

Detecting dynamical causality via intervened reservoir computing



Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

In this manuscript, the authors have designed a novel method to infer causal relations within complex systems by implementing the concept of controlled trials with only data-driven techniques. The topic is quite important among various disciplines and the idea of controlled trials also has been central in human understanding of causality. The designed method successfully incorporates such idea and represents differences and advantages over existing methods. Numerical examples and several real-world datasets verify the effectiveness of the proposed method. Overall, the manuscript is well-organized and easy to follow, with persuasive experiments provided. However, there are still a few issues that need to be addressed before I can recommend publication of this manuscript, as follows.

1. Correct the erroneous cross-references to equations denoted as 'Eq.' or 'eq.'.
2. It is important for the authors to show how this method is affected by different levels of external noise perturbations, as noise always exists in the datasets.
3. To reproduce the results and for the convenient use of readers, I suggest the authors to provide necessary codes affiliated to the manuscript, which is quite important especially for articles in Communications Physics with high influence.
4. Some most recent computational methods in handling dynamic information by reservoir computing are missing. I suggest that the following relevant literature should be addressed.
---Autoreservoir computing for multistep ahead prediction based on the spatiotemporal information transformation. Nature Communications, 2020, 11(1): 1-15. DOI: 10.1038/s41467-020-18381-0.
---Earthquake alerting based on spatial geodetic data by spatiotemporal information transformation learning. Proceedings of the National Academy of Sciences, 2023, 120(37):e2302275120. DOI:10.1073/pnas.2302275120
5. The authors should clearly distinguish the variables/symbols used when describing the closed-loop and intervened-loop. In current version, variable x^{t+1} may cause confusion as it represents the input/output in both loops. This is also the case for the reservoir state r used in both loops.
6. The equation in Fig. 1 seems different with its definition in eq. (12), please carefully check.
7. The predicted output of RC, as in eq. (4), looks abnormal with the common use (i.e., $W_{out}^T r$), please check if the matrix multiplication is correct.
8. Section 2.4 describes that 'A filtering process...by setting a minimum threshold value $j_{threshold}$.' How this process is applied and to which variable? This should be more clearly described.
9. The authors should have a fair and detailed discussion on the limitations of the proposed method.

Reviewer #2 (Remarks to the Author):

Comments on manuscript "Detecting dynamical causality via intervened reservoir computing" by Zhao et al.

The authors proposed a causality detection method based on reservoir computer (RC). Their algorithm can identify causal direction between coupled time series, and reveal causal-relationship network among multiple time series. Their attempt to apply RC in causal-relationship network is new and interesting, because the RC does good to deal with nonlinear dynamics and computation efficiency. Their testing results on the coupled Lorenz systems and air-pollution dataset demonstrate the good performance of their method. But I still have concerns and suggestions in terms of the algorithm design and some necessary interpretations as below.

Major Comments:

1. In abstract, the authors claimed that their method is different from correlation-based methods, but they did not further explain which of the comparable methods they used (GC, CCM, TE) are correlation-based methods, and how their IRC method is different from the correlation-based methods in the sense of designing the algorithm. This is confusing, and the authors should clarify it.

At least the authors should clarify how their method is different from GC or CCM. In my observation for Sections 2.1 and 2.4, the main difference is that the IRC employs trajectory deviation indices (TDI) to evaluate the prediction skill of RC. But to me it still belongs to the framework of Granger causality method. If their method does not propose a framework totally independent on GC or CCM, they should point out which framework of GC and CCM their method follows and improves on, rather than vaguely claiming their method is different from correlation-based method. Otherwise this will mislead the readers from the community.

2. The second concern is about reconstructing the causal network. Refer to a recently popular causal-network algorithm (PCMCI), there are three crucial factors to influence the causal inference for multi variables: (i) the autocorrelation of the time series; (ii) indirect chain effect to induce fake causality between the pair of variables; (iii) the common driver effect to induce fake causality between the pair of variables. Thus the PCMCI particularly take the partial correlation procedure. Large amount of recent studies already demonstrate these three factors will do influence the correct inference of time series causality. How did the authors deal with these three factors? Or how well their methods can perform under the cases of these three factors? I did not find the relevant discussions in the manuscript, and the authors also did not mention PCMCI although they want to focus on causal network. Building a RC-based causal-network algorithm is promising, but these three factors are crucial for causal networks. Please add relevant discussions and investigations.

3. Normally the causality methods would introduce the suitable statistical significant test for the causality inference. But the authors did not mention that. Does the TDI include both of causality intensity and significance test? Please clarify it.

4. The analysis on El Nino indices is wired. Firstly, the authors did not introduce if the SST time series include seasonal cycle. Secondly, the authors employ the changing states of Walker circulation to explain the bidirectional causality of the El Nino indices. Their claim that Walker Circulation could be reversed is wrong. And the bidirectional causality of the El Nino indices could be not induced by the Walker circulation and ocean water shifts, it could be also induced by the wind-SST feedbacks. These explanations are misleading.

5. Normally when one wants to utilize a causality method correctly, it is import to know the limitations on data length and hyper-parameter settings. For the better development of the method in causality community, the authors should present the experiments on different data lengths, and show the shortest data length (i.e. time series length) for their method to provide reliable inference. Secondly, the authors should present how the hyper-parameters settings in RC will influence the causal inference, and introduce to the readers how to select suitable hyper-parameters to derive the reliable

causal inference.

Reviewer #3 (Remarks to the Author):

This paper proposed an innovative approach, named as IRC, to detect causality from observational time series data. How to make intervention without manipulating the underlined complex system is very important in the causality analysis. By comparing the closed-loop and intervened-loop of RC outputs, the proposed method considers imposing interventions in neural network replica of the original systems, which is an brilliant idea worth publishing. However, there are still several concerns should be addressed before accepting.

Major Concerns:

1. It is essential to distinguish the method with the traditional Granger causality, which has also been implemented with machine learning techniques, as the authors also cited these references ([9] and [10]). More detailed discussions need to be elaborated to clarify the distinction.
2. The method is not clearly stated. Particularly subsections 2.3 and 2.4, as the core parts of the method, do not provide consistent description of the method. For example, in Eqs.(9,10,11), how to obtain W_{out} ? W_{out} could only be obtained after training, while in subsection 2.3 there is no description of training. In Fig.1 there are three reservoirs, what's the relationship between the one ϕ in the closed-loop and the one ϕ in the intervened-loop? Are they the same one?

Minor Concerns:

Besides, the authors also should address the following concerns:

- 1) It should be explained how the hyperparameters, such as N_{evo} , reservoir size, are selected. Moreover, the robustness of the approach against moderate to large noise should also be discussed. Are systematic and observational noises playing same role in affecting the detection results?
- 2) The references to equations appear erroneous as "Eq. eq. (1)". Check all the mathematical expressions to avoid confusion.
- 3) Enrich the introduction to existing literatures, especially those approaches using machine learning to detect causality.
- 4) The authors only use continuous Lorenz system as benchmark. In previous literatures, discrete systems such as Logistic map are also commonly used to detect causality. Can the proposed approach be applied to such discrete-time systems?
- 5) The formatting of cited references varies, with some author names in full and others abbreviated.

Nevertheless, the paper provides inspiring and new thinking on the important question of detecting causality with only data available. Before publication, the above points should be discussed properly.

Reply to Reviewer #1

In this manuscript, the authors have designed a novel method to infer causal relations within complex systems by implementing the concept of controlled trials with only data-driven techniques. The topic is quite important among various disciplines and the idea of controlled trials also has been central in human understanding of causality. The designed method successfully incorporates such idea and represents differences and advantages over existing methods. Numerical examples and several real-world datasets verify the effectiveness of the proposed method. Overall, the manuscript is well-organized and easy to follow, with persuasive experiments provided. However, there are still a few issues that need to be addressed before I can recommend publication of this manuscript, as follows.

Reply. We thank the reviewer very much for your valuable comments and precious time to review our work. As presented in the following responses and the revised manuscript, we have addressed properly all your concerns, which render significant improvements of this work.

1. Correct the erroneous cross-references to equations denoted as ‘Eq.’ or ‘eq.’.

Reply. Thank you very much for your careful check on our manuscript. We have corrected all the erroneous use and carefully checked the text throughout the manuscript to enhance the readability of our paper.

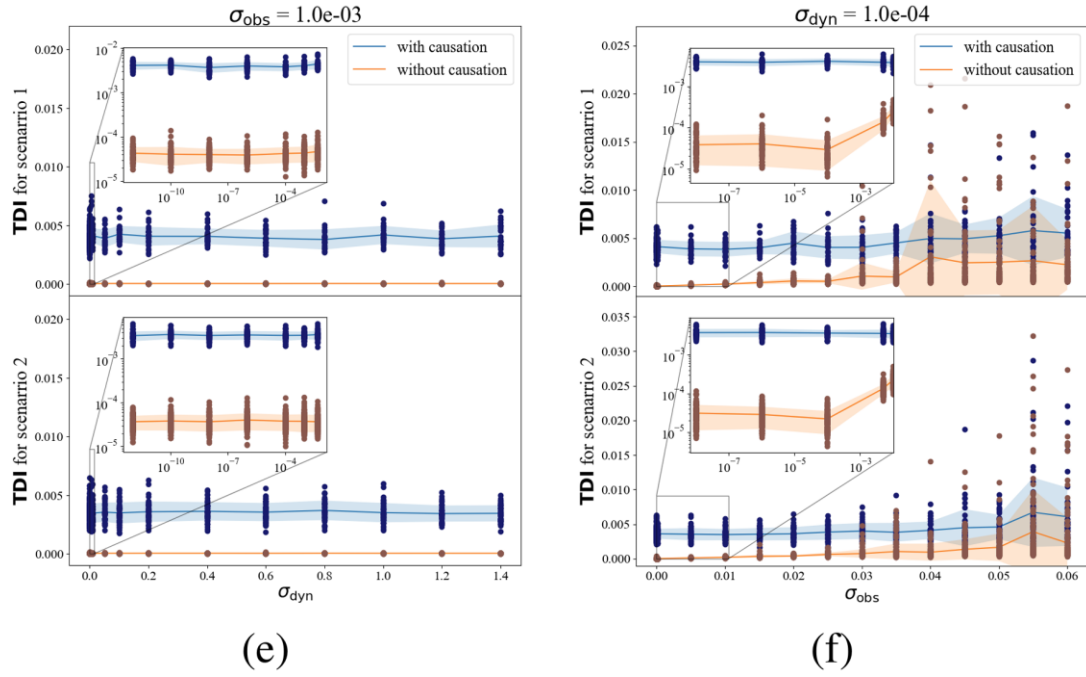
2. It is important for the authors to show how this method is affected by different levels of external noise perturbations, as noise always exists in the datasets.

Reply. Thank the reviewer for this important suggestion. In the revised manuscript, we supplement extensive experiments to show how the IRC method is affected by two kinds of noises, i.e., dynamical noise and observational noise. As shown in Fig. 6(e)(f), dynamical noise hardly affects the IRC performance regardless of its level, and observational noise gradually decreases the performance when its level increases. However, IRC only loses its efficacy when the observational noise is significantly large, demonstrating the robustness of our method against noises. The revisions are listed below for your quick reference.

3.1.3 Robustness to noise

We consider the effects of dynamical noise with variant σ_{dyn} ranging from 0 to 1.4 and observational noise with variant σ_{obs} ranging from 0 to 0.06 separately. Experiments show that the scale of dynamical noise has a negligible effect on IRC, whereas observational noise gradually decreases the performance when its level increases, as shown in Figs. 6(e) and 6(f). However, IRC only loses its efficacy when the observational noise is significantly large, which is as expected since excessive observational noise may obscure the dynamics of the unknown system, which is exactly what the IRC method tries to reproduce in the digital system, while dynamical noise does not.

--- from the revised Section 3.1.3



--- from the revised Figure 6

3. To reproduce the results and for the convenient use of readers, I suggest the authors to provide necessary codes affiliated to the manuscript, which is quite important especially for articles in Communications Physics with high influence.

Reply. In light of the reviewer's advice, we have provided all the necessary codes in <https://github.com/Jintong-CNS/Intervened-Reservoir-Computing> for the readers to reproduce the results in this manuscript.

4. Some most recent computational methods in handling dynamic information by reservoir computing are missing. I suggest that the following relevant literature should be addressed.

--- Autoreervoir computing for multistep ahead prediction based on the spatiotemporal information transformation. Nature Communications, 2020, 11(1): 1-15. DOI: 10.1038/s41467-020-18381-0.

--- Earthquake alerting based on spatial geodetic data by spatiotemporal information transformation learning. Proceedings of the National Academy of Sciences, 2023, 120(37): e2302275120. DOI: 10.1073/pnas.2302275120.

Reply. Thank the reviewer very much for providing us these useful references. We have discussed and cited them in the revised Introduction section of the manuscript and believe they will be informative for the readers to a better understanding.

5. The authors should clearly distinguish the variables/symbols used when describing the closed-loop and intervened-loop. In current version, variable x^{t+1} may cause confusion as it represents the input/output in both loops. This is also the case for the

reservoir state $\mathbf{r}$ used in both loops.

Reply. We are grateful to the reviewer for pointing out this crucial vagueness in our previous manuscript. We have reorganized all the variables/symbols and used subscripts “CL”/“ $j \rightarrow i$ ” to distinguish the variables used in the closed-loop and intervened-loop. We have also rewritten all the method description to enhance the clearness and readability of our manuscript. The revisions are appended below for your quick reference.

2.3 Intervened reservoir computing

Describing the causal relationships within a complex system as a causal network, the basic idea of IRC is that eliminating the influence of one node will lead to a (no) change in the subsequent evolutionary trajectory of another node if a causal link does (does not) exist. To compare whether it leads to a change, control and experimental results need to be generated using closed-loop and intervened-loop, respectively, as described later in this section. The two loop modes use the same RC parameters, i.e., $\mathbf{W}_{\text{in}}$, $\mathbf{W}$ and $\mathbf{W}_{\text{out}}$, where $\mathbf{W}_{\text{out}}$ is obtained by standard RC training directly using the observational time series as in Eq. (7).

2.3.1 Closed-loop

To compare the changes in subsequent evolution resulting from an intervention, an unintervened baseline sequence should first be generated as a control group, whose generation process can be considered a simple forecasting task, and is accomplished through a self-iterative process called closed-loop as shown in the blue dashed rectangle in Fig. 1. The process of generating the unintervened sequence $\{\mathbf{x}_{\text{CL}}^t\}_{t=0}^{N_{\text{evo}}}$ using closed-loop can be described as:

$$\begin{aligned}\mathbf{x}_{\text{CL}}^t &= \mathbf{W}_{\text{out}}[\mathbf{r}_{\text{CL}}^t; b_{\text{out}}], \\ \mathbf{r}_{\text{CL}}^{t+1} &= \tanh(\mathbf{W}_{\text{in}}[\mathbf{x}_{\text{CL}}^t; b_{\text{in}}] + \mathbf{W}\mathbf{r}_{\text{CL}}^t),\end{aligned}\quad (8)$$

where $\mathbf{r}_{\text{CL}}^t$ denotes the reservoir state at time t under closed-loop and the input used to update the reservoir is the output of the previous iteration. The starting point of the sequence is randomly selected among the testing data and denoted as $t = 0$, followed by N_{idle} idle iterations. The last result of the idle iteration serves as the initial state of the closed-loop.

2.3.2 Intervened-loop

The key to designing the structure of the intervened neural network, as well as the identification of causal influence (from node x_j to node x_i for example), is to minimize the difference between the two predicted sequences in the absence of causal influence while maximize the difference in the presence of causal influence. Based on this idea, we present the intervened-loop to simulate the elimination of the causal effect from x_j to x_i and to acquire intervened sequences, as shown in the loop process boxed by the red dashed box in Fig. 1. To implement the intervened-loop, it is necessary to employ two RC structures called aiding structure (AS) and intervened structure (IS), whose reservoir states are relatively independent of each other. The intervened structure has an input with node x_j artificially set to zero, denoting an intervention of eliminating its influence, and the output provided to the final result retains only node x_i . The aiding structure requires an unaltered input and the output provided to the final result keeps all elements except node x_i . Precisely, obtaining the intervened sequence $\{\mathbf{x}_{j \rightarrow i}^t\}_{t=0}^{N_{\text{evo}}}$ through the intervened-loop can be described as:

$$\begin{aligned}\mathbf{x}_{j \rightarrow i}^t &= \mathbf{0}_i \otimes \mathbf{W}_{\text{out}}[\mathbf{r}_{j \rightarrow i, \text{AS}}^t; b_{\text{out}}] + \mathbf{1}_i \otimes \mathbf{W}_{\text{out}}[\mathbf{r}_{j \rightarrow i, \text{IS}}^t; b_{\text{out}}], \\ \mathbf{r}_{j \rightarrow i, \text{AS}}^{t+1} &= \tanh(\mathbf{W}_{\text{in}}[\mathbf{x}_{j \rightarrow i}^t; b_{\text{in}}] + \mathbf{W}\mathbf{r}_{j \rightarrow i, \text{AS}}^t), \\ \mathbf{r}_{j \rightarrow i, \text{IS}}^{t+1} &= \tanh(\mathbf{W}_{\text{in}}[\mathbf{0}_j \otimes \mathbf{x}_{j \rightarrow i}^t; b_{\text{in}}] + \mathbf{W}\mathbf{r}_{j \rightarrow i, \text{IS}}^t),\end{aligned}\quad (9)$$

where $\mathbf{0}_k$ denotes a vector whose k -th dimension is zero and all other dimensions are one, $\mathbf{1}_k$ denotes a vector whose k -th dimension is one and all other dimensions are zero, $\otimes$ denotes element-wise product, $\mathbf{r}_{j \rightarrow i, \text{AS}}^t$ and $\mathbf{r}_{j \rightarrow i, \text{IS}}^t$ denote the reservoir state of aiding structure and intervened structure at time t , respectively, and $j \rightarrow i$ indicates that the sequence is obtained by eliminating the causal influence from node x_j to node x_i . In the intervened structure, $\mathbf{0}_j \otimes \mathbf{x}_{j \rightarrow i}^t$

achieves the evolution of the system unaffected by node x_j by adjusting the j -th dimension of the input to zero. And $\mathbf{1}_i \otimes \mathbf{W}_{\text{out}}[\mathbf{r}_{j \rightarrow i, \text{IS}}^t; b_{\text{out}}]$ restricts the result of intervened structure to be used only to produce the i -th dimension of current iteration. Thus, the intervened-loop simulates the artificial elimination and only the elimination of causal influences directed from node x_j to node x_i , regardless of its actual existence.

The intervened-loop provides distinguishable results for different cases with or without causation. The two structures share the same network parameters $\mathbf{W}_{\text{in}}, \mathbf{W}, \mathbf{W}_{\text{out}}$ and the initial reservoir state with closed-loop, i.e., $\mathbf{r}_{j \rightarrow i, \text{AS}}^0 = \mathbf{r}_{j \rightarrow i, \text{IS}}^0 = \mathbf{r}_{\text{CL}}^0$. Ideally, when there is no causal influence from node x_j to node x_i , the output of both structures $\mathbf{W}_{\text{out}} \mathbf{r}_{j \rightarrow i, \text{AS}}^t$ and $\mathbf{W}_{\text{out}} \mathbf{r}_{j \rightarrow i, \text{IS}}^t$ should yield highly similar results for node x_i . In this case, the intervened-loop is considered equivalent to the closed-loop thus ensuring the minimization of the difference produced by both the intervened-loop and the closed-loop. When the causal influence exists, x_i in the result of the first iteration should undergo a shift of a certain magnitude compared to the original trajectory. In the following iteration, the impact on x_i due to the artificial intervention will accumulate, while the slight change in x_i leads to a subsequent change in other variables. After sufficient iterations, the sequence generated by the intervened-loop will exhibit considerable inconsistency in comparison to that from the closed-loop, which comes from the elimination of the causal influence from node x_j to node x_i at each iteration.

--- from the revised Section 2.3

6. The equation in Fig. 1 seems different with its definition in eq. (12), please carefully check.

Reply. We have improved the figure and text to make sure their consistency in the revised manuscript.

7. The predicted output of RC, as in eq. (4), looks abnormal with the common use (i.e., $\mathbf{W}_{\text{out}} \{ \mathbf{r} \}$), please check if the matrix multiplication is correct.

Reply. We have corrected and unified all the equations/expressions as with their common use in the revised manuscript.

8. Section 2.4 describes that ‘A filtering process ... by setting a minimum threshold value $j_{\text{threshold}}$.’ How this process is applied and to which variable? This should be more clearly described.

Reply. Thank the reviewer for this important comment. In fact, this filtering process is applied to $\bar{x}_j$ to maintain the stability of the values and to prevent the detrimental effects caused by dividing by a value too close to zero. The revisions are listed below for your quick reference.

“... A filtering process is used by setting a minimum threshold value $j_{\text{threshold}}$ for $\bar{x}_j$ to maintain the stability of the values and to prevent the detrimental effects caused by dividing by a value too close to zero.”

--- from the revised Section 2.4

9. The authors should have a fair and detailed discussion on the limitations of the proposed method.

Reply. Thank the reviewer for this advice. We have supplemented discussions to the limitations of the proposed method accordingly in the Conclusion section. The

revisions are listed below for your quick reference.

“... For example, the application of IRC presupposes that RC or other neural networks are suitable for reproducing the dynamics of the target system, the time complexity of calculating the TDI rises with the square of the network size, dense networks are difficult to be accurately inferred, and like many existing methods, the choice of thresholds remains an imperfectly solved problem. ...”

--- from the revised Section 4

Thank again the reviewer for your tremendous effort and precious time to review our work. We sincerely accept the reviewer's advices and have revised the manuscript accordingly.

Reply to Reviewer #2

The authors proposed a causality detection method based on reservoir computer (RC). Their algorithm can identify causal direction between coupled time series, and reveal causal-relationship network among multiple time series. Their attempt to apply RC in causal-relationship network is new and interesting, because the RC does good to deal with nonlinear dynamics and computation efficiency. Their testing results on the coupled Lorenz systems and air-pollution dataset demonstrate the good performance of their method. But I still have concerns and suggestions in terms of the algorithm design and some necessary interpretations as below.

Reply. We are grateful to the reviewer's positive recognition on our work, as well as to your precious time and constructive comments in reviewing our manuscript. We have addressed properly all the concerns in the revised manuscript and in the following responses, according to your valuable suggestions.

1. In abstract, the authors claimed that their method is different from correlation-based methods, but they did not further explain which of the comparable methods they used (GC, CCM, TE) are correlation-based methods, and how their IRC method is different from the correlation-based methods in the sense of designing the algorithm. This is confusing, and the authors should clarify it. At least the authors should clarify how their method is different from GC or CCM. In my observation for Sections 2.1 and 2.4, the main difference is that the IRC employs trajectory deviation indices (TDI) to evaluate the prediction skill of RC. But to me it still belongs to the framework of Granger causality method. If their method does not propose a framework totally independent on GC or CCM, they should point out which framework of GC and CCM their method follows and improves on, rather than vaguely claiming their method is different from correlation-based method. Otherwise, this will mislead the readers from the community.

Reply. We thank the reviewer very much for this important comment. We agree with the reviewer that the term “correlation-based methods” is not quite accurate here. In fact, we would like to express the difference between our IRC method and the existing methods (including GC, TE, CCM) on that the latter detects causality via exploiting the relationship among multiple time series WITHOUT active intervention, which lies in the center of current causality concept. And our IRC method *for the first time* implements the idea of controlled trials via constructing digital twins of the underlying systems to detect causality. Specifically, in light of your comment, we explain the difference between GC, CCM and IRC in the sense of algorithm designing in the revised manuscript, that is, GC is based on the principle that if X causes Y, then incorporating X's past information improves the prediction of Y's future. This principle requires a separability condition of the time series, usually not satisfied in real-world nonlinear systems, which is then addressed by the CCM method based on phase space reconstruction. CCM confirms the existence of causation from X to Y when the neighbors of a point in the reconstructed manifold M_Y are still the neighbors

of the point with same time index in M_X , under the cross map. However, our IRC method admits active intervention and confirms X causing Y if an intervention on X results in subsequent change in Y's future, representing completely different idea with GC and CCM. Moreover, GC is implemented by comparing two different autoregression predictors, one of which includes the information of the cause while the other one does not. And IRC is implemented by observing the deviation of ONE predictor, fully replicating the target system, when intervention is imposed, also representing different designing with GC. We have further clarified these differences and rewritten Section 2 in the revised manuscript. The revisions are listed below for your quick reference.

2.3 Intervened reservoir computing

Describing the causal relationships within a complex system as a causal network, the basic idea of IRC is that eliminating the influence of one node will lead to a (no) change in the subsequent evolutionary trajectory of another node if a causal link does (does not) exist. To compare whether it leads to a change, control and experimental results need to be generated using closed-loop and intervened-loop, respectively, as described later in this section. The two loop modes use the same RC parameters, i.e., W_{in} , W and W_{out} , where W_{out} is obtained by standard RC training directly using the observational time series as in Eq. (7).

2.3.1 Closed-loop

To compare the changes in subsequent evolution resulting from an intervention, an unintervened baseline sequence should first be generated as a control group, whose generation process can be considered a simple forecasting task, and is accomplished through a self-iterative process called closed-loop as shown in the blue dashed rectangle in Fig. 1. The process of generating the unintervened sequence $\{x_{CL}^t\}_{t=0}^{N_{evo}}$ using closed-loop can be described as:

$$\begin{aligned} x_{CL}^t &= W_{out}[r_{CL}^t; b_{out}], \\ r_{CL}^{t+1} &= \tanh(W_{in}[x_{CL}^t; b_{in}] + W r_{CL}^t), \end{aligned} \quad (8)$$

where r_{CL}^t denotes the reservoir state at time t under closed-loop and the input used to update the reservoir is the output of the previous iteration. The starting point of the sequence is randomly selected among the testing data and denoted as $t = 0$, followed by N_{idle} idle iterations. The last result of the idle iteration serves as the initial state of the closed-loop.

2.3.2 Intervened-loop

The key to designing the structure of the intervened neural network, as well as the identification of causal influence (from node x_j to node x_i for example), is to minimize the difference between the two predicted sequences in the absence of causal influence while maximize the difference in the presence of causal influence. Based on this idea, we present the intervened-loop to simulate the elimination of the causal effect from x_j to x_i and to acquire intervened sequences, as shown in the loop process boxed by the red dashed box in Fig. 1. To implement the intervened-loop, it is necessary to employ two RC structures called aiding structure (AS) and intervened structure (IS), whose reservoir states are relatively independent of each other. The intervened structure has an input with node x_j artificially set to zero, denoting an intervention of eliminating its influence, and the output provided to the final result retains only node x_i . The aiding structure requires an unaltered input and the output provided to the final result keeps all elements except node x_i . Precisely, obtaining the intervened sequence $\{x_{j \rightarrow i}^t\}_{t=0}^{N_{evo}}$ through the intervened-loop can be described as:

$$\begin{aligned} x_{j \rightarrow i}^t &= \mathbf{0}_i \otimes W_{out}[r_{j \rightarrow i, AS}^t; b_{out}] + \mathbf{1}_i \otimes W_{out}[r_{j \rightarrow i, IS}^t; b_{out}], \\ r_{j \rightarrow i, AS}^{t+1} &= \tanh(W_{in}[x_{j \rightarrow i}^t; b_{in}] + W r_{j \rightarrow i, AS}^t), \\ r_{j \rightarrow i, IS}^{t+1} &= \tanh(W_{in}[\mathbf{0}_j \otimes x_{j \rightarrow i}^t; b_{in}] + W r_{j \rightarrow i, IS}^t), \end{aligned} \quad (9)$$

where $\mathbf{0}_k$ denotes a vector whose k -th dimension is zero and all other dimensions are one, $\mathbf{1}_k$ denotes a vector whose k -th dimension is one and all other dimensions are zero, $\otimes$ denotes element-wise product, $r_{j \rightarrow i, AS}^t$ and $r_{j \rightarrow i, IS}^t$ denote the reservoir state of aiding structure and intervened structure at time t , respectively, and $j \rightarrow i$ indicates that the sequence is obtained by eliminating the causal influence from node x_j to node x_i . In the intervened structure, $\mathbf{0}_j \otimes x_{j \rightarrow i}^t$

achieves the evolution of the system unaffected by node x_j by adjusting the j -th dimension of the input to zero. And $\mathbf{1}_i \otimes \mathbf{W}_{\text{out}}[\mathbf{r}_{j \rightarrow i, \text{IS}}^t; b_{\text{out}}]$ restricts the result of intervened structure to be used only to produce the i -th dimension of current iteration. Thus, the intervened-loop simulates the artificial elimination and only the elimination of causal influences directed from node x_j to node x_i , regardless of its actual existence.

The intervened-loop provides distinguishable results for different cases with or without causation. The two structures share the same network parameters $\mathbf{W}_{\text{in}}, \mathbf{W}, \mathbf{W}_{\text{out}}$ and the initial reservoir state with closed-loop, i.e., $\mathbf{r}_{j \rightarrow i, \text{AS}}^0 = \mathbf{r}_{j \rightarrow i, \text{IS}}^0 = \mathbf{r}_{\text{CL}}^0$. Ideally, when there is no causal influence from node x_j to node x_i , the output of both structures $\mathbf{W}_{\text{out}} \mathbf{r}_{j \rightarrow i, \text{AS}}^t$ and $\mathbf{W}_{\text{out}} \mathbf{r}_{j \rightarrow i, \text{IS}}^t$ should yield highly similar results for node x_i . In this case, the intervened-loop is considered equivalent to the closed-loop thus ensuring the minimization of the difference produced by both the intervened-loop and the closed-loop. When the causal influence exists, x_i in the result of the first iteration should undergo a shift of a certain magnitude compared to the original trajectory. In the following iteration, the impact on x_i due to the artificial intervention will accumulate, while the slight change in x_i leads to a subsequent change in other variables. After sufficient iterations, the sequence generated by the intervened-loop will exhibit considerable inconsistency in comparison to that from the closed-loop, which comes from the elimination of the causal influence from node x_j to node x_i at each iteration.

--- from the revised Section 2.3

3 Discussion

Unlike existing methods (including GC, TE, CCM) that detect causality via exploiting relationships among multiple time series, IRC, though similarly utilizing sequence forecasting as a tool, is fundamentally different. Firstly, IRC implements the idea of controlled trials via constructing digital twins of the underlying systems to detect causality. GC is based on the principle that if X causes Y , then incorporating X 's past information improves the prediction of Y 's future. This principle requires a separability condition of the time series, usually not satisfied in real-world nonlinear systems, which is then addressed by the CCM method based on phase space reconstruction. CCM confirms the existence of causation from X to Y when the neighbors of a point in the reconstructed manifold M_Y are still the neighbors of the point with same time index in M_X , under the cross map. However, our IRC method admits active intervention and confirms X causing Y if an intervention on X results in subsequent change in Y 's future, representing completely different idea with GC and CCM. Moreover, GC is implemented by comparing two different autoregression predictors, ranging from the original linear predictor to more recent neural network predictors, one of which includes the information of the cause while the other one does not. And IRC is implemented by observing the deviation of one predictor, fully replicating the target system, when intervention is imposed, also representing different designing with GC.

--- from the revised Section 3

2. The second concern is about reconstructing the causal network. Refer to a recently popular causal-network algorithm (PCMCI), there are three crucial factors to influence the causal inference for multi variables: (i) the autocorrelation of the time series; (ii) indirect chain effect to induce fake causality between the pair of variables; (iii) the common driver effect to induce fake causality between the pair of variables. Thus, the PCMCI particularly take the partial correlation procedure. Large amount of recent studies already demonstrate these three factors will do influence the correct inference of time series causality. How did the authors deal with these three factors? Or how well their methods can perform under the cases of these three factors? I did not find the relevant discussions in the manuscript, and the authors also did not mention PCMCI although they want to focus on causal network. Building a RC-based causal-network algorithm is promising, but these three factors are crucial for causal networks. Please add relevant discussions and investigations.

Reply. Thank you very much for this important question. We totally agree with the reviewer that autocorrelation, indirect chain effect, and common driver effect play essential roles in causal network reconstruction. To test the performance of our IRC method under the influence of these factors, we provide additional experiments in the

revised version of manuscript. Firstly, we use benchmark model with strong autocorrelation to show that our IRC still achieves good performance in distinguishing cases with or without causation, as shown in Fig. 6(l). We also design benchmark network structures containing indirect chain and common driver, and perform extensive numerical experiments to demonstrate that IRC successfully avoids the influence of these factors. Such performance can be attributed to the underlying principle of our IRC method that causation is confirmed via the subsequent changes under active intervention, which is in accordance with the common definition of causality, but is not only based on exploiting the relationship among multiple time series. We have supplemented detailed discussions on these points and referred to the celebrated work of PCMCI in the revised manuscript. The revisions are listed below for your quick reference.

3.1 Further experiments

Recent studies have shown that indirect chain effect and common driver effect to induce spurious causality are critical factors affecting the reconstruction of multivariate causal networks [32]. We use coupled Lorenz systems with the following two causal networks to more intuitively validate the ability of IRC to identify spurious causality and determine the direction of causal loops, as well as to discuss the effect of different parameters on the TDI results:

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{A}_2 = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}. \quad (14)$$

A visualization is shown in Fig. 6(a)

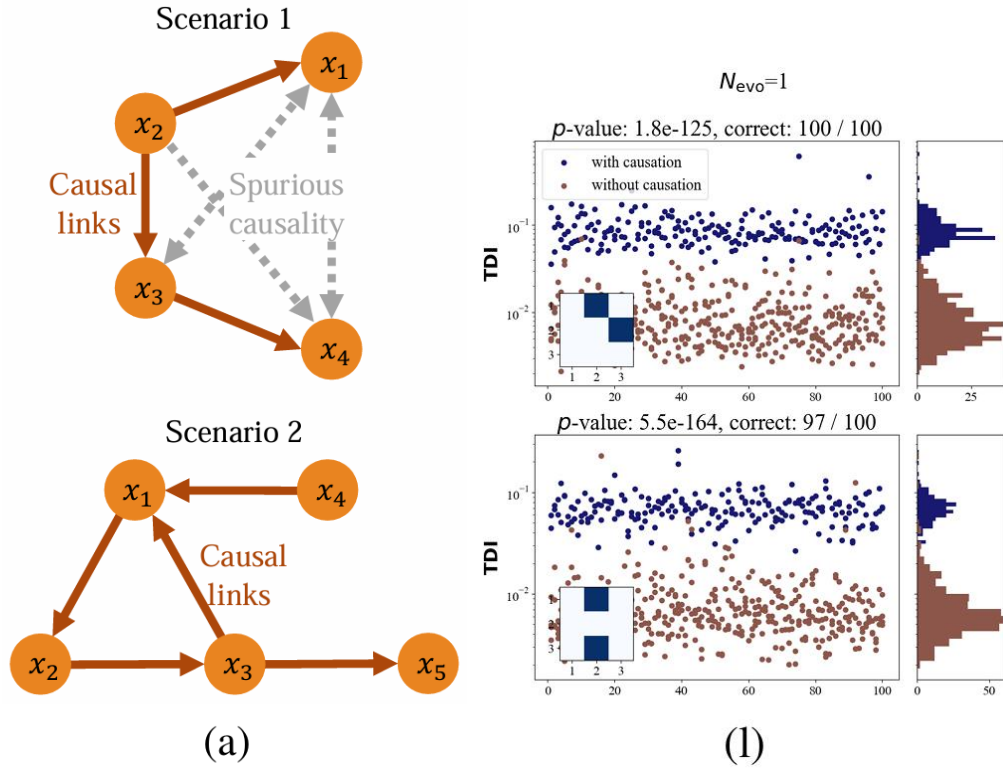
--- from the revised Section 3.1

Moreover, for a strong autocorrelation case, we use the following model:

$$x_k^{t+1} = \alpha_k x_k^t + c \sum_{l=1}^N a_{kl} x_l^t + \varepsilon_k^t, \quad (17)$$

where autocorrelation α_k is set to 0.8, coupling strength c is set to 0.3, $\mathbf{A} = [a_{ij}](i, j = 1, 2, \dots, N)$ represents the connectivity within the network and ε_k^t is independent white noise with zero mean and standard deviation 0.5. Similar to Logistic map, in scenarios $\mathbf{A}_3$ and $\mathbf{A}_4$, as shown in Fig. 6(1) the IRC approach yields accurate and distinguishable $\text{TDI}_{j \rightarrow i}$ that are consistent with the ground truth of the existence of causality between nodes, free from spurious causality and strong autocorrelation.

--- from the revised Section 3.1.5



--- from the revised Figure 6

3. Normally the causality methods would introduce the suitable statistical significant test for the causality inference. But the authors did not mention that. Does the TDI include both of causality intensity and significance test? Please clarify it.

Reply. Thank the reviewer for this valuable suggestion. We have supplemented T-test results to show whether a statistical significance exists between pairs of nodes with or without causal relation, with p -values provided in the experiments. From the results, TDI can reflect the causality intensity with statistical significance. We have provided explanations to the significance test and listed the p -values in the corresponding experimental results. Some of the revisions are listed below for your quick reference.

“... We also use the T-test to examine whether a statistical significance exists between pairs of nodes with or without causal relation, with p -value results shown in Table 1.”

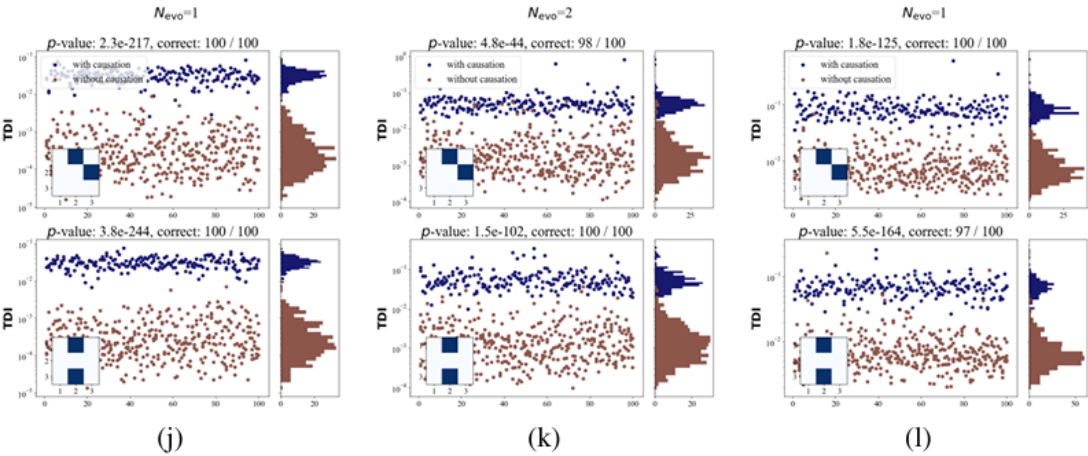
--- from the revised Section 3.1.2

*“... The results are shown in Figs. 6(j) and 6(k), with p -values and correct numbers given in the subtitle and **A** in the lower left. ...”*

--- from the revised Section 3.1.5

Table 1: For the coupled Lorenz systems in Fig. 4 the p -values associated with categories of different causal strengths.

categories	$1 \leftrightarrow 0.5$	$0.5 \leftrightarrow 0$	$1 \leftrightarrow 0$
$N = 5$	2.92×10^{-30}	7.26×10^{-77}	9.68×10^{-74}
$N = 10$	2.15×10^{-125}	approach zero	approach zero
$N = 20$	9.53×10^{-294}	approach zero	approach zero



--- from the revised Figure 6

4. The analysis on El Nino indices is weird. Firstly, the authors did not introduce if the SST time series include seasonal cycle. Secondly, the authors employ the changing states of Walker circulation to explain the bidirectional causality of the El Nino indices. Their claim that Walker Circulation could be reversed is wrong. And the bidirectional causality of the El Nino indices could be not induced by the Walker circulation and ocean water shifts; it could be also induced by the wind-SST feedbacks. These explanations are misleading.

Reply. We thank the reviewer very much for this important comment. The SST time series used in this work is the PLAIN monthly area averaged SST data calculated from the commonly used dataset HadISST1 in discovering the El Niño, whose description is shown below.

Data Set Name	Description
HadISST1	Globally complete SST, reconstructed using RSOI and blended with variance-corrected in situ data, includes bias-adjusted AVHRR SST and homogenized sea ice analysis. 1° area, 1871 onward (described here).

--- from *Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century* (doi: 10.1029/2002JD002670)

We also thank the reviewer for pointing out the vagueness in explaining the detected bidirectional causality. We have further explored literatures in Meteorology and

refined the explanations. According to the introduction in <https://oceanservice.noaa.gov/facts/ninonina.html>, during normal conditions in the Pacific Ocean, trade winds blow west along the equator, taking warm water from South America towards Asia. And during El Niño, trade winds weaken. Warm water is pushed back east, toward the west coast of the Americas. Moreover, during La Niña events, trade winds are even stronger than usual, pushing more warm water toward Asia. These facts provide evidences to the detected bidirectional causality via IRC, demonstrating its effectiveness. We hope these explanations can enhance the stringency of our work. We have revised the text in the manuscript accordingly, which is appended below for your quick reference.

As depicted in Fig. 5(a), the ENSO phenomenon is studied by monitoring the SST in four regions: Nino 1+2, Nino 3, Nino 3.4, and Nino 4. In this experiment, plain monthly area averaged SST data calculated from the HadISST1 for these regions are collected from January 1870 to December 2022, yielding a total of 1836×4 samples from the National Oceanic and Atmospheric Administration (NOAA) website, https://psl.noaa.gov/gcos_wgsp/Timeseries/, and the IRC approach is used to reconstruct the causal relationships between these four sets of time series data. An interpolation method is applied to improve the smoothness of the data.

During normal conditions, trade winds blow west along the equator in the Pacific Ocean, carrying warm water from South America towards Asia. In some years though, during El Niño, trade winds weaken and the warm surface water is pushed back east, toward the west coast of the Americas. La Niña has the opposite effect of El Niño, during which trade winds are even stronger than usual, pushing more warm water toward Asia. Thus, a bidirectional causal chain

between the neighboring Niño regions is formed. The reconstructed causal network and its visualization are presented in Fig. 5(a), where cells with $TDI_{j \rightarrow i}$ greater than 0.02 are boxed. Intuitively, each region has two causal connections back and forth with neighboring regions. Despite the non-constant direction of flow at the ocean surface, our result shows consistency to previous study [24] and is well-matched with the conclusions of existing related studies on El Niño.

--- from the revised Section 2.6

5. Normally when one wants to utilize a causality method correctly, it is important to know the limitations on data length and hyper-parameter settings. For the better development of the method in causality community, the authors should present the experiments on different data lengths, and show the shortest data length (i.e., time series length) for their method to provide reliable inference. Secondly, the authors should present how the hyper-parameters settings in RC will influence the causal inference, and introduce to the readers how to select suitable hyper-parameters to derive the reliable causal inference.

Reply. We are grateful to the reviewer's valuable suggestions. We agree the data length and the hyperparameters are important factors determining the applicability of the method. In light of this advice, we supplement sufficient numerical experiments testing the performance of IRC under different sampling intervals and data lengths. Results show that with the increasing of data length, IRC performance becomes better and gradually stabilizes, which is consistent with the intuition. Under different sampling intervals, the required data length for reliable inference differs (in Fig. 6(g)(h), ~4000 for sampling interval 0.01 and ~2000 for sampling interval 0.02). However, such required lengths are relatively short in reconstructing the underlying systems' dynamics and causality inference.

Moreover, we supplement experiments on how the hyperparameter, reservoir size,

influence the causal inference results. As shown in Fig. 6(i), with the increasing of reservoir size, IRC performance becomes better and gradually stabilizes. We further verify a criterion for selecting this hyperparameter that an N -dimensional network requires $\sim 40N$ reservoir size, which is computationally efficient. For hyperparameters input scaling σ_{in} and spectral radius ρ , we use K-Fold Cross Validation and Grid Search to determine them, as presented in *Robust optimization and validation of echo state networks for learning chaotic dynamics* (doi: 10.1016/j.neunet.2021.05.004). Other hyperparameters, such as the reservoir network's connectivity d , are not quite relevant to the performance, so we directly follow the settings from literatures. We have also provided public codes for the easy use of the readers.

We have revised the above points in the manuscript and appended them below for your quick reference.

3.1.4 What time series length is sufficient

Results show that with the increasing of data length, IRC performance becomes better and gradually stabilizes, which is consistent with the intuition. Meanwhile, we find the sufficient length to be closely related to the sampling interval τ . In other words, the time span of the sequence needs to be long enough to present the entire dynamic of the unknown system. When τ is 0.01, a sequence length N_{tr} of about 4000 is needed for training to get a desired result, while when τ is 0.02, the required length is halved to about 2000, as shown in Figs. 6(g) and 6(h). On the other hand, consistent with intuition, **TDI** is also related to τ since **TDI** portrays changes resulting from the elimination of causality. Therefore, absolute values of **TDI** are not comparable across systems on different time scales.

--- from the revised Section 3.1.4

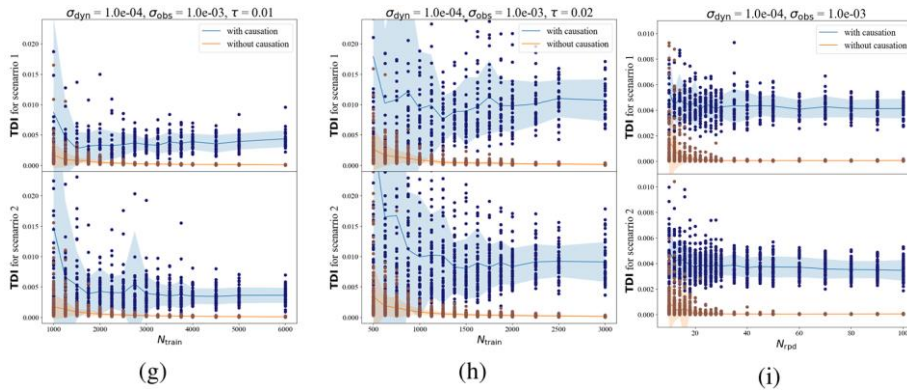
3.1.6 Choice of other parameters

To make the trained RC reproduce the dynamics of the target system as much as possible, it is necessary to choose suitable hyperparameters to improve the fitting performance of the RC. We use K-Fold Cross Validation and Grid Search [19] in a Lorenz system datasets to determine σ_{in} and ρ . In continuous (discrete) system experiments σ_{in} is set to 0.5(1). $\rho = 0.9$, $b_{\text{in}} = b_{\text{out}} = 1$, $\beta = 1 \times 10^{-4}$, and $d = 2$ are used in all experiments.

As for reservoir size N_r , RC fitting complex systems with higher dimensions are supposed to have larger reservoir sizes. We allocate N_{rpd} reservoir nodes for each dimension, i.e., $N_r = N_u N_{\text{rpd}}$. For the coupled Lorenz system, an N_{rpd} of about 40 works well enough, as shown in Fig. 6(i).

Practically, we use N_{net} of RCs with different initializations to randomly select N_{rep} starting positions in the test dataset respectively, and average each result as the final **TDI** result, thus reducing the influence of internal structure of the reservoir and different starting positions on the predicted sequences. Larger N_{net} and N_{rep} give more stable IRC outcomes but also lead to higher computational requirements. In calculating the **TDI** in Fig. 2 Fig. 3 Fig. 4 Fig. 5(a) and Fig. 5(b), N_{net} is set to 1, 5, 5, 25 and 30, respectively, and N_{rep} is set to 20, 20, 30, 20 and 15, respectively.

--- from the revised Section 3.1.6



--- from the revised Figure 6

We would like to thank again the reviewer for the tremendous effort and precious time to review our work, providing us with very insightful comments and suggestions which made our work improved substantially.

Reply to Reviewer #3

This paper proposed an innovative approach, named as IRC, to detect causality from observational time series data. How to make intervention without manipulating the underlined complex system is very important in the causality analysis. By comparing the closed-loop and intervened-loop of RC outputs, the proposed method considers imposing interventions in neural network replica of the original systems, which is a brilliant idea worth publishing. However, there are still several concerns should be addressed before accepting.

Reply. We sincerely appreciate the reviewer's highly positive recognition of our work, as well as the constructive comments and precious time dedicated to reviewing our manuscript. We are pleased to state that we have addressed properly all the concerns in the revised manuscript and in the following responses, according to your valuable suggestions.

Major Concerns:

1. It is essential to distinguish the method with the traditional Granger causality, which has also been implemented with machine learning techniques, as the authors also cited these references ([9] and [10]). More detailed discussions need to be elaborated to clarify the distinction.

Reply. Thank the reviewer for this important comment. In fact, the traditional Granger causality detects causation via exploiting the relationship among multiple time series WITHOUT active intervention, which lies in the center of current causality concept. And our IRC method *for the first time* implements the idea of controlled trials via constructing digital twins of the underlying systems to detect causality. Specifically, in light of your comment, we explain the difference between GC and IRC in the revised manuscript, that is, GC is based on the principle that if X causes Y, then incorporating X's past information improves the prediction of Y's future. However, our IRC method admits active intervention and confirms X causing Y if an intervention on X results in subsequent change in Y's future, representing completely different idea with GC. Moreover, GC is implemented by comparing two different autoregression predictors, one of which includes the information of the cause while the other one does not. And IRC is implemented by observing the deviation of ONE predictor, fully replicating the target system, when intervention is imposed, also representing different designing with GC. Though machine learning techniques have been incorporated into GC to replace the autoregression predictors recently, they still fall into the traditional framework of GC. We have further clarified these differences (also with CCM) in the revised manuscript. The revisions are listed below for your quick reference.

3 Discussion

Unlike existing methods (including GC, TE, CCM) that detect causality via exploiting relationships among multiple time series, IRC, though similarly utilizing sequence forecasting as a tool, is fundamentally different. Firstly, IRC implements the idea of controlled trials via constructing digital twins of the underlying systems to detect causality. GC is based on the principle that if X causes Y , then incorporating X 's past information improves the prediction of Y 's future. This principle requires a separability condition of the time series, usually not satisfied in real-world nonlinear systems, which is then addressed by the CCM method based on phase space reconstruction. CCM confirms the existence of causation from X to Y when the neighbors of a point in the reconstructed manifold M_Y are still the neighbors of the point with same time index in M_X , under the cross map. However, our IRC method admits active intervention and confirms X causing Y if an intervention on X results in subsequent change in Y 's future, representing completely different idea with GC and CCM. Moreover, GC is implemented by comparing two different autoregression predictors, ranging from the original linear predictor to more recent neural network predictors, one of which includes the information of the cause while the other one does not. And IRC is implemented by observing the deviation of one predictor, fully replicating the target system, when intervention is imposed, also representing different designing with GC.

--- from the revised Section 3

2. The method is not clearly stated. Particularly subsections 2.3 and 2.4, as the core parts of the method, do not provide consistent description of the method. For example, in Eqs. (9,10,11), how to obtain W_{out} ? W_{out} could only be obtained after training, while in subsection 2.3 there is no description of training. In Fig. 1 there are three reservoirs, what's the relationship between the one r in the closed-loop and the one r in the intervened-loop? Are they the same one?

Reply. We are grateful to the reviewer for pointing out this crucial vagueness in our previous manuscript. We have reorganized all the variables/symbols and used subscripts “CL”/“j $\rightarrow$ i” to distinguish the variables used in the closed-loop and intervened-loop. Moreover, the matrix W_{out} is obtained by standard RC training directly using the observational time series as in Equation (7), and is consistently used in all loops. We have rewritten all the method description and redrawn the figures to enhance the clearness and readability of our manuscript. The revisions are appended below for your quick reference.

2.3 Intervened reservoir computing

Describing the causal relationships within a complex system as a causal network, the basic idea of IRC is that eliminating the influence of one node will lead to a (no) change in the subsequent evolutionary trajectory of another node if a causal link does (does not) exist. To compare whether it leads to a change, control and experimental results need to be generated using closed-loop and intervened-loop, respectively, as described later in this section. The two loop modes use the same RC parameters, i.e., W_{in} , W and W_{out} , where W_{out} is obtained by standard RC training directly using the observational time series as in Eq. (7).

2.3.1 Closed-loop

To compare the changes in subsequent evolution resulting from an intervention, an unintervened baseline sequence should first be generated as a control group, whose generation process can be considered a simple forecasting task, and is accomplished through a self-iterative process called closed-loop as shown in the blue dashed rectangle in Fig. 1. The process of generating the unintervened sequence $\{x_{CL}^t\}_{t=0}^{N_{evo}}$ using closed-loop can be described as:

$$\begin{aligned} x_{CL}^t &= W_{out}[r_{CL}^t; b_{out}], \\ r_{CL}^{t+1} &= \tanh(W_{in}[x_{CL}^t; b_{in}] + W r_{CL}^t), \end{aligned} \quad (8)$$

where r_{CL}^t denotes the reservoir state at time t under closed-loop and the input used to update the reservoir is the output of the previous iteration. The starting point of the sequence is randomly selected among the testing data and denoted as $t = 0$, followed by N_{idle} idle iterations. The last result of the idle iteration serves as the initial state of the closed-loop.

2.3.2 Intervened-loop

The key to designing the structure of the intervened neural network, as well as the identification of causal influence (from node x_j to node x_i for example), is to minimize the difference between the two predicted sequences in the absence of causal influence while maximize the difference in the presence of causal influence. Based on this idea, we present the intervened-loop to simulate the elimination of the causal effect from x_j to x_i and to acquire intervened sequences, as shown in the loop process boxed by the red dashed box in Fig. 1. To implement the intervened-loop, it is necessary to employ two RC structures called aiding structure (AS) and intervened structure (IS), whose reservoir states are relatively independent of each other. The intervened structure has an input with node x_j artificially set to zero, denoting an intervention of eliminating its influence, and the output provided to the final result retains only node x_i . The aiding structure requires an unaltered input and the output provided to the final result keeps all elements except node x_i . Precisely, obtaining the intervened sequence $\{x_{j \rightarrow i}^t\}_{t=0}^{N_{evo}}$ through the intervened-loop can be described as:

$$\begin{aligned} x_{j \rightarrow i}^t &= \mathbf{0}_i \otimes \mathbf{W}_{out}[r_{j \rightarrow i, AS}^t; b_{out}] + \mathbf{1}_i \otimes \mathbf{W}_{out}[r_{j \rightarrow i, IS}^t; b_{out}], \\ r_{j \rightarrow i, AS}^{t+1} &= \tanh(\mathbf{W}_{in}[x_{j \rightarrow i}^t; b_{in}] + \mathbf{W}r_{j \rightarrow i, AS}^t), \\ r_{j \rightarrow i, IS}^{t+1} &= \tanh(\mathbf{W}_{in}[\mathbf{0}_j \otimes x_{j \rightarrow i}^t; b_{in}] + \mathbf{W}r_{j \rightarrow i, IS}^t), \end{aligned} \quad (9)$$

where $\mathbf{0}_k$ denotes a vector whose k -th dimension is zero and all other dimensions are one, $\mathbf{1}_k$ denotes a vector whose k -th dimension is one and all other dimensions are zero, $\otimes$ denotes element-wise product, $r_{j \rightarrow i, AS}^t$ and $r_{j \rightarrow i, IS}^t$ denote the reservoir state of aiding structure and intervened structure at time t , respectively, and $j \rightarrow i$ indicates that the sequence is obtained by eliminating the causal influence from node x_j to node x_i . In the intervened structure, $\mathbf{0}_j \otimes x_{j \rightarrow i}^t$

achieves the evolution of the system unaffected by node x_j by adjusting the j -th dimension of the input to zero. And $\mathbf{1}_i \otimes \mathbf{W}_{out}[r_{j \rightarrow i, IS}^t; b_{out}]$ restricts the result of intervened structure to be used only to produce the i -th dimension of current iteration. Thus, the intervened-loop simulates the artificial elimination and only the elimination of causal influences directed from node x_j to node x_i , regardless of its actual existence.

The intervened-loop provides distinguishable results for different cases with or without causation. The two structures share the same network parameters $\mathbf{W}_{in}$, $\mathbf{W}$, $\mathbf{W}_{out}$ and the initial reservoir state with closed-loop, i.e., $r_{j \rightarrow i, AS}^0 = r_{j \rightarrow i, IS}^0 = r_{CL}^0$. Ideally, when there is no causal influence from node x_j to node x_i , the output of both structures $\mathbf{W}_{out}r_{j \rightarrow i, AS}^t$ and $\mathbf{W}_{out}r_{j \rightarrow i, IS}^t$ should yield highly similar results for node x_i . In this case, the intervened-loop is considered equivalent to the closed-loop thus ensuring the minimization of the difference produced by both the intervened-loop and the closed-loop. When the causal influence exists, x_i in the result of the first iteration should undergo a shift of a certain magnitude compared to the original trajectory. In the following iteration, the impact on x_i due to the artificial intervention will accumulate, while the slight change in x_i leads to a subsequent change in other variables. After sufficient iterations, the sequence generated by the intervened-loop will exhibit considerable inconsistency in comparison to that from the closed-loop, which comes from the elimination of the causal influence from node x_j to node x_i at each iteration.

--- from the revised Section 2.3

2.4 Reconstruction of causal networks

The predicted intervened sequences need to be generated and compared with the unintervened sequence for each pair of nodes (x_i, x_j) , $i, j = 1, 2, \dots, N, i \neq j$ to reconstruct the entire causal network. Considering that the consequences of removing the causal influence of node x_j are closely related to the absolute value of x_j , to fairly compare the trajectory differences between pairs of nodes, the trajectory deviation indices $\mathbf{TDI} = [\mathbf{TDI}_{j \rightarrow i}]$ are calculated to minimize the influence of varying x_j values for all pairs of nodes through the following equation,

$$\begin{aligned} \bar{x}_j &= \frac{1}{N_{evo}} \sum_{t=1}^{N_{evo}} x_{CL, j}^t, \\ \mathbf{TDI}_{j \rightarrow i} &= \frac{1}{\bar{x}_j N_{evo}} \sum_{t=1}^{N_{evo}} (x_{CL, i}^t - x_{j \rightarrow i, i}^t)^2, \end{aligned} \quad (10)$$

where $x_{CL, k}^t$ denotes the k -th dimension of x_{CL}^t and $x_{j \rightarrow i, k}^t$ denotes the k -th dimension of $x_{j \rightarrow i}^t$. A filtering process is used by setting a minimum threshold value $j_{threshold}$ for $\bar{x}_j$ to maintain the stability of the values and to prevent the detrimental effects caused by dividing by a value too close to zero.

--- from the revised Section 2.4

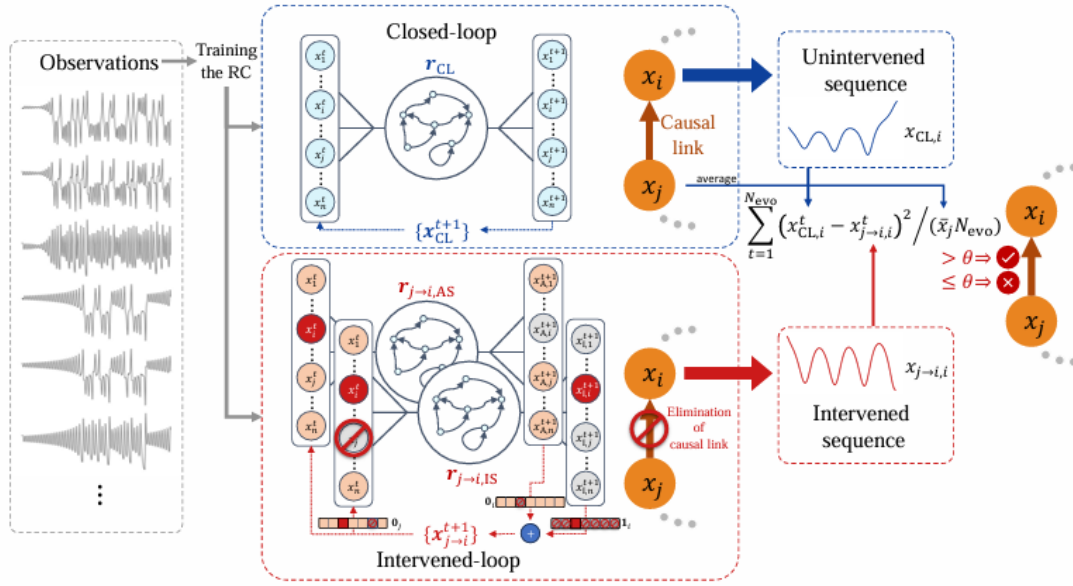


Figure 1: The overall architecture of IRC. The high-dimensional observations of the target system are used to train the RC, which generates predicted trajectories through two different modes (closed-loop and intervened-loop), respectively. The causal links within the target system can be identified by comparing the differences in the predicted trajectories. The TDI calculated using Eq. (10) measures these differences. For better viewing, $x_{A,k}^t$ and $x_{I,k}^t$ are used here to denote the k -th dimension of the output of aiding structure and intervened structure, respectively.

Minor Concerns:

Besides, the authors also should address the following concerns:

1) It should be explained how the hyperparameters, such as N_{evo} , reservoir size, are selected. Moreover, the robustness of the approach against moderate to large noise should also be discussed. Are systematic and observational noises playing same role in affecting the detection results?

Reply. Thank the reviewer very much for these questions. We have supplemented the experiments on how N_{evo} affects the performance of the IRC method. As shown in Fig. 6(b)(c), when no observational noise exists, IRC always has a good performance regardless of the choice of N_{evo} values. But when observational noise exists, there is a proper range of N_{evo} values that is not excessively small or large to support the satisfactory performance of IRC. Moreover, we supplement experiments on how the hyperparameter, reservoir size, influence the causal inference results. As shown in Fig. 6(i), with the increasing of reservoir size, IRC performance becomes better and gradually stabilizes. We further verify a criterion for selecting this hyperparameter that an N -dimensional network requires $\sim 40N$ reservoir size, which is computationally efficient.

In the revised manuscript, we also supplement extensive experiments to show how the IRC method is affected by two kinds of noises, i.e., dynamical noise and observational noise. As shown in Fig. 6(e)(f), dynamical noise hardly affects the IRC performance regardless of its level, and observational noise gradually decreases the performance when its level increases. However, IRC only loses its efficacy when the observational noise is significantly large, demonstrating the robustness of our method against noises.

The revisions are listed below for your quick reference.

3.1.1 Choice of a suitable N_{evo}

Overall, the mean and standard deviation of the $\text{TDI}_{j \rightarrow i}$ with causation increase as N_{evo} increases, while those of the $\text{TDI}_{j \rightarrow i}$ without causation remain stable, providing a larger possible range of feasible threshold. But when observational noise exists, there is a proper range of N_{evo} values that is not excessively small or large to support the satisfactory performance of IRC, as shown in Figs. 6(b) and 6(c). The baseline prediction error increases when the observation noise is large, at which point a smaller number of iterations ($N_{\text{evo}} = 1, 2, 3, 4$) is not effective in distinguishing between the deviation caused by the intervention and the error from the prediction itself, which are both on the order of 10^{-4} . This also illustrates the necessity of enhancing the influence of interventions through intervened-loop design. A larger number of iterations is also not preferred since larger standard deviations create difficulties in distinguishing different coupling strengths within a single system on the one hand, and the accumulation of prediction errors is more likely to lead to extreme values ($N_{\text{evo}} = 17, 18, 19, 20$) on the other hand. In calculating the TDI in Fig. 2, Fig. 3, Fig. 4, Fig. 5(a) and Fig. 5(b), N_{evo} is set to 20, 10, 10, 5 and 4, respectively.

--- from the revised Section 3.1.1

3.1.6 Choice of other parameters

To make the trained RC reproduce the dynamics of the target system as much as possible, it is necessary to choose suitable hyperparameters to improve the fitting performance of the RC. We use K-Fold Cross Validation and Grid Search [19] in a Lorenz system datasets to determine σ_{in} and ρ . In continuous (discrete) system experiments σ_{in} is set to 0.5(1). $\rho = 0.9$, $b_{\text{in}} = b_{\text{out}} = 1$, $\beta = 1 \times 10^{-4}$, and $d = 2$ are used in all experiments.

As for reservoir size N_r , RC fitting complex systems with higher dimensions are supposed to have larger reservoir sizes. We allocate N_{rpd} reservoir nodes for each dimension, i.e., $N_r = N_u N_{\text{rpd}}$. For the coupled Lorenz system, an N_{rpd} of about 40 works well enough, as shown in Fig. 6(f).

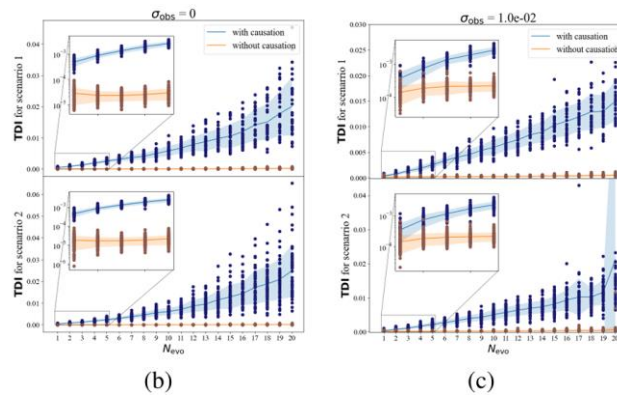
Practically, we use N_{net} of RCs with different initializations to randomly select N_{rep} starting positions in the test dataset respectively, and average each result as the final TDI result, thus reducing the influence of internal structure of the reservoir and different starting positions on the predicted sequences. Larger N_{net} and N_{rep} give more stable IRC outcomes but also lead to higher computational requirements. In calculating the TDI in Fig. 2, Fig. 3, Fig. 4, Fig. 5(a) and Fig. 5(b), N_{net} is set to 1, 5, 5, 25 and 30, respectively, and N_{rep} is set to 20, 20, 30, 20 and 15, respectively.

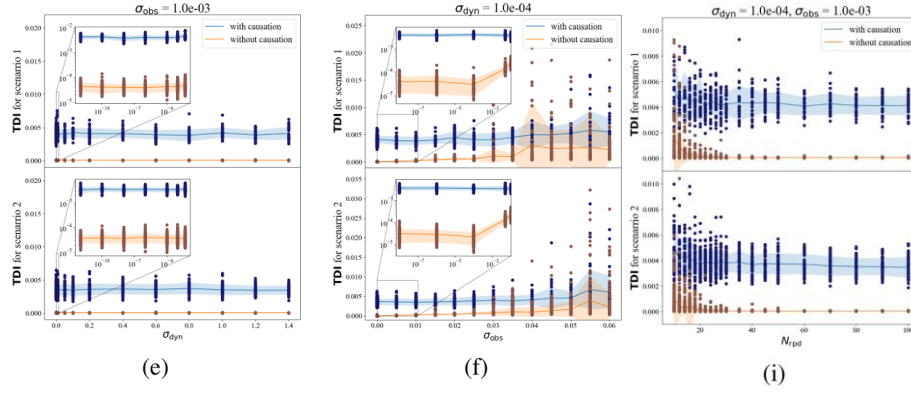
--- from the revised Section 3.1.6

3.1.3 Robustness to noise

We consider the effects of dynamical noise with variant σ_{dyn} ranging from 0 to 1.4 and observational noise with variant σ_{obs} ranging from 0 to 0.06 separately. Experiments show that the scale of dynamical noise has a negligible effect on IRC, whereas observational noise gradually decreases the performance when its level increases, as shown in Figs. 6(e) and 6(f). However, IRC only loses its efficacy when the observational noise is significantly large, which is as expected since excessive observational noise may obscure the dynamics of the unknown system, which is exactly what the IRC method tries to reproduce in the digital system, while dynamical noise does not.

--- from the revised Section 3.1.3





--- from the revised Figure 6

2) The references to equations appear erroneous as “Eq. eq. (1)”. Check all the mathematical expressions to avoid confusion.

Reply. Thank you very much for your careful check on our manuscript. We have corrected all the erroneous use and carefully checked the text throughout the manuscript to enhance the readability of our paper.

3) Enrich the introduction to existing literatures, especially those approaches using machine learning to detect causality.

Reply. Thanks for the suggestion. We have enriched the discussions to existing literatures accordingly. The revisions are listed below for your quick reference.

“...Researchers have focused on the improvement of the above three approaches, including incorporating the latest machine learning insights. Reservoir computing Granger (RCG) [9] uses reservoir computing to replace the autoregression in GC, and considerably solves the false-positive and false-negative problems in causality detection. Reservoir computing causality (RCC) [10] combines the ideas of GC and CCM to improve the ability to detect causal directions within real-world data by utilizing the positivity and negativity of the relative time lags of the two sequences (representing the past or the future) at optimal predictive performance. Echo state GC (ES-GC) [11] modifies and optimizes the RC formulation to model multivariate signal coupling and achieves internal decoupling using multiple reservoirs....”

--- from the revised Section 1

4) The authors only use continuous Lorenz system as benchmark. In previous literatures, discrete systems such as Logistic map are also commonly used to detect causality. Can the proposed approach be applied to such discrete-time systems?

Reply. We thank the reviewer for this advice. We confirm the applicability of our IRC method to discrete-time systems by supplementing experiments on coupled

Logistic maps. The results and corresponding revisions are listed below for your quick reference.

3.1.5 Application to discrete-time systems

We use coupled Logistic maps to generate simulated data, whose k -th node is described by:

$$x_k^{t+1} = x_k^t \left[\gamma_k - \gamma_k x_k^t - c \sum_{l=1}^N a_{kl} (x_l^t - x_k^t) \right] + \varepsilon_k^t, \quad (15)$$

where $c = 0.3$ is the coupling constant, adjacency matrix $\mathbf{A} = [a_{ij}](i, j = 1, 2, \dots, N)$ represents the connectivity within the network and ε_k^t is independent white noise with zero mean and standard deviation 0.005. We generate 100

sets of time series with length $N_{tr} = 5000$ in each of the following two scenarios:

$$\mathbf{A}_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{A}_4 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad (16)$$

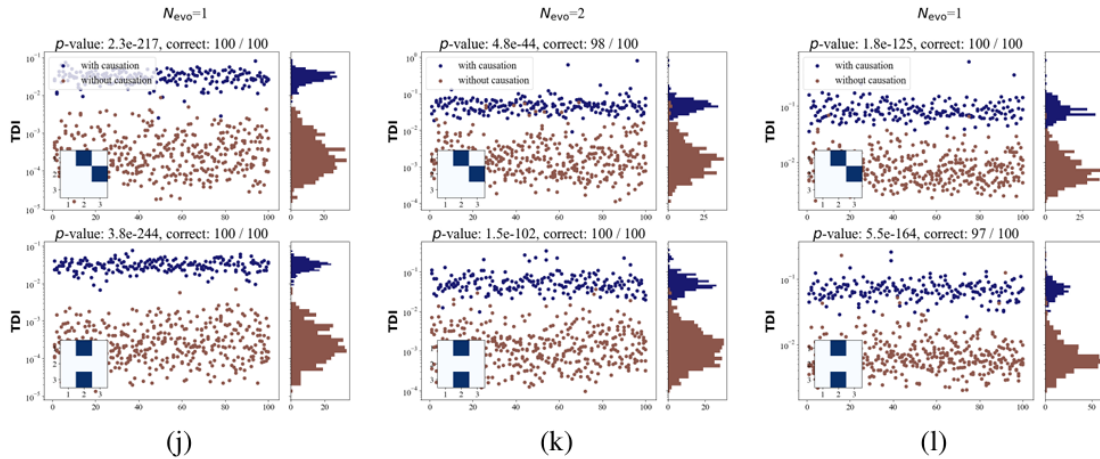
and $\gamma_1 = 3.7, \gamma_2 = 3.72, \gamma_3 = 3.78$. For each pair of nodes, $\text{TDI}_{j \rightarrow i}$ is categorized as with and without causal links based on ground truth a_{ij} . Although RC is not preferred for fitting discrete systems, the IRC approach still gives fairly correct classifications with small prediction lengths N_{evo} . Correct here means that the minimum value of the $\text{TDI}_{j \rightarrow i}$ corresponding to all pairs of nodes with causal link is greater than the maximum value of those without causal link in a single simulation. The results are shown in Figs. 6(j) and 6(k) with p -values and correct numbers given in the subtitle and $\mathbf{A}$ in the lower left.

Moreover, for a strong autocorrelation case, we use the following model:

$$x_k^{t+1} = \alpha_k x_k^t + c \sum_{l=1}^N a_{kl} x_l^t + \varepsilon_k^t, \quad (17)$$

where autocorrelation α_k is set to 0.8, coupling strength c is set to 0.3, $\mathbf{A} = [a_{ij}](i, j = 1, 2, \dots, N)$ represents the connectivity within the network and ε_k^t is independent white noise with zero mean and standard deviation 0.5. Similar to Logistic map, in scenarios $\mathbf{A}_3$ and $\mathbf{A}_4$, as shown in Fig. 6(l), the IRC approach yields accurate and distinguishable $\text{TDI}_{j \rightarrow i}$ that are consistent with the ground truth of the existence of causality between nodes, free from spurious causality and strong autocorrelation.

--- from the revised Section 3.1.5



--- from the revised Figure 6

5) The formatting of cited references varies, with some author names in full and others abbreviated.

Reply. We have checked and formatted all the cited reference carefully and make sure their consistency.

Nevertheless, the paper provides inspiring and new thinking on the important question of detecting causality with only data available. Before publication, the above points should be discussed properly.

Reply. We would like to express our gratitude to the reviewer again for your tremendous effort and precious time dedicated to reviewing our work. We sincerely appreciate your insightful suggestions, which have substantially improved our work.

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

The authors have well addressed all my concerns. I recommend accepting and publishing the revised manuscript as it is.

Reviewer #2 (Remarks to the Author):

Please see the attachment.

Reviewer #3 (Remarks to the Author):

All my concerns have been answered in the revised manuscript. I am glad to recommend publication as it is.