Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2019-63-RC2, 2020

Earth System Dynamics
Discussions

Interactive comment on "Reconstructing coupled time series in climate systems by machine learning" *by* Yu Huang et al.

Anonymous Referee #2

Received and published: 7 February 2020

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.

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. 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. 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), 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). The same goes for the description of Pearson's correlation coefficient, its description should be shifted from the Results to the Methodology section. 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. The conclusion section should be shortened.

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.

Specific Points:

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

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.

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.

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.

Line 105: Typographical error: it should be "Learning" not "Leaning".

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.

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

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

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).

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.

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

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.

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

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.

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*.

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

Figure 8: It is missing the R2 and p-value of LSTM. The behavior of LSTM should also

be evaluated in the same manner.

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

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2019-63, 2019.