nature portfolio

Peer Review File

Self-supervised Reservoir Computing with Spatial-temporal Encoding for Identifying Critical Transitions

Corresponding Author: Dr Na Yang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors propose a novel and potentially impactful framework for detecting critical transitions and identifying corresponding bifurcation types in high-dimensional dynamical systems via STI. This topic is both timely and of significant relevance to researchers in complex systems and data-driven modeling. I believe the work has strong potential for high impact within the modeling community. The manuscript is generally well written and clearly structured; but I do think the following issues needs to be revised/addressed before the paper can be accepted.

Comment 1: The manuscript proposes a self-supervised approach of reservoir computing, stARC. Can the authors clarify what constitutes the fundamental advance of stARC beyond traditional reservoir computing? In what ways does the integration of STI (or the delay embedding transformation) lead to improved performance or robustness??

Comment 2: Given the availability of modern deep learning architectures for time-series modeling, what motivated the choice of reservoir computing as the underlying architecture of stARC, and what advantages does it offer for the problem of critical transition detection?

Comment 3: Extracting a one-dimensional representative variable is an appealing strategy for the dynamical analysis of high-dimensional nonlinear systems. However, does such an ultra-reduction raise concerns about potential information loss, and how does the proposed framework ensure reliable identification of the corresponding bifurcation types?

Comment 4: The manuscript employs a Bayesian online change point detection (BOCD) approach to identify critical transitions, with methodological details provided in the Supplementary Information. However, the role and assumptions of BOCD within the main framework are not sufficiently articulated in the main text. It would be helpful for the authors to briefly clarify in the main manuscript the key assumptions underlying the use of BOCD and how the detected change points should be interpreted in relation to early warning signals of critical transitions, with technical details remaining in the Supplementary Information.

Comment 5: The manuscript discusses the limitations of deep neural networks and compares stARC with several baseline methods. However, a wide range of unsupervised or weakly supervised approaches for modeling dynamical systems have been proposed in recent years. It would be helpful for the authors to review more representation methods which reflect recent advances in this area. If possible, additional comparisons with dynamical systems–oriented approaches, such as Koopman- or DMD-based methods, could further strengthen the evaluation.

Comment 6: The framework is demonstrated for several codimension-1 bifurcations, including transcritical, period-doubling and Neimark-Sacker bifurcation. How does the proposed approach extend to other types of bifurcations, and what are the main challenges in detecting or distinguishing them within the stARC framework?

Comment 7: The manuscript discusses the use of embedding theory to generate the latent representation. However, further clarification is needed to improve the accessibility of the method. Specifically, how does the proposed STI representation differ from other gradient-based learning approaches in terms of representation capacity, interpretability, or applicability to dynamical systems?

Comment 8: The manuscript investigates the effects of window length and noise strength on the performance of stARC to assess its robustness. In addition, the authors emphasize the incorporation of neighborhood information in constructing the representative variable. However, it remains unclear to what extent this neighborhood information is necessary for the effectiveness of stARC. An explicit comparison with a variant that excludes neighborhood information would help to isolate its contribution and more clearly demonstrate the effectiveness of the proposed design.

Comment 9: In classical dynamical systems theory, bifurcations are defined with respect to changes in control parameters. However, in the analysis of real-world data, such parameters are often unobserved. To avoid potential ambiguity, it would be helpful for the authors to clarify how the notion of bifurcation is interpreted in the absence of explicit parameter variation, and how this interpretation relates to the concept of observed bifurcations used in the manuscript.

Minor ones:
Comment 10: The manuscript analyzes three real-world datasets; however, only two of them are acknowledged in the Acknowledgements section. Please ensure that all datasets used in the study are appropriately acknowledged.

Comment 11: In section "Delay Embedding Constraint Based on Spatial-to-Temporal Information Transformation", the column vectors in Z do not need to be separated by commas. Please consider revising the notation for consistency with standard matrix representations.

Comment 12: Given the number of symbols and abbreviations introduced in the manuscript, a summary table of key notations and variables, provided either in the main text or in the Supplementary Information, would improve readability, particularly for readers from different backgrounds.

Comment 13: The notation introduced in Fig. 1 is not explicitly defined in the main text, where only its components are mentioned. Please clarify this notation for consistency.

Comment 14: The readability of Fig. 4 could be improved by increasing the font size of the legend.

Comment 15: In Fig. S1, increasing the line thickness in the bifurcation subplots would improve visual clarity.

Comment 16: In Supplementary Note S4, Eqs. S11–S12 appear to use L to denote the embedding dimension. For consistency and to avoid confusion with the neighbor number L defined in the main text, it would be preferable to use l consistently.

(Remarks on code availability)
The code is clean, including data generation and algorithms with reservoir computing to predict tipping and bifurcation types. The README file is clear and instructive. The code is easy to run.

Reviewer #2

(Remarks to the Author)
The paper entitled ``Self-supervised Reservoir Computing with Spatial-Temporal Encoding for Identifying Critical Transition'' proposes a novel approach for detecting early-warning signals to predict critical transitions using a reservoir computing framework. In this method, a high-dimensional input vector (x \in R^n) over (m) time steps produces high-dimensional reservoir states (r \in R^N) at each time step. An auxiliary variable (\Phi(x)) is then constructed for each reservoir state as a linear combination of (l) neighboring reservoir states. To compute the output weight matrix (W_out), the authors formulate a min–max optimization problem with quadratic constraints and derive an analytical solution using the Lagrange multiplier technique. Using (W_out) and the auxiliary variable (\Phi(x)), they compute the output (z), which is subsequently used to detect early-warning signals for predicting a possible future critical transition as well as identifying the type of bifurcation associated with that transition.
The paper addresses an interesting problem and proposes a clever solution. However, I have several concerns that must be addressed before I can recommend the manuscript for publication.

1. The representation of the paper should be enhanced significantly. The main issue is the introduction of many unnecessary and frequently changing variables/notations. In its current form, the presentation is confusing and difficult to follow. For example, in scientific and engineering literature, scalars are typically denoted by lowercase italic letters (e.g., \beta, \lambda), vectors by bold lowercase letters (e.g., \bf{x}, \bf{y}), and matrices by uppercase letters (e.g., A, W). Consistency in notation is essential: variables should be clearly defined and used consistently throughout the paper. For example:

• In the current manuscript, the notation is inconsistent and confusing. The variable z (lowercase z) is used as a scalar (one-dimensional quantity), \bf{Z}^t (bold uppercase Z) is an l-dimensional vector containing the values of z for time t to t+l-1. Z (uppercase Z) is an l x m matrix. \bf{z} (bold lowercase) is used as a m+l-1 dimensional vector containing the values of z for time t to t+m+l-2. Although these meanings can be inferred by following the equations, the most confusing part is reading the paper. For example, the statement: ``Thus, the one-dimensional representation \bf{z} = (z^t,z^1,⋯,z^t+m-1,z^t+m,⋯,z^t+m+l-2)′ can be easily obtained by expanding Z''. What do you mean by `` the one-dimensional representation'', as the quantity is clearly a vector!! Or what do you mean by ``can be easily obtained by expanding Z''? How can a vector be obtained by expanding a matrix??

• In line 175, the authors wrote: ``…vector formed by only one representative variable \bf{z} but at several time points t,t+1,⋯,t+l-1.'' Here, the symbol \bf{z} should be z, since it refers to a scalar (one-dimensional quantity!!).

• In line 690, the author addressed \bf{Z}^t as a one-dimensional output vector. What does it mean by a one-dimensional vector??

• Following the paper notation, the superscript refers to time, such as x^t means the value of variable x in time t. Is the function \psi in Eq. (9) a time-varying function?? What is the different between \psi, \psi^2, …, \psi^l-1 functions??
• In line 657, what is d_0? Following the definition of d_i, this quantity should be zero, which would cause all associated u variables to be zero as well!!

2. Due to the quadratic constraint, the optimization in (20) is not necessarily convex, and therefore, obtaining the global solution is nontrivial. Which means satisfying the KKT conditions (taking the derivatives with respect to the variables and the Lagrangian multiplier and setting them to zero) only guarantees that you have found a local optimum or a saddle point, not necessarily the global optimum. This observation partially explains the authors' use of averaging techniques when constructing the Hankel matrix. However, the conditions under which this averaging provides a good approximation are not discussed. For example, what happens if the elements along each anti-diagonal of \tilde{Z} differ significantly from one another? Does this require modifying the reservoir structure and rerunning the simulations? Alternatively, is there a theoretical bound beyond which the reshaped eigenvector of \Phi(x) no longer provides a reliable approximation of W_out?

3. The manuscript does not clearly explain how many neighbors L are required to obtain reliable results.

4. In line 237, the authors refer readers to methods to find out about the change point of sd(\bf{Z}^t). But the information is not provided in the method!

5. Figure captions should be self-explanatory. For example, in Figure 1, Panel B, it is unclear why the variable z is divided into two shaded regions. Similarly, the meaning of the yellow star in Panel C is not explained.

6. The current manuscript could be made more accessible to a broader readers by providing intuitive explanations in addition to mathematical formulations.

7. I recommend embedding the main results into the literature, that is, not only citing related work but addresses the novelties and remaining open questions.

In summary, the paper addresses an interesting problem and presents mathematically rigorous results, but the presentation requires substantial improvement. I encourage the authors to move beyond abstract mathematics and provide a clearer discussion of the broader implications and interpretations of their results for the interdisciplinary dynamical systems community. In addition, the limitations of the proposed framework, particularly its applicability to data with more complex bifurcation types, should be clearly stated.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have solved all the questions I had, So I propose to accept this manuscript.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have thoroughly addressed my previous concerns and have substantially improved the manuscript. I have no further questions or comments and recommend the paper for publication.

(Remarks on code availability)

Response to the Comments of Reviewer #1

Reviewer's Comment 1: The authors propose a novel and potentially impactful framework for detecting critical transitions and identifying corresponding bifurcation types in high-dimensional dynamical systems via STI. This topic is both timely and of significant relevance to researchers in complex systems and data-driven modeling. I believe the work has strong potential for high impact within the modeling community. The manuscript is generally well written and clearly structured; but I do think the following issues needs to be revised/addressed before the paper can be accepted.

Authors' Response: Thank you very much for the positive comments and constructive suggestions. We appreciate your efforts in reviewing our manuscript. Those comments are all valuable and helpful for revising and improving our paper. We have studied the comments carefully and made revisions. Our point-by-point responses are listed as follows.

Reviewer's Comment 2: The manuscript proposes a self-supervised approach of reservoir computing, stARC. Can the authors clarify what constitutes the fundamental advance of stARC beyond traditional reservoir computing? In what ways does the integration of STI (or the delay embedding transformation) lead to improved performance or robustness?

Authors' Response: Thank you for your suggestion. In this work, the proposed stARC consists of a reservoir network with two linear layers. It shares similarities with traditional reservoir computing (tRC), particularly in configurations with recurrence neurons, both featuring evolution dynamics and linear readout (Table R1).

Table R1. Comparative overview of stARC and tRC.

	stARC	tRC
Input	$\mathbf{X}^t$	
Driven Function	$\mathbf{r}^t = \alpha\mathbf{r}^{t-1} + (1-\alpha)f(W_{\text{in}}\mathbf{X}^t + W_{\text{r}}\mathbf{r}^{t-1} + \mathbf{b})$	
Output	$\mathbf{Z}^t = W_{\text{out}}\mathbf{r}^t$, where $\mathbf{Z}^t = (z^t, z^{t+1}, \dots, z^{t+l-1})'$ is a latent representation, l denotes the delay-embedding dimension.	$\mathbf{Y}^t = W_{\text{out}}\mathbf{r}^t$, where $\mathbf{Y}^t = \mathbf{X}^{t+1}$.
Target	Dynamical dimensional reduction	Prediction

Training	Parameter	W_{out}	
	Type	Self-supervised learning	Supervised learning
	Loss Function	$\underset{W_{\text{out}}}{\text{minimize}} - (1-\theta)\sum_{i=1}^{l} \Vert \mathbf{W}_i \Phi(X) \Vert_2^2$ $+\theta \sum_{i=1}^{l-1} \Vert \mathbf{W}_i P - \mathbf{W}_{i+1} Q \Vert_2^2$	$\underset{W_{\text{out}}}{\text{minimize}} \Vert \widetilde{\mathbf{Y}^t} - \mathbf{Y}^t \Vert_2^2$
	Method	Lagrange multipliers	Linear regression

As summarized in Table R1, stARC and tRC differ notably in terms of their targets and training approaches.

i) tRC is mainly designed for state prediction in high-dimensional dynamical systems. In contrast, stARC is formulated to extract a low-dimensional latent representation that facilitates downstream applications. Such a representation enables the detection of critical transitions and the identification of their underlying bifurcation types, extending substantially the scope of tRC prediction.

ii) tRC operates in a supervised framework, where the readout layer is optimized via linear regression (e.g., ordinary least squares or ridge regression) to minimize the prediction error with respect to predefined target outputs. Instead of fitting external targets, stARC employs the STI transformation to enforce embedding consistency and obtain the representation variable that captures system fluctuations (as detailed in Supplementary Figure S2). Solving the STI equations ensures that the framework operates in a self-supervised manner, thereby enabling the direct extraction of intrinsic dynamical characteristics from the system.

iii) The representative variable of stARC is analytically obtained from an embedding-based optimization problem by solving a characteristic equation, allowing the reservoir readout matrix W_{out} to be solved in a rapid, and unsupervised manner. This analytical scheme eliminates reliance on initial values of parameters for iterative training, which yields a stable and accurate result, and contributes to the robustness of the overall stARC framework.

Reviewer's Comment 3: Given the availability of modern deep learning architectures for time-series modeling, what motivated the choice of reservoir computing as the underlying architecture of stARC, and what advantages does it offer for the problem of critical transition detection?

Authors' Response: Thank you for your valuable and insightful comment. While modern deep learning architectures have achieved remarkable performance in time-series modeling, the choice of reservoir computing in stARC is primarily motivated by the following three reasons.

(1) In many real-world nonstationary dynamical systems, only recent, short-term observations are available, resulting in limited sample sizes. However, the performance of many machine learning methods depends on sufficiently large and well-structured training datasets. By contrast, reservoir computing (RC) employs a fixed recurrent reservoir with randomly initialized connections and optimizes only a linear readout layer, thus substantially reducing the number of trainable parameters, while preserving expressive temporal representations. This structural property mitigates overfitting and improves robustness under limited and even short-term data conditions.

(2) RC, as an extension of recurrent neural network (RNN) architectures, is particularly well suited for temporal and sequential data due to its intrinsic recurrent structure. The fixed reservoir generates high-dimensional nonlinear dynamics driven by input history, effectively constructing a temporal embedding of the input sequence. Under appropriate conditions (e.g., echo state property), the reservoir maintains fading memory, allowing stable yet excellent modeling of time-dependent systems.

(3) The readout structure of RC ($\mathbf{Y}^t = W_{\text{out}}\mathbf{r}^t$) is aligned with the STI formulation ($\mathbf{Z}^t = W_{\text{out}}\Phi(\mathbf{X}^t)$), making their integration particularly natural. The reservoir serves as a nonlinear dynamical mechanism that embeds the original system into a higher-dimensional state space, thereby generating rich nonlinear dynamics. The STI transformation subsequently maps these high-dimensional reservoir states into an ultra-low-dimensional temporal representation, and is designed to preserve key dynamical information of the original high-dimensional space. In other words, the STI transformation has been formally derived to map the spatial structure of high-dimensional observations onto the temporal evolution of a single representative variable, which is roughly viewed as a reduced representation related to the center-manifold behavior near the tipping point; i.e., this 1-dimensional representative sequence may be adopted to capture the critical state behaviors.

Furthermore, this low-dimensional representation provides a principled basis for applying dynamical analysis tools to identify critical transitions and characterize associated bifurcation types. In this sense, the RC architecture offers a structural advantage within stARC by enabling expressive dynamical embeddings while preserving analytical tractability.

Reviewer's Comment 4: Extracting a one-dimensional representative variable is an appealing strategy for the dynamical analysis of high-dimensional nonlinear systems. However, does such an ultra-reduction raise concerns about potential information loss, and how does the proposed framework ensure reliable identification of the corresponding bifurcation types?

Authors' Response: Thank you for your valuable comment. Theoretically, based on the generalized Takens' embedding theory, the dynamics of a system can be topologically reconstructed from a one-dimensional representative variable by the delay embedding scheme without information loss if $l > 2d > 0$, where d is the box-counting dimension of the attractor for the original system and l represents the embedding dimension.

Computationally, for real-world datasets, such an ultra-low reduction may introduce approximation error because the model optimizes two loss terms simultaneously. In some datasets, the optimization can satisfy both terms well; however, in other cases, a trade-off may arise, potentially limiting the accuracy with which the reduced representation captures the full nonlinear dynamics. Importantly, rather than predicting the overall dynamics, our method aims to identify critical transition dynamics, which actually explores such inaccuracy (e.g., strong fluctuations).

As discussed before, this 1-dimensional representation can be roughly considered to represent the center manifold near the tipping point. Building on this representation, in a manner similar to DEV[1] combined with autoregression, we can identify the corresponding bifurcation types based on the local Jacobian inferred by this 1-dimensional representation. We have clarified this point in the revised manuscript and added the relevant references and explanations (including DEV) accordingly (Page 32).

1. Grziwotz, F. *et al.* Anticipating the occurrence and type of critical transitions. *Sci. Adv.* **9**, eabq4558 (2023).

Reviewer's Comment 5: The manuscript employs a Bayesian online change point detection (BOCD) approach to identify critical transitions, with methodological details provided in the Supplementary Information. However, the role and assumptions of BOCD within the main framework are not sufficiently articulated in the main text. It would be helpful for the authors to briefly clarify in the main manuscript the key assumptions underlying the use of BOCD and how the detected change points should be interpreted in relation to early warning signals of critical transitions, with technical details remaining in the Supplementary Information.

Authors' Response: Thank you for your suggestion. We have added a brief clarification in the Methods section ("Detecting Early Warning Signals of Critical Transitions", Page 31) of the revised manuscript regarding the role of BOCD in our framework, its key assumptions, and the interpretation of the detected change points. In brief, BOCD is used as an online probabilistic detector under a piecewise-stationary assumption. The Std($\mathbf{z}$) inferred change points are treated as potential early-warning moments indicating a shift in the system's dynamical state; algorithmic details are provided in the Supplementary Information.

Reviewer's Comment 6: The manuscript discusses the limitations of deep neural networks and compares stARC with several baseline methods. However, a wide range of unsupervised or weakly supervised approaches for modeling dynamical systems have been proposed in recent years. It would be helpful for the authors to review more representation methods which reflect recent advances in this area. If possible, additional comparisons with dynamical systems-oriented approaches, such as Koopman- or DMD-based methods, could further strengthen the evaluation.

Authors' Response: Thank you for your valuable and insightful review. We have now cited more representative studies and included three more prominent papers about the directed anisotropic diffusion map[1], CEBRA[2], CANDyMan[3], and MARBLE[4] in the Introduction section of the revised manuscript (Page 3). Besides, we address the DMD/Koopman-based representation methods in the Discussion section (Page 22). We have clarified the differences and relations between these state-of- the-art representation techniques and our method in both the Introduction and Discussion sections.

Moreover, to strengthen the critical evaluation of our framework, we have conducted additional comparative experiments with DMD/Koopman-based approaches (see Table 2 in the revised manuscript, Supplementary Note S5 and S8). Specifically, we applied the Ridge DMD and Neural Koopman approaches to the same benchmark systems and evaluated their ability to detect critical transitions and identify the corresponding bifurcations using analogous settings. From this comparison, we mainly find that stARC outperforms the compared baselines under the tested settings both in EWS detection and bifurcation identification.

Table R2. Performance comparison between stARC and two representation-based methods. Performance metrics include EWS detection accuracy, bifurcation identification accuracy (ACC), and true positive rate (TPR). An EWS detection is counted as correct if a signal is detected before the bifurcation point in bifurcating samples and absent in non-bifurcating samples. Each method was tested on 400 randomly generated samples per system (200 bifurcating and 200 non-bifurcating) under identical experimental settings.

<table>
<tr><th>Dataset</th><th colspan="2">Metrics</th><th>stARC</th><th>Ridge DMD</th><th>Neural Koopman</th></tr>
<tr><td rowspan="3">Coupled Lorenz</td><td colspan="2">Accuracy of EWS detection</td><td>96.54%</td><td>79.25%</td><td>87.00%</td></tr>
<tr><td rowspan="2">Bifurcation Type</td><td>ACC</td><td>97.75%</td><td>79.25%</td><td>87.00%</td></tr>
<tr><td>TPR</td><td>96.00%</td><td>92.50%</td><td>95.75%</td></tr>
<tr><td rowspan="3">Coupled Hénon</td><td colspan="2">Accuracy of EWS detection</td><td>94.75%</td><td>61.00%</td><td>52.50%</td></tr>
<tr><td rowspan="2">Bifurcation Type</td><td>ACC</td><td>97.75%</td><td>61.25%</td><td>96.50%</td></tr>
<tr><td>TPR</td><td>99.50%</td><td>22.50%</td><td>98.50%</td></tr>
<tr><td rowspan="3">Coupled ADVP</td><td colspan="2">Accuracy of EWS detection</td><td>96.15%</td><td>66.25%</td><td>71.75%</td></tr>
<tr><td rowspan="2">Bifurcation Type</td><td>ACC</td><td>99.25%</td><td>66.25%</td><td>71.75%</td></tr>
<tr><td>TPR</td><td>98.50%</td><td>99.00%</td><td>98.00%</td></tr>
</table>

1. Feng, L., Gao, T., Xiao, W. & Duan, J. Early warning indicators via latent stochastic dynamical systems. *Chaos* **34**, 031101 (2024).
2. Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360–368 (2023).
3. Floryan, D. & Graham, M. D. Data-driven discovery of intrinsic dynamics. *Nat Mach Intell* **4**, 1113–1120 (2022).

4. Gosztolai, A., Peach, R. L., Arnaudon, A., Barahona, M. & Vandergheynst, P. MARBLE: interpretable representations of neural population dynamics using geometric deep learning. *Nat Methods* **22**, 612–620 (2025).

<u>Reviewer's Comment 7:</u> The framework is demonstrated for several codimension-1 bifurcations, including transcritical, period-doubling and Neimark-Sacker bifurcation. How does the proposed approach extend to other types of bifurcations, and what are the main challenges in detecting or distinguishing them within the stARC framework?

<u>Authors' Response:</u> Thank you for your valuable suggestion. First, stARC performs spatial-to-temporal information (STI) transformation via a reservoir structure, by encoding high-dimensional spatial data into the dynamics of a single representative variable. Then, we locally linearize the reconstructed one-dimensional dynamics around its equilibrium and compute the associated eigenvalues of the Jacobian, which allow us to distinguish different bifurcation types underlying the detected critical transition, including transcritical, period-doubling, and even Neimark-Sacker bifurcations.

However, eigenvalue-based criteria alone may not be sufficient to discriminate more intricate bifurcation scenarios, such as global or higher-codimension bifurcations. In such cases, richer spectral or nonlinear diagnostic measures may be required. Extending the current framework toward a more comprehensive spectral characterization of complex bifurcation structures is an important direction for future work. We have addressed this point in the Discussion section of the revised manuscript "While effective in distinguishing transcritical, period-doubling, and Neimark-Sacker bifurcations, … Extending the framework to additional bifurcation scenarios will require more refined spectral criteria." in Page 24.

<u>Reviewer's Comment 8:</u> The manuscript discusses the use of embedding theory to generate the latent representation. However, further clarification is needed to improve the accessibility of the method. Specifically, how does the proposed STI representation differ from other gradient-based learning approaches in terms of representation capacity, interpretability, or applicability to dynamical systems?

<u>Authors' Response:</u> Thank you for your insightful comment. To clarify how the proposed STI representation differs from gradient-based learning approaches, we have added a conceptual comparison shown in Table R3 and S2. In brief, the STI representation in stARC is not obtained through backpropagation-based optimization toward a predefined task; instead, it is analytically derived from the system dynamics through a characteristic equation. This design provides (i) dynamical interpretability, since critical transitions and bifurcation types are inferred from intrinsic dynamical properties (via local linearization/eigenvalue estimation) rather than purely loss-based optimization; (ii) practical applicability under limited or short/nonstationary data conditions, because in our implementation only the linear-readout parameter is optimized and a large labeled training set is not required; and (iii) explicit capacity for dynamical analysis, as the representation is constructed to support downstream critical transition prediction and bifurcation identification, whereas gradient-based models typically learn black-box representations tailored to prediction tasks only.

Table R3. Conceptual comparison between stARC and gradient-based learning approaches.

	stARC	Gradient-based Learning Approaches
Interpretability	Bifurcation analysis with theoretical foundation	Task-optimized black-box
Identification Principle	Intrinsic dynamical properties (Eigenvalue solution)	Loss-based classification
Data Requirement	Short unlabeled time series	Typically large/long training datasets
Training Mechanism	Analytical solution (self-supervised)	Gradient-based optimization (iterative backpropagation)
Robustness under limited data	Stable under short/nonstationary data	Sensitive to noise
Capacity	Prediction of critical transition; identification of bifurcation types	Prediction of critical transition; identification of bifurcation types
Applicability	Transcritical, period-doubling, and Neimark-Sacker bifurcations	Dependent on training datasets

Reviewer's Comment 9: The manuscript investigates the effects of window length and noise strength on the performance of stARC to assess its robustness. In addition, the authors emphasize the incorporation of neighborhood information in constructing the representative variable. However, it remains unclear to what extent this neighborhood information is necessary for the effectiveness of stARC. An explicit comparison with a variant that excludes neighborhood information would help to isolate its contribution and more clearly demonstrate the effectiveness of the proposed design.

Authors' Response: Thank you for your valuable suggestion. In stARC, the neighborhood structure is introduced to preserve the local manifold geometry and promote dynamical consistency of the reduced representation. This helps decrease noise-induced fluctuations and stabilizes the inferred one-dimensional dynamics (as is also reflected by increased persistence of recurrence/extreme events based on extreme value theory[1]). The choice of L follows a locality-stability trade-off consistent with manifold learning theory.

To clarify the necessity of incorporating neighborhood information into stARC, we conducted additional ablation experiments by removing the neighbor-related components from the framework and comparing performance against the full model. We further evaluated sensitivity to the neighborhood size by systematically varying the number of neighbors. The results are summarized in Table R4 (see also in Supplementary Table S4) and described in detail in Supplementary Note S5. Overall, these experiments show that incorporating neighborhood information yields a clear performance improvement, while stARC remains robust across a broad range of neighborhood settings.

Table R4. Bifurcation identification of stARC under different neighbor settings. Bifurcation identification performance is evaluated by accuracy (ACC) and true positive rate (TPR). For each simulated system, 200 bifurcating and 200 non-bifurcating samples were randomly generated. Results are reported for noise strengths $\sigma = 0.1$, 0.2, and 0.5, and for different neighborhood settings, including no neighbors (i.e., $L = 0$) and $L = 5$, 10, 15.

System	Noise Strength σ	Metrics	Exclude Neighbors ($L = 0$)	Neighbor Number L		
				$L = 5$	$L = 10$	$L = 15$
Coupled Lorenz	$\sigma = 0.1$	ACC	67.50%	97.25%	97.75%	98.25%
		TPR	37.00%	96.00%	96.00%	96.50%

	$\sigma = 0.2$	ACC	58.00%	98.50%	97.25%	98.00%
		TPR	22.00%	97.50%	95.92%	96.00%
	$\sigma = 0.5$	ACC	67.00%	93.00%	95.75%	96.25%
		TPR	36.00%	90.00%	91.50%	92.50%
Coupled Hénon	$\sigma = 0.1$	ACC	87.50%	97.75%	97.75%	98.50%
		TPR	75.00%	96.00%	99.50%	99.00%
	$\sigma = 0.2$	ACC	86.00%	96.75%	97.00%	97.00%
		TPR	72.50%	94.00%	94.00%	95.75%
	$\sigma = 0.5$	ACC	71.00%	95.00%	95.75%	99.50%
		TPR	42.75%	90.00%	91.50%	99.00%
Coupled ADVP	$\sigma = 0.1$	ACC	97.00%	97.50%	99.25%	98.50%
		TPR	94.50%	95.00%	98.50%	97.00%
	$\sigma = 0.2$	ACC	97.00%	99.00%	99.50%	99.50%
		TPR	94.00%	98.00%	99.00%	99.50%
	$\sigma = 0.5$	ACC	81.00%	97.25%	99.50%	98.75%
		TPR	62.00%	96.50%	99.50%	97.50%

1. Fisher, R. A. & Tippett, L. H. C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society* **24**, 180–190 (1928).

Reviewer's Comment 10: In classical dynamical systems theory, bifurcations are defined with respect to changes in control parameters. However, in the analysis of real-world data, such parameters are often unobserved. To avoid potential ambiguity, it would be helpful for the authors to clarify how the notion of bifurcation is interpreted in the absence of explicit parameter variation, and how this interpretation relates to the concept of observed bifurcations used in the manuscript.

Authors' Response: Thank you for your helpful and important suggestion. We agree that in classical bifurcation theory, bifurcations are defined with respect to changes in an explicit control parameter. In our synthetic benchmarks, the control parameter is known by construction, and the corresponding critical threshold can be determined directly from the governing equations. However, for real-world datasets, the underlying control parameter(s) are typically unobserved.

In this work, a bifurcation is therefore interpreted operationally through the local dynamical behavior inferred from the reconstructed time series, rather than through direct parameter variation. This interpretation is commonly adopted in data-driven dynamical systems analysis, where local stability changes can still be inferred from observed dynamics. Specifically, we infer the bifurcation type from the reconstructed one-dimensional dynamics by local linearization and estimating the dominant eigenvalue of Jacobian.

To avoid ambiguity, we also clarify the meaning of "observed bifurcation" in the revised manuscript. Here, "observed bifurcation" refers to an empirically defined reference point based on a qualitative state change in the observed system behavior (e.g., the onset of a limit cycle, a period-doubling pattern, or a visible shift of a stable equilibrium). For example, in the ADVP system, an early-warning signal emerges before a clear limit cycle is observed, indicating that the system has begun to lose stability even though no obvious state change is yet visible. Its bifurcation type is then inferred as Neimark-Sacker, which is subsequently supported by the later emergence of sustained oscillations (See Fig. 2). Accordingly, we define the time point at which such a qualitative state change becomes observable as the "observed bifurcation point" for evaluation purposes.

The same operational definition is used for real-world datasets: we set the "observed bifurcation point" based on the first observable qualitative change in system behavior, and we interpret the corresponding bifurcation type according to the observed pattern (oscillation/period-doubling/equilibrium shift) together with the stability signature inferred from the reconstructed one-dimensional dynamics. We have revised the manuscript in Section "Performances of stARC on Real-world Time Series" (Page 14) to make these definitions explicit and distinguish parametric bifurcations (simulation systems) from operationally inferred bifurcations (empirical/observational analysis).

<u>Reviewer's Comment 11:</u> The manuscript analyzes three real-world datasets; however, only two of them are acknowledged in the Acknowledgements section. Please ensure that all datasets used in the study are appropriately acknowledged.

<u>Authors' Response:</u> Thank you for your comment. The source data of fish community data is available from https://doi.org/10.6073/PASTA/C08B808FE0CA65AE30B65B7D780F037F and the source of chicken heart data has now been clearly specified as https://github.com/ThomasMBury/dl_discrete_bifurcation. In response to this comment, we have explicitly stated the data source and properly acknowledged it in the Acknowledgements section in accordance with Nature Communications guidelines. In addition, we have documented the dataset provenance in the code repository to ensure full compliance with data transparency requirements of the journal.

Reviewer's Comment 12: In section "Delay Embedding Constraint Based on Spatial-to-Temporal Information Transformation", the column vectors in Z do not need to be separated by commas. Please consider revising the notation for consistency with standard matrix representations.

Authors' Response: Thank you for your correction. We have corrected the notation to ensure consistency in the revised manuscript (see Page 7).

Reviewer's Comment 13: Given the number of symbols and abbreviations introduced in the manuscript, a summary table of key notations and variables, provided either in the main text or in the Supplementary Information, would improve readability, particularly for readers from different backgrounds.

Authors' Response: Thank you for your suggestion. We have compiled a comprehensive summary of the parameters and variables involved in the stARC framework and incorporated it into the revised Supplementary Information to enhance clarity and facilitate reproducibility (Table R5, see also Supplementary Table S1).

Table R5. Summary of parameters and variables in stARC framework.

Symbols	Dimensions	Explanation
Ψ	$n \to l+1$	STI transformation function
R	$n \to N$	Function, a given reservoir network whose weights are randomly given
Φ	$n \to N$	Function, a structural readout regarding each window
$\mathbf{X}^t$	n	Vector, observed states of high-dimensional systems at time point t
$\mathbf{Z}^t$	l	Representative vector, delay embedding of the original system

		state at time point t
$\mathbf{z}$	$m+l-1$	Representative vector, extended dynamics of the original system within each window
$\mathbf{z}^t$	E	Vector, delayed dynamics of z for a local linear approximation, and $E < l$
z	1	Variable, the representative variable obtained by stARC
Z	$l \times m$	Matrix, low-dimensional Hankel matrix
$\mathbf{w}$	$L+1$	Vector, structural weights of state-neighbors
$\mathbf{r}^t$	N	Vector, reservoir state at time point t
$\Phi(\mathbf{X}^t)$	N	Vector, reservoir readout of stARC at time point t
$\mathbf{b}$	N	Vector, node bias vector
W_{in}	$N \times n$	Input matrix, defining the weights from the input layer to the reservoir
W_{r}	$N \times N$	Reservoir adjacency matrix, representing the reservoir's internal connectivity
W_{out}	$l \times N$	Readout matrix, the to-be-solved weights of $\Phi(\mathbf{X}^t)$
X	$n \times m$	Matrix, the observed high-dimensional time series with m time points
$\mathbb{R}^n$	n	n-dimensional Euclidean space
m	1	Scalar, the window size
l	1	Scalar, delay embedding dimension
d	1	Scalar, box-counting dimension
t	1	Scalar, time point
n	1	Scalar, the dimension of the observed variables in time-series
N	1	Scalar, the dimension of the reservoir states (i.e., the nodes of reservoir network)
L	1	Scalar, the number of neighbors
θ	1	Scalar, regularization parameter in stARC loss function
x_k^t	1	Scalar, the value of the k-th variable at time point t
y^t	1	Scalar, the value of the target variable y at time point t
$\circ$		Function composition operation
$'$		Transpose of a vector
Std		Standard deviation calculation

Reviewer's Comment 14: The notation introduced in Fig. 1 is not explicitly defined in the main text, where only its components are mentioned. Please clarify this notation for consistency.

Authors' Response: Thank you for your helpful comment. The notation used in Fig. 1 has now been clearly defined in the revised manuscript to ensure consistency and avoid ambiguity (Page 28). We have also revised the Fig.1 caption to further improve readability.

Reviewer's Comment 15: The readability of Fig. 4 could be improved by increasing the font size of the legend.

Authors' Response: Thank you for your helpful suggestion. The legend font size in Fig. 4 has been enlarged in the revised manuscript to enhance readability.

Reviewer's Comment 16: In Fig. S1, increasing the line thickness in the bifurcation subplots would improve visual clarity.

Authors' Response: Thank you for your helpful suggestion. The line thickness in the bifurcation subplots of Fig. S1 has been adjusted in the revised Supplementary Information to improve clarity.

Reviewer's Comment 17: In Supplementary Note S4, Eqs. S11-S12 appear to use L to denote the embedding dimension. For consistency and to avoid confusion with the neighbor number L defined in the main text, it would be preferable to use l consistently.

Authors' Response: Thank you for your suggestion. The embedding dimension in Eqs. S11-S12 has been consistently denoted by l in the revised Supplementary Information to avoid ambiguity with the neighbor number L defined in the main text. Moreover, the notation has been standardized throughout the manuscript to maintain consistency.

Response to the Comments of Reviewer #2

Reviewer's Comment 1: The paper entitled ``Self-supervised Reservoir Computing with Spatial-Temporal Encoding for Identifying Critical Transition" proposes a novel approach for detecting early-warning signals to predict critical transitions using a reservoir computing framework. In this method, a high-dimensional input vector (x \in R^n) over (m) time steps produces high-dimensional reservoir states (r \in R^N) at each time step. An auxiliary variable (\Phi(x)) is then constructed for each reservoir state as a linear combination of (l) neighboring reservoir states. To compute the output weight matrix (W_out), the authors formulate a min–max optimization problem with quadratic constraints and derive an analytical solution using the Lagrange multiplier technique. Using (W_out) and the auxiliary variable (\Phi(x)), they compute the output (z), which is subsequently used to detect early-warning signals for predicting a possible future critical transition as well as identifying the type of bifurcation associated with that transition.

The paper addresses an interesting problem and proposes a clever solution. However, I have several concerns that must be addressed before I can recommend the manuscript for publication.

Authors' Response: Thank you for your positive and valuable comments. We really appreciate your efforts in reviewing our manuscript. We have revised the manuscript accordingly. Our point-by-point responses are detailed below.

Reviewer's Comment 2: The representation of the paper should be enhanced significantly. The main issue is the introduction of many unnecessary and frequently changing variables/notations. In its current form, the presentation is confusing and difficult to follow. For example, in scientific and engineering literature, scalars are typically denoted by lowercase italic letters (e.g., \beta, \lambda), vectors by bold lowercase letters (e.g., \bf{x}, \bf{y}), and matrices by uppercase letters (e.g., A, W). Consistency in notation is essential: variables should be clearly defined and used consistently throughout the paper. For example:

• In the current manuscript, the notation is inconsistent and confusing. The variable z (lowercase z) is used as a scalar (one-dimensional quantity), \bf{Z}^t (bold uppercase Z) is an l-dimensional vector containing the values of z for time t to t+l-1. Z (uppercase Z) is an l x m matrix. \bf{z} (bold lowercase) is used as a m+l-1 dimensional vector containing the values of z for time t to t+m+l-2. Although these meanings can be inferred by following the equations, the most confusing part is reading the paper. For example, the statement: ``Thus, the one-dimensional representation \bf{z} = (z^t,z^1,⋯,z^t+m-1,z^t+m,⋯,z^t+m+l-2)′ can be easily obtained by expanding Z''. What do you mean by `` the one-dimensional representation'', as the quantity is clearly a vector!! Or what do you mean by ``can be easily obtained by expanding Z''? How can a vector be obtained by expanding a matrix??

• In line 175, the authors wrote: "…vector formed by only one representative variable \bf{z} but at several time points t,t+1,⋯,t+l-1." Here, the symbol \bf{z} should be z, since it refers to a scalar (one-dimensional quantity!!).

• In line 690, the author addressed \bf{Z}^t as a one-dimensional output vector. What does it mean by a one-dimensional vector??

• Following the paper notation, the superscript refers to time, such as x^t means the value of variable x in time t. Is the function \psi in Eq. (9) a time-varying function?? What is the different between \psi, \psi^2, …, \psi^l-1 functions??

• In line 657, what is d_0? Following the definition of d_i, this quantity should be zero, which would cause all associated u variables to be zero as well!!

Authors' Response: Thank you for these detailed and valuable suggestions. We agree that the notation in the original manuscript was not sufficiently clear and could make the presentation difficult to follow. In response, we have thoroughly revised the notation throughout the manuscript to improve clarity, consistency, and conformity with standard scientific conventions.

Specifically, we have systematically distinguished scalars, vectors, and matrices using consistent notations and revised the corresponding definitions in the main text. We have also corrected the ambiguous or inaccurate expressions pointed out by the reviewer, including the description of the "one-dimensional representation," the distinction among z, $\mathbf{Z}^t$, Z, and $\mathbf{z}$, the interpretation of ψ, ψ^2, … , ψ^{l-1}, and the definition of d_0. Some details are demonstrated as below:

(1) We refined the description of the "one-dimensional representation" as the delay-embedding sequence of the representative variable z rather than a single scalar value, thereby eliminating potential misunderstanding. The ambiguous expressions such as "expanding Z" have been replaced by "Thus, the extended vector $\mathbf{z} = (z^1, z^2, \cdots, z^m, z^{m+1}, \cdots, z^{m+l-1})'$ can be easily obtained from Hankel matrix Z" (Page

30).

(2) The symbol $\mathbf{z}$ in "… formed by only one representative variable $\mathbf{z}$ but at several time points $t,\ t+1,\ \cdots\ ,\ t+l-1$." has been replaced by z (Page 7), and the usage of these notations has been consistently clarified throughout the revised manuscript.

(3) $\mathbf{Z}^t$, an l-dimensional vector, denotes the delay-embedding sequence $(z^t, z^{t+1}, \ldots, z^{t+l-1})$ of the one-dimensional representative variable z. The description of z, $\mathbf{z}$, and $\mathbf{Z}^t$ have been clarified in the revised manuscript and Table R6.

(4) The ψ in Eq. (9) is not a time-varying function, and ψ^i $(i > 1)$ denotes the i-fold composition of ψ. To avoid ambiguity, the denotation ψ^i $(i > 1)$ has been clarified in the revised manuscript (Page 25).

(5) The $d_0 = 0$ is the Euclidean distance between the central state $\mathbf{X}^t$ and itself, which is consistent with the definition of d_i. We revised the definition of the neighboring weight as $w_i = u_i / \sum_{j=0}^{L} u_j$, where $u_i = e^{-d_i/\bar{d}}$ and $\bar{d} = \sum_{j=0}^{L} d_j/(L+1)$ denotes the average Euclidean distance (Page 28).

Moreover, the notations have been standardized throughout the manuscript to ensure consistency and to reduce unnecessary switching between symbols (Table R6, see also Supplementary Table S1). We hope that these revisions have substantially improved the readability and accessibility of the paper.

Table R6. Summary of parameters and variables in stARC framework.

Symbols	Dimensions	Explanation
Ψ	$n \to l+1$	STI transformation function
R	$n \to N$	Function, a given reservoir network whose weights are randomly given
Φ	$n \to N$	Function, a structural readout regarding each window
$\mathbf{X}^t$	n	Vector, observed states of high-dimensional systems at time point t
$\mathbf{Z}^t$	l	Representative vector, delay embedding of the original system state at time point t
$\mathbf{z}$	$m+l-1$	Representative vector, extended dynamics of the original system within each window
$\mathbf{z}^t$	E	Vector, delayed dynamics of z for a local linear approximation, and $E < l$
z	1	Variable, the representative variable obtained by stARC
Z	$l \times m$	Matrix, low-dimensional Hankel matrix
$\mathbf{w}$	$L+1$	Vector, structural weights of state-neighbors
$\mathbf{r}^t$	N	Vector, reservoir state at time point t
$\Phi(\mathbf{X}^t)$	N	Vector, reservoir readout of stARC at time point t
$\mathbf{b}$	N	Vector, node bias vector
W_{in}	$N \times n$	Input matrix, defining the weights from the input layer to the reservoir
W_{r}	$N \times N$	Reservoir adjacency matrix, representing the reservoir's internal

		connectivity
W_{out}	$l \times N$	Readout matrix, the to-be-solved weights of $\Phi(\boldsymbol{X}^t)$
X	$n \times m$	Matrix, the observed high-dimensional time series with m time points
$\mathbb{R}^n$	n	n-dimensional Euclidean space
m	1	Scalar, the window size
l	1	Scalar, delay embedding dimension
d	1	Scalar, box-counting dimension
t	1	Scalar, time point
n	1	Scalar, the dimension of the observed variables in time-series
N	1	Scalar, the dimension of the reservoir states (i.e., the nodes of reservoir network)
L	1	Scalar, the number of neighbors
θ	1	Scalar, regularization parameter in stARC loss function
x_k^t	1	Scalar, the value of the k-th variable at time point t
y^t	1	Scalar, the value of the target variable y at time point t
$\circ$		Function composition operation
$'$		Transpose of a vector
Std		Standard deviation calculation

Reviewer's Comment 3: Due to the quadratic constraint, the optimization in (20) is not necessarily convex, and therefore, obtaining the global solution is nontrivial. Which means satisfying the KKT conditions (taking the derivatives with respect to the variables and the Lagrangian multiplier and setting them to zero) only guarantees that you have found a local optimum or a saddle point, not necessarily the global optimum. This observation partially explains the authors' use of averaging techniques when constructing the Hankel matrix. However, the conditions under which this averaging provides a good approximation are not discussed. For example, what happens if the elements along each anti-diagonal of \tilde{Z} differ significantly from one another? Does this require modifying the reservoir structure and rerunning the simulations? Alternatively, is there a theoretical bound beyond which the reshaped eigenvector of \Phi(x) no longer provides a reliable approximation of W_out?

Authors' Response: Thank you for this insightful comment which indeed raises a deeper theoretical consideration of our proposed framework. i) We agree that satisfying the KKT conditions only guarantees a local optimum or a saddle point, rather than a global optimum. In our framework, the optimization is reformulated via a Lagrangian construction into an eigenvalue problem, and the solution is obtained analytically by solving the associated characteristic equation of a data-derived matrix. Importantly, this closed-form formulation avoids iterative parameter initialization and optimization, thereby reducing the instability associated with iterative training and mitigating the risk of convergence to poor local solutions. However, we do not claim global optimality in general, as the problem does not necessarily admit a convex or Rayleigh-quotient structure. Instead, the analytical solution provides a stable and well-defined representation that is effective for capturing critical-transition dynamics in practice.

ii) The approximate anti-diagonal consistency of $\tilde{Z}$ can be enforced by increasing the regularization weight θ $(0 \leq \theta \leq 1)$ of the explicit delay-embedding constraint term in the stARC loss (Eq. (5) and Eq. (19), Fig. S2). This provides a flexible penalty regulator to control embedding consistency within the optimization framework, without modifying the reservoir architecture or re-running simulations. In the extreme case where $\theta = 1$, the optimization is governed solely by the delay-embedding constraint (the second term in Eq. (5) and Eq. (19), "Embedding constraint" in Fig. S2), while the objective of maximizing the variance of $\tilde{Z}$ is no longer enforced. Although this setting may improve anti-diagonal consistency of $\tilde{Z}$, the resulting one-dimensional representation becomes less sensitive to dynamical instability and thus loses its effectiveness for critical transition detection. Conversely, when θ is too small, the delay-embedding constraint becomes insufficiently enforced, so that the elements along each anti-diagonal of $\tilde{Z}$ may differ significantly from one another. In this case, the optimization becomes dominated by the fluctuation term (the first term in Eq. (5) and Eq. (19), "Projection variance" in Fig. S2) and tends to degenerate toward a PCA-like low-dimensional approximation, which primarily captures variance structure rather than preserving the underlying dynamics in a one-dimensional representation. Therefore, θ should lie within an intermediate range in which delay-embedding consistency is preserved while instability-related fluctuations remain detectable.

However, for datasets exhibiting substantial heterogeneity or stochasticity, it may be difficult to find a suitable value of θ, because the two terms controlled by θ can become inherently difficult to balance. On the one hand, if all eigenvalues γ of $H(\Phi(X))\mathbf{V} = \gamma\mathbf{V}$ are negative, stARC may have reduced effectiveness for critical transitions. This can be understood from Eq. (5) and Eq. (19), whose optimization is equivalent to maximizing the Rayleigh quotient of $H(\Phi(X))$. Since the maximum Rayleigh quotient is given by its largest eigenvalue, a fully negative spectrum implies that the optimal objective value also remains negative, i.e., $(1-\theta)\sum_{i=1}^{l} \| \mathbf{W}_i \Phi(X) \|_2^2 - \theta \sum_{i=1}^{l-1} \| \mathbf{W}_i P - \mathbf{W}_{i+1} Q \|_2^2 < 0$. This means that the fluctuation term is smaller than the delay-embedding consistency term. Since the latter term is already constrained to remain close to 0, the fluctuation term is driven even closer to zero, thereby losing its ability to detect critical-transition signals. On the other hand, when θ decreases, the delay-embedding consistency cannot be sufficiently satisfied, which in turn compromises the accuracy of bifurcation-type identification.

As shown in Table R7 (see also in Supplementary Table S5), the bifurcation-identification performance of stARC depends on θ in a system-specific manner. For the coupled Lorenz system, stARC remains robust for relatively large θ values ($\theta > 0.8$), whereas the performance degrades at $\theta = 0.6$. By contrast, for the coupled Hénon and ADVP systems, high ACC and TPR are achieved over an intermediate range ($0.6 \leq \theta \leq 0.9$), while performance drops sharply at $\theta = 0.99$. These results indicate that θ should be chosen within a system-dependent range that balances delay-embedding consistency and fluctuation sensitivity. An important direction for future work is to explore how to adaptively select an appropriate θ for a given dataset from a theoretical perspective, and how to improve the flexibility and applicability of stARC by integrating AI-based methods. We have clarified this point in the Discussion section of the revised manuscript (Page 24).

Table R7. Bifurcation identification of stARC with different regularization parameters. Each method was evaluated on 400 randomly generated samples per system (200 bifurcating and 200 non-bifurcating) under consistent experimental settings. The performance is evaluated by accuracy (ACC) and true positive rate (TPR).

Dataset	**Metrics**	$\theta = 0.6$	$\theta = 0.8$	$\theta = 0.9$	$\theta = 0.99$
Coupled Lorenz	ACC	82.50%	97.75%	96.00%	94.50%
	TPR	66.00%	96.00%	92.00%	89.00%
Coupled Hénon	ACC	95.50%	97.75%	98.25%	60.25%
	TPR	97.00%	99.50%	99.00%	12.00%
Coupled ADVP	ACC	99.50%	99.25%	99.00%	50.05%
	TPR	99.00%	98.50%	98.00%	1.00%

<u>Reviewer's Comment 4:</u> The manuscript does not clearly explain how many neighbors L are required to obtain reliable results.

<u>Authors' Response:</u> Thank you for this valuable comment. The neighbor number L controls the extent of local spatial interaction incorporated into the stARC framework. From a theoretical perspective, the choice of neighborhood size L is governed by a trade-off between preserving local manifold geometry and achieving statistical stability. A too small L leads to high variance and unstable representations, whereas a too large L oversmooths the dynamics and obscures local nonlinear structures. To ensure consistency, the neighborhood size should increase with sample size while remaining sufficiently small relative to the total number of samples, e.g., 5-15%. We have clarified this point in Methods of the revised manuscript (Page 28).

To assess its influence, we conducted a detailed additional robustness analysis by varying L and noise strength σ over a range of values (see Table R8 and Supplementary Note S5). Our results indicate that the performance of stARC remains stable within a certain range of neighborhood sizes, suggesting that the framework is robust across a reasonable range of L.

Table R8. Bifurcation identification of stARC under different neighbor settings. Bifurcation identification performance is evaluated by accuracy (ACC) and true positive rate (TPR). For each simulated system, 200 bifurcating and 200 non-bifurcating samples were randomly generated. Results are reported with noise strengths $\sigma = 0.1,\ 0.2$, and 0.5, and different neighborhood settings, including no neighbors (i.e., $L = 0$) and $L = 5,\ 10,\ 15$.

System	Noise Strength σ	Metrics	Exclude Neighbors ($L = 0$)	Neighbor Number L		
				$L = 5$	$L = 10$	$L = 15$
Coupled Lorenz	$\sigma = 0.1$	ACC	67.50%	97.25%	97.75%	98.25%
		TPR	37.00%	96.00%	96.00%	96.50%
	$\sigma = 0.2$	ACC	58.00%	98.50%	97.25%	98.00%
		TPR	22.00%	97.50%	95.92%	96.00%
	$\sigma = 0.5$	ACC	67.00%	93.00%	95.75%	96.25%
		TPR	36.00%	90.00%	91.50%	92.50%
Coupled		ACC	87.50%	97.75%	97.75%	98.50%

Hénon	$\sigma = 0.1$	TPR	75.00%	96.00%	99.50%	99.00%
	$\sigma = 0.2$	ACC	86.00%	96.75%	97.00%	97.00%
		TPR	72.50%	94.00%	94.00%	95.75%
	$\sigma = 0.5$	ACC	71.00%	95.00%	95.75%	99.50%
		TPR	42.75%	90.00%	91.50%	99.00%
Coupled ADVP	$\sigma = 0.1$	ACC	97.00%	97.50%	99.25%	98.50%
		TPR	94.50%	95.00%	98.50%	97.00%
	$\sigma = 0.2$	ACC	97.00%	99.00%	99.50%	99.50%
		TPR	94.00%	98.00%	99.00%	99.50%
	$\sigma = 0.5$	ACC	81.00%	97.25%	99.50%	98.75%
		TPR	62.00%	96.50%	99.50%	97.50%

Reviewer's Comment 5: In line 237, the authors refer readers to methods to find out about the change point of sd(\bf{Z}^t). But the information is not provided in the method!

Authors' Response: Thank you for pointing out this issue. The detailed procedure for identifying the change point of $\mathrm{Std}(\mathbf{z})$ (previously denoted as $\mathrm{sd}(\boldsymbol{Z}^t)$) was originally provided in the Supplementary Information, but we agree that the corresponding method should also be briefly summarized in the main text. In response, we have now explicitly described the procedure for identifying the change point of $\mathrm{Std}(\mathbf{z})$ in the Methods section using the Bayesian online change point detection (BOCD) framework (Page 31). The key assumptions and main formulas of the BOCD approach have also been incorporated into the revised manuscript to improve clarity.

Reviewer's Comment 6: Figure captions should be self-explanatory. For example, in Figure 1, Panel B, it is unclear why the variable z is divided into two shaded regions. Similarly, the meaning of the yellow star in Panel C is not explained.

Authors' Response: Thank you for your helpful comment. To address this issue, we have expanded the caption of Fig. 1 to ensure it is self-explanatory. Specifically, in Panel B, we now explicitly explain that the two shaded regions represent the past information and the future information, respectively. In Panel C, the yellow star is now clearly defined as the detected early warning signal. These revisions improve the clarity of the figures.

Reviewer's Comment 7: The current manuscript could be made more accessible to a broader reader by providing intuitive explanations in addition to mathematical formulations.

Authors' Response: Thank you for your constructive suggestion. To improve accessibility and broaden the readership of the manuscript, we have added more intuitive explanations alongside the mathematical formulations in the revised version. In particular, we have expanded the conceptual descriptions in Section Methods and clarified the underlying dynamical intuition behind the stARC framework. For example, we explain Eq. (1) in Page 7 "In this way, high-dimensional spatial information is transformed into temporal dynamics, motivating the term spatial-to-temporal information transformation (STI)", and Eq. (18) in Page 28 "$\bar{d} = \sum_{j=0}^{L} d_j/(L+1)$ denotes the average Euclidean distance" in the revised manuscript. These additions aim to make the framework more accessible, while preserving mathematical precision.

Reviewer's Comment 8: I recommend embedding the main results into the literature, that is, not only citing related work but addresses the novelties and remaining open questions.

Authors' Response: Thank you for this insightful suggestion. We have revised the Introduction and Discussion sections to better position our main results within the existing literature. Specifically, we now clarify (i) the key novelties of stARC relative to existing representation- and prediction-based approaches, and (ii) remaining open challenges in critical transition detection for high-dimensional systems. In particular, we have expanded the discussion of prior studies on machine-learning-based early warning methods and Koopman/DMD methods, and clarified how the integration of STI-based embedding with reservoir computing differs from these existing frameworks and addresses a methodological gap (Page 22). We have also expanded the Discussion to outline open questions related to higher-codimension bifurcation identification, robustness under extreme nonstationarity, and an adaptive selection of θ for the effectiveness of stARC for a specific dataset (Page 24). These revisions help place our contributions more clearly within the broader literature.

Reviewer's Comment 9: In summary, the paper addresses an interesting problem and presents mathematically rigorous results, but the presentation requires substantial improvement. I encourage the authors to move beyond abstract mathematics and provide a clearer discussion of the broader implications and interpretations of their results for the interdisciplinary dynamical systems community. In addition, the limitations of the proposed framework, particularly its applicability to data with more complex bifurcation types, should be clearly stated.

Authors' Response: Thank you for this constructive comment. We agree that clearer interpretation and broader contextualization of the results would improve the accessibility of the manuscript to the interdisciplinary dynamical systems community.

In response, we have revised the Introduction, Results and Discussion sections to better explain the broader implications of our framework. In particular, we now emphasize how the proposed STI transformation and reservoir computing together provide a bridge between data-driven representation learning and classical dynamical-systems analysis, enabling a dynamical interpretation of critical transitions.

We have also expanded the Discussion to more clearly state the limitations of the current framework. In particular, we now discuss the challenges associated with identifying more complex bifurcation scenarios (e.g., higher-codimension or global bifurcations) and the potential need for richer spectral or nonlinear identification in such cases. These additions help clarify the scope of applicability of stARC and highlight directions for future research.

Response to the Comments of Reviewer #1

Reviewer's Comment 1: The authors have solved all the questions I had, So I propose to accept this manuscript.

Authors' Response: Thank you very much for your positive comments. We appreciate your time and consideration.

Response to the Comments of Reviewer #2

Reviewer's Comment 1: The authors have thoroughly addressed my previous concerns and have substantially improved the manuscript. I have no further questions or comments and recommend the paper for publication.

Authors' Response: Thank you very much for your positive assessment of our manuscript. We are grateful for your time and consideration.