**Reply to the comments of Anonymous Referee #2:**

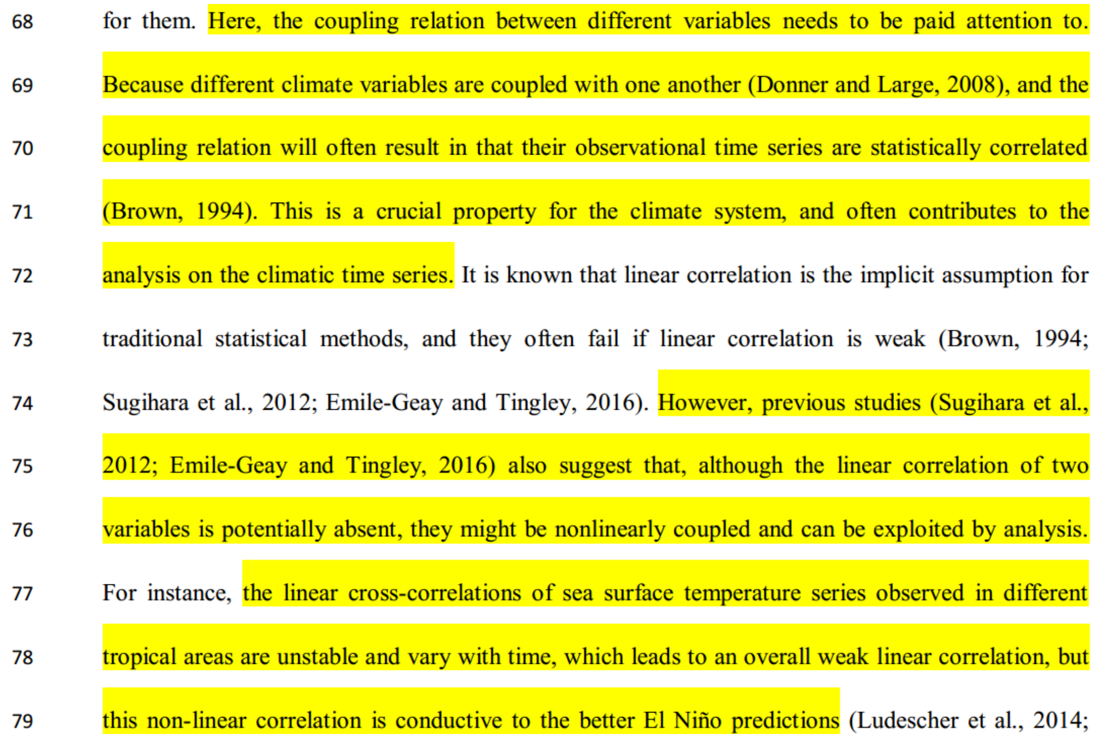
**The comments of Anonymous Referee #2:**

1. This manuscript investigates the feasibility of using Machine Learning (ML) algorithm for the reconstruction of a time series with the help of a coupled time series. The study also examines the ability of an ML algorithm to represent the coupling strength of a system. The reconstruction analysis investigates three ML algorithms: Back Propagation (BP), Long Short-Term Memory (LSTM), and Reservoir Computing (RC). The study also investigates the influence of type of coupling (linear or non-linear) on the performance of ML algorithm. This is achieved by using a simple linear system, a simple non-linear system (Lorenz-63), a high-dimensional non-linear system (Lorenz-96), and a real-world system (coupling between Tropical surface air temperature and Northern Hemisphere surface air temperature). The linearity is measured using Pearson's correlation coefficient while the non-linearity is measure using Convergent Cross Map ping Causality index (CCM). The influence of the direction of coupling and coupling strength, and the number of explanatory variables on the accuracy of reconstruction of different ML algorithms is also examined. The performance evaluation of ML algorithms found that RC is most suitable for the reconstruction of non-linearly coupled time series. The work is scientifically sound and I see a lot of value in this work. Especially in the future applications of ML algorithms for reconstruction of coupled time series and in understanding the influence of coupling mechanisms on the behavior of ML algorithm. However, the presentation of the work in its current form is very confusing and diverts the attention of the reader from the importance of the work. The manuscript has errors related to English too which need to be corrected. Please find my major suggestions on the manuscript below.

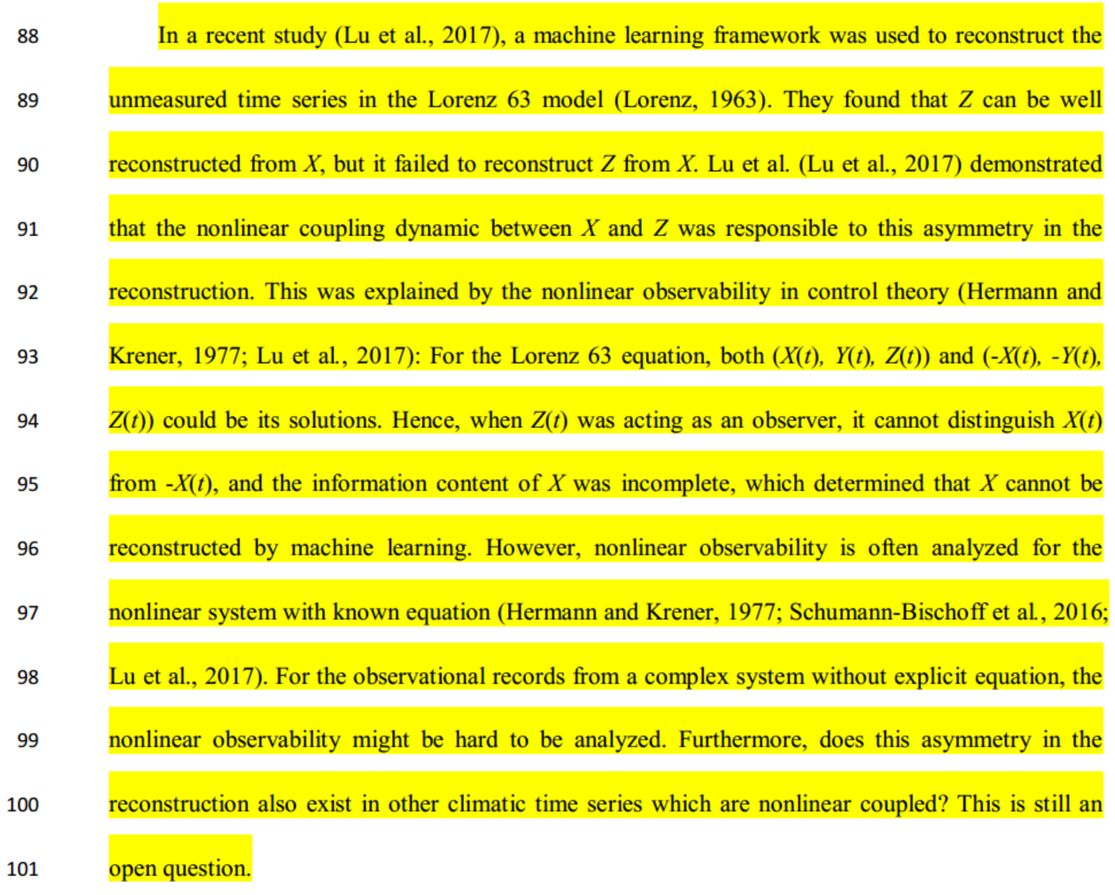
**Response:** Thanks for your thoughtful comments and suggestions! The suggestions are very helpful for improving our manuscript, and we will carefully revise the manuscript according to these suggestions.

2. The abstract talks about the reconstruction of a time series of a coupled system from its other coupled counter-parts. However, the introduction is not representing it intuitively. I would suggest the authors to focus on the problem of reconstruction of a time series and build the importance of coupling mechanism, importance of linear and non-linear coupling around the time series reconstruction.

**Response:** Thank you! In the introduction, we will focus more on the importance of coupling mechanism to the time series reconstruction, and the importance of linear and non-linear correlations. Some of our modification in the introduction is shown by the following screenshot:



for them. Here, the coupling relation between different variables needs to be paid attention to. Because different climate variables are coupled with one another (Donner and Large, 2008), and the coupling relation will often result in that their observational time series are statistically correlated (Brown, 1994). This is a crucial property for the climate system, and often contributes to the analysis on the climatic time series. It is known that linear correlation is the implicit assumption for traditional statistical methods, and they often fail if linear correlation is weak (Brown, 1994; Sugihara et al., 2012; Emile-Geay and Tingley, 2016). However, previous studies (Sugihara et al., 2012; Emile-Geay and Tingley, 2016) also suggest 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;



In a recent study (Lu et al., 2017), a machine learning framework was used to reconstruct the unmeasured time series in the Lorenz 63 model (Lorenz, 1963). They found that $Z$ can be well reconstructed from $X$, but it failed to reconstruct $Z$ from $X$. Lu et al. (Lu et al., 2017) demonstrated that the nonlinear coupling dynamic between $X$ and $Z$ was responsible to 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. Hence, when $Z(t)$ was acting as an observer, it cannot distinguish $X(t)$ from $-X(t)$, and the information content of $X$ was incomplete, which determined that $X$ cannot be reconstructed by machine learning. However, nonlinear observability is often analyzed for the nonlinear system with known equation (Hermann and Krener, 1977; Schumann-Bischoff et al., 2016; Lu et al., 2017). For the observational records from a complex system without explicit equation, the nonlinear observability might be hard to be analyzed. Furthermore, does this asymmetry in the reconstruction also exist in other climatic time series which are nonlinear coupled? This is still an open question.

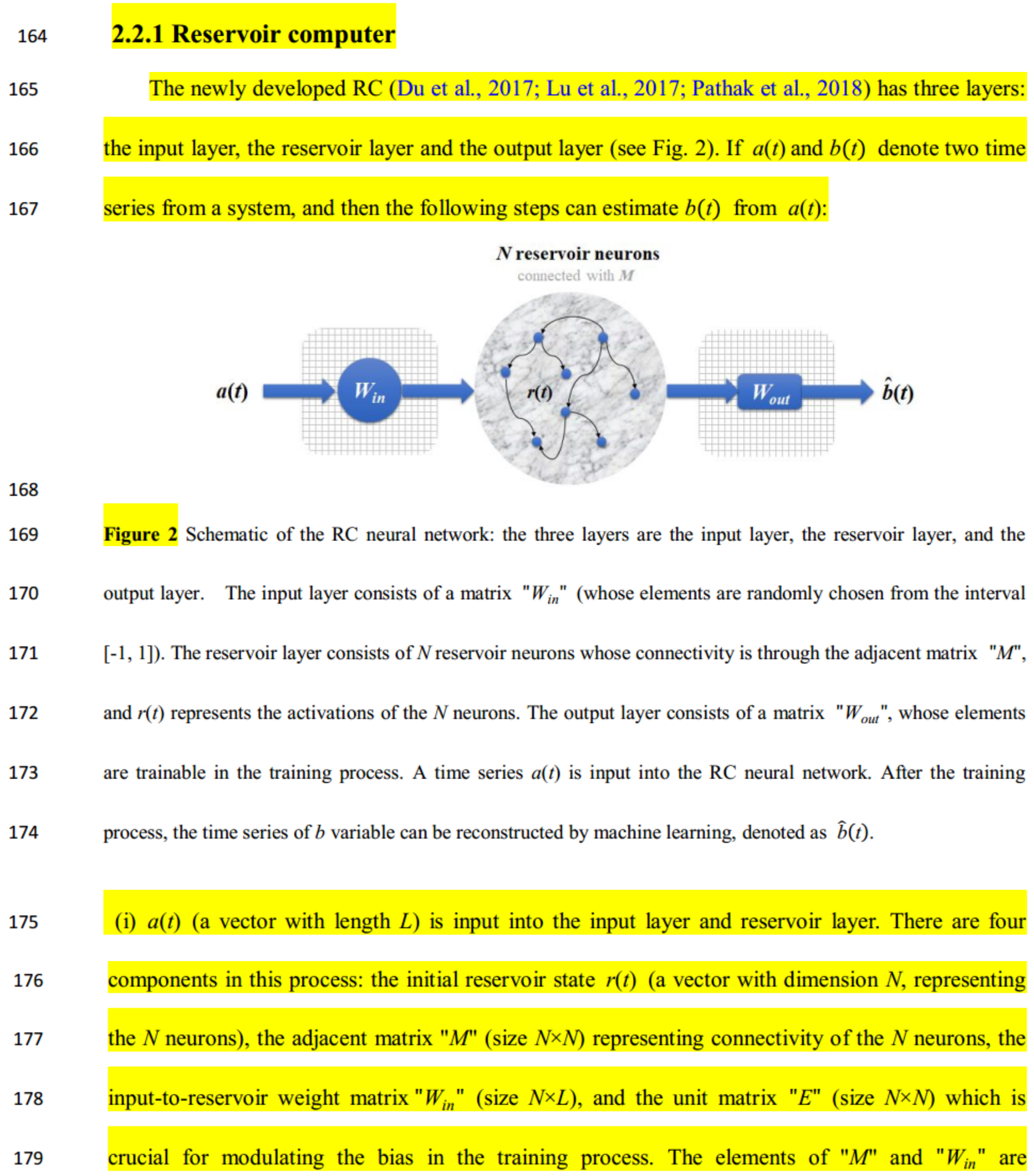
**3.** The Methodology section does not seem to have a description of BP and LSTM in it, in as much detail as stated for RC. I would suggest the authors to incorporate the description of BP and LSTM too, as it will help the readers to better understand the behavior of the algorithms.

**Response:** Thank you! We will add more detailed description of BP and LSTM into the revised manuscript.

But the algorithms of BP are much more complicated than that of RC, and there are too many equations (about 15 mathematical equations) for their algorithms so that the article will be not concise. We will carefully introduce the key steps for BP, and the relevant references will be cited for the steps.

**Especially, we will highlight the crucial differences in algorithms among RC, BP and LSTM, and this might be very helpful for understanding the application results of them.**

Our modification for the neural network algorithms are shown by the following screenshot:



### 2.2.1 Reservoir computer

The newly developed 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. The elements of "$M$" and "$W_{in}$" are randomly chosen from a uniform distribution in [–1, 1], and we set $N = 1000$ here (we found this can yield the good performance). These components are associated with 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 $\alpha$ is the ridge regression coefficient, whose values will be determined after the training.

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

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

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). The crucial difference between the BP and the RC neural networks is as follows: unlike RC, all the neuron state of the BP neural network is independent on the temporal variation of time series, 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 training 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.

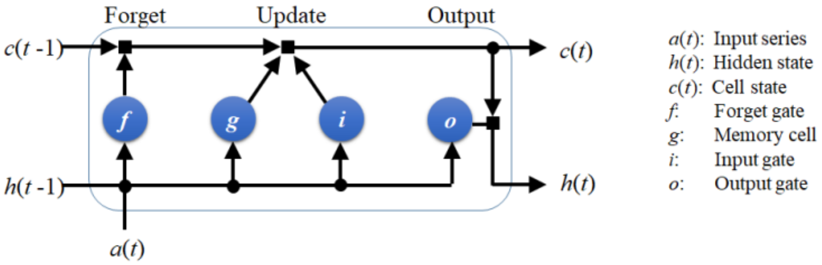


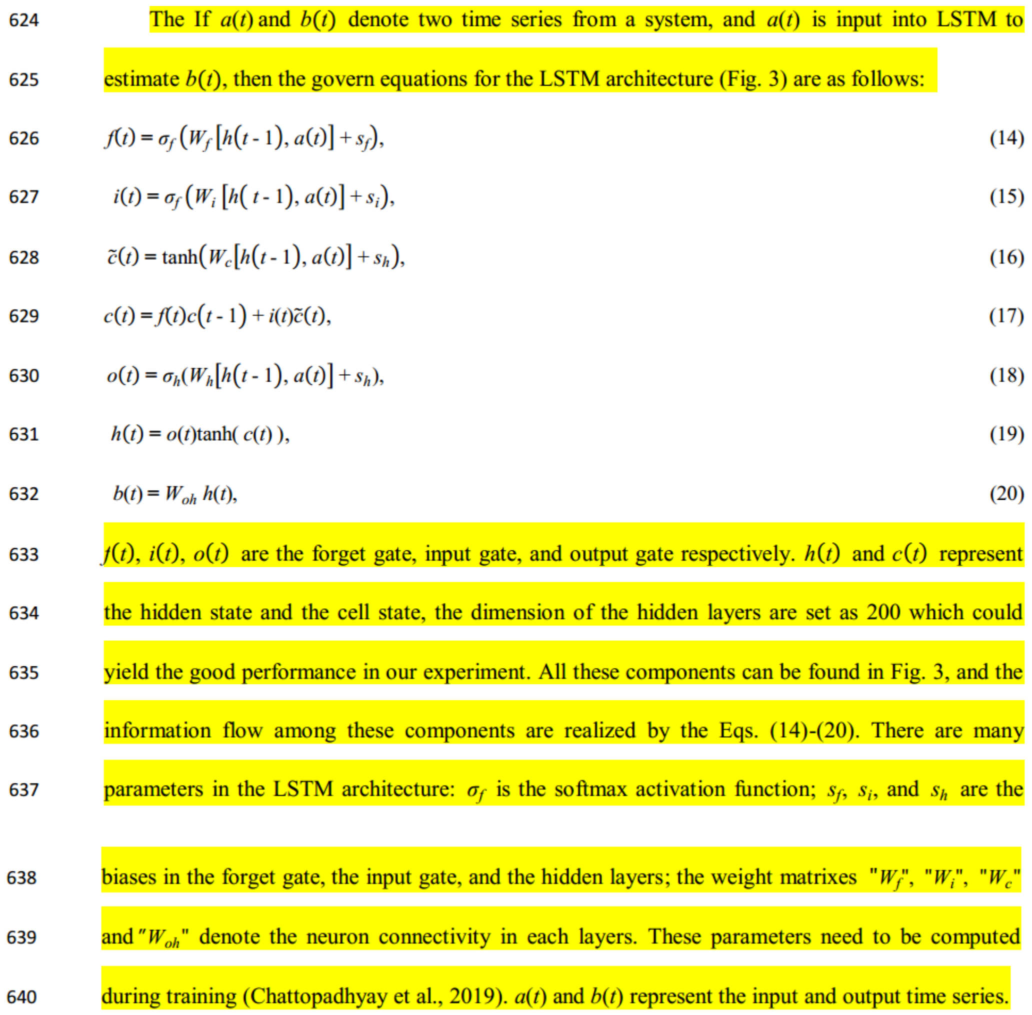
**Figure 3** Schematic of the LSTM architecture. LSTM has a memory cell, input gate, output gate, and a forget gate to control the information of the previous time to flow into the neural network.

The crucial improvement of LSTM on the traditional recurrent neural network, 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 neural state of LSTM has ability to track the temporal evolution of time series (Chattopadhyay et al., 2019; Kratzert 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 demonstrated.

## 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:

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

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

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

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

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

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

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

$f(t)$, $i(t)$, $o(t)$ 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: $\sigma_f$ 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.

**4.** The CCM method has been introduced in the Results section. It should be introduced in the Methodology section. In the discussion of CCM method, relate it with the direction of reconstruction as well (explanatory variable to reconstructed variable)

**Response:** Thank you! We will add the description of the CCM algorithm into the method part of the revised manuscript, and also relate it with the direction of reconstruction. Our modification is shown by the following screenshot:

### 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 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 the 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 ($\tau$) can be determined through the false nearest neighbor algorithm (Hegger and Kantz, 1999).

ii) Estimating the weight parameter $w_i$ denoting 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.

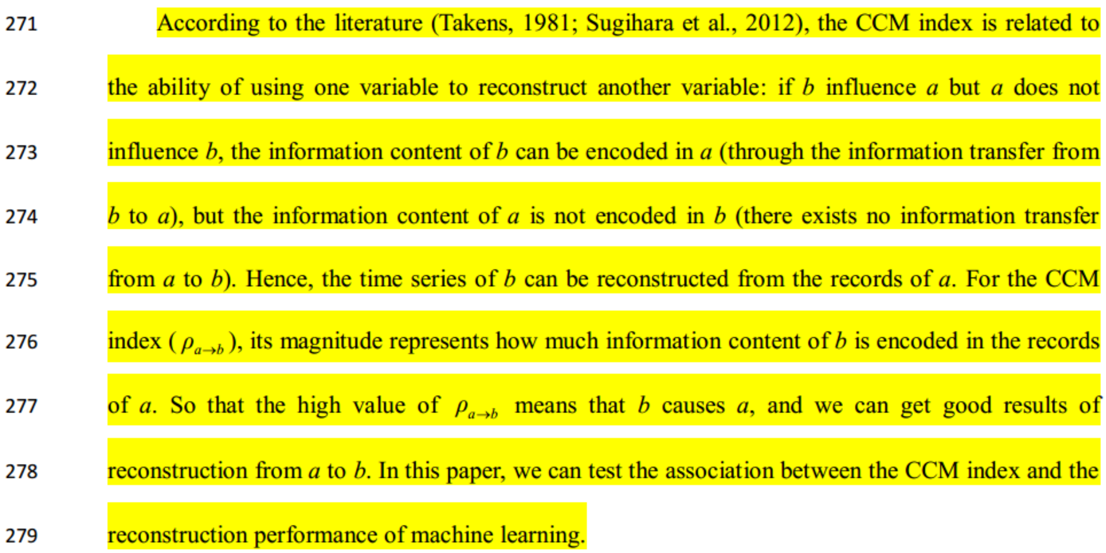
iii) Cross mapping the value of $b(t)$ by

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

$\hat{b}(t)$ 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:

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

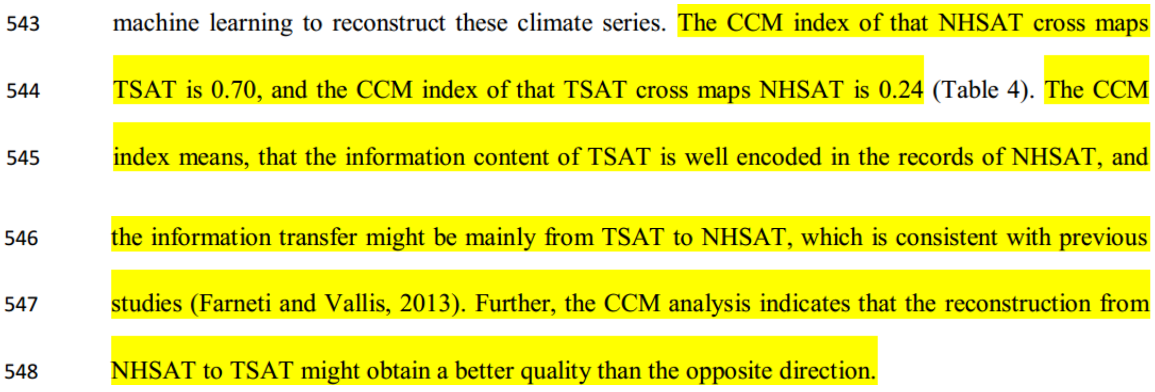
The cross mapping skill from $b$ to $a$ is also measured according to the above steps, marked as $\rho_{b \to a}$ . *Sugihara et al.* and *Tsonis et al.* defined the causal inference from $\rho_{a \to b}$ and $\rho_{b \to a}$ like that: (i) if $\rho_{a \to b}$ is convergent when $L$ is increased, and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



According to the literature (Takens, 1981; Sugihara et al., 2012), 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*). Hence, 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*. So that the high value 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 can test the association between the CCM index and the reconstruction performance of machine learning.

5. Otherwise it is a little confusing to relate the notation of with its notation when it is being applied and shown in the Results section (Line number 462-463).

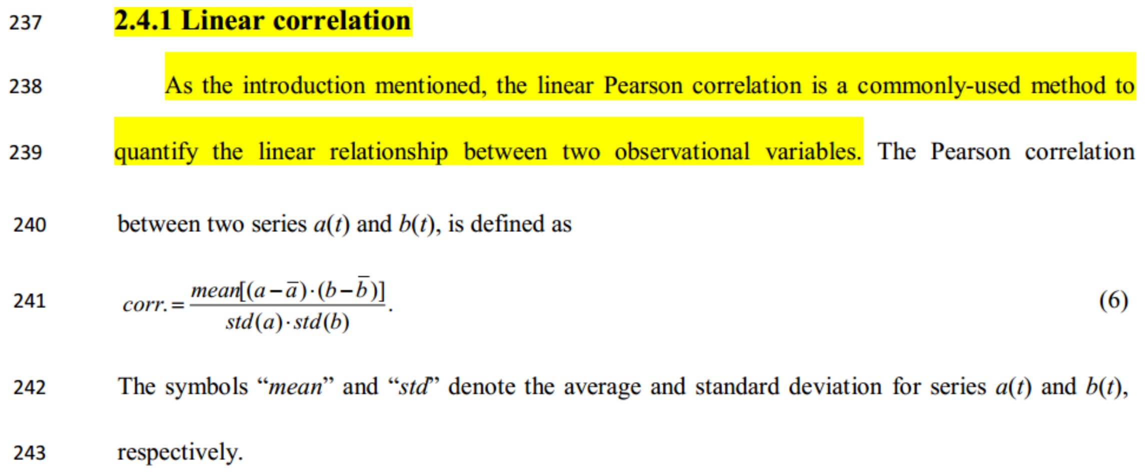
**Response:** Thank you! We will modify this narration, and improve such narration thoroughly in the revised manuscript. Our modification is shown by the following screenshot:



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.

6. The same goes for the description of Pearson’s correlation coefficient, its description should be shifted from the Results to the Methodology section.

**Response:** Thank you! We will move the description of Pearson’s correlation to the method in the revised manuscript. Our modification is shown by the following screenshot:



### 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

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

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

7. The flow of the Results section is hard to follow. **The Results section just lists the author’s observations, from the Figures and Tables, and does not provide any insights into those observations. For example, line number 329 - 330 states that BP and LSTM* are not sensitive to non-linear coupling, but no explanation is given as to why this is so.** The authors should provide more insight into the observed behavior of the ML algorithms mentioned in the Results section.

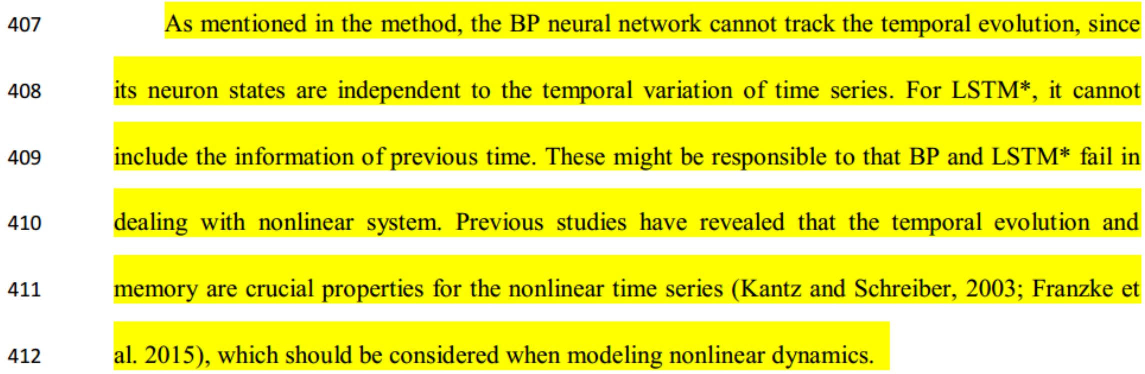
**Response:** Thank you! We will provide more insight into the observed behavior of the ML algorithms mentioned in the Results section. For the analysis on other results, we will also pay more attention to this.

For the results of that BP and LSTM* are not sensitive to non-linear coupling, their algorithms might be responsible to this. **When analyzing their algorithm, we can find that the BP neural network cannot track the temporal evolution, because its neuron states are independent to the temporal variation of time series. For LSTM*, it cannot include the information of previous time. Previous studies have revealed that the temporal evolution and memory are crucial properties for the nonlinear time series** [1, 2]**,** which should be considered when modeling nonlinear dynamics. But the algorithms of RC and LSTM have made improvements on these issues (we have added these contents into the method part of the revised manuscript).

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

[2] 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.

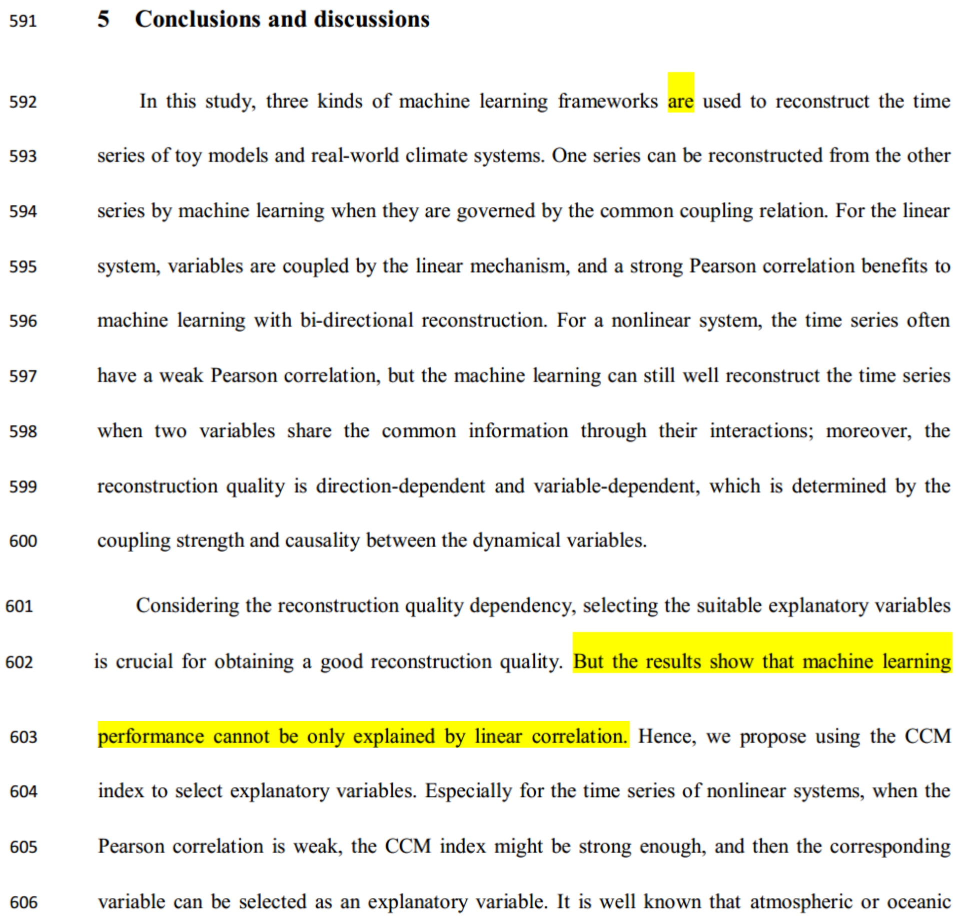
Our modification is shown by the following screenshot:



As mentioned in the method, the BP neural network cannot track the temporal evolution, since its neuron states are independent to the temporal variation of time series. For LSTM*, it cannot include the information of previous time. These might be responsible to that BP and LSTM* fail in dealing with nonlinear system. Previous studies have revealed that the temporal evolution and memory are crucial properties for the nonlinear time series (Kantz and Schreiber, 2003; Franzke et al. 2015), which should be considered when modeling nonlinear dynamics.

8. The conclusion section should be shortened.

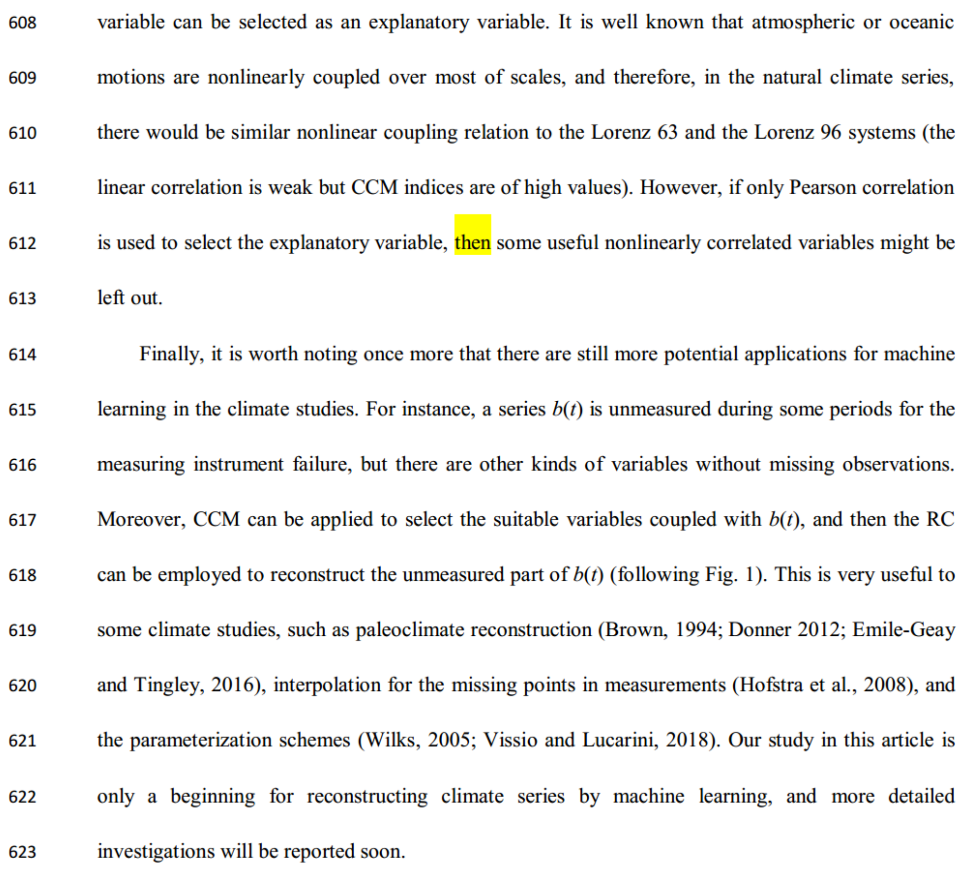
**Response:** Thanks for your suggestion. We will shorten the length of the conclusion, and move part of the discussion into the results part. Our modification for the conclusion is shown by the following screenshot:



### 5 Conclusions and discussions

In this study, three kinds of machine learning frameworks 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 correlation benefits to machine learning with bi-directional reconstruction. For a nonlinear system, the time series often have a weak Pearson correlation, but the machine learning can still well reconstruct the time series when two variables share the common information through their interactions; 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 Pearson correlation is weak, the CCM index might be strong enough, and then the corresponding variable can be selected as an explanatory variable. It is well known that atmospheric or oceanic



variable can be selected as an explanatory variable. It is well known that atmospheric or oceanic motions are nonlinearly coupled over most of 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 linear correlation is weak but CCM indices are of high values). However, if only Pearson correlation is used to select the explanatory variable, then some useful nonlinearly correlated variables might be left out.

Finally, it is worth noting once more that there are still more 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 the RC can be employed to reconstruct the unmeasured part of $b(t)$ (following Fig. 1). This is very useful to 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.

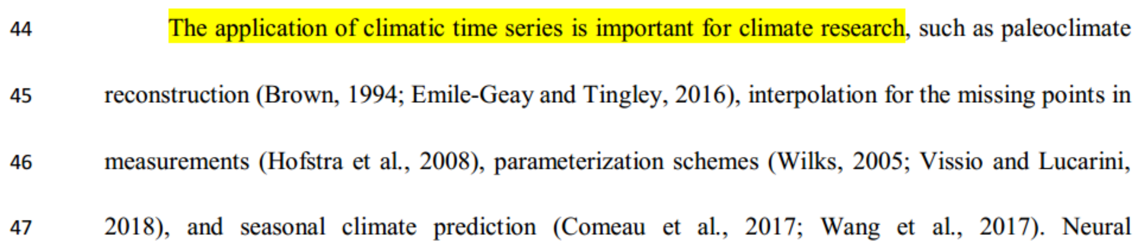
9. Although the work is interesting and has a lot of future scope, the above concerns prevents me from recommending this work for publication in its current form. I hope the authors would incorporate the suggestions and rewrite the manuscript.

**Response:** Thanks for your comments and suggestions! We will carefully improve the detail descriptions, and recognize most of parts according to your suggestions.

Specific Points:

10. Lines 43-46: The climate problems mentioned here are actually applications of climate data.

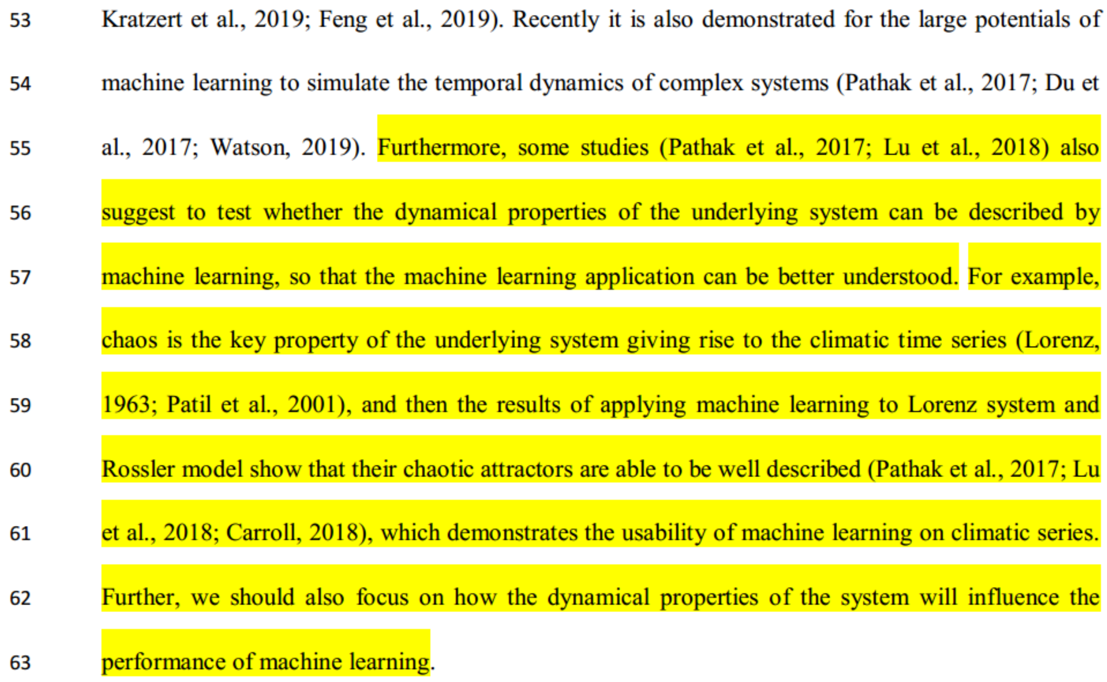
**Response:** Thank you! We will modify this narration. Our modification is shown by the following screenshot:



The application of climatic time series is important for climate research, such as paleoclimate reconstruction (Brown, 1994; Emile-Geay and Tingley, 2016), interpolation for the missing points in measurements (Hofstra et al., 2008), parameterization schemes (Wilks, 2005; Vissio and Lucarini, 2018), and seasonal climate prediction (Comeau et al., 2017; Wang et al., 2017). Neural

11. Lines 52-54: Re-write this sentences to make it intuitive. For example, this line: "...while the physics of systems is suggested for consideration" feels like it refers to the study by Watson, 2019, where neural network based algorithm is used to augment a physics based model to improve its performance. However, this is not clear from the text.

**Response:** Thank you! We will modify this narration. Our modification is shown by the following screenshot:



Kratzert et al., 2019; Feng et al., 2019). Recently it is also demonstrated for the large potentials of machine learning to simulate the temporal dynamics of complex systems (Pathak et al., 2017; Du et al., 2017; Watson, 2019). Furthermore, some studies (Pathak et al., 2017; Lu et al., 2018) also suggest to test whether the dynamical properties of the underlying system can be described by machine learning, so that the machine learning application can be better understood. For example, chaos is the key property of the underlying system giving rise to the climatic time series (Lorenz, 1963; Patil et al., 2001), and then the results of applying machine learning to Lorenz system and Rossler model show that their chaotic attractors are able to be well described (Pathak et al., 2017; Lu et al., 2018; Carroll, 2018), which demonstrates the usability of machine learning on climatic series. Further, we should also focus on how the dynamical properties of the system will influence the performance of machine learning.

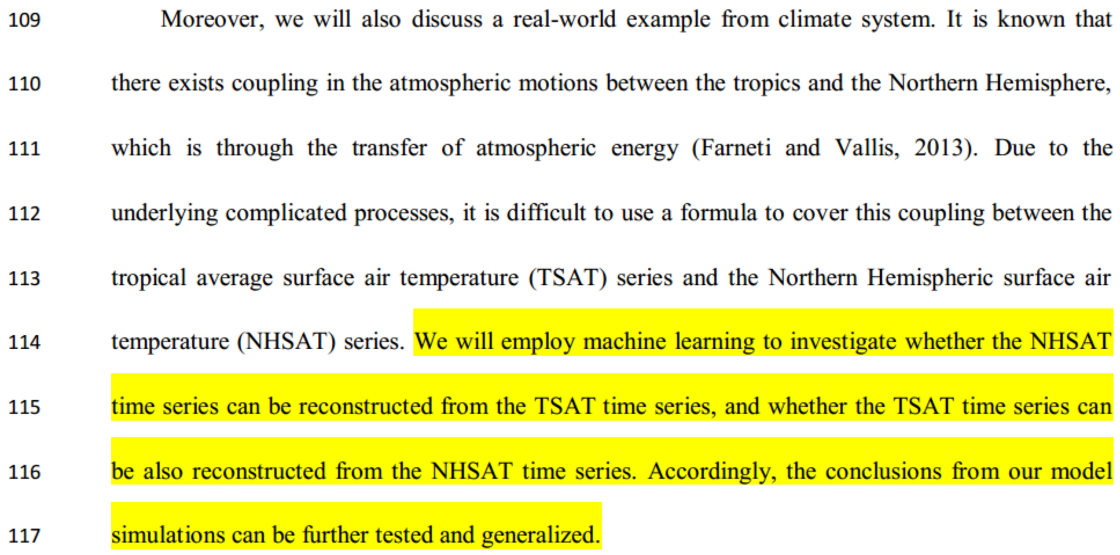
12. Lines 63-64: The statement infers that, since linear correlation is an intrinsic assumption of traditional statistical methods, cross-correlation analysis should be carried out for investigating the performance of ML algorithms. This is not a valid reasoning, as the approach of ML algorithms and traditional statistical methods are very different.

**Response:** Thank you! We will modify this narration. Our modification is shown by the following screenshot:

Applying machine learning to climatic series attracts much attention, but it is still unclear what can be learnt by machine learning during the training process, and what is the key factor determining the performance of machine learning applied to climatic time series. This is crucial for investigating why machine learning performs not well with some datasets, and how to improve the performance for them. Here, the coupling relation between different variables needs to be paid attention to. Because different climate variables are coupled with one another (Donner and Large, 2008), and the coupling relation will often result in that their observational time series are statistically correlated (Brown, 1994). This is a crucial property for the climate system, and often contributes to the analysis on the climatic time series. It is known that linear correlation is the implicit assumption for traditional statistical methods, and they often fail if linear correlation is weak (Brown, 1994; Sugihara et al., 2012; Emile-Geay and Tingley, 2016). However, previous studies (Sugihara et al., 2012; Emile-Geay and Tingley, 2016) also suggest 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 can contribute to 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.

13. Lines 83-87: This part should be there in the Results section. However, this line can be modified to be a hypothesis the authors are trying to check.

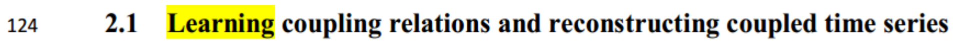
**Response:** Thank you! We will modify this narration. This part has been modified to be a hypothesis in the introduction. Our modification is shown by the following screenshot:



Moreover, we will also discuss a real-world example from climate system. It is known that there exists coupling in the atmospheric motions between the tropics and the Northern Hemisphere, which is through the transfer of atmospheric energy (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 will employ machine learning 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.

14. Line 105: Typographical error: it should be “Learning” not “Leaning”.

**Response:** Thank you! We will modify this typographical error. We will also inspect the manuscript to avoid the any typographical error. Our modification is shown by the following screenshot:



**2.1 Learning coupling relations and reconstructing coupled time series**

15. Figure 1: The big black arrow used to represent (3), is confusing in the sense that the reconstructed time series from the testing stage is being compared with the time series from the training stage, which is not the case.

**Response:** Thank you! We will modify this figure. Our modification is shown by the following screenshot:

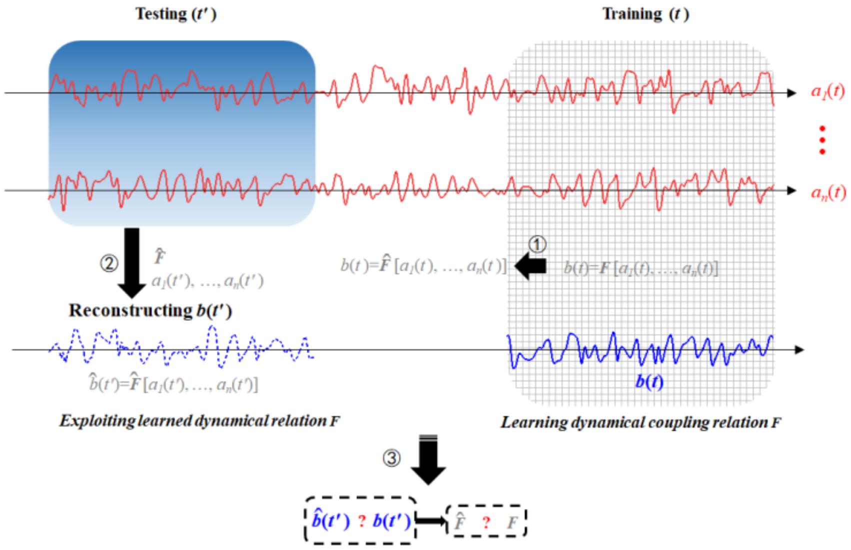

Testing ($t'$)
Training ($t$)
$a_1(t)$
$a_n(t)$
①
$b(t)=\hat{F}[a_1(t), \ldots, a_n(t)]$
$b(t)=F[a_1(t), \ldots, a_n(t)]$
②
$\hat{F}$
$a_1(t'), \ldots, a_n(t')$
Reconstructing $b(t')$
$\hat{b}(t')=\hat{F}[a_1(t'), \ldots, a_n(t')]$
$b(t)$
Exploiting learned dynamical relation F
Learning dynamical coupling relation F
③
$\hat{b}(t')$ ? $b(t')$
$\hat{F}$ ? $F$

**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.

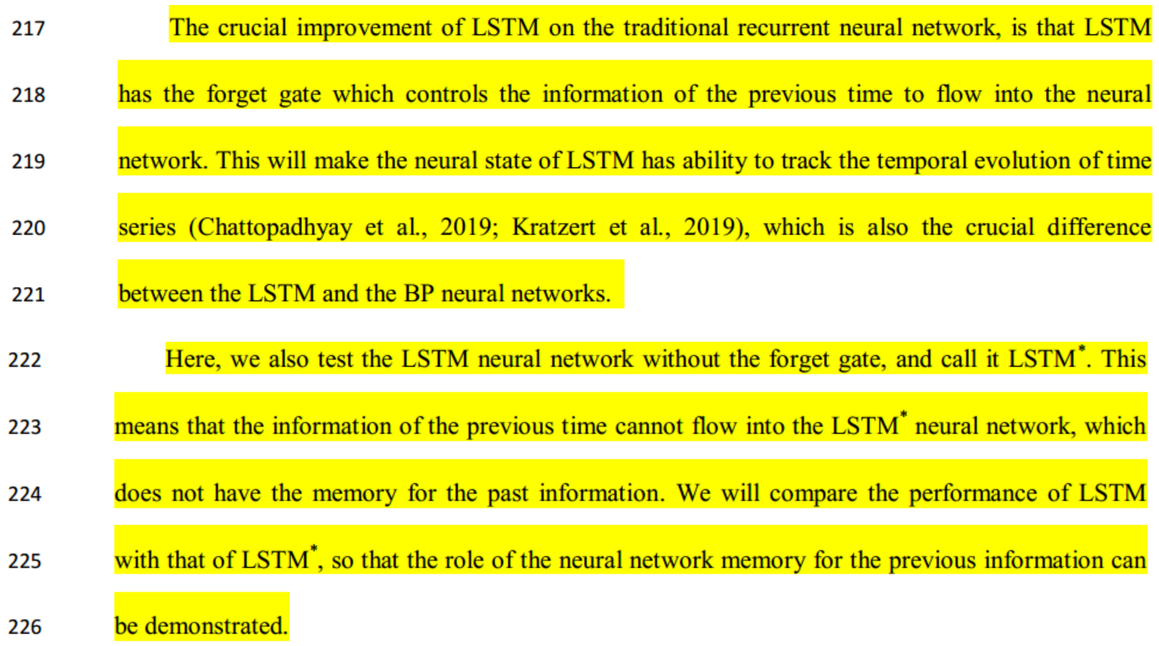
16. Lines 182-183: Mention clearly why an analysis of LSTM* reconstructed time series is required.

**Response:** Thank you! We will modify this narration.

The crucial improvement of LSTM on the traditional recurrent neural network, is that LSTM has the **forget gate** which controls the information of the previous time to flow into the neural network. This also make the neural state of LSTM has ability to track the temporal evolution, which is also the crucial difference between LSTM and 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 demonstrated.**

Our modification is shown by the following screenshot:

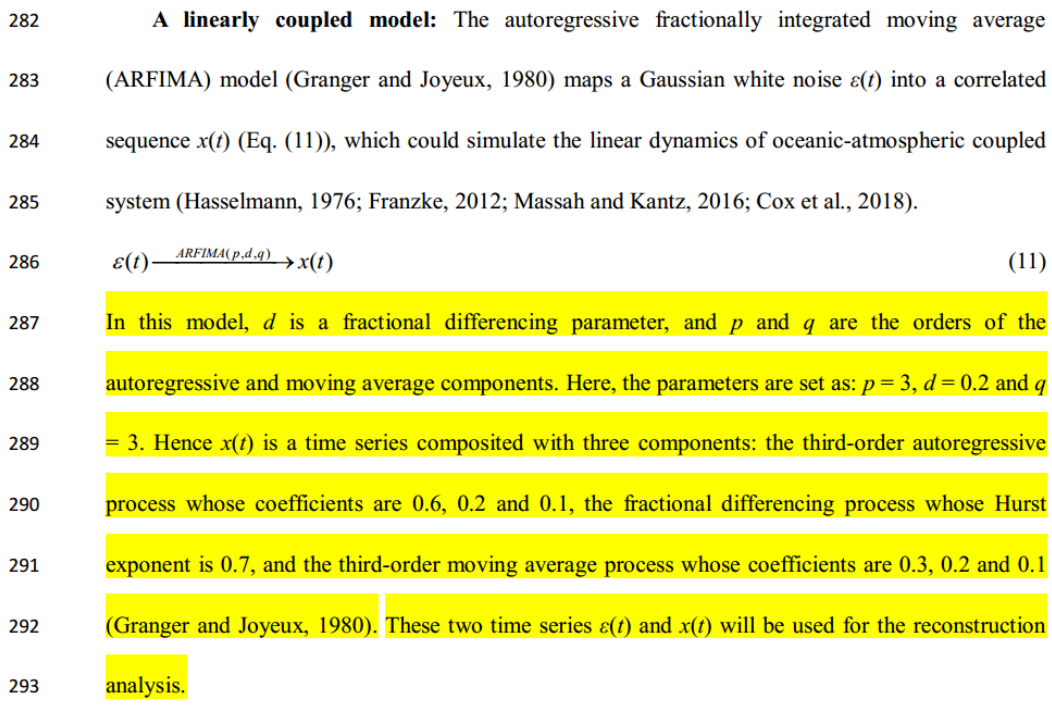


The crucial improvement of LSTM on the traditional recurrent neural network, is that LSTM has the forget gate which controls the information of the previous time to flow into the neural network. This will make the neural state of LSTM has ability to track the temporal evolution of time series (Chattopadhyay et al., 2019; Kratzert 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 demonstrated.

17. Lines 201-203: The introduction of the parameters, p, d, and q is not proper and causes confusion. Rewrite the sentence.

**Response:** Thank you! We will modify this narration. Our modification is shown by the following screenshot:



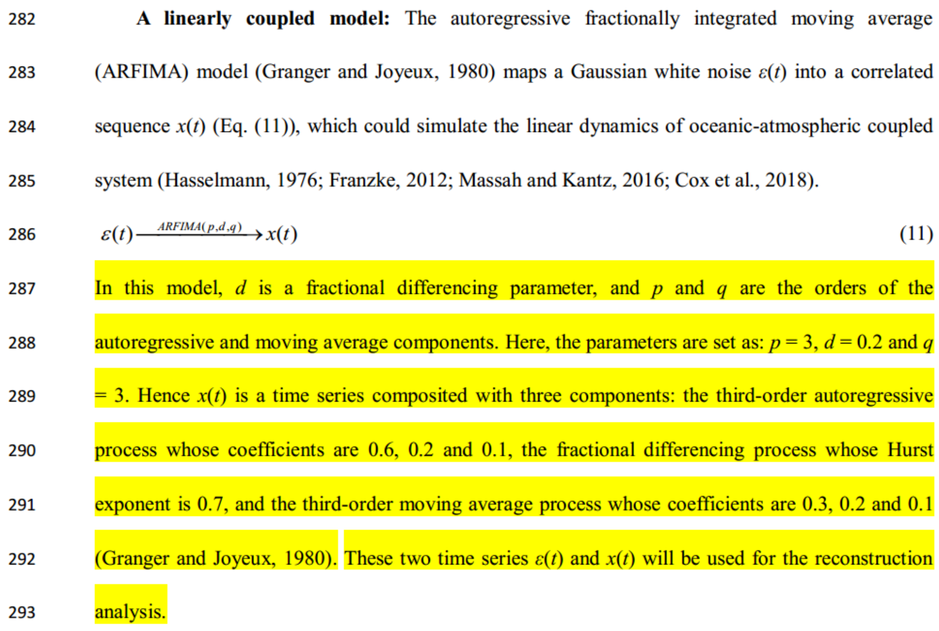
**A linearly coupled model:** The autoregressive fractionally integrated moving average (ARFIMA) model (Granger and Joyeux, 1980) maps a Gaussian white noise *ε*(*t*) into a correlated sequence *x*(*t*) (Eq. (11)), which could simulate the linear dynamics of oceanic-atmospheric coupled system (Hasselmann, 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 *ε*(*t*) and *x*(*t*) will be used for the reconstruction analysis.

18. Lines 205-206: x(t) and the Gaussian noise () time series are the two time series being used for the coupled analysis. This has to be mentioned clearly in the text. This comment goes for all the cases of coupled time series being used (non-linear, higher order non-linear, real world scenario).

**Response:** Thank you! We will mention this information for all the used data in the revised manuscript. Our modification is shown by the following screenshot:



**A linearly coupled model:** The autoregressive fractionally integrated moving average (ARFIMA) model (Granger and Joyeux, 1980) maps a Gaussian white noise *ε*(*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 *ε*(*t*) and *x*(*t*) will be used for the reconstruction analysis.

**A nonlinearly coupled model:** The Lorenz 63 (in the following referred as to Lorenz63) 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 ($\sigma$, $\mu$, $B$) = (10, 28, 8/3), the state in the system is chaotic. We employed the Runge-Kutta integrator of the fourth order to acquire the series output from Lorenz63. The time steps were 0.01. The time series $X(t)$ and $Z(t)$ will be used for the reconstruction analysis.

**A high-dimensional model:** The two-layer Lorenz 96 (in the following referred as to Lorenz96) model (Lorenz, 1996) is a high-dimensional chaotic system, which is generally employed 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

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

In the first layer of the Lorenz 96 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 scale parameter $\theta$ controls the scale separation of the two layers. When $\theta > 1$, processes in the second layer will be slower than processes in the first layer because the increment of $Y_{k,\,j}$ is decreased by the term of $\theta$. The time scale of $Y_{k,\,j}$ can be also close to that of $X_k$ by modulating the value of $\theta$; especially, the coupling strength will be amplified when $\theta$ is much smaller than 1. The Runge-Kutta integrator of the fourth order and periodic boundary condition are adopted (that is: $X_0 = X_K$ and $X_{K+1} = X_1$ ; $Y_{k,\,0} = Y_{k\text{-}1,\,J}$ 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 $Y_{1,\,1}(t)$ will be used for the reconstruction analysis.

## 3.2 Real-world climatic time series

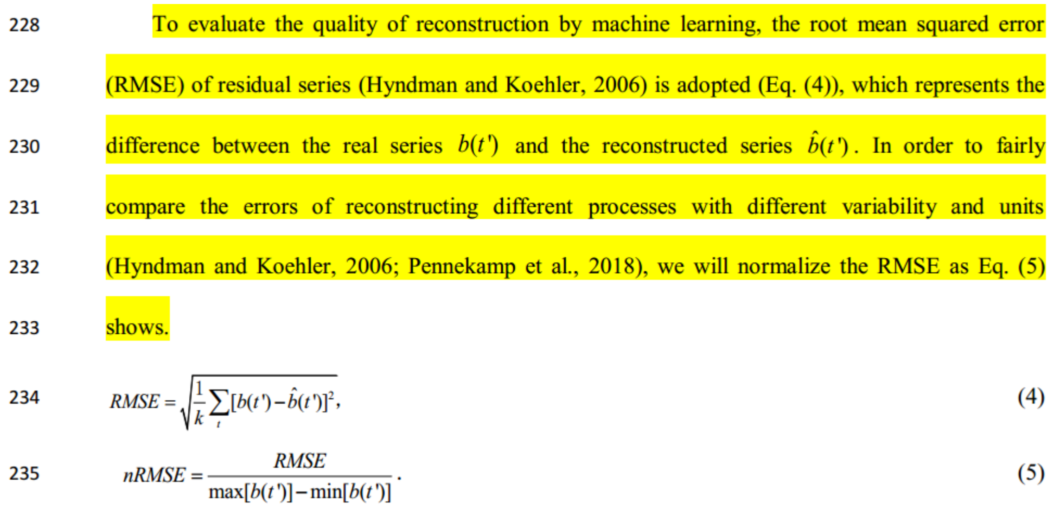
TSAT, NHSAT and the Nino3.4 index are chosen to represent real-world climatic time series, and they will be used for the reconstruction analysis. The original data is from National Centers for Environmental Prediction (https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis2.html) and KNMI Climate Explorer (http://climexp.knmi.nl). The series of TSAT and NHSAT are from the

19. Lines 236-237: The time series are being standardized (mean is zero and standard deviation is one) before being used in the reconstruction analysis. Explain why are they standardized.

**Response:** Thank you! We will explain for this processing of standardization.

For the time series that come from different processes, they might have different variability and units. In order to avoid the disturbance given by such different variability and units, we select to standardize all the time series with uniform mean value and variance.

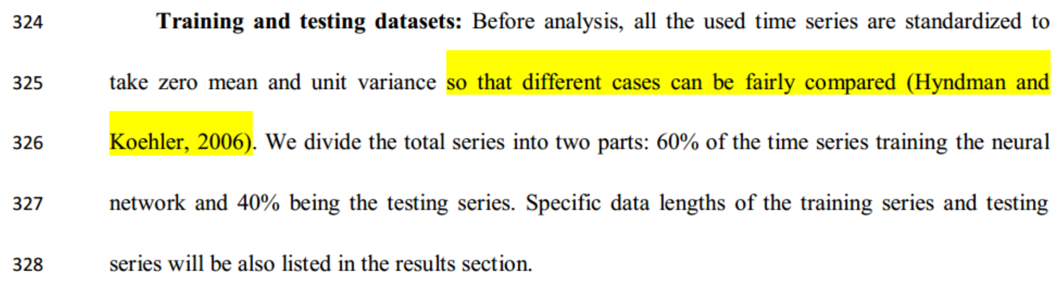
Our modification is shown by the following screenshot:



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), we will normalize the RMSE as Eq. (5) shows.

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

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



**Training and testing datasets:** Before analysis, all the used time series are standardized to take zero mean and unit variance so that different cases can be fairly compared (Hyndman and Koehler, 2006). 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.

20. Lines 275-277: Incorporate the plots for LSTM* in Figure 3c and 3d.

**Response:** Thank you! We will add the results of LSTM$^*$ into the corresponding figures. Our modification is shown by the following screenshot:

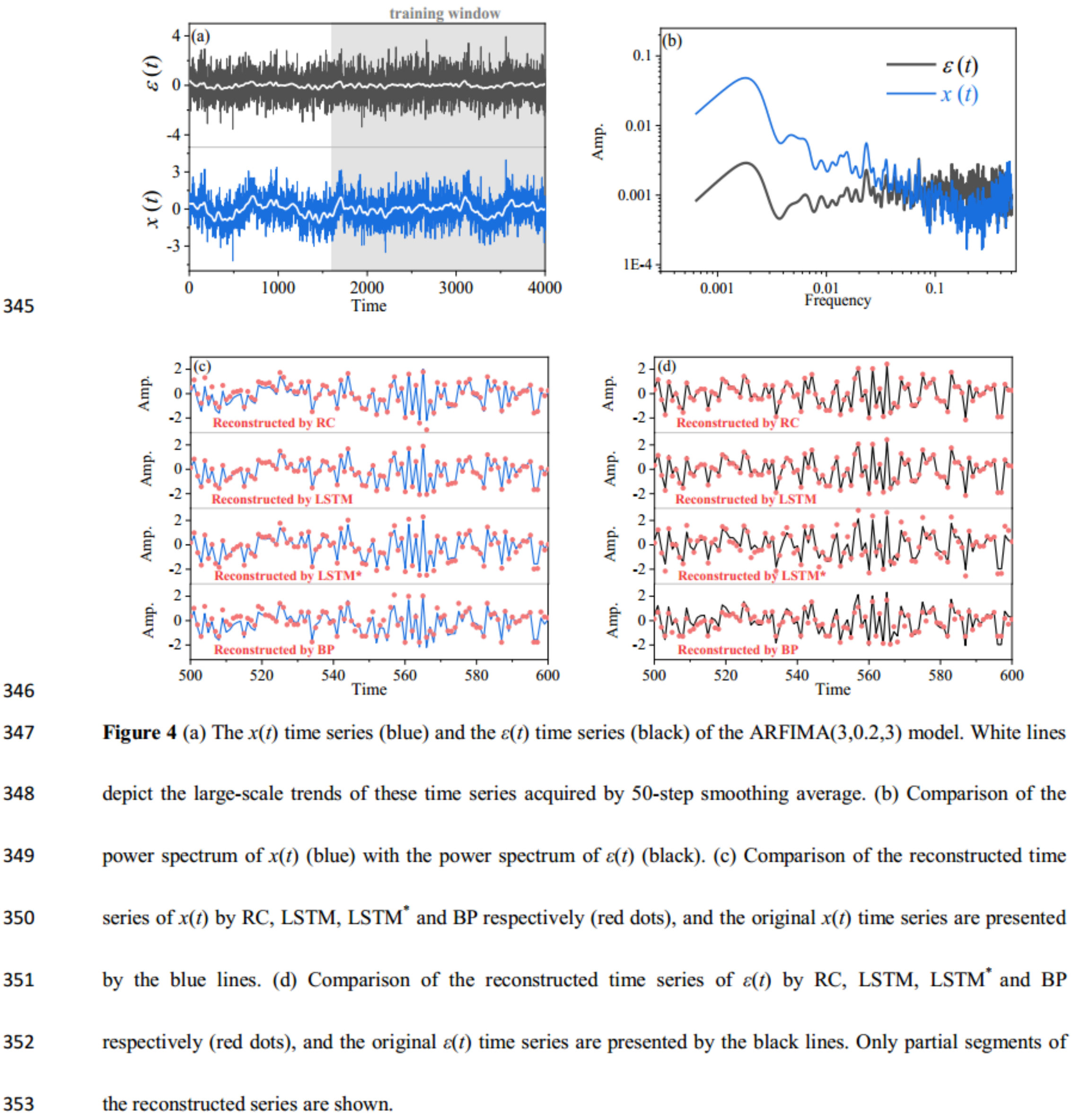


**Figure 4** (a) The $x(t)$ time series (blue) and the $\varepsilon(t)$ time series (black) of the ARFIMA(3,0.2,3) model. White lines depict the large-scale trends of these time series acquired by 50-step smoothing average. (b) Comparison of the power spectrum of $x(t)$ (blue) with the power spectrum 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.

21. Lines 286-297: The information about convergent cross mapping (CCM) should be introduced in the methodology section in detail. Are there other methods for estimating non-linear correlation or causality between two time-series. If so, why CCM was specifically used.

**Response:** Thank you! We will move the detailed description of CCM to the method part.

Apart from CCM, the Granger method [1] and transfer entropy [2] can be also used to measure the causality. However, it has been demonstrated that the Granger causality cannot measure the causality or coupling in nonlinear systems [3]. Transfer entropy can be an alternative choice to measure the nonlinear coupling. But the index value of transfer entropy often ranges from 0 to 3 [4], while the CCM index always ranges from 0 to 1, so that it is often hard to judge if transfer entropy is strong or weak. In previous studies [5], the CCM index has been successfully used to measure the nonlinear coupling strength and causality in many kinds of complex systems. However, it is worth to make comparisons for CCM, transfer entropy and machine learning performance in the future study.

[1] Granger C. W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424-438, 1969.

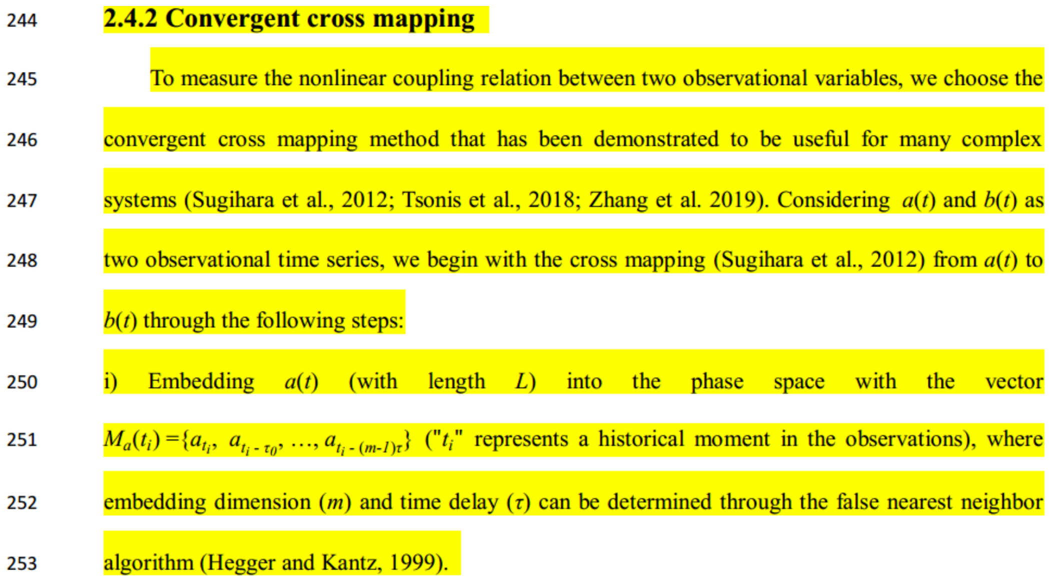
[2] Schreiber T.: Measuring information transfer. Phys Rev Lett 85(2), 461, 2000.

[3] Malevergne Y., Sornette D.: Extreme financial risks: From dependence to risk management. Springer Science & Business Media, 2006.

[4] Paluš, M.: Multiscale atmospheric dynamics: cross-frequency phase-amplitude coupling in the air temperature. Phys Rev Lett, 112(7), 078702, 2014.

[5] 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.

Our modification is shown by the following screenshot:



**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 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 the 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 ($\tau$) can be determined through the false nearest neighbor algorithm (Hegger and Kantz, 1999).

ii) Estimating the weight parameter $w_i$ denoting 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.* defined the causal inference from $\rho_{a \to b}$ 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 value, 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 value, then the causal relationship between $a$ and $b$ is bidirectional ($a$ and $b$ cause each other). In our study, all the values of CCM indices are measured when they are convergent with the data length.

According to the literature (Takens, 1981; Sugihara et al., 2012), 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 $b$). Hence, 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 $a$. So that the high value 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 can test the association between the CCM index and the reconstruction performance of machine learning.

22. Lines 390-392: Explain the decrease in LSTM nRMSE with an increase in CCM. As, this behavior is contradictory to the LSTM's nRMSE behavior in the other cases.

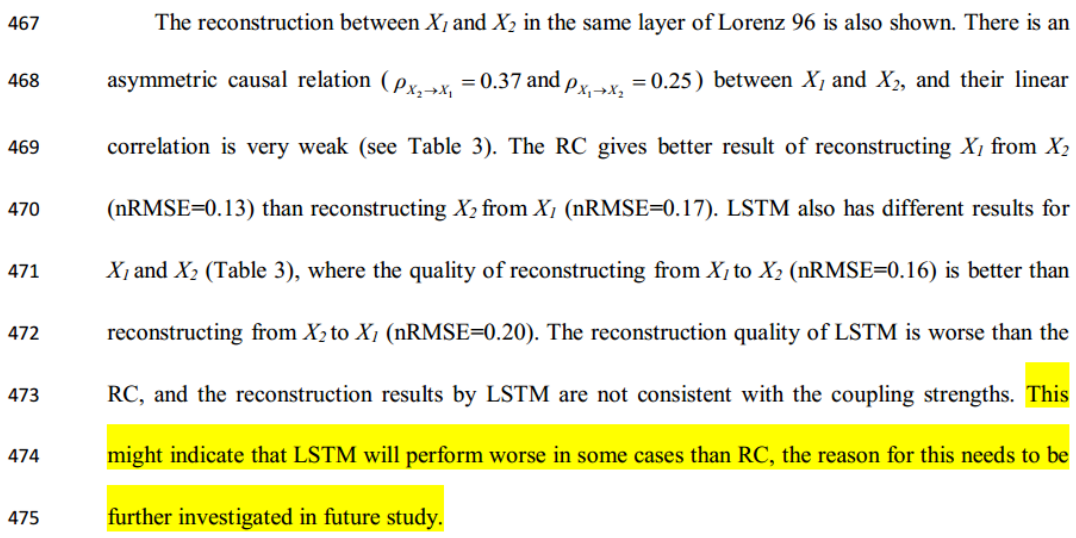
**Response:** Thank you! We will supplement the explanation for this.

For all cases of RC results, when the CCM index is increasing, the nRMSE will be decreasing. Likewise, for most cases of LSTM results, when the CCM index is increasing, the nRMSE will be decreasing.

But in this case for LSTM, the relation between CCM and nRMSE is not like the normal cases. The reason might be that the used time series ($X_1$ and $X_2$ of Lorenz 96 system) have the time-varying local mean values (i. e. in the previous time period, the local mean value of time series is 0, and then in the next time period, the local mean value of time series is 0.5), and this influences the performance of LSTM.

We found that the time-varying mean values in time series tend to impact the performance of LSTM. For example, in a time series, at the previous time period, the local mean value of time series is 0, and then at the next time period, the local mean value of time series is 0.5. In this case, LSTM tends to perform badly, and the nRMSE might be increased. **The reason might be that the LSTM algorithm always requires incorporating the time-series values at previous time points (the memory for past time points), and then the varied local mean value of time series will easily influence the results of LSTM.**

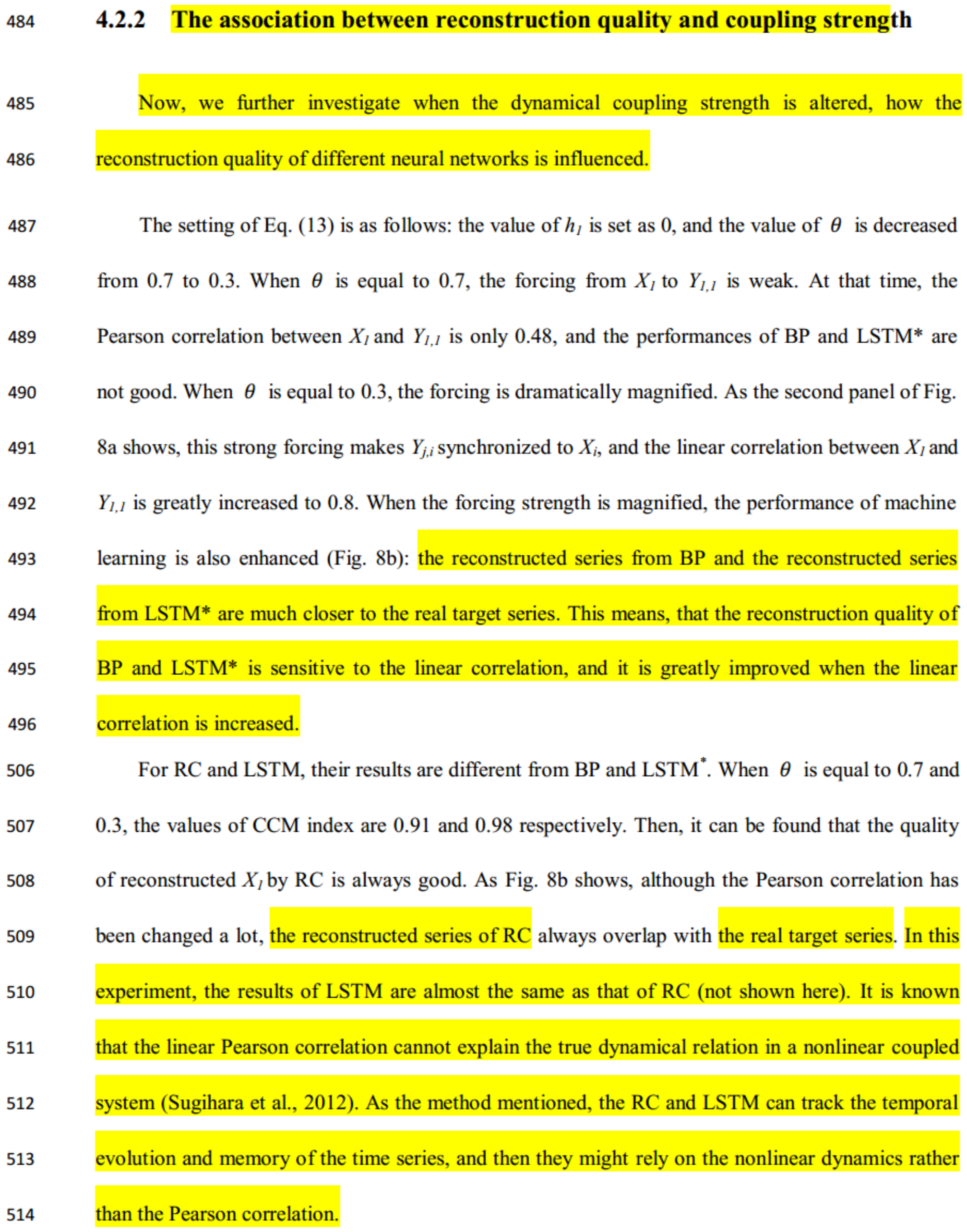
However, we have not been able to ensure that this is the only reason. More investigations are needed in the further study. Our modification is shown by the following screenshot:



The reconstruction between $X_1$ and $X_2$ in the same layer of Lorenz 96 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). The reconstruction quality of LSTM is worse than the RC, and the reconstruction results by LSTM are not consistent with the coupling strengths. This might indicate that LSTM will perform worse in some cases than RC, the reason for this needs to be further investigated in future study.

24. Lines 407-408: Explain how did the authors arrive at this statement. RC and LSTM performed better than LSTM* and BP in the linearly coupled system. And BP and LSTM* were not part of the analysis of the high dimensional lorenz-96 analysis. However, this statement can be the conclusion of this section, which shows the sensitivity of RC and LSTM to different coupling strength.

**Response:** Thank you! We will modify this narration. In our previous manuscript, the expected meaning of this statement was not a conclusion, but was used to open the topic of this subsection. Our modification for this part is shown by the following screenshot:



### 4.2.2 The association between reconstruction quality and coupling strength

Now, we further investigate when the dynamical coupling strength is altered, how the reconstruction quality of different neural networks is influenced.

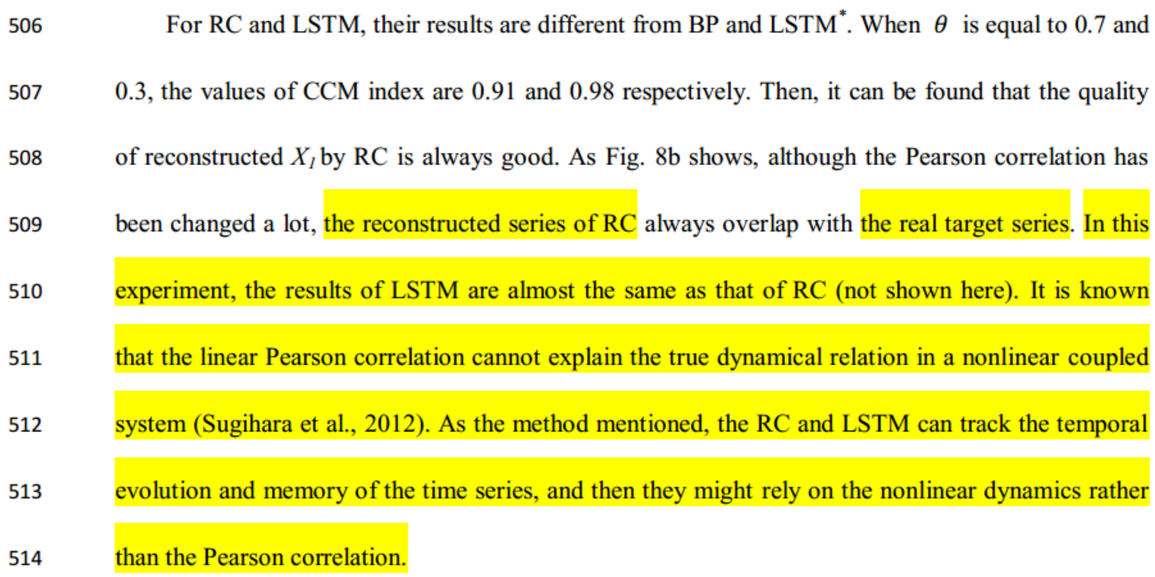
The setting of Eq. (13) is as follows: the value of $h_1$ is set as 0, and the value of $\theta$ is decreased from 0.7 to 0.3. When $\theta$ is equal to 0.7, the forcing from $X_1$ to $Y_{1,1}$ is weak. At that time, the Pearson correlation between $X_1$ and $Y_{1,1}$ is only 0.48, and the performances of BP and LSTM* are not good. When $\theta$ is equal to 0.3, the forcing is dramatically magnified. As the second panel of Fig. 8a shows, this strong forcing makes $Y_{j,i}$ synchronized to $X_i$, and the linear 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. 8b): the reconstructed series from BP and the reconstructed series from LSTM* are much closer to the real target series. This means, that the reconstruction quality of BP and LSTM* is sensitive to the linear correlation, and it is greatly improved when the linear correlation is increased.

For RC and LSTM, their results are different from BP and LSTM$^*$. When $\theta$ is equal to 0.7 and 0.3, the values of CCM index are 0.91 and 0.98 respectively. Then, it can be found that the quality of reconstructed $X_1$ by RC is always good. As Fig. 8b shows, although the Pearson correlation has been changed a lot, the reconstructed series of RC always overlap with the real target series. In this experiment, the results of LSTM are almost the same as that of RC (not shown here). It is known that the linear Pearson correlation cannot explain the true dynamical relation in a nonlinear coupled system (Sugihara et al., 2012). As the method mentioned, the RC and LSTM can track the temporal evolution and memory of the time series, and then they might rely on the nonlinear dynamics rather than the Pearson correlation.

25. Lines 416-420: Examine LSTM for its behavior with change in θ, like the one done for the behavior of LSTM*. This will probably give more insight into the behavior of LSTM*.

**Response:** Thank you! In this case of reconstructing $X_1$ from $Y_{1,1}$ (Lorenz 96 system), all the results of LSTM and RC are almost overlapped with each other. We will supplement the results of LSTM in the revised manuscript.

Our modification for this part is shown by the following screenshot:



For RC and LSTM, their results are different from BP and LSTM$^*$. When $\theta$ is equal to 0.7 and 0.3, the values of CCM index are 0.91 and 0.98 respectively. Then, it can be found that the quality of reconstructed $X_l$ by RC is always good. As Fig. 8b shows, although the Pearson correlation has been changed a lot, the reconstructed series of RC always overlap with the real target series. In this experiment, the results of LSTM are almost the same as that of RC (not shown here). It is known that the linear Pearson correlation cannot explain the true dynamical relation in a nonlinear coupled system (Sugihara et al., 2012). As the method mentioned, the RC and LSTM can track the temporal evolution and memory of the time series, and then they might rely on the nonlinear dynamics rather than the Pearson correlation.

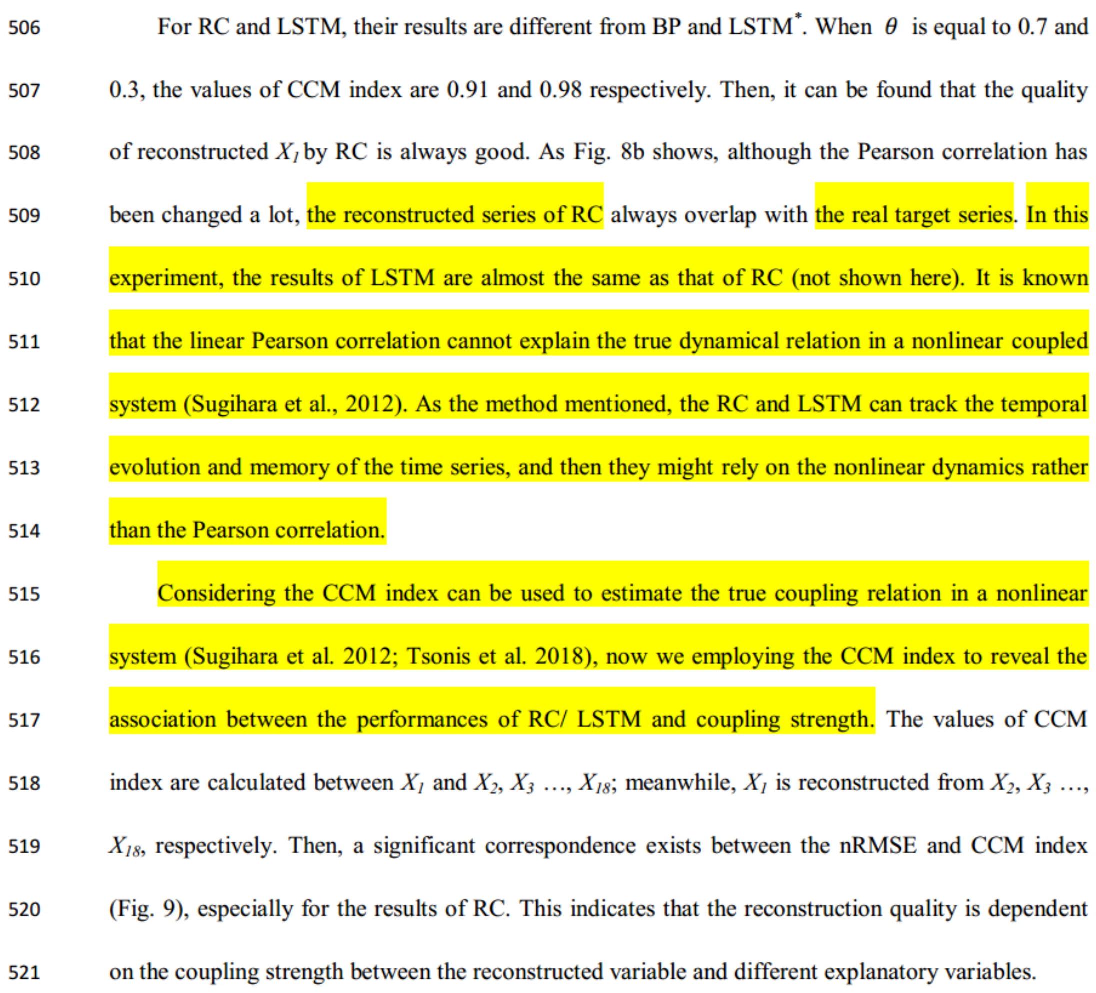
26. Line 430: Why is RC not sensitive to Pearson’s correlation.

**Response:** Thank you! Here the RC was applied to the nonlinear Lorenz 96 system. It is known that the linear Pearson correlation cannot explain the true dynamical relation in a nonlinear coupled system [1-2]. As the method mentioned, the RC and LSTM can track the temporal evolution and memory of the time series, and then they might rely on the nonlinear dynamics rather than the Pearson correlation.

[1] Malevergne Y., Sornette D.: Extreme financial risks: From dependence to risk management. Springer Science & Business Media, 2006.

[2] 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.

We will add some words to explain such phenomenon, which is shown by the following screenshot:



For RC and LSTM, their results are different from BP and LSTM$^*$. When $\theta$ is equal to 0.7 and 0.3, the values of CCM index are 0.91 and 0.98 respectively. Then, it can be found that the quality of reconstructed $X_1$ by RC is always good. As Fig. 8b shows, although the Pearson correlation has been changed a lot, the reconstructed series of RC always overlap with the real target series. In this experiment, the results of LSTM are almost the same as that of RC (not shown here). It is known that the linear Pearson correlation cannot explain the true dynamical relation in a nonlinear coupled system (Sugihara et al., 2012). As the method mentioned, the RC and LSTM can track the temporal evolution and memory of the time series, and then they might rely on the nonlinear dynamics rather than the Pearson correlation.

Considering the CCM index can be used to estimate the true coupling relation in a nonlinear system (Sugihara et al. 2012; Tsonis et al. 2018), now we employing the CCM index to reveal the association between the performances of RC/ LSTM and coupling strength. 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. Then, a significant correspondence exists between the nRMSE and CCM index (Fig. 9), especially for the results of RC. This indicates that the reconstruction quality is dependent on the coupling strength between the reconstructed variable and different explanatory variables.

27. Figure 8: It is missing the R2 and p-value of LSTM. The behavior of LSTM should also be evaluated in the same manner.

**Response:** Thank you! We will add the results of LSTM into this figure. Our modification is shown by the following screenshot:

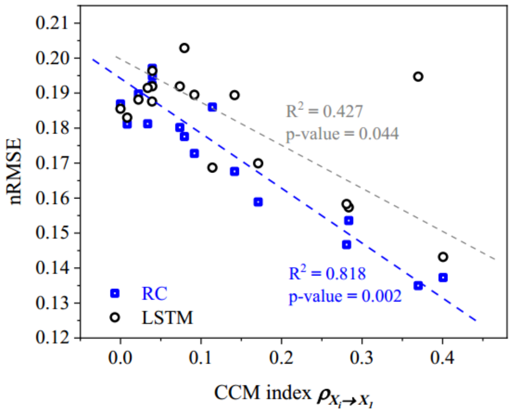


**Figure 9** Scatter plot of nRMSE values and CCM index values. The blue boxes are results of the RC machine learning, and the black cycles are results of the LSTM machine learning. The blue and grey dashed lines are the fitted linear trends of the blue boxes and black cycles respectively, and these two dependency trends are both significant because their p-values are both smaller than 0.05.

28. Lines 472-473: What do you mean by unstable variance, elaborate.

**Response:** Thank you! We will supplement the explanation for this.

For the real-world time series (such as the time series in figure R1), the local mean value and the local variance of the time series, are often time-varying. For example, in a time series, at the previous time period, the local mean value of time series is 0, and then at the next time period, the local mean value of time series is 0.5; at the previous time period, the local variance of time series is 1, and then at the next time period, the local variance of time series is 1.5.

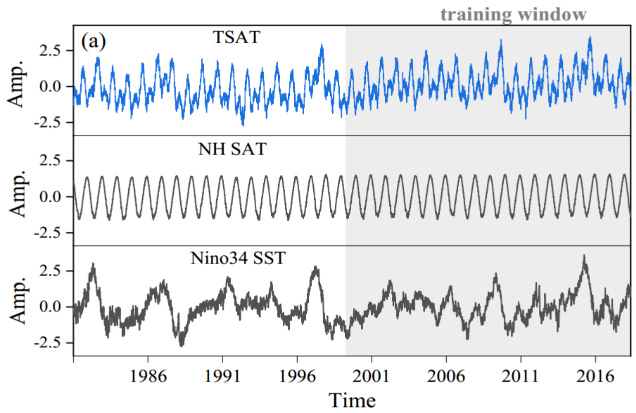
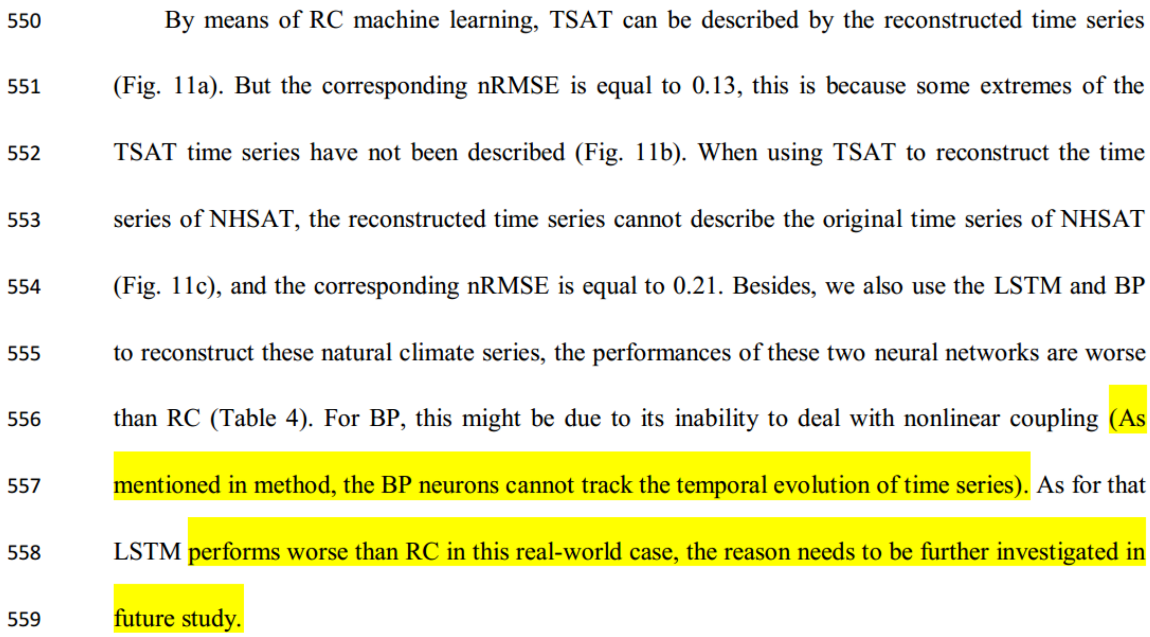


Figure R1: Daily time series of the Tropical surface air temperature, the Northern Hemispheric surface aire temperature, and the Nino 3.4 index.

We found that the time-varying local mean value and local variance in time series tend to impact the performance of LSTM. In this case, LSTM tends to perform badly, and the nRMSE might be increased.

**The reason might be that the LSTM algorithm always requires incorporating the time-series values in previous time points (the memory for past time points), and then the varied local mean value of time series will easily influence the results of LSTM. Likewise, the varied local mean value of time series will also influence the results of LSTM.**

However, we have not been able to ensure that this is the only reason. More investigations are needed in the future study. Our modification in this part is shown by the following screenshot:



By means of RC machine learning, TSAT can be described by the reconstructed time series (Fig. 11a). But the corresponding nRMSE is equal to 0.13, this is because some extremes of the TSAT time series have not been 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 the 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 time series). As for that LSTM performs worse than RC in this real-world case, the reason needs to be further investigated in future study.