Reply to the comments of Anonymous Referee #1:

The comments of Anonymous Referee #1:

1. This manuscript investigates the potentialities of reconstructing time series using machine learning (ML) techniques. This approach is applied on a set of simple systems, and then applied to the interaction between the Tropical surface temperature and the Northern extra-tropical surface temperature. Different configurations of the machine learning approaches are explored, the reservoir computing, the long short-term memory, but also a simplified version of the latter and back-propagation. The authors use the correlation (for linear systems) and the convergent cross mapping (for nonlinear systems), CCM, as tools to evaluate the ability of the machine learning approaches to reproduce the original time series.

   Although I find the idea of putting in parallel the CCM with the ability of reconstructing time series based on ML very interesting, the description of the tools and the results is confusing, the presentation is quite poor and many details on the approaches used are missing.

Response: Thanks for your comments and suggestions! We will carefully improve the description and details of the methods, including the machine learning framework and the CCM theory. **And then, the results and conclusions in the paper are correct. The confusion of Anonymous Referee #1 is the relationship between reconstruction direction and the CCM dependence, and this confusion is mainly induced by the lack of description of the CCM theory.** We will carefully introduce the CCM theory in the revised manuscript, so that the results part can be better understood. Meanwhile, we will also carefully improve the manuscript according to your specific comments and suggestions.

In addition, we also would like to summarize the contributions of this work with the following plain language:

i) Investigating how to better apply machine learning to the reconstruction of climate time series (under different coupling dynamics of climate systems), which might be very useful for some important climate problems such as **paleoclimate reconstruction**, **interpolation for the missing points in measurements** and **parameterization schemes**. For instance, for the records of proxy data (tree ring or ice core), we might obtain the data from the historical and current period. For the records of climatic variable like air temperature, we might only obtain the data from the current period. At that time, the conclusions of this paper will be useful to reconstruct the historical data of climatic variable.

ii) We proposed to use nonlinear causality coefficient to select explanatory variable, which is demonstrated more effective than the Pearson correlation.

iii) Revealing that the reconstruction quality is direction-dependent for two nonlinearly coupled variables: for example, the tropical average surface temperature can be well reconstructed from the average Northern Hemispheric surface temperature, but the average Northern Hemispheric surface temperature cannot be reconstructed from the tropical average surface temperature. Then we explain the reasons and how to deal with such issues. This might be an important suggestion for the future application of data-driven approach to geoscience.

2. My first main point is the confusion present in the notation of input/output and the notion of directional dependence. Let me clarify my point by considering Table 2 in which the results for the Lorenz 3-variable system are displayed. The first column indicates the input of the ML approach (also indicated a(t)), the second the output of the ML (also indicated b(t)), while the fourth represents the CCM dependence. The later, as defined at lines 291-297, has high values if b(t) influence a(t). So according to that table if b(t) is influencing a(t) I should get good results of fitting from a(t) to b(t). I am really confused with this claim.

Response: Thanks for your comments and suggestions. The results of Table 2 is correct: the Lorenz-X can be used to reconstruct the Lorenz-Z, but the Lorenz-Z cannot be used to reconstruct the Lorenz-X, which can be also seen in the previous literature of *Lu et al. 2017*[1]. In the paper of *Lu et al. 2017*[1], they used the “nonlinear observability” of the controlled system theory to explain such phenomenon. However, the “nonlinear observability” introduced in *Lu et al. 2017*[1] is only usable in the system with known mathematical equation, here we employ the CCM coefficient which does not rely on any known equation.

According to the literature [2-6], **the claim about the relationship of the CCM dependence and reconstruction direction, is correct and accurate:** if *b* influence *a* but *a* does not influence *b*, the information of *b* can be shared with *a* (through the information transfer from *b* to *a*), but *a* ’s information cannot be shared with *b* (there exists no information transfer from *a* to *b*). Hence, the records of *a* will be encoded with the information of *b*, and the time series of *b* can be recovered from the records of *a*.

As the above mentioned, the information transfer induced by causal influence, is the reason of that if *b* influence *a* and then *a* can reconstruct *b*. Further, according to *Sugihara et al. 2012* [4-6], for the CCM index ($\rho_{a\to b}$), its computation is using a phase-space model [6] to estimate the values of *b* from *a*’s records. And then the magnitude of $\rho_{a\to b}$ represents: when using *a*'s records to recover the values of *b*, how well the quality is. Hence, the magnitude of $\rho_{a\to b}$ also represents how much information of *b* is encoded in *a*'s records.

Sugihara et al. 2012 [4-6] ever suggested that the reconstruction direction is opposite to the causal dependence direction: when $\rho_{a\to b}$ is high, this means that *b* causes *a*, and we can get good results of reconstruction from *a* to *b*.

In the previous manuscript, the above description about the CCM theory is not fully presented, so that it might take confusion to the understanding the results of Tables 2 and 4. But the results about Tables 2 and 4 are really correct. **We will carefully improve the description of the CCM theory [4-6], and add the necessary description of the CCM computational algorithm,** so that the results of the CCM and reconstruction quality will be better understood.

[1] 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.

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

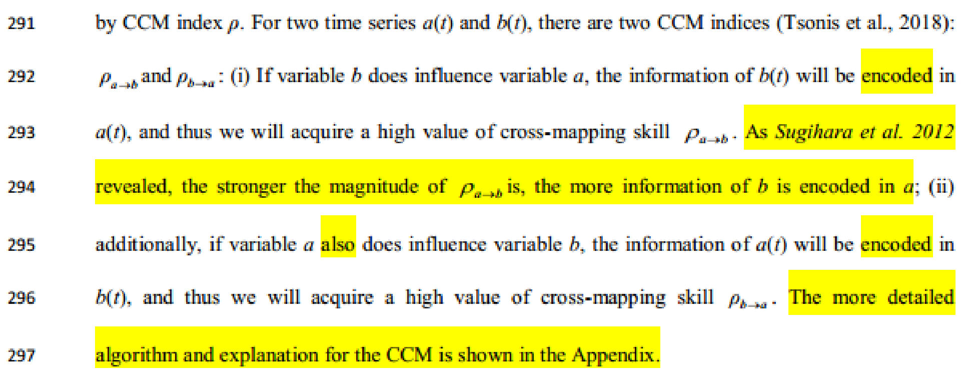
[3] Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. Physics Reports, 441(1), 1-46, 2007.

[4] 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.

[5] 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.

[6] 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.

Additionally, we will modify the sentences in lines 291-297 of the previous manuscript, as the following screenshot shows:



by CCM index ρ. For two time series $a(t)$ and $b(t)$, there are two CCM indices (Tsonis et al., 2018): $\rho_{a\to b}$ and $\rho_{b\to a}$: (i) If variable b does influence variable a, the information of $b(t)$ will be encoded in $a(t)$, and thus we will acquire a high value of cross-mapping skill $\rho_{a\to b}$. As *Sugihara et al. 2012* revealed, the stronger the magnitude of $\rho_{a\to b}$ is, the more information of b is encoded in a; (ii) additionally, if variable a also does influence variable b, the information of $a(t)$ will be encoded in $b(t)$, and thus we will acquire a high value of cross-mapping skill $\rho_{b\to a}$. The more detailed algorithm and explanation for the CCM is shown in the Appendix.

3. I have the same problem with the other tables, and in particular with Table 4 which is even more confusing when related with the discussion in the text. In the table it is indicated that TSAT influences strongly NHSAT but then the ML modeling is done from NHSAT to TSAT. This is what is claimed at lines 463-464, while in the conclusion it is said (line 542) that the TSAT is mainly influencing the NHSAT. I hope this is just a matter of confused notation but I am not sure and I strongly recommend the authors to revisit carefully their notations and interpretation carefully.

Response: Thanks for your comments and suggestions. We have inspected the results and conclusions, and the results and conclusions about Table 4 are correct. **Sugihara et al. 2012 [1] ever suggested that the reconstruction direction is opposite to the causal dependence direction.** The confusion about the relationship between reconstruction direction and the CCM dependence, is induced by the lack of description of the CCM theory in the previous manuscript.

Firstly, we can comprehend the CCM index according to the literature [1-4]: if *b* does influence *a* (*a* and *b* are two arbitrary variables), and then the information of *b* can be shared with *a* (through the information transfer from *b* to *a*). Hence, the records of *a* will be encoded with the information of *b*, and the time series of *b* can be recovered from the records of *a*. At that time, the CCM coefficient $\rho_{a\to b}$ denotes: when using *a*'s records to recover the values of *b*, how well the quality is. Likewise, the magnitude of $\rho_{a\to b}$ represents how much information of *b* is encoded in the records of *a*.

Then, in our results about using NHSAT to reconstruct TSAT, the CCM index that NHSAT cross maps TSAT is of high value (0.7). This suggests that the NHSAT's records are able to recover the values of TSAT, which stems from that the information of TSAT is encoded in NHSAT. But the CCM index that TSAT cross maps NHSAT is of high value (0.24). According to the CCM theory, we know that the influence from NHSAT to TSAT, is not strong as the influence from TSAT to NHSAT, which also consists with the real dynamical process revealed by previous literature [6].

Finally, the information transfer inferred from the CCM suggests that: when employing Reservoir Computing to reconstruct TSAT from the NHSAT's records, the reconstruction quality will be better than reconstruct NHSAT from the TSAT's records. And our results are really consisting with the suggestion of CCM.

We will carefully improve the description of the CCM theory [1, 4, 5], and add the necessary description of the CCM computational algorithm, so that the results of the CCM and reconstruction quality will be better understood.

[1] 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.

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

[3] Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. Physics Reports, 441(1), 1-46, 2007.

[4] 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.

[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.

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

4. A second important concern is the way the ML is used. In Figure 2 there are three parts but it seems to me that the ML system is composed of the two first ones, the third one being the application of the optimized system to new input data. So **It should be worth to split both and also to clarify the details of the Machine Learning underlying structure, number of nodes, number of layers (if any)**… **Details on the different ML systems used are necessary. A detailed description is also missing for the CCM method**.

Response: Thanks for your comments and suggestions**.** By means of the first two components shown in Figure 1*, the $a(t)$ is trained and then $\psi[r^*(t)]$ is obtained. In this procedure, the value of $\psi[r^*(t)]$ is already very close to the value of $b(t)$. Then, if $\psi[r^*(t)]$ is feedback to function "*f*" and "ψ", this repetitive operation might make the value of $\psi[r^*(t)]$ more close to the value of $b(t)$. Actually we also found this repetitive operation no longer influenced the results. This is to say, that the third component shown in Figure 1* might be redundant in this reconstruction framework, and the first two components are enough.

The Reservoir Computer framework used in our work is developed in *Lu et al.* 2017 [1]. In *Lu et al.* 2017 [1], the Reservoir Computer framework only has the first two components shown in Figure 1*. We have tested the third component (a repetitive operation for the first two components) did not influence the results, and the first two components were enough. In the revised manuscript, we will carefully improve the diagram and the description of Reservoir computer according to the introduction in *Lu et al.* 2017 [1].

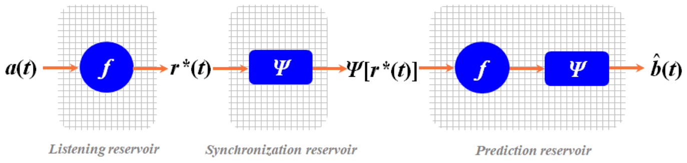
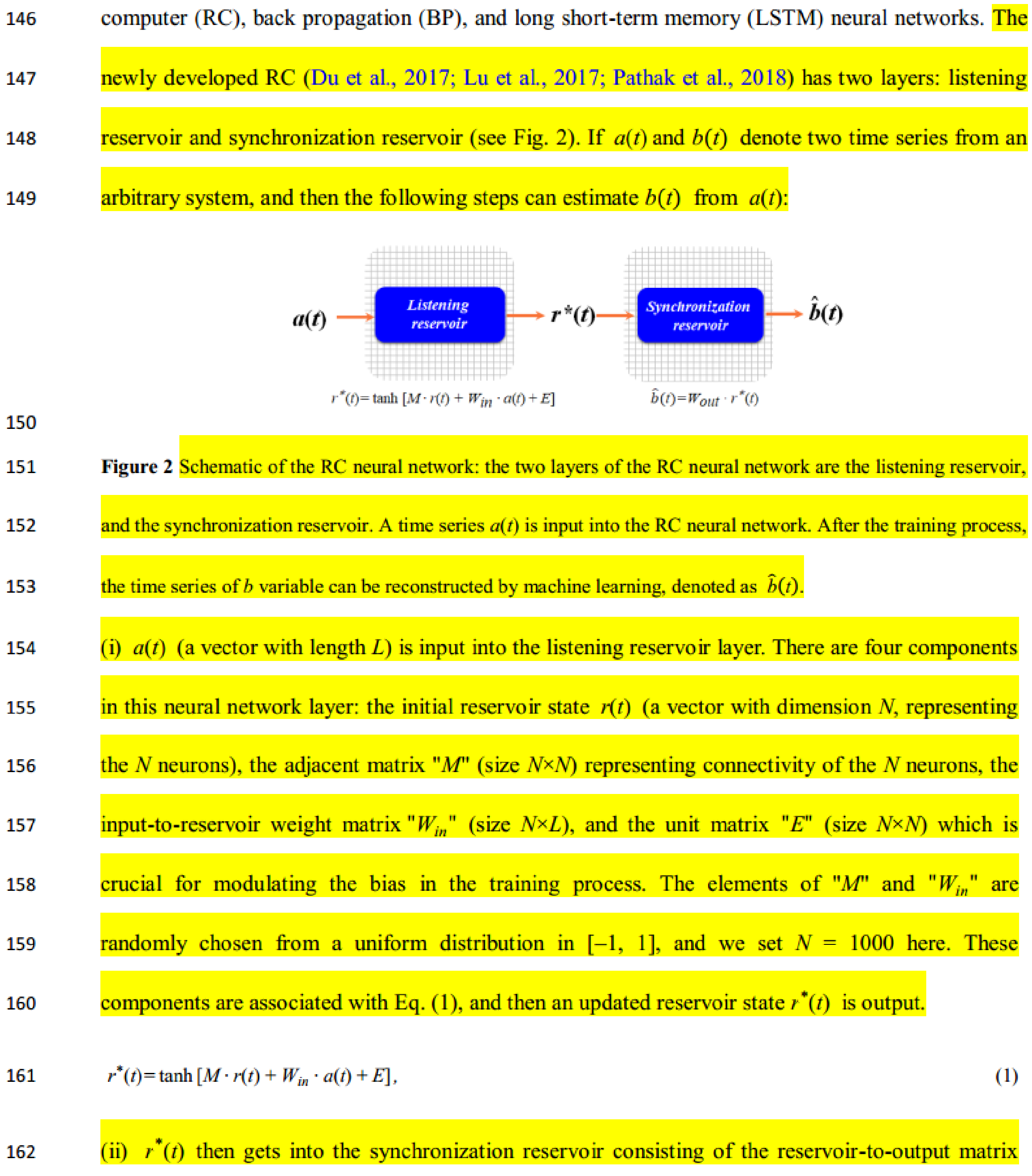


Figure 1* The schematic of Reservoir computer in the previous manuscript (we will revised this figure in the revised manuscript).

[1] 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.

Then, we will improve the detail description of Reservoir Computer, including the structure, number of nodes, number of layers, and so on. As the following screenshot shows:



computer (RC), back propagation (BP), and long short-term memory (LSTM) neural networks. The newly developed RC (Du et al., 2017; Lu et al., 2017; Pathak et al., 2018) has two layers: listening reservoir and synchronization reservoir (see Fig. 2). If $a(t)$ and $b(t)$ denote two time series from an arbitrary system, and then the following steps can estimate $b(t)$ from $a(t)$:

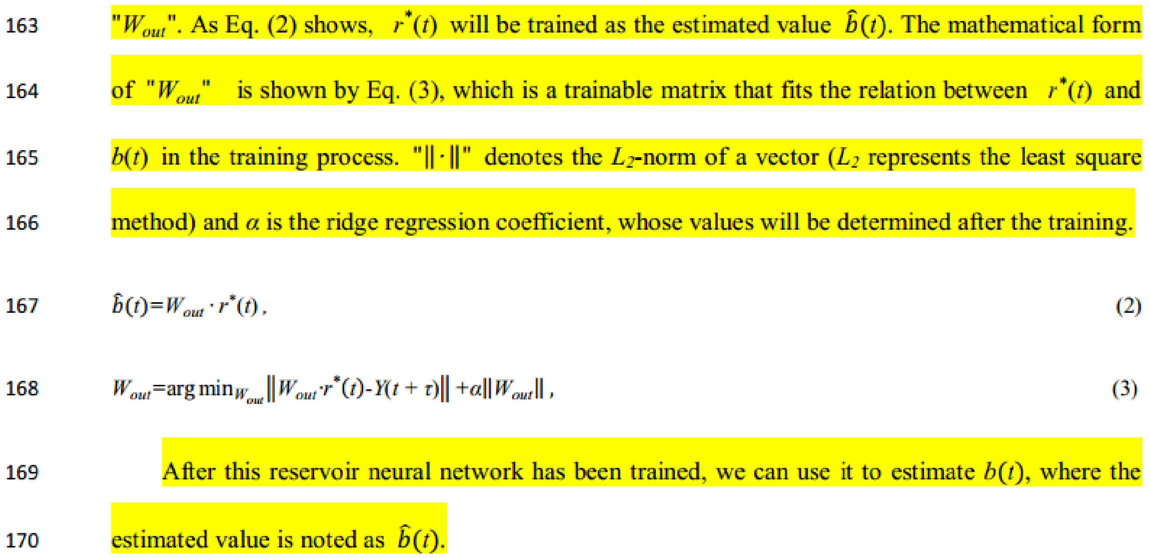
a(t)
Listening reservoir
r*(t)
Synchronization reservoir
b̂(t)
r*(t)= tanh [M · r(t) + W_in · a(t) + E]
b̂(t)=W_out · r*(t)

Figure 2 Schematic of the RC neural network: the two layers of the RC neural network are the listening reservoir, and the synchronization reservoir. 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 listening reservoir layer. There are four components in this neural network layer: the initial reservoir state $r(t)$ (a vector with dimension N, representing the N neurons), the adjacent matrix "M" (size N×N) representing connectivity of the N neurons, the input-to-reservoir weight matrix "W_{in}" (size N×L), and the unit matrix "E" (size N×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. These components are associated with Eq. (1), and then an updated reservoir state $r^*(t)$ is output.

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

(ii) $r^*(t)$ then gets into the synchronization reservoir consisting of the reservoir-to-output matrix



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

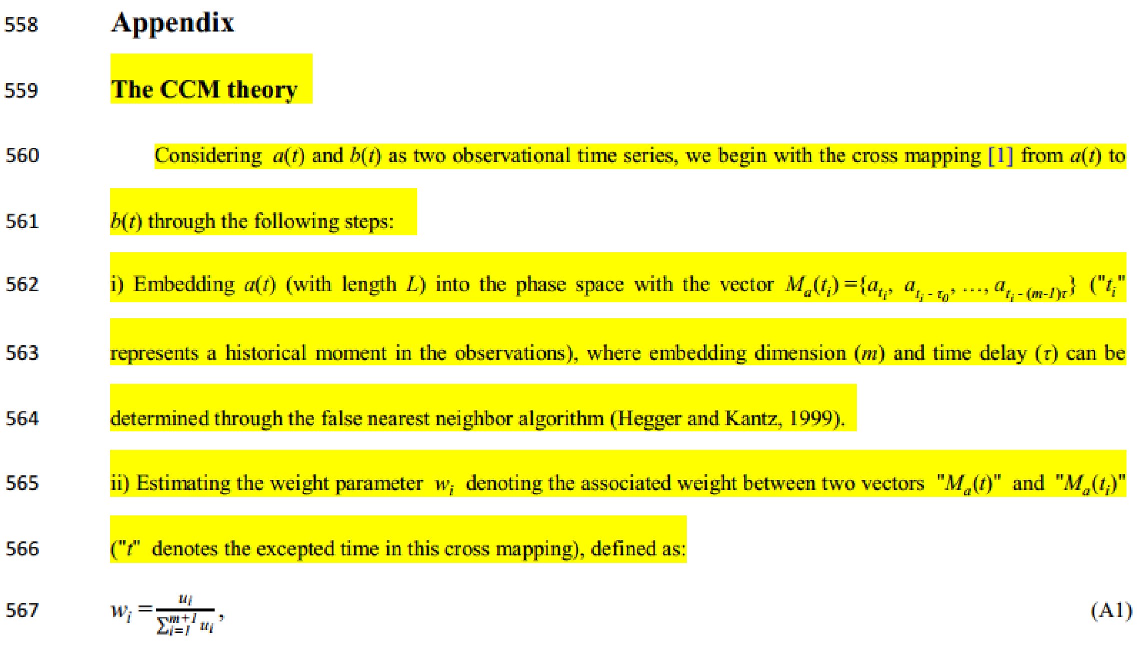
$$\hat{b}(t)=W_{out}\cdot r^*(t)\,, \tag{2}$$

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

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

For the details of LSTM and BP, since both of them have been widely used and well-known in many fields, and in recent years the Matlab language turns them into products for ease of use. Their underlying structures and usage guideline are open access in https://ww2.mathworks.cn/help/deeplearning. We will add the details of parameter setting in the revised manuscript.

Additionally, we will add the CCM computational algorithm into the revised manuscript, as the following screenshot shows:



Appendix

The CCM theory

Considering $a(t)$ and $b(t)$ as two observational time series, we begin with the cross mapping [1] from $a(t)$ to $b(t)$ through the following steps:

i) Embedding $a(t)$ (with length L) into the phase space with the vector $M_a(t_i)=\{a_{t_i},\ a_{t_i-\tau_0},\ \dots,\ a_{t_i-(m-1)\tau}\}$ ("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 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:

$$w_i=\frac{u_i}{\sum_{i=1}^{m+1}u_i}, \tag{A1}$$

$$u_i = exp\{- \frac{d\,[M_a(t),\, M_a(t_i)]}{d\,[M_a(t),\, M_a(t_1)]}\}, \qquad \text{(A2)}$$

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

$$\hat{b}(t) = \sum_{i=1}^{m+1} w_i b(t_i). \qquad \text{(A3)}$$

$\hat{b}(t)$ 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 Eq. (A4) shows:

$$\rho_{a \to b} = corr.\,[b(t),\ \hat{b}(t)]. \qquad \text{(A4)}$$

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): if b influence a but a does not influence b, the information of b can be shared with a (through the information transfer from b to a), but the information of a cannot be shared with b (there exists no information transfer from a to b). Hence, the records of a will be encoded with the information of b, and the time series of b can be recovered from the records of a. For the CCM index ($\rho_{a \to b}$), its magnitude represents how much information 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.

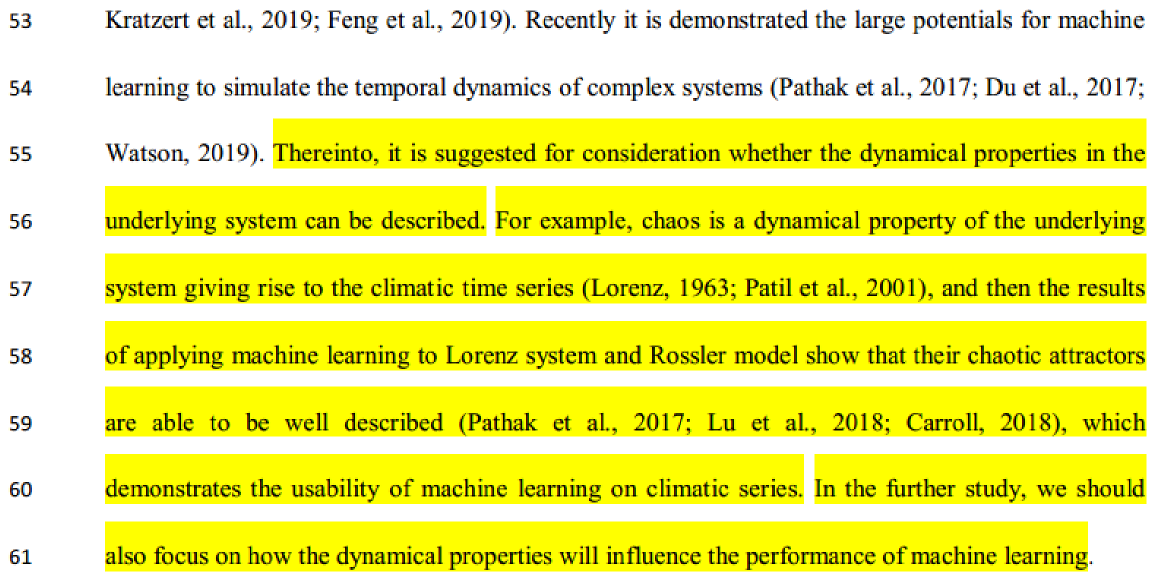
5. These two main problems prevent me to recommend publication of this manuscript at this stage although the main question addressed is very interesting (CCM vs ML). A considerable effort of clarification and rewriting is necessary.

Response: Thanks for your comments and suggestions! According to your above suggestions, we will carefully work on the more detailed clarification and rewriting for the machine learning method and the CCM theory, so that the relationship between CCM and machine learning can be better presented. And then, results and conclusions will be better understood.

More specific points:

6. Line 54: What does mean “wile physics of systems is suggested for consideration”? Please rephrase.

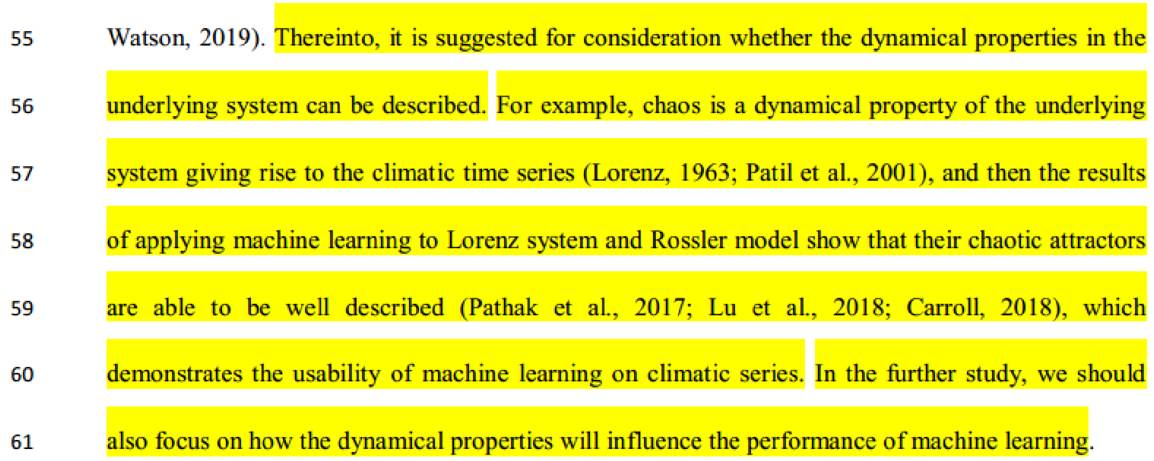
Response: Thank you! The excepted meaning is that: we should focus on whether the dynamical properties in the underlying system can be described, and how the dynamical properties will influence the performance of machine learning. We will revise these sentences as the following screenshot:



Kratzert et al., 2019; Feng et al., 2019). Recently it is demonstrated the large potentials for machine learning to simulate the temporal dynamics of complex systems (Pathak et al., 2017; Du et al., 2017; Watson, 2019). Thereinto, it is suggested for consideration whether the dynamical properties in the underlying system can be described. For example, chaos is a dynamical 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. In the further study, we should also focus on how the dynamical properties will influence the performance of machine learning.

7. Lines 57-58. You probably meant that: sensitivity to initial conditions is a property of the underlying system giving rise to the climate time series. Chaos theory is a framework in which this type of dynamics can be described. Please rephrase.

Response: Thank you! We will carefully rephrase these sentences, as the following screenshot shows:



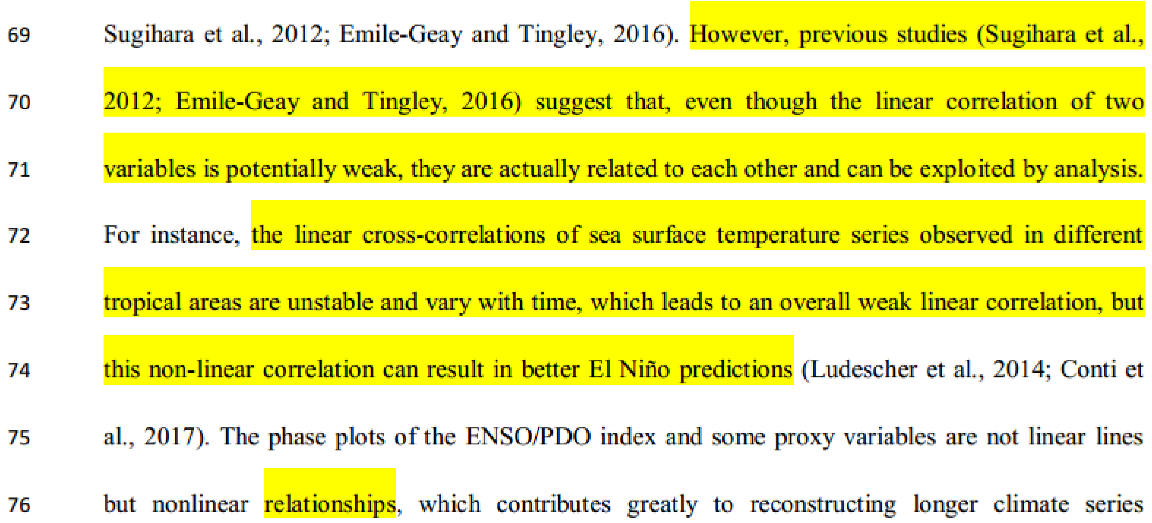
Watson, 2019). Thereinto, it is suggested for consideration whether the dynamical properties in the underlying system can be described. For example, chaos is a dynamical property of the underlying system giving rise to the climatic time series (Lorenz, 1963; Patil et al., 2001), and then the results 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. In the further study, we should also focus on how the dynamical properties will influence the performance of machine learning.

8. Line 67. What is nonlinear correlation? I think that this is not an appropriate terminology. Please revisit your manuscript with that in mind.

Response: Thank you! We will carefully rephrase the explanation of “nonlinear correlation” in the revised manuscript.

Here the excepted meaning of “nonlinear correlation” is that: for two variables from a common system, their time series might have dynamical relationship with each other. Sometimes the linear Pearson correlation of these two time series is weak or even equal to zero, but by means of some other statistical measurement their relationship can be quantified. At that time, such relationship whose linear correlation is potentially weak, is regarded as nonlinear correlation.

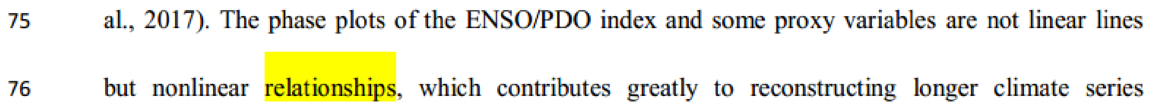
We will modify the sentences as the following screenshot:



Sugihara et al., 2012; Emile-Geay and Tingley, 2016). However, previous studies (Sugihara et al., 2012; Emile-Geay and Tingley, 2016) suggest that, even though the linear correlation of two variables is potentially weak, they are actually related to each other 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 can result in better El Niño predictions (Ludescher et al., 2014; Conti et al., 2017). The phase plots of the ENSO/PDO index and some proxy variables are not linear lines but nonlinear relationships, which contributes greatly to reconstructing longer climate series

9. Line 72. You speak about “trajectories”. Maybe this is more “relationships”.

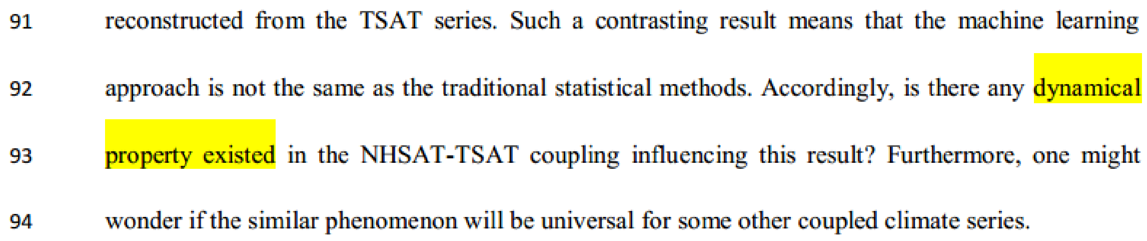
Response: Thank you! We will revise this word in the manuscript, as the screenshot shows:



al., 2017). The phase plots of the ENSO/PDO index and some proxy variables are not linear lines but nonlinear relationships, which contributes greatly to reconstructing longer climate series

10. Line 87. “hided”?

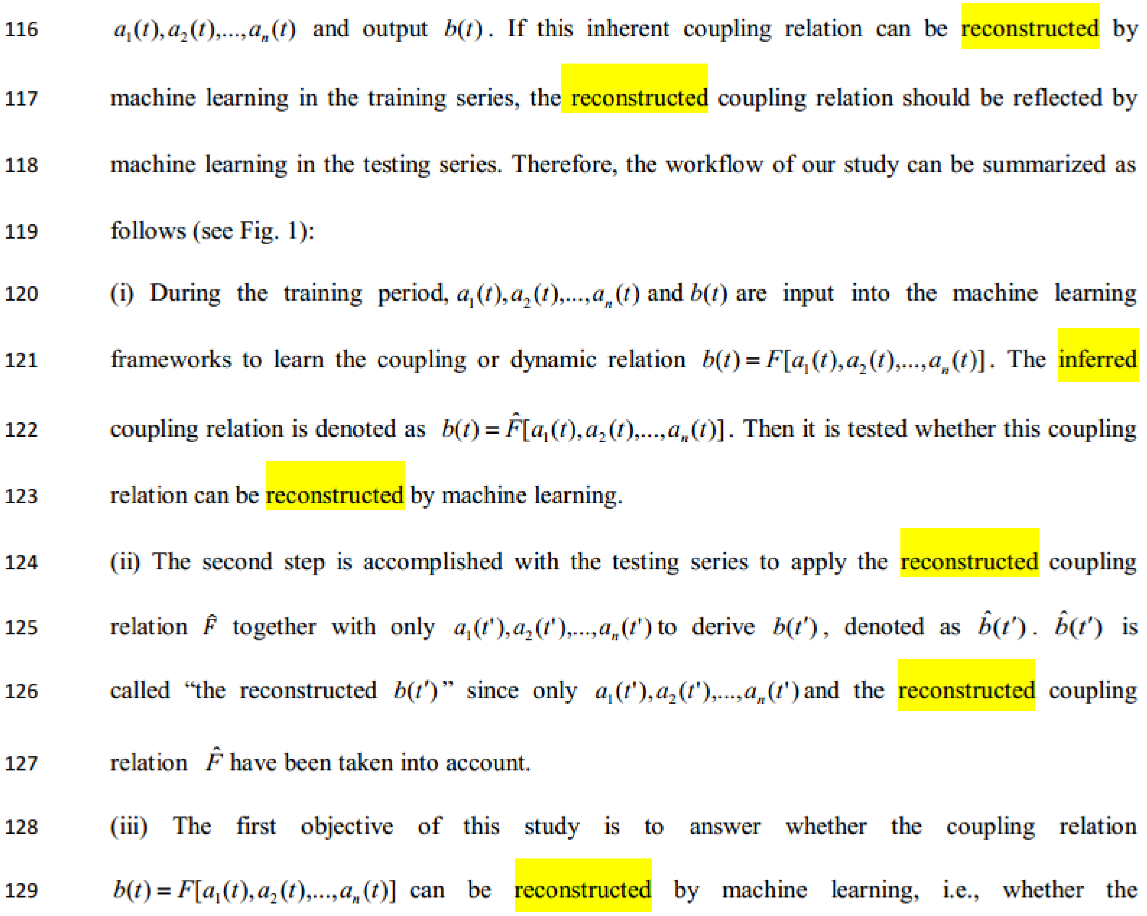
Response: Thank you! We will revise this word in the manuscript, as the screenshot shows:



reconstructed from the TSAT series. Such a contrasting result means that the machine learning approach is not the same as the traditional statistical methods. Accordingly, is there any dynamical property existed in the NHSAT-TSAT coupling influencing this result? Furthermore, one might wonder if the similar phenomenon will be universal for some other coupled climate series.

11. Line 111. “learnt” should probably be “reconstructed”.

Response: Thank you! We will revise this word in the manuscript, as the screenshot shows:



$a_1(t),a_2(t),...,a_n(t)$ and output $b(t)$. If this inherent coupling relation can be reconstructed by machine learning in the training series, the reconstructed coupling relation should be reflected by machine learning in the testing series. Therefore, the workflow of our study can be summarized as follows (see Fig. 1):

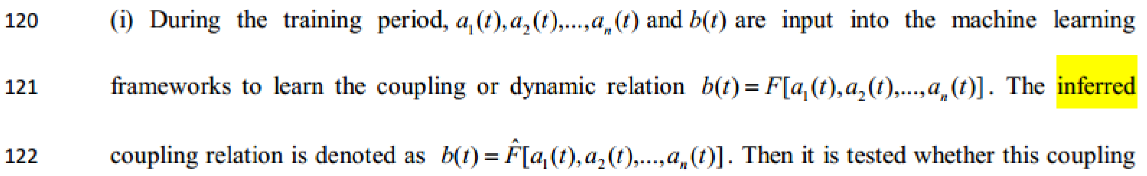
(i) During the training period, $a_1(t),a_2(t),...,a_n(t)$ and $b(t)$ are input into the machine learning frameworks to learn the coupling or dynamic relation $b(t)=F[a_1(t),a_2(t),...,a_n(t)]$. The inferred coupling relation is denoted as $b(t)=\hat{F}[a_1(t),a_2(t),...,a_n(t)]$. Then it is tested whether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

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

(iii) The first objective of this study is to answer whether the coupling relation $b(t)=F[a_1(t),a_2(t),...,a_n(t)]$ can be reconstructed by machine learning, i.e., whether the

12. Line 115. “learnt” is probably “estimated” or “inferred”.

Response: Thank you! We will revise this word in the manuscript, as the screenshot shows:



(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

13. Figure 1. Why putting the training after the testing? It does not look natural (and also confusing).

Response: Thanks for your suggestions. Such arrangement is due to the consideration of reconstructing climate records. We are inspired by that it is often necessary to reconstruct the historical records for climate variables.

For instance, as Figure 2* shows, for the records of proxy data (tree ring or ice core, labeled as $a(t)$ in Figure 2*), we might obtain the data from the historical and current period. For the records of climatic variable like air temperature (labeled as $b(t)$ in Figure 2*), we might only obtain the data from the current period. At that time, the data-driven approach (such linear regression) is often applied to fit the relation between proxy data ($a(t)$) and air temperature ($b(t)$) through their current observational data, and then the historical proxy data and the fitted relationship can be used to reconstruct the historical records of air temperature.

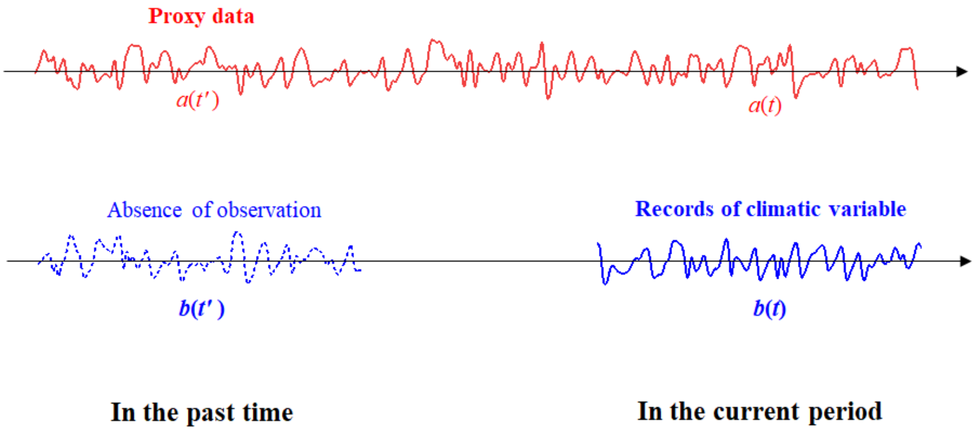


Figure 2* The blue solid line denotes the observational records of climatic variable (labeled as $b(t)$) in current period. The blue dashed line denotes that the records of climatic variable are absence of observation in the past time. The red solid line denotes the proxy data (labeled as $a(t)$) in both of current period and past time.

The above reconstruction scheme is also very useful for some important climate problems such as **paleoclimate reconstruction** [1], **interpolation for the missing points in measurements** [2] and **parameterization schemes** [3]. Our study is motivated by investigating how to better apply machine learning to the reconstruction of climate time series (under different coupling dynamics of climate systems).

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

[2] 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.

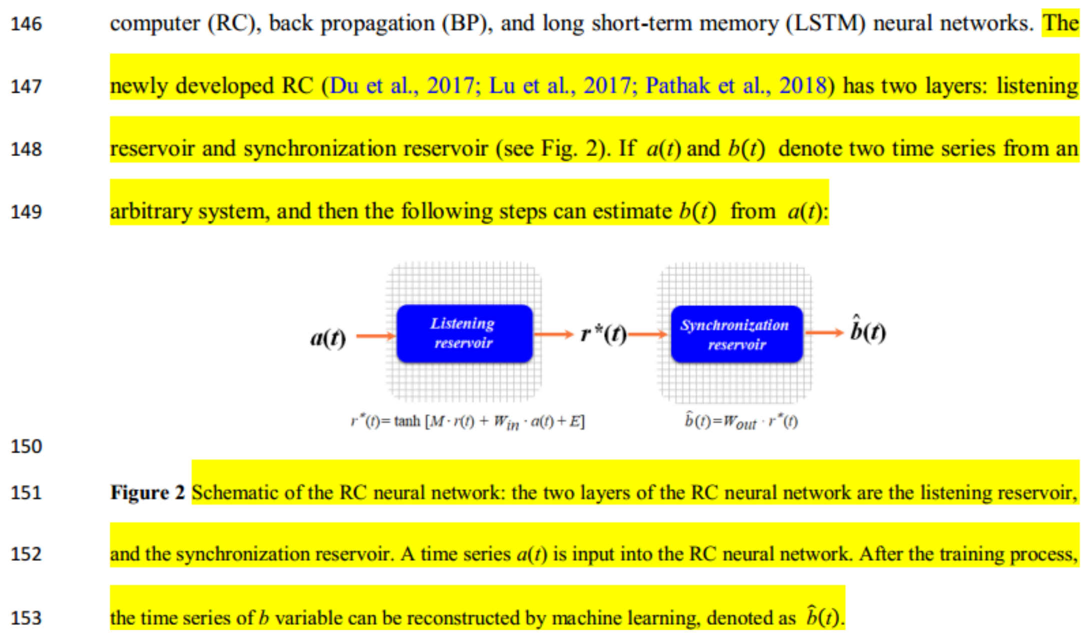
[3] Vissio, G., Lucarini, V.: A proof of concept for scale‐adaptive parameterizations: the case of the Lorenz 96 model. Q. J. Roy. Meteor. Soc., 144(710), 63-75, 2018.

14. Lines 175-178. Quite confusing. Please clarify the way prediction is done. I think that the presentation of the ML approach should be completely revisited.

Response: Thanks for your suggestions. We will thoroughly rewrite this part about the machine learning framework, and detail description of Reservoir Computer, including the structure, number of nodes, number of layers will be clearly presented.

The Reservoir Computer framework used in our work is developed in *Lu et al.* 2017 [1]. And we will refer the introduction in *Lu et al.* 2017 [1] to modify the description. Our modified version will be as the screen shot shows:

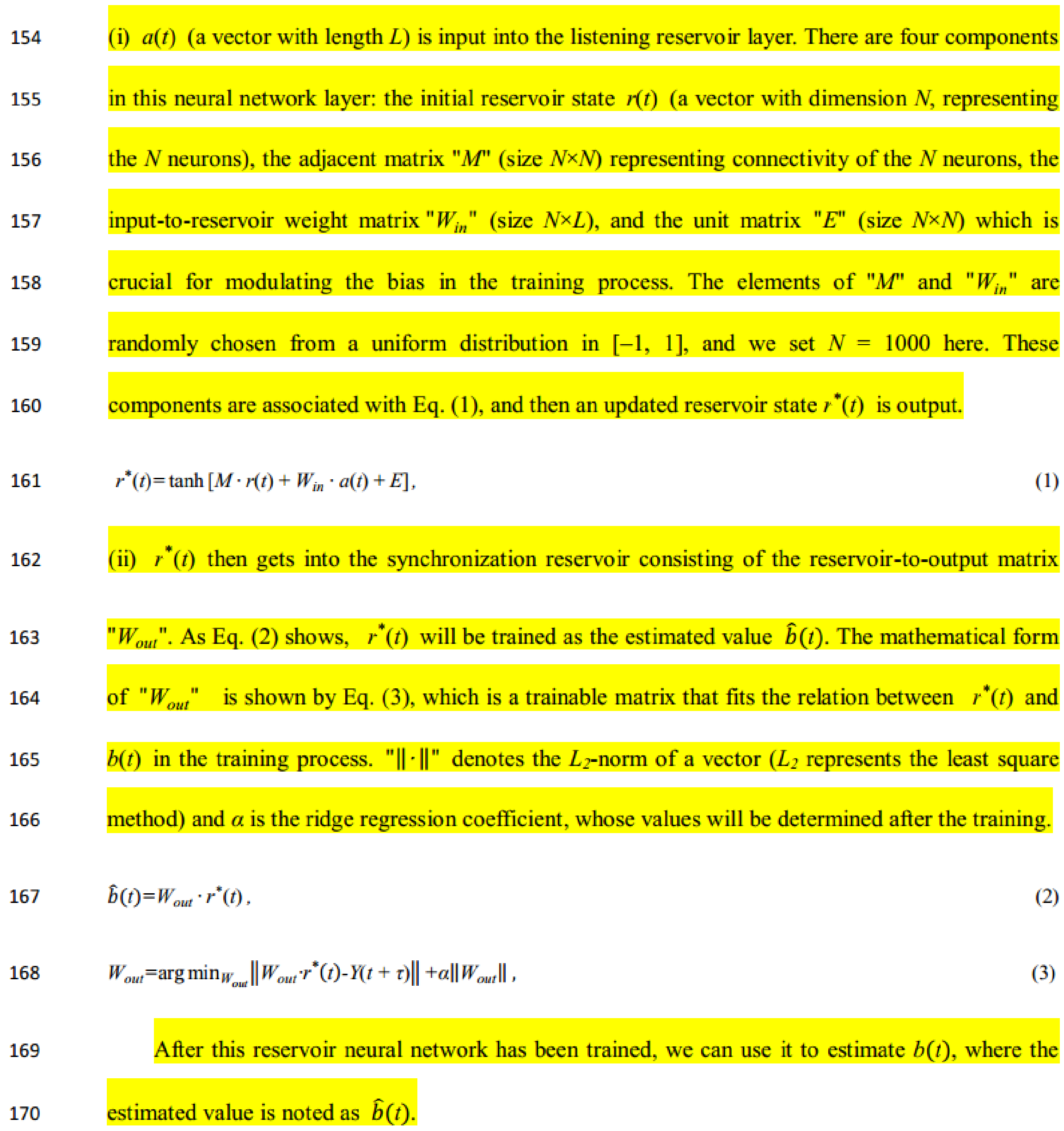
[1] Lu Z, Pathak J, Hunt B, Girvan M, Brockett R, Ott E. Reservoir observers: Model-free inference of unmeasured variables in chaotic systems. Chaos 27(4), 041102 (2017).



computer (RC), back propagation (BP), and long short-term memory (LSTM) neural networks. The newly developed RC (Du et al., 2017; Lu et al., 2017; Pathak et al., 2018) has two layers: listening reservoir and synchronization reservoir (see Fig. 2). If $a(t)$ and $b(t)$ denote two time series from an arbitrary system, and then the following steps can estimate $b(t)$ from $a(t)$:

$a(t)$
Listening reservoir
$r^*(t)$
Synchronization reservoir
$\hat{b}(t)$
$r^*(t)=\tanh[M\cdot r(t)+W_{in}\cdot a(t)+E]$
$\hat{b}(t)=W_{out}\cdot r^*(t)$

Figure 2 Schematic of the RC neural network: the two layers of the RC neural network are the listening reservoir, and the synchronization reservoir. 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 listening reservoir layer. There are four components in this neural network layer: the initial reservoir state $r(t)$ (a vector with dimension N, representing the N neurons), the adjacent matrix "M" (size N×N) representing connectivity of the N neurons, the input-to-reservoir weight matrix "W_{in}" (size N×L), and the unit matrix "E" (size N×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. 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 synchronization reservoir consisting of the reservoir-to-output matrix "W_{out}". As Eq. (2) shows, $r^*(t)$ will be trained as the estimated value $\hat{b}(t)$. The mathematical form of "W_{out}" is shown by Eq. (3), which is a trainable matrix that fits the relation between $r^*(t)$ and $b(t)$ in the training process. "$\|\cdot\|$" denotes the L_2-norm of a vector (L_2 represents the least square method) and α is the ridge regression coefficient, whose values will be determined after the training.

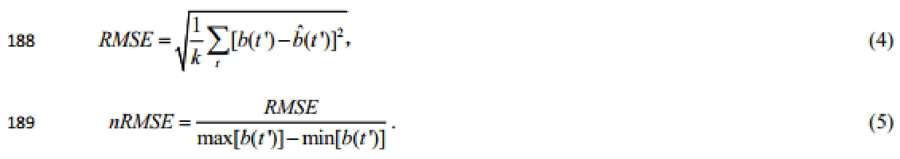
$$\hat{b}(t)=W_{out}\cdot r^*(t), \tag{2}$$

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

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

15. Line 191. Why using this measure and why 0.1 is a good threshold? These should be detailed.

Response: Thanks for your suggestions. Normalizing the RMSE is to compare the time series with different variability and unit [1, 2]. For instance, the time series of x_1 and x_2 in Figure 3* are both with zero mean and unit variance, but the extreme values of x_2 are much stranger than of x_1. It is revealed [1, 2] that such difference will interfere in the fair comparison of the RMSE. In order to avoid such interference induced by the extreme values, we are suggested to normalize the RMSE with the max distribution range of the original data [1, 2], as equation (5) shows.



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

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

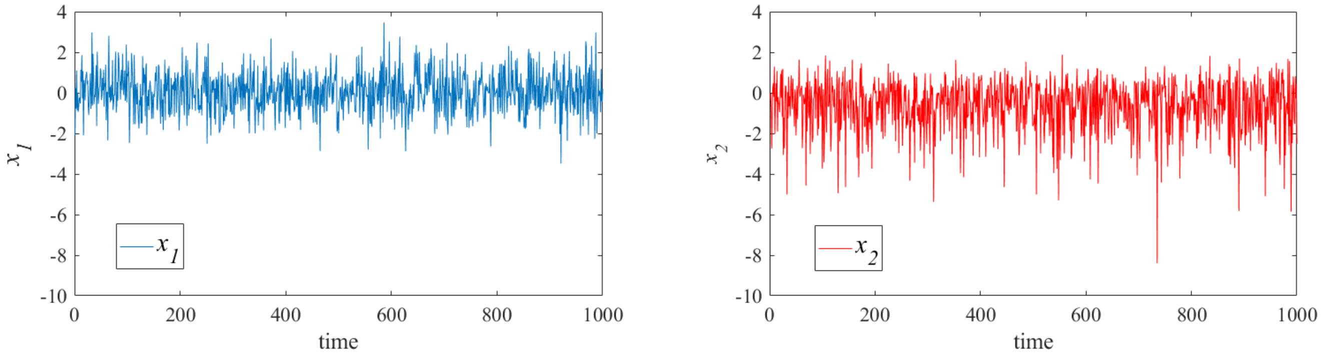


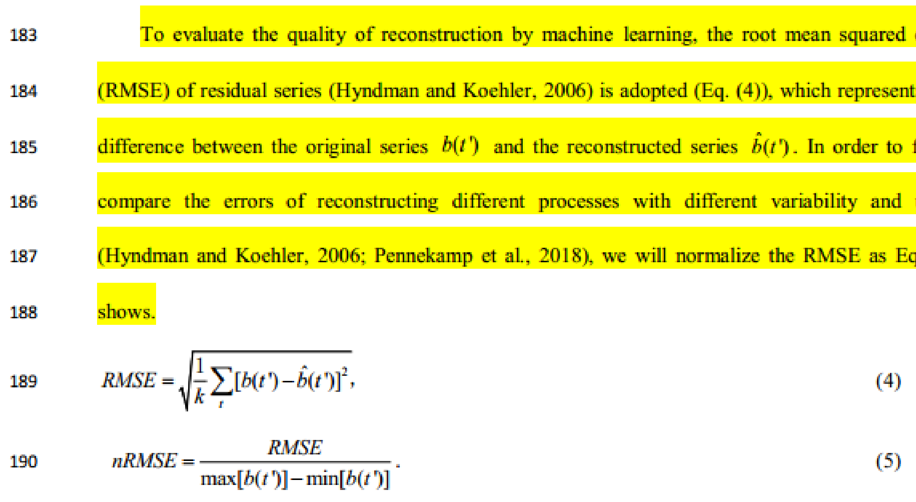
Figure 3* The standardized time series of x_1(blue) and x_2 (red) with zero mean and unit variance. The x_1 is a random time series with Gaussian probability distribution, and x_2 is a random time series with extreme probability distribution.

"nRMSE = 0.1" means that the RMSE occupies 10% of the max distribution range of the original data, and this is a tolerable level of the bias [1, 2]. In the figures of comparing reconstructed series with real series, we can observe that when the reconstructed series is close to the real series in curves, the corresponding nRMSE is less than 0.1.

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

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

We will carefully explain the meaning of nRMSE and its threshold in the revised manuscript, as the following screenshot shows:



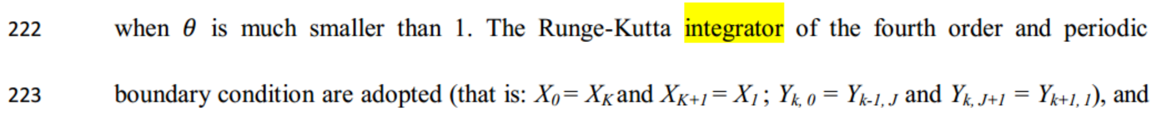
To evaluate the quality of reconstruction by machine learning, the root mean squared error (RMSE) of residual series (Hyndman and Koehler, 2006) is adopted (Eq. (4)), which represents the difference between the original series $b(t')$ and the reconstructed series $\hat{b}(t')$. In order to fairly compare 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.

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

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

16. Line 212. Runge-Kutta integral? What does it mean? Maybe “integrator”?

Response: Thanks for your suggestions. We will revise this word in the manuscript, as the screenshot shows:



when θ 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 $X_{K+1} = X_1$; $Y_{k,0} = Y_{k-1,J}$ and $Y_{k,J+1} = Y_{k+1,1}$), and

17. Section 2.4.2. Please give more details on the way average is done, and whether the seasonality is removed and how?

This also open the question on how the parameters of the ML are changing as a function of the season. There is not enough details on how the datasets are handled.

Response: Thanks for your suggestions. We will improve the details on the way average is done in the manuscript.

The seasonality was not removed, and this did not influence the parameters of the machine learning. The reasons are as the following shows:

Firstly, literature [1-4] has revealed that seasonal cycle of air temperature is time-varying (especially for the mid-latitude regions [1] and tropics [2]), and the existing methods are often hard to thoroughly remove such time-varying seasonal cycle [4]. So that removing seasonality might take some controversial and unknown bias for the results [5].

[1] Paluš, M., Novotná, D., Tichavský, P.: Shifts of seasons at the European mid‐latitudes: Natural fluctuations correlated with the North Atlantic Oscillation. Geophysical research letters, 32(12), 2005.

[2] Qian, C., Wu, Z., Fu, C., Wang, D.: On changing El Niño: A view from time-varying annual cycle, interannual variability, and mean state. Journal of Climate, 24(24), 6486-6500, 2011.

[3] Jajcay, N., Hlinka, J., Kravtsov, S., Tsonis, A. A., Paluš, M.: Time scales of the European surface air temperature variability: The role of the 7–8 year cycle. Geophysical Research Letters, 43(2), 902-909, 2016.

[4] Deng, Q., Nian, D., Fu, Z.: The impact of inter-annual variability of annual cycle on long-term persistence of surface air temperature in long historical records. Climate dynamics, 50(3-4), 1091-1100, 2018.

[5] Theiler, J., Eubank, S.: Don’t bleach chaotic data. Chaos: An Interdisciplinary Journal of Nonlinear Science, 3(4), 771-782, 1993.

Secondly, if focusing on the application in reconstructing regional temperature [6-8], the annual variability will be the most important and commonly concerned. At that time, the seasonality is not necessary to be removed. And as the Figure 4* shows, the annual variability of reconstructed series is really close to the real series. If we remove the seasonality, it might take with some unknown bias [4-5].

[6] Van Engelen, A. F., Buisman, J., Jnsen, F.: A millennium of weather, winds and water in the low countries. In History and climate (pp. 101-124). Springer, Boston, MA, 2001.

[7] Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M., Karlen, W.: 2,000-year Northern Hemisphere temperature reconstruction. IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series, 19, 2005.

[8] Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ni, F.: Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. Science, 326(5957), 1256-1260, 2009.

Thirdly, when employing neural network approach, it is a common step to divide the data into training data and testing data. Then the training data is used to train the parameters of neural network. After the training process is accomplished, the parameters of neural network will be determined and fixed. And then, the trained neural network will be used in the testing data, and they will be not changed any more.

Fourthly, if dividing the time series into different seasons, and respectively reconstructing them in different seasons, the parameters of machine learning might be changing in different seasons. However, after dividing these daily time series into different seasons, the data length will be not long enough to accomplish the machine learning approach, which might take the large bias to the results. So, we did not divide the time series according to different seasons, and the seasonality will not influence the parameters of machine learning changing with the season.

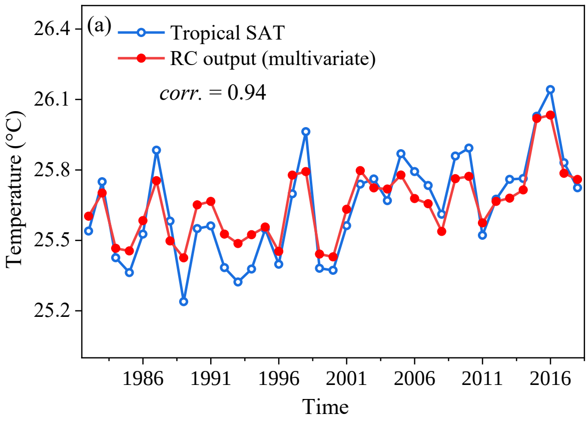
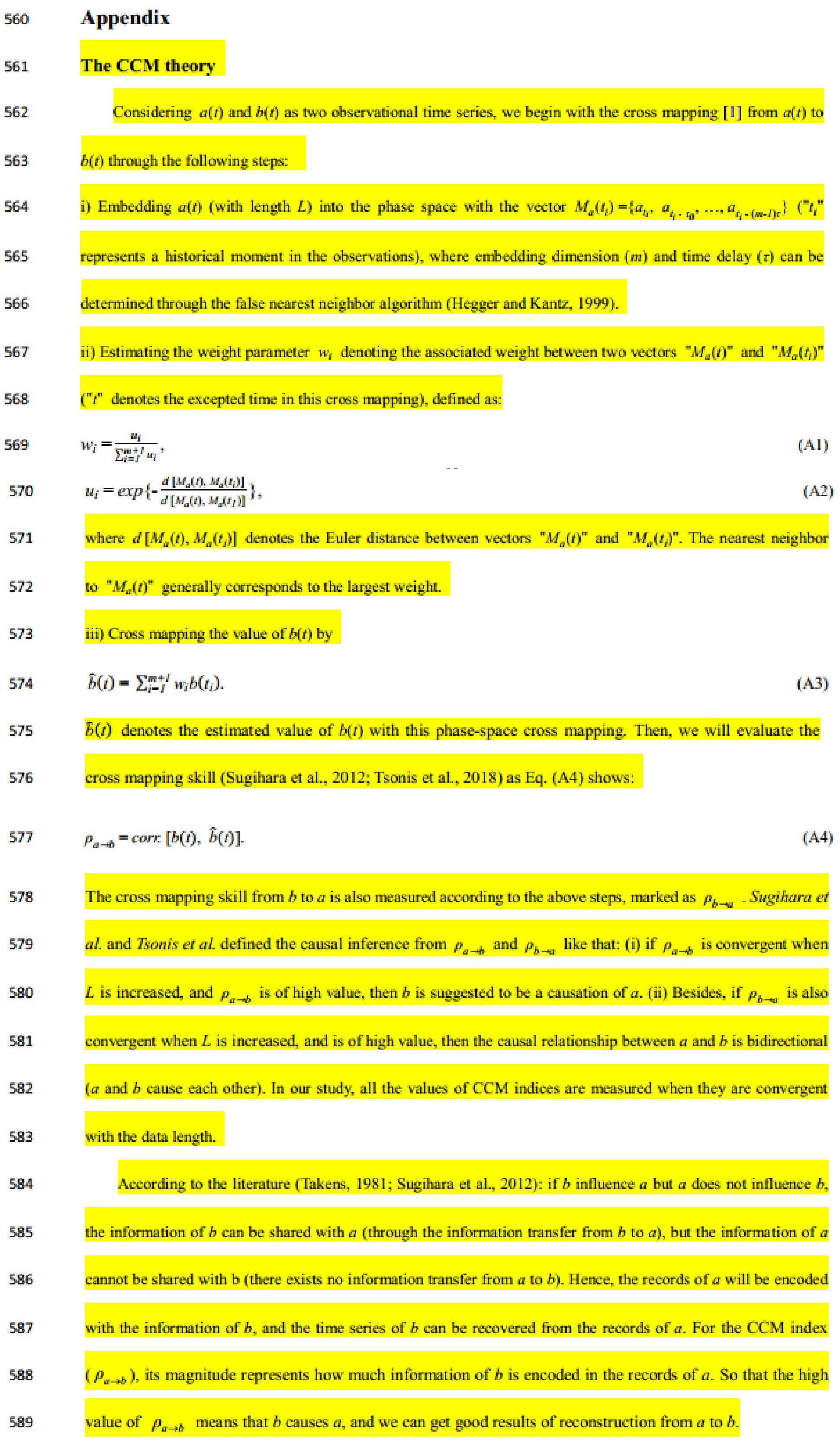


Figure 4* Comparison between the annual mean values of reconstructed TSAT (red) and the annual mean values of original TSAT (blue).

18. Lines 295-296. Sugihara (1994). This reference does not exist in the reference list. What is “empirical dynamics model? Much more information is needed on the way it is used. Embedding dimension and so on.

Response: Thanks for your suggestions. We will revise this part in the manuscript, as the screenshot shows:



Appendix

The CCM theory

Considering $a(t)$ and $b(t)$ as two observational time series, we begin with the cross mapping [1] 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 (τ) 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:

$$w_i=\frac{u_i}{\sum_{i=1}^{m+1} u_i}, \tag{A1}$$

$$u_i=exp\{-\frac{d\,[M_a(t),\ M_a(t_i)]}{d\,[M_a(t),\ M_a(t_1)]}\}, \tag{A2}$$

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

$$\hat{b}(t) = \sum_{i=1}^{m+1} w_i b(t_i). \tag{A3}$$

$\hat{b}(t)$ 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 Eq. (A4) shows:

$$\rho_{a\to b}=corr.\ [b(t),\ \hat{b}(t)]. \tag{A4}$$

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.

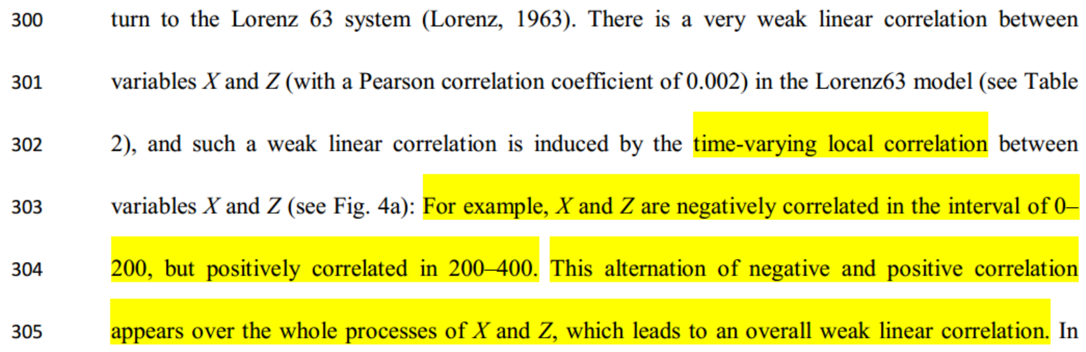
According to the literature (Takens, 1981; Sugihara et al., 2012): if b influence a but a does not influence b, the information of b can be shared with a (through the information transfer from b to a), but the information of a cannot be shared with b (there exists no information transfer from a to b). Hence, the records of a will be encoded with the information of b, and the time series of b can be recovered from the records of a. For the CCM index ($\rho_{a\to b}$), its magnitude represents how much information 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.

19. Line 302. What is “unstable local correlation”. What is this?

Response: Thank you! The expected meaning of "unstable local correlation" is that the local Pearson correlation between two variables is time-varying. As the Figure 5*(a) shows, the time series of X and Z are sometimes positively correlated but sometimes nonlinear correlated at different regimes. Hence, the overall Pearson correlation between X and Z is very weak. Such time-varying local Pearson correlation is suggested to be universal in nonlinear dynamical systems [1].

[1] 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 modify the word in the revised manuscript for better understanding, as the following screenshot shows:



turn to the Lorenz 63 system (Lorenz, 1963). There is a very weak linear correlation between variables X and Z (with a Pearson correlation coefficient of 0.002) in the Lorenz63 model (see Table 2), and such a weak linear correlation is induced by the time-varying local correlation between variables X and Z (see Fig. 4a): For example, X and Z are negatively correlated in the interval of 0–200, but positively correlated in 200–400. This alternation of negative and positive correlation appears over the whole processes of X and Z, which leads to an overall weak linear correlation. In

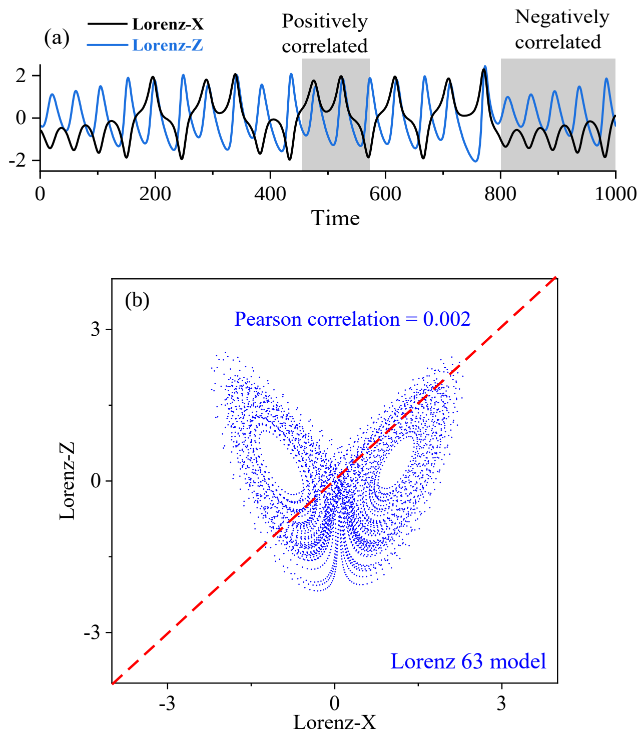


Figure 5* (a) The X time series (black) and the Z time series (blue) of the Lorenz 63 system. (b) Scatter plot of X time series and Z time series of the Lorenz 63 model (blue dots).

20. Table 2. As already mentioned in my main comment, very confusing. Please modify.

Response: Thanks for your suggestions. The results and conclusion of Table 2 is correct (see also *Lu et al. 2017*[1]), and this confusion is induced by the lack description of the CCM theory. After the CCM theory is well explained in the manuscript, the result can be better understood.

[1] 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).

21. Figure 6. Some typos in titles. Also where is panel (d)? Is it (c)?

Response: Thank you! We will revise this typo in the manuscript, as the screenshot shows:

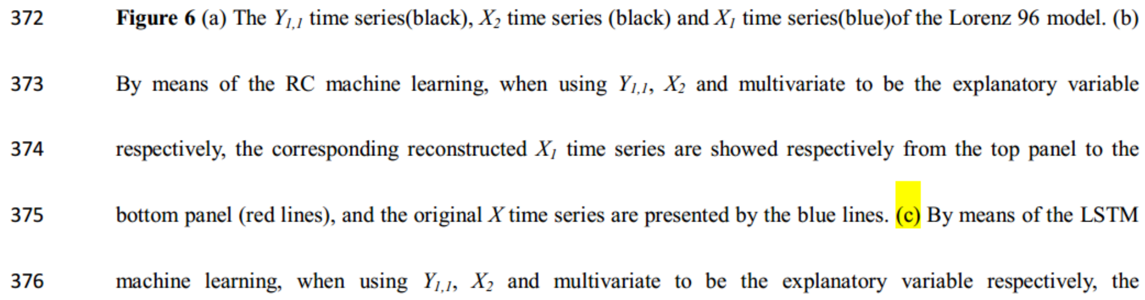


Figure 6 (a) The $Y_{1,1}$ time series(black), X_2 time series (black) and X_1 time series(blue)of the Lorenz 96 model. (b) By means of the RC machine learning, when using $Y_{1,1}$, X_2 and multivariate to be the explanatory variable respectively, the corresponding reconstructed X_1 time series are showed respectively from the top panel to the bottom panel (red lines), and the original X time series are presented by the blue lines. (c) By means of the LSTM machine learning, when using $Y_{1,1}$, X_2 and multivariate to be the explanatory variable respectively, the

22. Table 3 and Fig 6. Why not using a multivariate CCM to compare with the ML fitting with multiple predictors?

Response: Many thanks for your suggestions! The multi-variable CCM analysis might be useful and promising, but first of all we need to know which variable is able to become the explanatory variable. Similar to the multi-variable regression analysis, if we do not know the Pearson correlation between the target variable with every potential explanatory variable, the multi-variable regression will easily suffer from the overfitting problem.

Considering the potential **overfitting problem** and **common-driver problem** [1-2], the comparison between the multi-variable CCM and the multi-variable machine learning **absolutely deserves a further investigation**. This might occupy too many words and figures in the manuscript, so that the presentation of the main and original ideal might be influenced. In the future study, we will consider a thorough investigation for the comparison between the multi-variable CCM and the multi-variable machine learning.

[1] Runge, J., Heitzig, J., Petoukhov, V., Kurths, J.: Escaping the curse of dimensionality in estimating multivariate transfer entropy. Physical review letters, 108(25), 258701, 2012.

[2] Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., van Nes, E. H.: Inferring

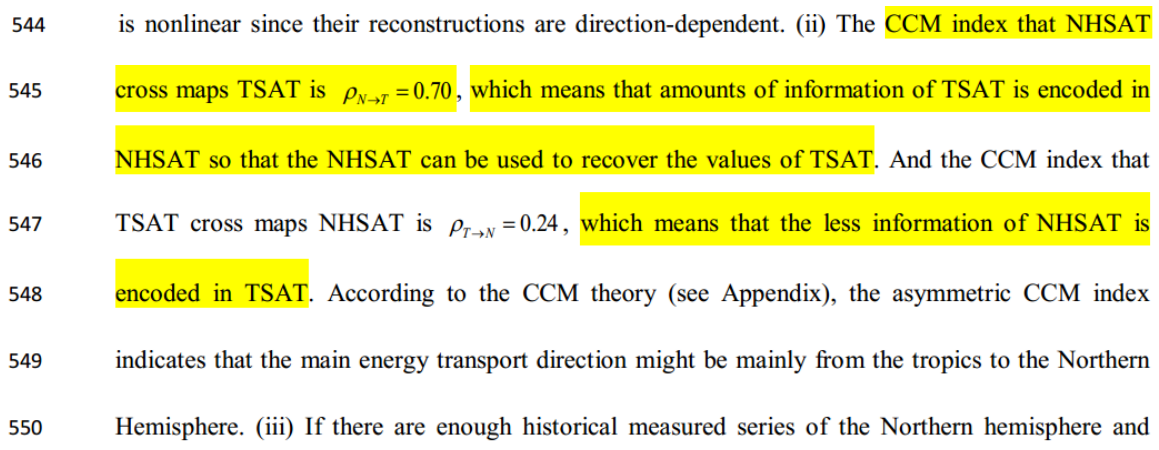
causation from time series in Earth system sciences. Nature communications, 10(1), 1-13, 2019.

23. Lines 536-543. Really confusing. What is influencing what? TSAT or NHSAT?

Response: Thanks for your suggestions. The excepted meaning is that TSAT influences NHSAT, which can be explained by that the energy is transferred from the tropical climate system to the Northern Hemispheric climate system [1].

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

We will improve the description as the following shows:



is nonlinear since their reconstructions are direction-dependent. (ii) The CCM index that NHSAT cross maps TSAT is $\rho_{N\to T} = 0.70$, which means that amounts of information of TSAT is encoded in NHSAT so that the NHSAT can be used to recover the values of TSAT. And the CCM index that TSAT cross maps NHSAT is $\rho_{T\to N} = 0.24$, which means that the less information of NHSAT is encoded in TSAT. According to the CCM theory (see Appendix), the asymmetric CCM index indicates that the main energy transport direction might be mainly from the tropics to the Northern Hemisphere. (iii) If there are enough historical measured series of the Northern hemisphere and

24. I have also noted many typographical errors, and the manuscript will benefit for a careful reading by the authors and by an English native speaker to rephrase some sentences.

Response: Thanks for your suggestions. We will carefully inspect the manuscript, and later than we will also invite a colleague of our field speaking native English to improve some sentences.