables also influences the reconstruction quality. If more than one explanatory variable in the same layer is used, 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 the lower error than those from unit-variable reconstruction.

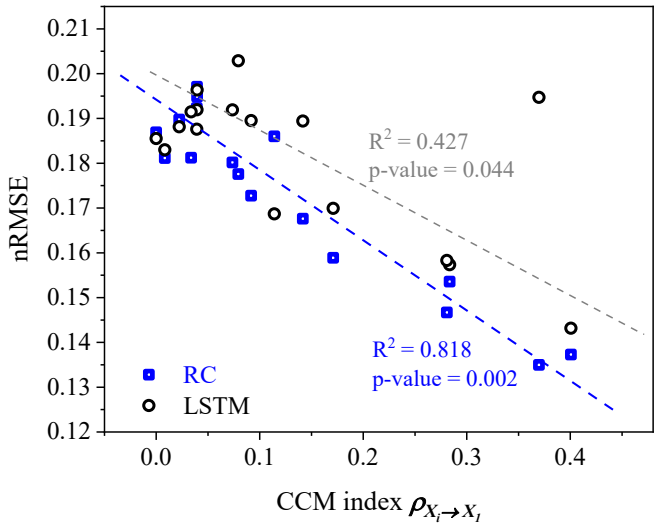

Figure 8 Scatter plot of nRMSE values and the CCM index values. Blue and grey dashed lines are the fitted linear trends for the scatters.

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 performance of RC and LSTM. The values of the CCM index are calculated between X_1 and X_2, X_3 …, X_{18} respectively; meanwhile, X_1 is reconstructed from X_2, X_3 …, X_{18} respectively. We find a

497 significant correspondence between the nRMSE and the CCM index (Fig. 8), for both results of RC
498 and LSTM. Here we only use a simple LSTM architecture, and there are many other variants of this
499 architecture where the abnormal point of LSTM in Fig. 8 might be reduced. The result of Fig. 8
500 reveals the robust association between the CCM index and reconstruction quality in the machine
501 learning frameworks of RC and LSTM. For other machine learning methods, such association
502 deserves further investigation.

503 ### 4.2.3 Performance of BP and LSTM* in Lorenz 96 system

504 In nonlinear systems, the performance of reconstruction through BP and LSTM* is much worse
505 than that of RC and LSTM (Fig. 5). Here we present a simple experiment, to illustrate what might
506 influence the performance of BP and LSTM* in a nonlinear system.

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

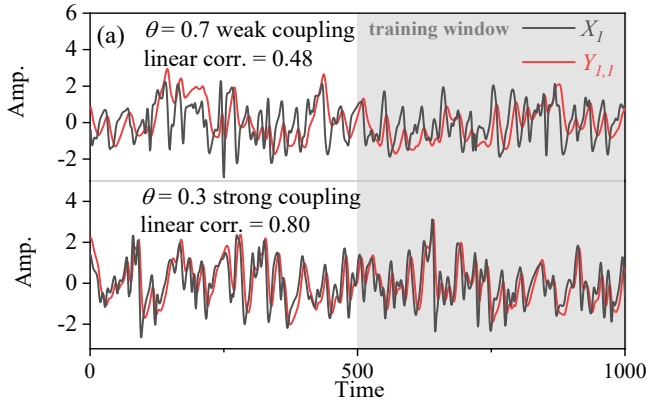

Figure 9 Influence of strong nonlinear coupling on linear Pearson correlation and machine learning performance. (a) Comparison of the linear correlation with different coupling strength. (b) Comparison of the machine learning performance with different coupling strength. The black lines are the real series; the reconstructed series by RC (green lines), LSTM*(blue lines) and BP (red dots) are shown respectively (here the results of LSTM are overlapped with that of RC).

However, RC and LSTM are not restricted to the Pearson correlation in this nonlinear system. When θ is altered from 0.7 to 0.3, although the Pearson correlation is changed a lot, the values of the CCM index are kept consistently above 0.9. For all values of θ, RC is able to equally produce a good quality reconstruction of X_1. Fig. 9b shows that the reconstructed series through RC and LSTM always overlap with the real time series. These results indicate that the performance of both RC and LSTM is sensitive to the value of CCM index, which is in line with the results given in section 4.2.2.

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.

The daily NHSAT and TSAT time series are shown in Fig. 10a. There are quite different temporal patterns in the NHSAT and TSAT series, with 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 method fail to reconstruct one series from the other. Meanwhile, there is no explicit physical expression that can transform TSAT and NHSAT to each other. Now we try to use machine learning to learn their coupling between them and then 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. This finding 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 that from the opposite direction.

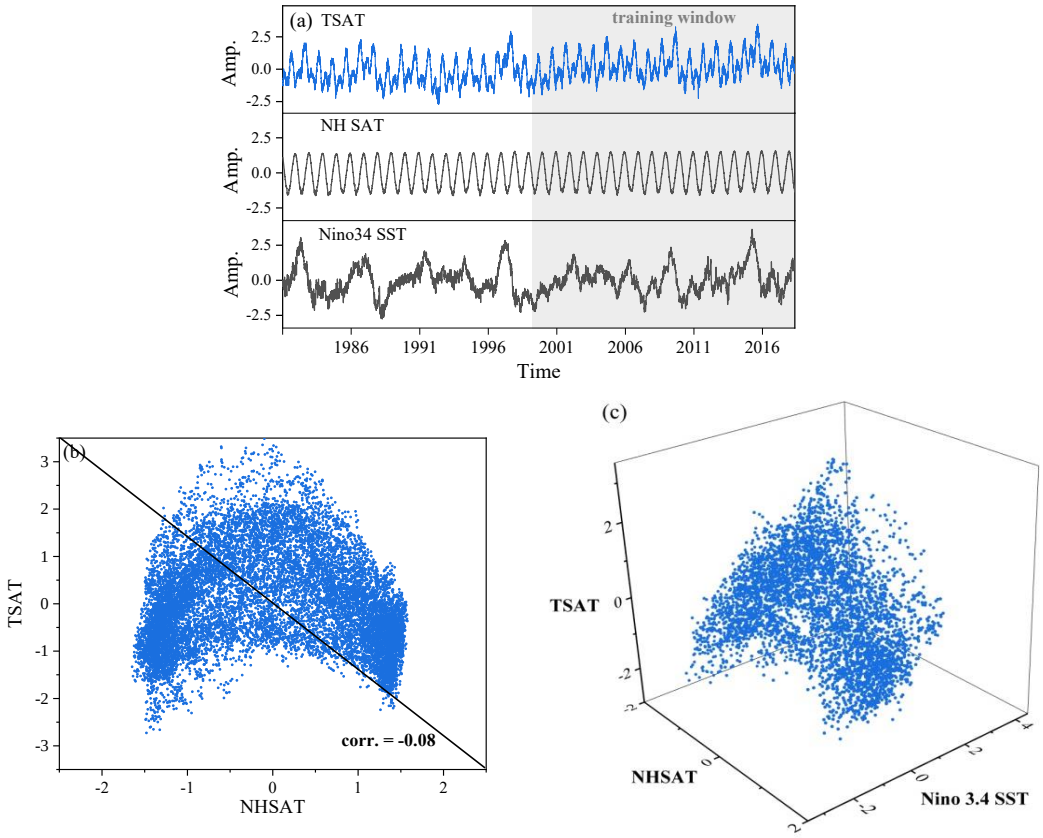

Figure 10 (a) Daily time series of TSAT, NHSAT and Nino 3.4 index. (b) Scatter plot of normalized NHSAT and normalized TSAT. (c) Three-dimensional scatter plot of normalized NHSAT, normalized TSAT and normalized Nino 3.4 SST.

The results validate 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. This nRMSE 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 real 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 worse

performance may be due to its inability to deal with nonlinear coupling. LSTM performs worse than RC in this real-world case might be induced by the used simple variant of LSTM architecture.

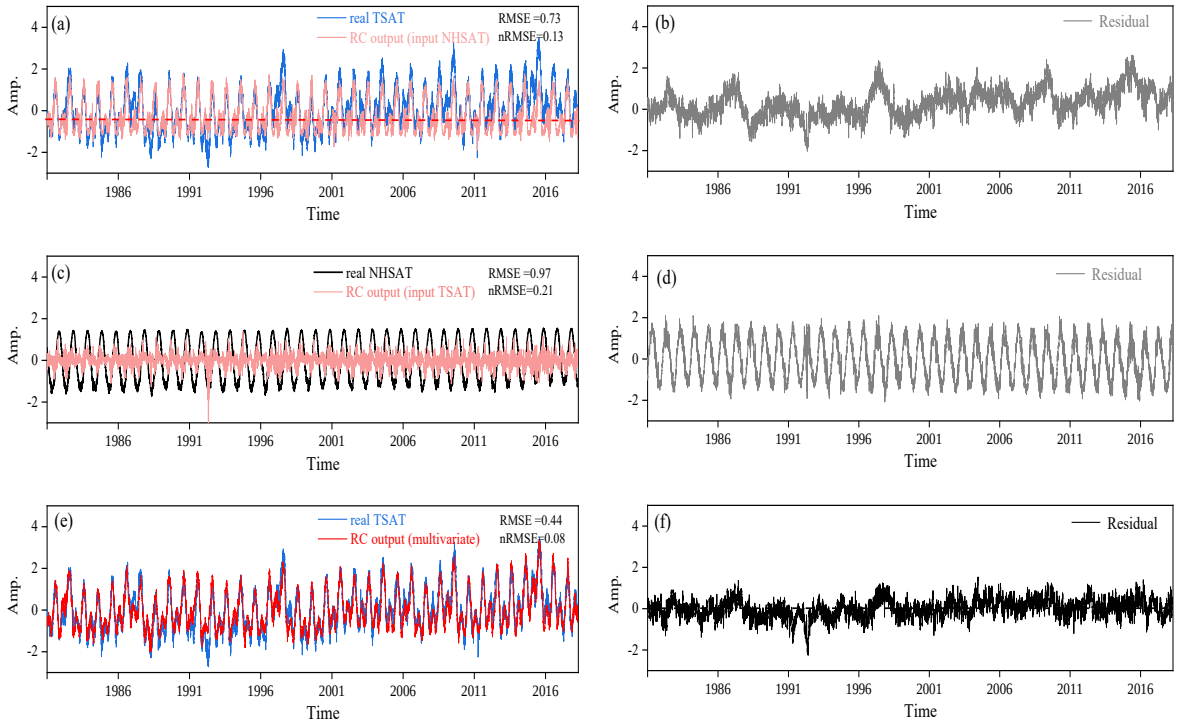

Figure 11 (a) Reconstructed TSAT time series from NHSAT, and its residual series (b); (c) Reconstructed NHSAT time series from TSAT, and its residual series (d); (e) Reconstructed TSAT time series from NHSAT and Nino3.4 index, and its residual series (f).

We can further improve the reconstruction quality of TSAT. Considering that the tropical climate system interacts not only with the Northern Hemisphere climate system, we can use the information of other systems to improve the reconstruction. Looking at the time series of Nino 3.4 index (Fig. 10a), some of its extremes occur at the same time intervals as the extremes of TSAT. Moreover, when Nino 3.4 index is included into the scatter plot (Fig. 10c), a nonlinear attractor structure is revealed. We combine NHSAT with Nino 3.4 index to reconstruct the time series of TSAT through RC. The reconstructed TSAT (Fig. 11e) is much closer to the real TSAT series, and

576 the corresponding nRMSE has been reduced to 0.08.

577 **Table 4** Details of temperature records' reconstruction

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

578 Finally, we make a further comparison between the real TSAT and the reconstructed TSAT: (i)
579 the annual variations of the reconstructed TSAT are close to those of the real TSAT (Fig. 12a). (ii)
580 The power spectra of TSAT and the reconstructed TSAT are compared in Fig. 12b, and the main
581 deviation occurs at the frequency bands of 0-15 days. The reason might be that the local weather
582 processes are not input into this RC reconstruction. This conjecture can be further confirmed
583 through a red-noise test with response time 15 days for the residual series (this red-noise test is the
584 same as the method used in Roe, 2009). All data points of the residual series lie within the
585 confidence intervals (Fig. 12c). This means that the residual is possibly induced by local weather
586 processes, and this information is not input into RC for the reconstruction.

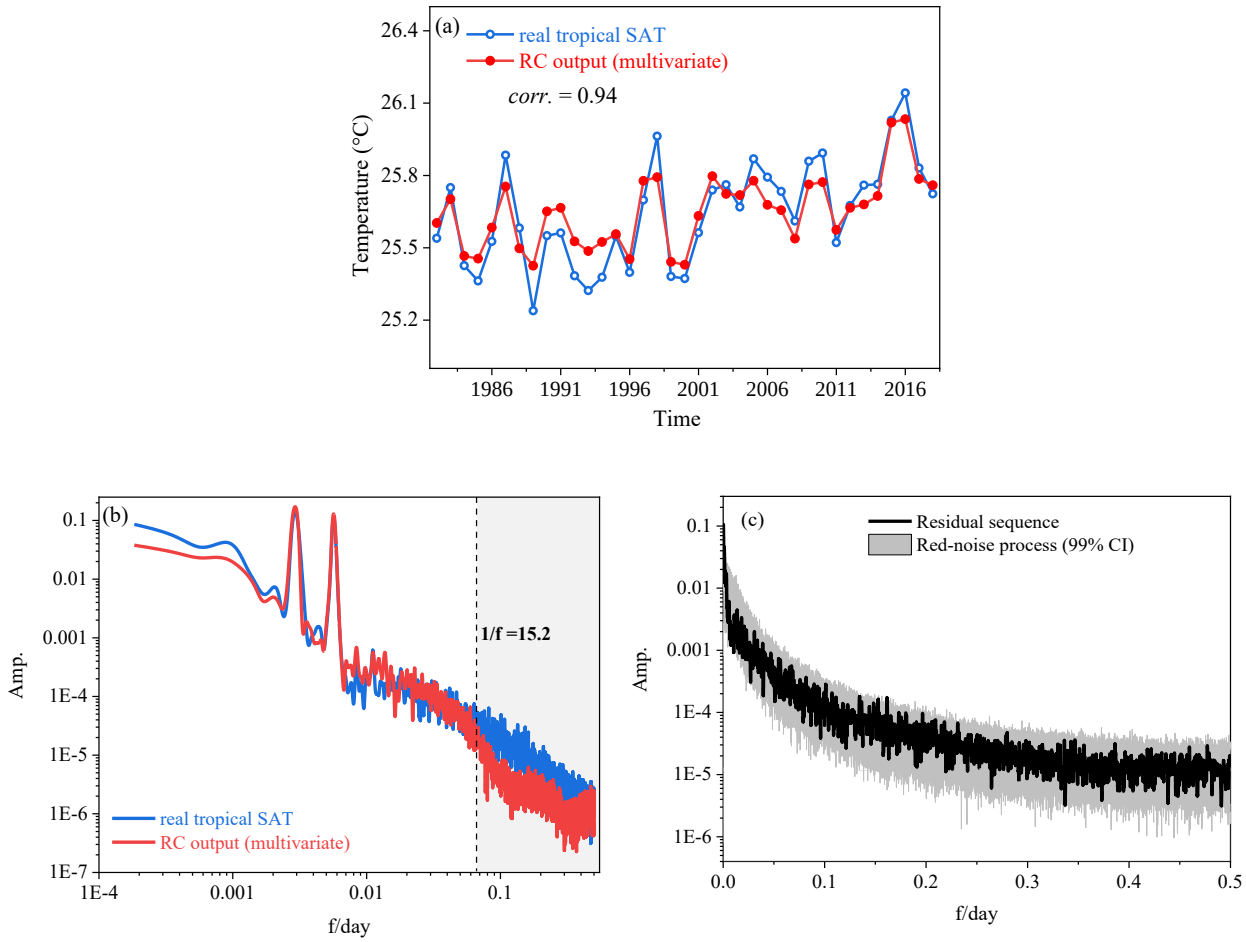

Figure 12 (a) Comparison of the annual mean values between the reconstructed TSAT and real TSAT. (b) Comparison of the power spectrum between the reconstructed TSAT and real TSAT. (c) A red-noise test for residual series.

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 through the linear mechanism, and a large Pearson coefficient can benefit to machine learning with bi-directional reconstruction. For a nonlinear system, the coupled time series

598 often have a small Pearson coefficient, but machine learning can still well reconstruct the time series
599 when the CCM index is strong; moreover, the reconstruction quality is direction-dependent and
600 variable-dependent, which is determined by the coupling strength and causality between the
601 dynamical variables.

602 Choosing suitable explanatory variables is crucial for obtaining a good reconstruction quality.
603 But the results show that machine learning performance cannot be explained only by linear
604 correlation. In this study, we suggest to use the CCM index to select explanatory variables.
605 Especially for the time series of nonlinear systems, the strong CCM index can be taken as a
606 benchmark to select an explanatory variable. When the CCM index is higher than 0.5 in this study,
607 then the nRMSE is often smaller than 0.1 with the reconstructed series very close to the real series in
608 the presented results. Thus, the CCM index higher than 0.5 may be considered as a criterion for
609 choosing appropriate explanatory variables. It is well known that atmospheric or oceanic motions
610 are nonlinearly coupled over most of time scales, and therefore, in the natural climate series, there
611 would be similar nonlinear coupling relation as found in the Lorenz 63 and the Lorenz 96 systems
612 (the weak Pearson correlation but the high CCM coefficient). If only Pearson coefficient is used to
613 select the explanatory variable, then some useful nonlinearly correlated variables may be left out.

614 Finally, it is worth noting the potential application for machine learning in the climate studies.
615 For instance, a series $b(t)$ is unmeasured during some periods for the measuring instrument failure,
616 but there are other kinds of variables without missing observations. Then, CCM can be applied to
617 select the suitable variables coupled with $b(t)$, and RC or LSTM can be employed to reconstruct the
618 unmeasured part of $b(t)$ (following Fig. 1). This is useful for some climate studies, such as
619 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

Governing equations for the LSTM neural network

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

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

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

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

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

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

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

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

$f(t)$, $i(t)$, and $o(t)$ denote 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 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"

641 and $''W_{oh}''$ denote the neuron connectivity in each layers. These parameters need to be computed

642 during training (Chattopadhyay et al., 2020). $a(t)$ and $\hat{b}(t)$ represent the input and output time series.

643

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 editor, the two anonymous reviewers and Dr. Zhixin Lu for their constructive suggestions. We also thank Dr. Christian L.E. Franzke and Dr. Naiming Yuan for their in-depth and helpful discussions. We acknowledge the supports from National Natural Science Foundation of China through Grants (No. 41675049 and No. 41975059).

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., and Subramanian, D.: Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network, Nonlin. Processes Geophys., 27, 373–389, 2020.

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.: NeuralHydrology – Interpreting LSTMs in Hydrology. In: Samek W., Montavon G., Vedaldi A., Hansen L., Müller KR. (eds) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol 11700. Springer, Cham, 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

724 next El Niño. P. Natl. Acad. Sci., 111(6), 2064-2066, 2014.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

758 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.