Fundamental Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

Spatiotemporal information conversion machine for time-series forecasting

Hao Peng [a], Pei Chen [a,*], Rui Liu [a,b,*], Luonan Chen [c,d,e,f,*]

[a] School of Mathematics, South China University of Technology, Guangzhou 510640, China
[b] Pazhou Lab, Guangzhou 510330, China
[c] Key Laboratory of Systems Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China
[d] Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China
[e] Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, Guangdong 519031, China
[f] West China Biomedical Big Data Center, Med-X center for informatics, West China Hospital, Sichuan University, Chengdu 610041, China

ARTICLE INFO

Article history:
Received 14 July 2022
Received in revised form 19 November 2022
Accepted 16 December 2022
Available online xxx

Keywords:
Spatiotemporal information conversion network
Robust time-series forecasting
High-dimensional time series
Takens' embedding theory
Causal inference

ABSTRACT

Making time-series forecasting in a robust way is a difficult task only based on the observed data of a nonlinear system. In this work, a neural network computing framework, the spatiotemporal information conversion machine (STICM), was developed to efficiently and accurately render a forecasting of a time series by employing a spatial-temporal information (STI) transformation. STICM combines the advantages of both the STI equation and the temporal convolutional network, which maps the high-dimensional/spatial data to the future temporal values of a target variable, thus naturally providing the forecasting of the target variable. From the observed variables, the STICM also infers the causal factors of the target variable in the sense of Granger causality, which are in turn selected as effective spatial information to improve the robustness of time-series forecasting. The STICM was successfully applied to both benchmark systems and real-world datasets, all of which show superior and robust performance in time-series forecasting, even when the data were perturbed by noise. From both theoretical and computational viewpoints, the STICM has great potential in practical applications in artificial intelligence (AI) or as a model-free method based only on the observed data, and also opens a new way to explore the observed high-dimensional data in a dynamical manner for machine learning.

1. Introduction

It is still difficult to render forecasting of a nonlinear dynamical system based on time-series data due to its complicated nonlinearity and insufficient information regarding future dynamics. Actually, great efforts have been devoted to solving this challenging problem [1–3]. A number of methods, including statistical regression (*e.g.*, autoregressive integrated moving average (ARIMA) [4], robust regression [5]), exponential smoothing [6,7], and machine learning (*e.g.*, long-short-term-memory (LSTM) network [8,9]), were utilized in forecasting unknown states [10–13]. However, most of them cannot make satisfactory predictions regarding short-term time series due to insufficient information. To solve this problem, the auto-reservoir neural network (ARNN) [14]) was developed by using the semi-linearized spatial-temporal information (STI) transformation equation [14,15], which transforms high-dimensional information into temporal dynamics of any target variable, thus effectively extending the data size. However, this approach does not fully explore the nonlinearity of the STI equation from the observed data, which is essential for accurately predicting many complex systems. In addition, few existing approaches take spatial and temporal causal interactions of high-dimensional time-series data into consideration, which can compensate for insufficient data and provide reliable information to predict a complex dynamical system.

Under the condition that the steady state of a high-dimensional dynamical system is contained in a low-dimensional manifold, which is actually satisfied for most real-world systems, the STI transformation equation has theoretically been derived from delay embedding theory [16–18]. This equation can transform the spatial information of high-dimensional data into the temporal information of any target variable, thus equivalently expanding the sample size. Based on the STI transformation, the randomly distributed embedding (RDE) method was

* Corresponding authors.
E-mail addresses: chenpei@scut.edu.cn (P. Chen), scliurui@scut.edu.cn (R. Liu), lnchen@sibs.ac.cn (L. Chen).

https://doi.org/10.1016/j.fmre.2022.12.009

Please cite this article as: H. Peng, P. Chen, R. Liu et al., Spatiotemporal information conversion machine for time-series forecasting, Fundamental Research, https://doi.org/10.1016/j.fmre.2022.12.009

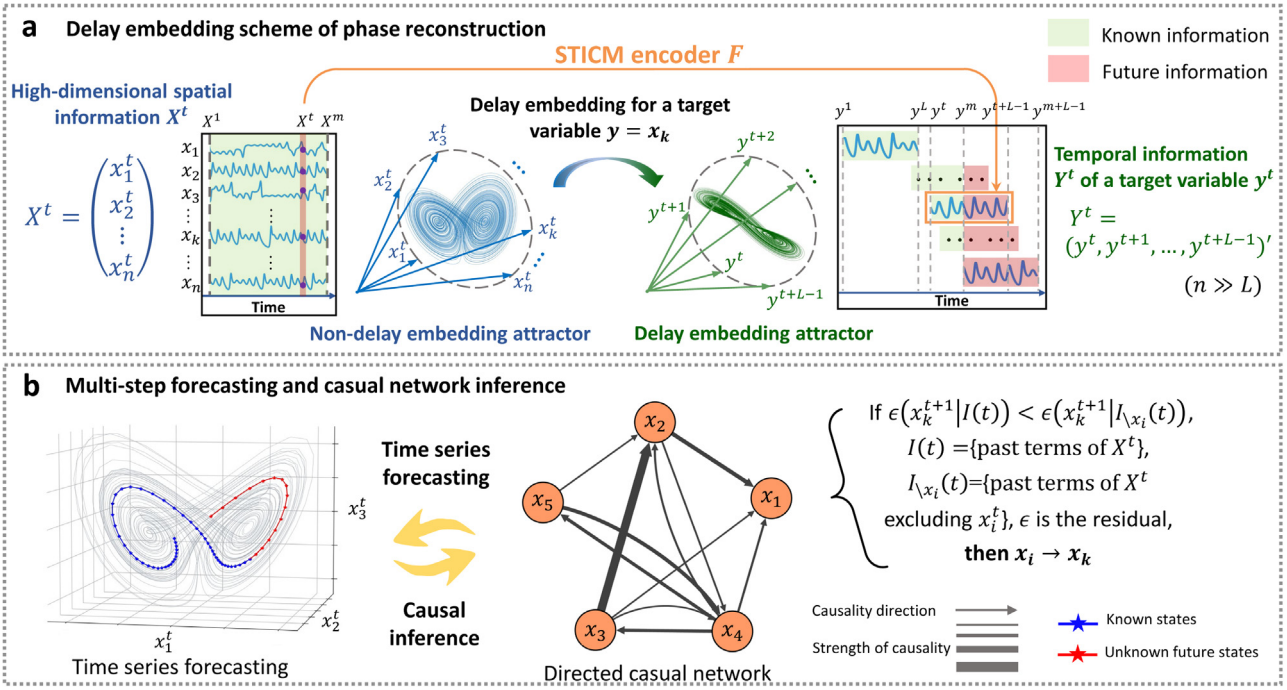

Fig. 1. High dimensional time series forecasting based on delay embedding scheme. (a) For a to-be-predicted/target variable y selected from the high-dimensional observables $\{x_1, x_2, \dots, x_n\}$, a temporal vector $\boldsymbol{Y}^t$ is constructed through a delay embedding scheme. The temporal vector $\boldsymbol{Y}^t$ is corresponding to an observed spatiotemporal matrix $[\boldsymbol{X}^{t-w}, \boldsymbol{X}^{t-w+1}, \dots, \boldsymbol{X}^t]$ via a nonlinear function F. (b) By inferring the causal relations and selecting the effective variables, the forecasting performance is considerably improved. Note that the mapping F is from a matrix $[\boldsymbol{X}^{t-w}, \boldsymbol{X}^{t-w+1}, \dots, \boldsymbol{X}^t]$ to a vector $\boldsymbol{Y}^t$.

proposed to predict the one-step-ahead value from high-dimensional time-course data by separately constructing multiple STI maps (or primary STI equations) to form the distribution of the predicted values [15]. Our recent auto-reservoir computing framework ARNN [14] achieves multistep-ahead forecasting based on a semi-linearized STI transformation; however, the nonlinear features and spatiotemporal causal relations of the observed high-dimensional variables have not yet been exploited, which restricts the forecasting performance in the sense of robustness and accuracy.

On the other hand, a temporal convolutional network (TCN) [19] was recently reported to outperform canonical recurrent neural networks (RNNs) [20–22], such as the LSTM network [8,9], and the gated recurrent units (GRU) [23], across a diverse range of sequence modelling tasks and datasets. Compared with RNNs, the TCN possesses advantages, including a longer effective memory length, a flexible receptive field size, stable gradients, a low memory requirement for training, variable-length inputs, and parallelism [19]. Besides, the TCN employs dilated convolution, which enables an exponentially large receptive field, to handle long sequences. However, the traditional TCN does not fully reveal the causal relations among high-dimensional variables and cannot make accurate multistep-ahead forecasting without future information/labels.

In this study, we propose a novel framework, *i.e.*, spatiotemporal information conversion machine (STICM), to achieve accurate and robust multistep-ahead forecasting with high-dimensional data, and explore the underlying causal relations among high-dimensional variables. The central idea is to represent both primary and conjugate STI equations in an autoencoder form (Figs. 1 and 2) by exploiting the advantages of the causal convolution and STI nonlinear transformation. Computationally, the STICM includes three basic processes: (1) the embedding scheme to reconstruct the phase space (Fig. 1a). (2) the STICM to realize the STI transformation (Fig. 2a, b). (3) effective/causal variable selection to forecast more accurately and robustly (Fig. 1b). In particular, we adopt both the primary and the conjugate forms of the STI equations to encode (through nonlinear function F) and decode (through the reverse function F^{-1}) the temporal dynamics from the high-dimensional data (Fig. 2a and Eq. 7) Through the STI equations, the STICM transforms the spatiotemporal information of high-dimensional data to the temporal/dynamical future values of a target variable. Given the time-course data of high-dimensional variables, the STICM trains the encoder F and decoder F^{-1} by taking both spatial and temporal information into consideration (Fig. 2a, b), thus equivalently expanding the data size on the target variable or naturally resulting in the future values of the target variable y. Moreover, by comparing the forecasting error, the STICM directly makes the Granger inference of causal factors on the target variable, which are in turn selected as the effective/spatial variables to significantly improve the forecasting robustness and accuracy of the target variable.

To validate the accuracy and robustness, STICM was applied to a series of representative mathematical models, *i.e.*, a 90-dimensional coupled Lorenz system [24] under different noise conditions. Furthermore, the STICM was applied to many real-world datasets in this study and predicted, *e.g.* (1) the daily number of cardiovascular inpatients in the major hospitals of Hong Kong [25,26], (2) the wind speed in Japan [27], (3) a ground meteorological dataset in the Houston, Galveston, and Brazoria areas [28], (4) the population of the plankton community isolated from the Baltic Sea [29,30], (5) the spread of COVID-19 in the Kanto region of Japan [31], (6) the traffic speed of multiple locations in Los Angeles [32]. The results show that the STICM achieves multistep-ahead forecasting that is better than the other seven existing methods in terms of accuracy and robustness. More descriptions of each compared method are illustrated in Supplementary Section 6. As a model-free method based only on the observed data, the STICM framework paves a new way to make multistep-ahead forecasting by incorporating the primary-conjugate STI equations into an autoencoder form. This framework exploits both the STI transformation and causal convolutional structure, thus is of great potential for practical applications in many scientific and engineering fields, and also opens a new way to dynamically explore high-dimensional information in machine learning.

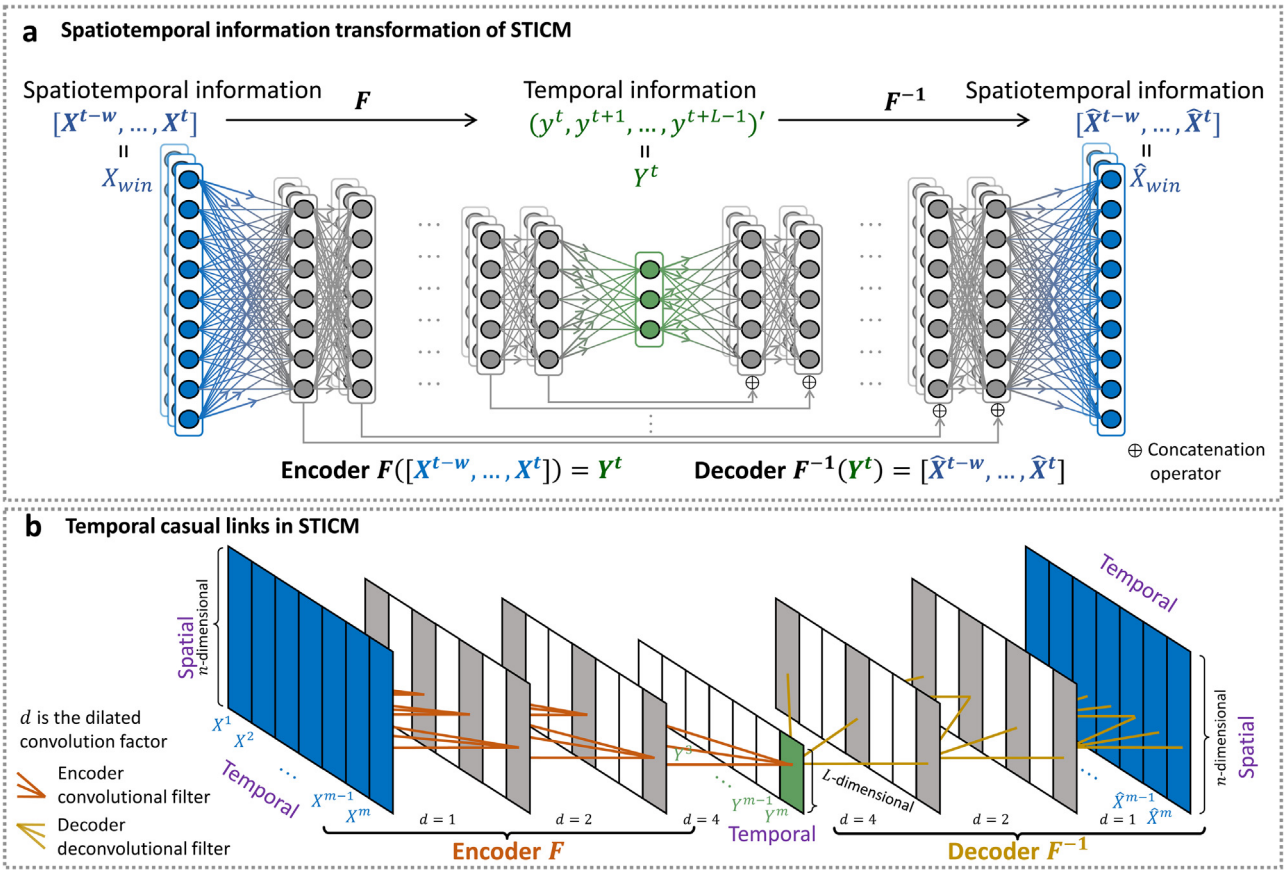

Fig. 2. Schematic illustration of the STICM framework. (a) The information flow of the STICM is similar to the autoencoder (AE) but is constrained by primary and conjugate STI equations. The primary STI equation represents the encoder, while the conjugate STI equation corresponds to the decoder. However, unlike AE, the low-dimensional/temporal code Y^t is mapped by the delay embedding scheme from the time series of a target variable y. (b) The encoder and decoder of STICM are implemented by a temporal convolutional network (TCN) structure and a temporal deconvolutional structure, respectively, through which the spatiotemporal matrix $[\boldsymbol{X}^1, \boldsymbol{X}^2, \ldots, \boldsymbol{X}^t]$ is input sequentially and mapped to $[\boldsymbol{Y}^1, \boldsymbol{Y}^2, \ldots, \boldsymbol{Y}^t]$.

2. Methods

The detailed description of the parameters and variables in STICM framework are summarized in Supplementary Table 3. The default hyperparameters for STICM on each dataset are summarized in Supplementary Table 5.

2.1. Delay embedding theorem for dynamical systems

Generally, the dynamics of a discrete-time dissipative system can be presented as

$$\mathbf{X}^{t+1} = \phi(\mathbf{X}^t) \tag{1}$$

where $\phi: \mathbb{R}^n \to \mathbb{R}^n$ represents a nonlinear function, whose n-dimensional variables are denoted as vector $\mathbf{X}^t = (x_1^t, x_2^t, \ldots, x_n^t)'$ with time superscript t and vector transpose symbol " $'$ ". The Takens' embedding theorem provides the following facts [16,17].

If $\mathcal{V} \subseteq \mathbb{R}^n$ is a compact attractor with the Minkowski dimension/box-counting dimension d, for a smooth diffeomorphism $\phi: \mathcal{V} \to \mathcal{V}$ and a smooth function $h: \mathcal{V} \to \mathbb{R}$, there is a generic property that the mapping $\Phi_{\phi,h}: \mathcal{V} \to \mathbb{R}^L$ is an embedding when $L > 2d$, that is,

$$\Phi_{\phi,h}(X) = \left(h(X), h\circ\phi(X), \ldots, h\circ\phi^{L-1}(X)\right)' \tag{2}$$

where symbol " $\circ$ " is the function composition operation. In particular, letting $X = \mathbf{X}^t$ and $h(\mathbf{X}^t) = y^t$ where $y^t \in \mathbb{R}$, then the mapping above has the following form with $\Phi_{\phi,h} = \Phi$ and

$$\Phi(\mathbf{X}^t) = (y^t, y^{t+1}, \ldots, y^{t+L-1})' = \mathbf{Y}^t \tag{3}$$

Moreover, since the embedding is one-to-one mapping, we can also derive its conjugate form $\Psi: \mathbb{R}^L \to \mathbb{R}^n$ as $\mathbf{X}^t = \Phi^{-1}(\mathbf{Y}^t) = \Psi(\mathbf{Y}^t)$ (Supplementary Section 1). Here $\mathbf{X}^t$ is an n-dimensional vector here, but sometimes it is used as D-dimensional variables ($D \le n$) in this work. The above theory can be summarized as the following spatiotemporal information (STI) transformation equation:

$$\begin{cases} \Phi(\mathbf{X}^t) = \mathbf{Y}^t \\ \mathbf{X}^t = \Psi(\mathbf{Y}^t) \end{cases} \tag{4}$$

where $\Phi: \mathbb{R}^D \to \mathbb{R}^L$ and $\Psi: \mathbb{R}^L \to \mathbb{R}^D$ are nonlinear differentiable functions satisfying $\Phi\circ\Psi = id$, symbol " $\circ$ " represents the function composition operation and id denotes the identity function.

Note that to use the causal convolution framework of TCN, we let $X = [\mathbf{X}^{t-w}, \mathbf{X}^{t-w+1}, \ldots, \mathbf{X}^t]$ in Eq. 3 for this work with map F (*i.e.*, Eq. 7), rather than $X = \mathbf{X}^t$ with map Φ.

2.2. STICM framework with STI transformation

For each observed high-dimensional/spatial state $\mathbf{X}^t = (x_1^t, x_2^t, \ldots, x_n^t)'$ with n variables with $t = 1, 2, \ldots, m$, a corresponding delayed/temporal vector $\mathbf{Y}^t = (y^t, y^{t+1}, \ldots, y^{t+L-1})'$ is constructed for one target variable y (*e.g.*, $y^t = x_k^t$) through a delay embedding scheme with parameter L as the embedding dimension satisfying $n > L > 1$ (Fig. 1a), where the symbol " $'$ " is the transpose of a vector. Specifically, the matrix X of the original measurable variables

$\{x_1, x_2, \dots, x_n\}$ is as follows:

$$X = \left[\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^m\right] = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^m \\ x_2^1 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^m \end{bmatrix}_{n\times m} \tag{5}$$

Through the delay embedding scheme, the matrix Y of the target variable $y = x_k$ is

$$Y = \begin{bmatrix} y^1 & y^2 & \cdots & y^m \\ y^2 & y^3 & \cdots & y^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ y^L & y^{L+1} & \cdots & y^{m+L-1} \end{bmatrix}_{L\times m} \tag{6}$$

where Y contains the unknown/future values $\{y^{m+1}, y^{m+2}, \dots, y^{m+L-1}\}$ in the lower-right corner (shadow area) of the target variable. It is clear that $\mathbf{X}^t$ is a known high-dimensional/spatial vector for multiple variables at one time point t, while $\mathbf{Y}^t$ is a temporal vector of one target y at multiple time points $t,\ t+1, \dots,\ t+L-1$.

Based on the generalized Takens' embedding theory [33], the dynamics of the original system can be topologically reconstructed from a delay embedding scheme if $L > 2d > 0$, where d is the Minkowski dimension of the attractor [16,17]. By combining the causal convolution structure and STI transformation, we developed a STICM framework, which provides multistep-ahead forecasting with dynamic causal inference among the observed variables on the basis of both the primary and conjugate STI equations (Fig. 2a). The known high-dimensional time series, *i.e.*, one sliding window matrix $X_{win} = [\mathbf{X}^{t-w}, \mathbf{X}^{t-w+1}, \dots, \mathbf{X}^t]$ with window size $w+1$ from the whole spatiotemporal matrix $X = [\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^m]$, is mapped to one delayed temporal vector $\mathbf{Y}^t$ for $t = 1, 2, \dots, m$, which actually forms the following STICM-based STI equation set:

$$\begin{cases} F\left(\left[\mathbf{X}^{t-w}, \mathbf{X}^{t-w+1}, \dots, \mathbf{X}^t\right]\right) = \mathbf{Y}^t \\ F^{-1}\left(\mathbf{Y}^t\right) = \left[\hat{\mathbf{X}}^{t-w}, \hat{\mathbf{X}}^{t-w+1}, \dots, \hat{\mathbf{X}}^t\right] \end{cases} \tag{7}$$

where the first formula is the primary equation with $F : \mathbb{R}^{n\times(w+1)} \to \mathbb{R}^L$ and the second formula is the conjugate equation with $F^{-1} : \mathbb{R}^L \to \mathbb{R}^{n\times(w+1)}$ (Fig. 2a), $\hat{\mathbf{X}}^t$ is the recovered vector of $\mathbf{X}^t$. Given m known states $\boldsymbol{X}^t$ $(t = 1, 2, \dots, m)$, there are $L-1$ future values of y, *i.e.*, $\{y^{m+1}, y^{m+2}, \dots, y^{m+L-1}\}$ in $\mathbf{Y}^t$ (Fig. 1a and Supplementary Fig. 1b). Matrix $[\mathbf{X}^{t-w}, \mathbf{X}^{t-w+1}, \dots, \mathbf{X}^t]$ of Eq. 7 is the known spatiotemporal information of n variables, and $\mathbf{Y}^t$ presents the temporal information of the target variable. In Eq. 7, the first and second equations are the primary and conjugate forms of the STI equations, respectively. The primary equation encodes one spatiotemporal matrix $[\mathbf{X}^{t-w}, \mathbf{X}^{t-w+1}, \dots, \mathbf{X}^t]$ to one temporal vector $\mathbf{Y}^t$, while the conjugate form decodes/recovers the encoded temporal information $\mathbf{Y}^t$ to the spatiotemporal information $[\hat{\mathbf{X}}^{t-w}, \hat{\mathbf{X}}^{t-w+1}, \dots, \hat{\mathbf{X}}^t]$. The STI equations (Eq. 7) hold when some generic conditions are satisfied according to the delay embedding theory [16,17]. Clearly, the properly determined function F is the key to solving the STICM-based STI equations (Eq. 7) for the high-dimensional input/matrix X and providing the future values $\{y^{m+1}, y^{m+2}, \dots, y^{m+L-1}\}$ of the target variable. The details of Takens' embedding theory and the STI equations are given in Supplementary Section 1 and Supplementary Section 2, respectively.

The dilated causal convolution layers are employed in the framework of STICM (Fig. 2b), that is, for input series $X = [\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^m]$ and a filter $g : \{0, \dots, k-1\} \to \mathbb{R}$, the dilated causal convolution operation G on element $\mathbf{X}^t$ is defined as

$$\left(\mathbf{X}^t\right) = \left(X *_d g\right)\left(\mathbf{X}^t\right) = \sum_{i=0}^{k-1} g(i) \cdot \mathbf{X}^{t-d\cdot i} \tag{8}$$

where d is the dilation factor, k is the filter size. Dilation is thus equivalent to introducing a fixed step between every two adjacent filter taps. A larger dilation enables an output at the top level to represent a wider range of inputs, thus effectively expanding the receptive field of a ConvNet. In this way, we construct the network structure for encoder F. Similarly, we adopted an inverse dilated convolution scheme in decoder F^{-1}, which is shown in Supplementary Section 4 in detail.

2.3. STICM algorithm

The determination of F and F^{-1} includes two main factors: 1) the semi-supervised training scheme and 2) the effective variable selection. This structure of STICM is capable of exploiting not only the input of spatial information but also the temporally intertwined information among the numerous variables of the complex dynamic system, thus significantly enhancing the forecasting robustness and accuracy. In this study, each layer of the encoder F and decoder F^{-1} is followed by the ReLU activation function. The STICM algorithm is carried out to uncover the to-be-predicted/future values $\{y^{m+1}, y^{m+2}, \dots, y^{m+L-1}\}$ of the target $y = x_k$ with the following procedure.

Step 1: Construct the STICM-based STI equation. Based on the delay embedding scheme, we construct the delay-embedded matrix of the target variable y as Eq. 6 with the columns $\mathbf{Y}^t = (y^t, y^{t+1}, \dots, y^{t+L-1})'$. Clearly, vectors $\{\mathbf{Y}^{m-L+2}, \dots, \mathbf{Y}^m\}$ contains the unknown/future values. The steady state or the attractor is generally constrained in a low-dimensional space for a high-dimensional dissipative system, which holds for most real-world systems. Assuming $F = (F_1, F_2, \dots, F_L)'$ and $L > 2d$ where d is the Minkowski dimension of the attractor, the primary form of the STICM-based STI equation set (Eq. 7) is

$$\begin{bmatrix} F_1\left(\left[\mathbf{X}^1\right]\right) & F_1\left(\left[\mathbf{X}^1, \mathbf{X}^2\right]\right) & \cdots & F_1\left(\left[\mathbf{X}^{m-w}, \dots, \mathbf{X}^m\right]\right) \\ F_2\left(\left[\mathbf{X}^1\right]\right) & F_2\left(\left[\mathbf{X}^1, \mathbf{X}^2\right]\right) & \cdots & F_2\left(\left[\mathbf{X}^{m-w}, \dots, \mathbf{X}^m\right]\right) \\ \vdots & \vdots & \ddots & \vdots \\ F_L\left(\left[\mathbf{X}^1\right]\right) & F_L\left(\left[\mathbf{X}^1, \mathbf{X}^2\right]\right) & \cdots & F_L\left(\left[\mathbf{X}^{m-w}, \dots, \mathbf{X}^m\right]\right) \end{bmatrix} = \begin{bmatrix} y^1 & y^2 & \cdots & y^m \\ y^2 & y^3 & \cdots & y^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ y^L & y^{L+1} & \cdots & y^{m+L-1} \end{bmatrix} \tag{9}$$

In a similar form of Eq. 9, we have the conjugate equation with F^{-1} (see Supplementary Eq. 7). Clearly, by simultaneously solving both the primary and conjugate STICM-based STI equations, the STICM provides a series of future values $\{y^{m+1}, y^{m+2}, \dots, y^{m+L-1}\}$, which is indeed the $(L-1)$-step-ahead forecasting.

Step 2: Train the STICM network. Because there are both known and unknown values in the delay embedding matrix Y, the STICM is trained in a semi-supervised manner. Specifically, the nonlinear mappings $F = (F_1, F_2, \dots, F_L)'$are fit via a "consistently self-constrained scheme" simultaneously for preserving the time consistency for the known and unknown values, thus maintaining the integrity of F. According to the framework of STICM, there are three high-level requirements for the network used in training.

Due to the delay-embedding nature in the output Y (as shown in Eq. 9), we have totally $m+L-3$ temporally self-constrained conditions

$$F_{j-1}\left(\left[\mathbf{X}^{t-w}, \mathbf{X}^{t-w+1}, \dots, \mathbf{X}^t\right]\right) = F_j\left(\left[\mathbf{X}^{t-w-1}, \mathbf{X}^{t-w}, \dots, \mathbf{X}^{t-1}\right]\right) \tag{10}$$

where $j \in \{2,\ 3, \dots,\ L\}$ and $\mathbf{X}^t = (x_1^t, x_2^t, \dots, x_n^t)'$ is a spatial sample at time point t. Among conditions Eq. 10, there are $m-1$ conditions for the determined states and $L-2$ conditions for future values. Clearly, these conditions constrain the training of STICM in terms of the temporal sequence of samples. For the target variable y, the estimated values of its delay embeddings in each iteration are obtained as follows

$$\hat{Y} = \begin{bmatrix} \left(\hat{y}^1\right)_1 & \left(\hat{y}^2\right)_1 & \cdots & \left(\hat{y}^m\right)_1 \\ \left(\hat{y}^2\right)_2 & \left(\hat{y}^3\right)_2 & \cdots & \left(\hat{y}^{m+1}\right)_2 \\ \vdots & \vdots & \ddots & \vdots \\ \left(\hat{y}^L\right)_L & \left(\hat{y}^{L+1}\right)_L & \cdots & \left(\hat{y}^{m+L-1}\right)_L \end{bmatrix} \tag{11}$$

where $\left(\hat{y}^t\right)_j$ $(t = 1,\ 2, \dots, m+L-1;\ j = 1,\ 2, \dots, L)$ is generated from the output of the j^{th} sub-mapping function F_j.

Through an auto perception procedure, the training or optimization of STICM is accomplished through a process of minimizing a loss function with three weighted mean-squared error components

$$\mathcal{L} = \lambda_1 \mathcal{L}_{DS} + \lambda_2 \mathcal{L}_{FC} + \lambda_3 \mathcal{L}_{REC} \tag{12}$$

In Eq. 12, the first part $\mathcal{L}_{DS}$ is a determined-state loss from the observed/known states $\{y^1, y^2, \dots, y^m\}$ of y, and is of the following form

$$\mathcal{L}_{DS} = \frac{1}{2mL - L^2 + L} \sum_{j=1}^{L} \sum_{t=j}^{m} \left(\left(\hat{y}^t\right)_j - y^t \right)^2 \tag{13}$$

where $\left(\hat{y}^t\right)_j$ $(t = 1, 2, \dots, m)$ is the estimation of $F_j([\mathbf{X}^{t-w}, \mathbf{X}^{t-w+1}, \dots, \mathbf{X}^t])$, and y^t $(t = 1, 2, \dots, m)$ is the known value of y. Loss $\mathcal{L}_{DS}$ is constructed from the differences between the estimations $\left(\hat{y}^t\right)_j$ and the observed values (ground truth) y^t for all past time points t $(t = 1, 2, \dots, m)$.

In Eq. 12, the second part $\mathcal{L}_{FC}$ is a future-consistency loss in terms of the future/unknown series $\{y^{m+1}, y^{m+2}, \dots, y^{m+L-1}\}$ of y, and has form

$$\mathcal{L}_{FC} = \frac{1}{L(L-1)} \sum_{j=2}^{L} \sum_{t=m+1}^{m+j-1} \left(\left(\hat{y}^t\right)_j - \text{mean}\left(\hat{y}^t\right) \right)^2 \tag{14}$$

where $\text{mean}(\hat{y}^t)$ denotes the average of all estimated future values of $\hat{y}^t$ in Eq. 11 that corresponds to the same future time point t $(t = m+1, m+2, \dots, m+L-1)$. Clearly, $\mathcal{L}_{FC}$ is constructed from the temporally self-constrained conditions in Eq. 10. An intuitive understanding of the future-consistency loss is that by minimizing $\mathcal{L}_{FC}$, it ensures that the outputs from different sub-mappings but corresponding to the same future time point t are identical, which preserves the temporal consistency of the outputs at the lower right corner of the delay embedding matrix $\hat{Y}$ during the training procedure.

In Eq. 12, the third part $\mathcal{L}_{rec}$ is a reconstruction loss in terms of the consistency of encoder and decoder, which is of the following form

$$\mathcal{L}_{rec} = \| X - \hat{X} \|_{\text{F}} \tag{15}$$

where $X = [\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^t]$, $\hat{X} = [\hat{\mathbf{X}}^1, \hat{\mathbf{X}}^2, \dots, \hat{\mathbf{X}}^t]$, and $\| \cdot \|_{\text{F}}$ is the Frobenius norm.

Based on the integration of the above three losses, the STICM is trained in a semi-supervised manner. The cooperation of future-consistency loss $\mathcal{L}_{FC}$ and determined-state loss $\mathcal{L}_{DS}$ helps to fit the nonlinear mapping $F = (F_1, F_2, \dots, F_L)'$. The reconstruction loss $\mathcal{L}_{rec}$ guarantees the consistency of encoder and decoder. After the convergence of the training process, the $m + L - 1$ to-be-predicted values $\{y^{m+1}, y^{m+2}, \dots, y^{m+L-1}\}$ can eventually be determined from the estimated matrix $\hat{Y}$, *i.e.*,

$$y^{m+i} = \text{mean}\left(\hat{y}^{m+i}\right) = \frac{1}{L-i} \sum_{j=i+1}^{L} \left(\hat{y}^{m+i}\right)_j \tag{16}$$

with $i = 1, 2, \dots, L-1$. The implementation of the deconvolution layer in decoder F^{-1} is similar and provided in Supplementary Section 4 and Supplementary Fig. 1c.

Step 3: Identify the causal/driving variables.

To decrease the noisy effect and boost the robustness on the forecasting results, we choose the most relevant variables to the target variable from the high-dimensional data. Given a time series of n-dimensional samples $(x_1^t, x_2^t, \dots, x_n^t)'_{t=1,2,\dots,m}$, we calculate the forecasting errors between the case "with an observable x_i" and the case "without x_i". Then, one can determine whether x_i is a causal/effective factor of the target variable y in the sense of Granger causality, thus improving the forecasting performance by selection or deletion of the variable.

First, a reference RMSE ϵ_r of the model trained by the original n-dimensional input was calculated as the normalized difference between original and predicted values, *i.e.*,

$$\epsilon_r = \text{RMSE}(y, \hat{y}|\Lambda) = \sqrt{\frac{\sum_{t=m}^{m+L-1} (y^t - \hat{y}^t|\Lambda^t)^2}{L-1}} \tag{17}$$

where y^t denotes the original value of the target variable, $\hat{y}^t$ denotes the predicted one, and Λ^t denotes the past terms of X^t. Subsequently, by excluding $x_i, i = 1, 2, \dots, n$ from the original data, the model is trained based on an $(n-1)$-dimensional input with a test RMSE ϵ_i,

$$\epsilon_i = \text{RMSE}(y, \hat{y}|\Lambda \backslash x_i) = \sqrt{\frac{\sum_{t=m}^{m+L-1} \left(y^t - \hat{y}^t|\Lambda^t \backslash x_i^t\right)^2}{L-1}} \tag{18}$$

where $\Lambda^t \backslash x_i^t$ denotes the past terms of X^t without x_i^t. Then, a causality error $\epsilon_{i,r}$ is obtained as

$$\epsilon_{i,r} = \epsilon_i - \epsilon_r \tag{19}$$

which denotes the influence of Granger causality from variable x_i to y. After ranking all $\epsilon_{i,c}$ $(i = 1, 2, \dots, n)$, we selected the spatial information of top q causal/effective variables as new input data, which are most relevant to the target variable y. By excluding irrelevant variables or noisy information, the final STICM is trained based on such a lower-dimensional input and enhances the forecasting performance in terms of both accuracy and robustness. The schematic illustration of this step can be found in Supplementary Fig. 1a. The time complexity analysis of this step is presented in Supplementary Section 7. The other details of the STICM algorithm are provided in Supplementary Section 3.

3. Results

3.1. Performance of the STICM on Lorenz models

To demonstrate the basic idea of the STICM method, the synthetic time-series datasets under multiple noise levels were generated from a benchmark nonlinear system, *i.e.*, the following 90-dimensional coupled Lorenz model $(n = 90)$ [24]

$$\dot{\mathbf{X}}(t) = G(\mathbf{X}(t); P) \tag{20}$$

where P is a parameter vector of the function set $G(\cdot)$with $\mathbf{X}(t) = (x_1^t, x_2^t, \dots, x_{90}^t)'$. The specific Lorenz system is presented in Supplementary Section 5.

3.1.1. Noise-free situation

We first apply the STICM to a noise-free Lorenz system (Eq. 20) with $m = 50$ and $L - 1 = 15$, *i.e.*, taking a time series of 50 steps as known information/input, and making a 15-step-ahead forecasting /output for the target variables. As demonstrated in Fig. 3, the STICM predicted the future values for both the single-wing (Fig. 3c, the observed and to-be-predicted series distributed in a single wing of the attractor) and cross-wing (Fig. 3d, the observed and to-be-predicted series distributed in two different wings of the attractor) cases. By randomly selecting three target variables y_1, y_2 and y_3 from $\{x_1, x_2, \dots, x_{90}\}$, the forecasting performances of the STICM on three-dimensional cases are presented in Fig. 3a and b. Notably, the predicted values (the red curves) for each target variable were obtained by the one-time forecasting; that is, the STICM provides an efficient way to obtain a whole horizon (15 steps) of future information. Clearly, on the basis of the 90-dimensional short-term time series, the STICM inferred the top 30 effective/causal variables of the targets and significantly improved the performance in both accuracy and robustness by applying the forecasting of the target with these 30 variables (Fig. 3c and d, Tables 1 and S1). Note that the training and forecasting of the STICM are based only on the observed data.

Here and below, to validate the effectiveness of the STICM, its forecasting performance was compared with seven representative methods, *i.e.*, the LSTM network [8,9], Holt's exponential smoothing (HES) [6,7], autoregression (AR) [34], autoregressive integrated moving average (ARIMA) [4], radial basis function network (RBFN) [35], multiview embedding (MVE) [36], and support vector regression (SVR) [37,38]. Additionally, it is from Table 1 that the STICM achieved better performances compared with other time-series forecasting methods on the

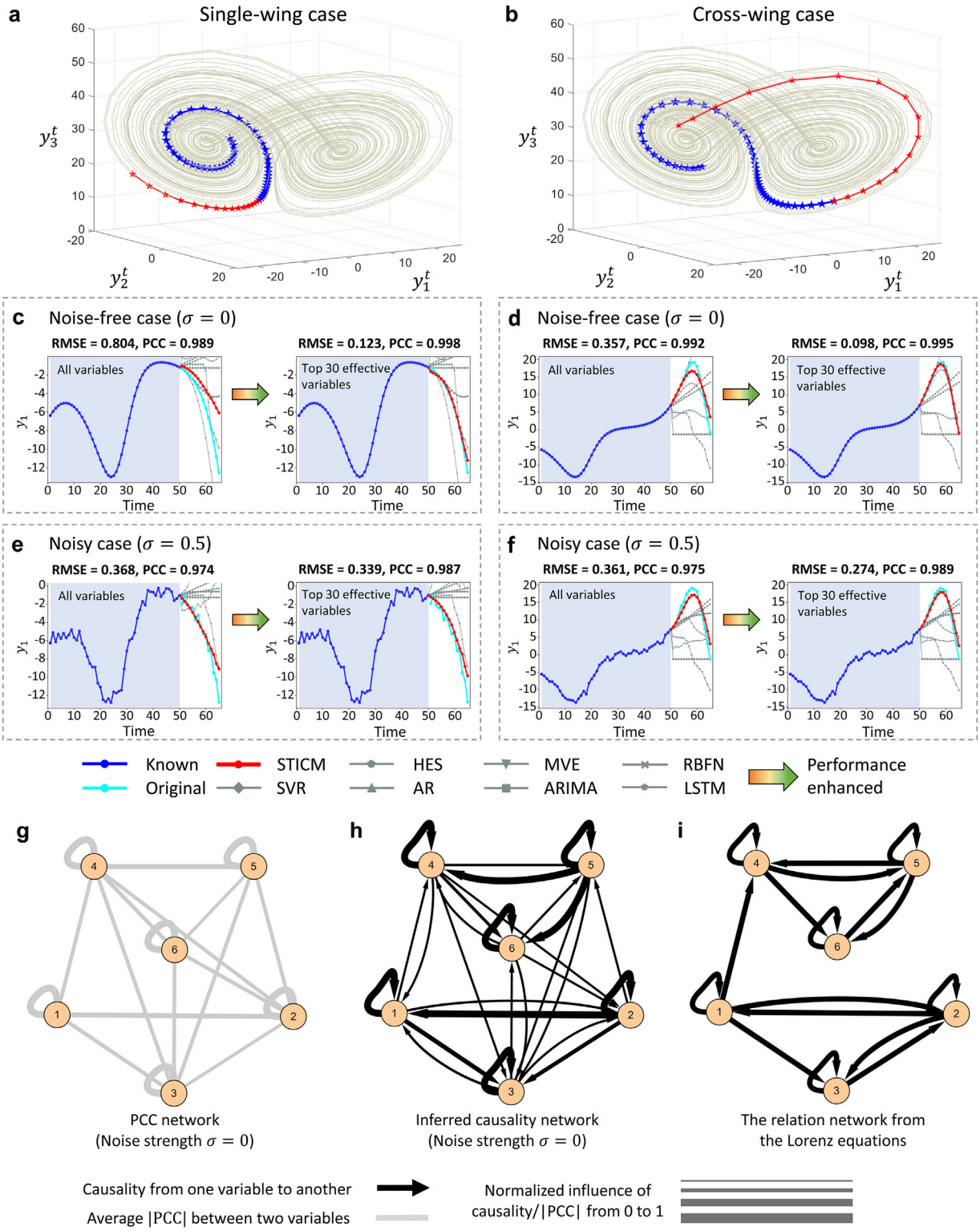

Fig. 3. The forecasting performance of the STICM on the high-dimensional Lorenz system. In noise-free or noisy situations, the time-series data were generated on the basis of the 90D coupled Lorenz system (Eq. 20). We randomly selected three targets y_1, y_2 and y_3 among variables $\{x_1, x_2, \ldots, x_{90}\}$. By applying the STICM with parameter $m = 50$ (*i.e.*, the length of the input series is 50), the 15-step-ahead forecasting ($L - 1 = 15$) were performed for y_1, y_2 and y_3, respectively. (a) The forecasting of the 3D system of y_1, y_2 and y_3 in the single-wing situation. (b) The forecasting of the 3D system in the cross-wing situation. (c) The forecasting of y_1 in a noise-free case of the single-wing situation. (d) The forecasting of y_1 in a noise-free case of the cross-wing situation. (e) The forecasting of target y_1 in a noisy case (with noise strength $\sigma = 0.5$) of the single-wing situation. (f) The forecasting of y_1 in a noisy case (with noise strength $\sigma = 0.5$) of the cross-wing situation. For each case, the forecasting are carried out based on all variables (the left panels of (b), (c), (d), and (e)) and based on the top 30 causal variables (the right panels of (b), (c), (d), and (e)). The PCC network (each edge is weighted with Pearson correlation coefficient) and causal relation network (each edge is weighted with Granger causality index) of the six selected effective variables in the noisy-free case (g) and (h), respectively. These two networks in the noisy case (with noise strength $\sigma = 0.5$) are illustrated in Supplementary Fig. 4. (i) The relation network of the six variables from the Lorenz equations.

Table 1
Comparison of the performance among eight forecasting methods.

Dataset	Metric[a]	Methods							
		STICM	MVE	AR	ARIMA	HES	LSTM	RBFN	SVR
Lorenz system	RMSE	**0.111**	1.498	1.546	1.685	1.564	0.806	1.798	2.024
(noise-free)	PCC	**0.997**	0.731	-0.66	0.297	-0.637	0.992	-0.419	0.193
Lorenz system with noise	RMSE	**0.307**	1.607	1.486	1.382	1.565	1.62	1.86	2.026
($\sigma = 0.5$)	PCC	**0.989**	0.711	-0.66	-0.339	-0.644	-0.15	0.29	0.218
Cardiovascular inpatients	RMSE	**0.228**	0.968	1.071	1.065	1.391	1.104	0.994	0.804
	PCC	**0.974**	0.467	0.351	0.366	-0.157	0.21	0.244	0.865
Plankton density	RMSE	**0.548**	1.669	1.441	0.776	2.408	3.647	3.728	2.84
	PCC	**0.917**	0.522	0.359	0.781	-0.372	0.377	-0.503	0.412
Wind speed	RMSE	**0.908**	2.632	1.348	3.28	5.144	2.243	1.985	2.384
	PCC	**0.942**	0.895	-0.28	0.817	0.417	-0.189	0.873	0.321
Traffic speed	RMSE	**0.66**	2.248	2.344	3.135	2.728	4.597	6.676	3.544
	PCC	**0.901**	0.359	0.044	0.162	-0.434	0.204	-0.181	0.265
Japan Covid-19	RMSE	**0.608**	2.311	2.553	4.148	2.819	4.031	3.48	6.16
transmission	PCC	**0.9**	0.012	0.049	-0.037	0.356	0.016	0.405	0.422
Meteorological data	RMSE	**0.811**	0.935	1.065	1.029	1.267	1.154	1.278	1.165
	PCC	**0.728**	0.324	0.015	0.093	-0.171	0.067	-0.053	0.341

[a] The performance metrics include the values of the root-mean-square error (RMSE) and the Pearson correlation coefficient (PCC). The RMSE was normalized by the standard deviation of the real data.

noise-free cases of the 90-dimensional Lorenz system; that is, the accuracy of the STICM is the best in terms of the Pearson correlation coefficient (PCC) and the root mean square error (RMSE). Specifically, the RMSEs decreased from 0.804 to 0.123 and from 0.357 to 0.098 for cases in Fig. 3c, d, respectively. As shown in Table 1, the STICM achieved the smallest RMSE 0.111 in the noise-free situation, while the best record of the other methods is 0.806. In addition, the inferred causality network among the six selected variables (Fig. 3h) is consistent with the direct causal relations from the original equations (Fig. 3g), and fully reveals intrinsic dynamic associations (including both direct and indirect causal relations) of the coupled Lorenz system comparing with the PCC network (Fig. 3g). Note that the direct causal relation from x_i to x_j in Fig. 3i is determined if x_i is one of the bases/independent variables of x_j in Supplementary Eq. 17. Moreover, the performances of eight time-series forecasting methods on the datasets without effective/causal variable selection are shown in Supplementary Table 1.

3.1.2. Additive noise situation

Second, the STICM was applied to the noisy cases of the 90D Lorenz system (Eq. 20) with additive white noise ($\sigma = 0.5$) to predict the same target variable, while $m = 50$, and $L - 1 = 15$. Specifically, the cross-wing case is exhibited in Fig. 3f, and the single-wing case is presented in Fig. 3e. After the selection of the top 30 effective/causal variables, the forecasting accuracy of the STICM improves significantly and is better than that of the other seven methods for both the single-wing and cross-wing cases (Table 1 and Supplementary Table 1). Specifically, the RMSEs of our proposed method decreased from 0.368 to 0.339 (Fig. 3e) and from 0.361 to 0.274 (Fig. 3f). The average RMSE (0.307) of STICM across all noisy cases is the best among all forecasting methods (Tables 1 and S1). Therefore, the STICM still predicts the future dynamics accurately when the system is perturbed by additive noise, which demonstrates the robustness property of the STICM framework.

The STICM achieves satisfactory performance even with noisy data compared with traditional approaches because of its two characteristics, that is, simultaneously solving both conjugated STI equations in Eq. 7, and effective variable selection among all observables.

3.2. The application of the STICM on real-world datasets

Predicting the future values of key variables by exploiting the relevant high-dimensional information is of great importance for studying complex systems forecasting potential risk. The STICM method was applied to the following various high-dimensional real-world datasets, and was also compared with seven existing methods. The detailed performances of all the forecasting methods are exhibited in Table 1. For each dataset, the specific settings and parameters are presented in Supplementary Table 2. The description of each dataset is also provided in Supplementary Section 5. To further demonstrate the generalization of the proposed method, the STICM is applied to the field of human pose prediction [39], and corresponding contents are presented in Supplementary Section 8 and Supplementary Table 6.

3.2.1. Cardiovascular inpatients forecasting

The first real-world dataset contains the number series of cardiovascular inpatients in major hospitals in Hong Kong and the indices series of air pollutants, *i.e.*, the daily concentrations of nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), respirable suspended particulate (Rspar), mean daily temperature, relative humidity, etc., which were obtained from air monitoring stations in Hong Kong from 1994 to 1997 [25]. As the previous study has reported the relevance between air pollutants and cardiovascular inpatients [40], the STICM was employed to predict daily cardiovascular disease admissions on the basis of a group of air pollutants (Fig. 4). Thus, for the 14-dimensional system ($n = 14$), the known time points were set as $m = 70$ (days) and the forecasting horizon as $L - 1 = 25$ (days). By inferring and selecting the top 11 effective variables, the forecasting accuracy of the STICM increases significantly and is better than that of the other methods (Tables 1 and S1). As shown in Fig. 4i, the STICM uncovers the causal relationship between the admissions of cardiovascular diseases and air pollutants, in accordance with the literatures [41–43]. The inferred causal relations among the air pollutants also agree with the chemical reactions (Table S4).

3.2.2. Plankton density forecasting

The STICM was then applied to a dataset collected in a long-term experiment with a marine plankton community isolated from the Baltic Sea from 1990 to 1997 [29,30,44], including the species abundance time series of herbivorous and predatory zooplankton species, several phytoplankton species, detritivores, and bacteria. These plankton species constructed a complex food web. As shown in Fig. 4e-4h, the STICM predicts the dynamic trend of the abundances of two target species (*Cyclopoids* and *Rotifers*), with parameter settings $n = 12$ (total 12 plankton species), $m = 18$ (the known abundance information of 18 steps), and $L - 1 = 6$ (6 step-ahead forecasting). By selecting the top 8 effective variables, the STICM achieves a higher forecasting accuracy, *i.e.*, RMSE = 0.542 and

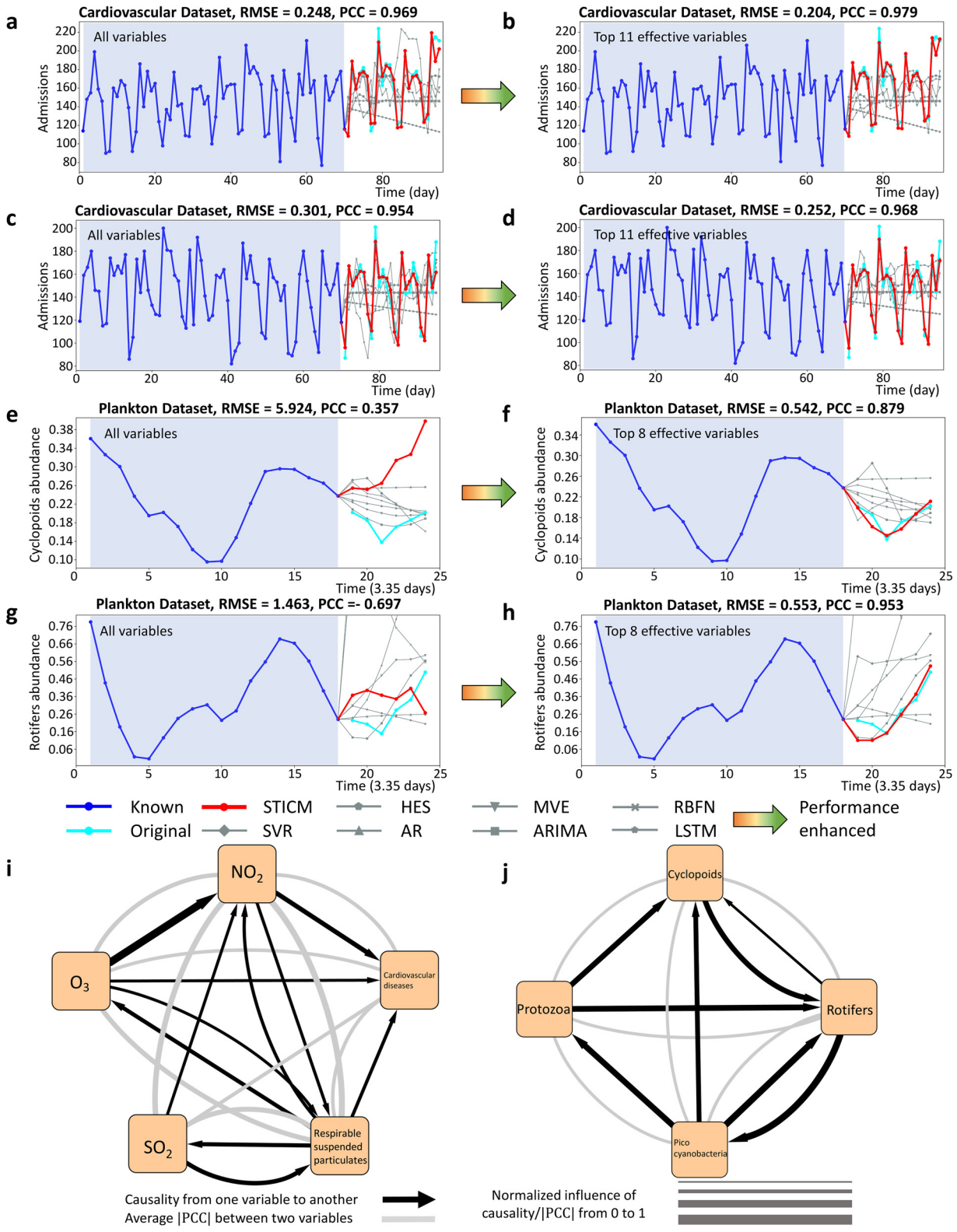

Fig. 4. Future state forecasting of cardiovascular admission and plankton abundance. For two periods (a)-(b) and (c)-(d), the STICM predicted the number of cardiovascular admissions based on the high-dimensional time series of air pollutant indices with known length $m = 70$ and forecasting horizon $L - 1 = 25$. For two target planktons, *i.e.*, *Cyclopoids* and *Rotifers*, the STICM predicted the dynamic change of their abundance based on the high-dimensional plankton dataset with known length $m = 18$ and forecasting horizon $L - 1 = 6$. By selecting the top 11 and top 8 effective variables for the cardiovascular admission dataset and plankton abundance dataset, respectively, the forecasting accuracy of the STICM increases significantly ((b), (d), (f), and (h)). The performances of the STICM and other methods are compared in (a)-(h). Based on the STICM, causal networks (i) and (j) were constructed to show the regulatory relationship among cardiovascular admission and air pollutants and that among the plankton, respectively.

PCC = 0.879 for cyclopoids and RMSE = 0.553 and PCC = 0.953 for *Rotifers*, than other methods (Tables 1 and S1). In addition, Fig. 4j depicts the inferred causal network among four species, *i.e., Rotifers, Cyclopoids, Pico cyanobacteria*, and *Protozoa*. Being consistent with the original food chain network, the causal network also contains other relations among these four species. For example, the links from *Cyclopoids* to *Rotifers* and from *Rotifers* to *Pico cyanobacteria* reveal the fact that the abundance of predators can influence that of the preys. The link from *Protozoa* to *Rotifers* reveals the competitive relation when they have the common predators and preys.

3.2.3. Wind speed forecasting

Wind speed is one of the weather variables with highly time-varying characteristics in nonlinear meteorological systems and is thus extremely difficult to predict. The wind speed dataset was collected from the Japan Methodological Agency [27]. Among the 155 wind stations distributed all around Japan, we selected one target station near Tokyo. As shown in Fig. 5, the STICM predicted the dynamics of the wind speed in the target station with parameter settings $n = 155$, $m = 64$, and $L - 1 = 26$ (Fig. 5a and c). After inferring and selecting the 70 most effective variables, the forecasting accuracy of the STICM increases significantly, as shown by the comparisons in Fig. 5b and d. Based on the effective variables, the forecasting of the STICM are better than those of the other methods (Tables 1 and S1). Long-term forecastings were also performed by selecting 70 top effective variables and are provided in Fig. 5e and f, from which the wind speed in the target station was continuously predicted for a whole season (3 months). The forecastings for more periods are provided in Supplementary Fig. 5. The inferred causal relations between the locations of top 50 effective variables and that of the target station are consistent with the corresponding monsoon-specific wind directions (Fig. 5g and h).

3.2.4. Traffic speed forecasting

The transportation system consisting of vehicles, roads and other transportation elements, can be considered as a high-dimensional complex system [45]. Meanwhile, intelligent inspection on such a system is of great importance to city management and development. However, due to the complexity of traffic dynamical systems, predicting the traffic flow precisely is full of challenges. STICM was applied to predict the traffic speed (mile/h), which was based on a dataset generated from $n = 207$ loop detectors in the 134-highway of California, USA. The traffic speed was recorded every five minutes from Mar 1st, 2012 to Jun 30th, 2012 [32]. In such a dynamic system, each loop detector was considered as a variable by which the traffic speed detected was mainly determined by the observed values from the nearest neighbor sensors. We selected four target sensors, which are the intersections of main roads (Target 1 is located at the intersection of San Diego Freeway and Ventura Freeway; Target 2 is located at the intersection of Hollywood Freeway and Ventura Freeway; Target 3 is located at the intersection of Glendale Freeway and Ventura Freeway; Target 4 is located at the intersection of Hollywood Freeway and Harbor Freeway). Consequently, 55 nearest-neighbor detectors of the target detector were selected to constitute a subsystem. By applying the STICM, the multistep forecasting ($L - 1 = 19$ time points ahead) of four target locations/sensors were obtained based on the neighbor 55 variables ($n = 55$, Fig. 6a, c, e, and g) and top 30 effective variables ($n = 30$, Fig. 6b, d, f, and h) with $m = 60$ time points. Based on the effective variables, the RMSEs of the predicted traffic speed on 19 time points significantly decreased, *i.e.*, from 1.757 (Fig. 6a) to 0.852 (Fig. 6b) for Target 1, from 1.551 (Fig. 6c) to 0.536 (Fig. 6d) for Target 2, from 2.207 (Fig. 6e) to 0.762 (Fig. 6f) for Target 3, and from 1.844 (Fig. 6g) to 0.489 (Fig. 6h) for Target 4. The forecasting results of the STICM are better than those of the other seven forecasting methods (Tables 1 and S1). Supplementary Movie S1 shows the dynamic change in the predicted traffic speed. As shown in Fig. 6i and j, most of the causal/effective detectors are distributed around each target detector.

3.2.5. Japan Covid-19 transmission forecasting

The pandemic of coronavirus disease 2019 (COVID-19) has posed a global threat to public health. To assist public health departments with their strategic planning, it is important to predict the spread of this infectious disease. The STICM provides a data-driven approach to predict the dynamic change in daily new cases of infectious disease. As shown in Fig. 7, the STICM predicted the number of COVID patients in several cities with severe epidemics in Japan [31,46], with parameter settings $m = 30$ and $L - 1 = 14$. Based on all 47 districts ($n = 47$), the forecasting of COVID-19 new cases of the six target districts are provided in Fig. 7a (Tokyo), 7c (Tochigi), and 7e (Gunma). After inferring and selecting the top 20 effective/causal districts in each target district, the STICM was predicted much more accurately than the other methods for the six districts (Fig. 7b (Tokyo), 7d (Tochigi), and 7f (Gunma)). The quantitative comparisons are provided in Tables 1 and S1. Fig. 7g presents the network of COVID-19 transmission in the Kanto region, Japan. The forecasting for more districts is provided in Supplementary Fig. 3.

3.2.6. Meteorological data forecasting

The last real-world dataset contains 72-dimensional ground meteorological data ($n = 72$) recorded per month in an area around Houston, Galveston, and Brazoria [28] from 1998 to 2004. As shown in Supplementary Fig. 2, the relative humidity and geopotential height were accurately predicted. For each target index, the STICM was applied to make a 17-step-ahead forecasting ($L - 1 = 17$) based on the former $m = 50$ steps of the 72-dimensional data. The forecasting results of the STICM are better than those predicted by other seven methods (Tables 1 and S1).

4. Discussion

Time-series forecasting is of great importance in a wide range of real-world applications. There are many interesting related works leveraging spatial information from data to enhance time-series forecasting [47–51]. In this work, we proposed the STICM framework to achieve the multistep-ahead forecasting with causal factor inference based on high-dimensional time series in a robust way. Through STICM, the spatiotemporal information of high-dimensional observables is transformed into the temporal information of a target variable on the basis of the delay embedding theory. That is, the primary STI form is an encoder which transforms the spatiotemporal matrix $[\mathbf{X}^{t-w}, \mathbf{X}^{t-w+1}, \dots, \mathbf{X}^{t}]$ to the temporal vector $\mathbf{Y}^t$ of a target variable by F, while the conjugate STI form recovers the temporal vector $\mathbf{Y}^t$ back to the original matrix $[\mathbf{X}^{t-w}, \mathbf{X}^{t-w+1}, \dots, \mathbf{X}^{t}]$ by F^{-1}. Training F and F^{-1} simultaneously in a semi-supervised manner, the STICM solves the STI equations and makes the forecasting highly robust, as shown in the applications. Clearly, the multiple future/unknown values $\{y^{m+1}, y^{m+2}, \dots, y^{m+L-1}\}$ are obtained concurrently by the STICM, indicating that the proposed method makes the multistep-ahead forecasting. The results of the ablation study on training loss (Eq. 12) are shown in Supplementary Section 9 and Supplementary Table 7, demonstrating that all the constraints in the primary and conjugate STICM-based STI equations contribute to producing accurate and robust time-series forecasting. Moreover, the STICM carries out causal inference based on Granger causality and thus identifies the causal/effective variables on the target variable. Causal inference enables a deep understanding of the intrinsic dynamics of the complex system, thus providing the interpretability of the STICM, and to a considerable extent, reducing the dimension. Thus, forecasting accuracy is enhanced by selecting the effective variables for forecasting. A series of applications show that the STICM achieves better performance than seven traditional forecasting approaches. However, there are limitations of Granger causality that it fails to reveal the real causality in some cases. In the future, we will explore the causality relationship from different perspectives to better investigate the intrinsic dynamics of a complex system.

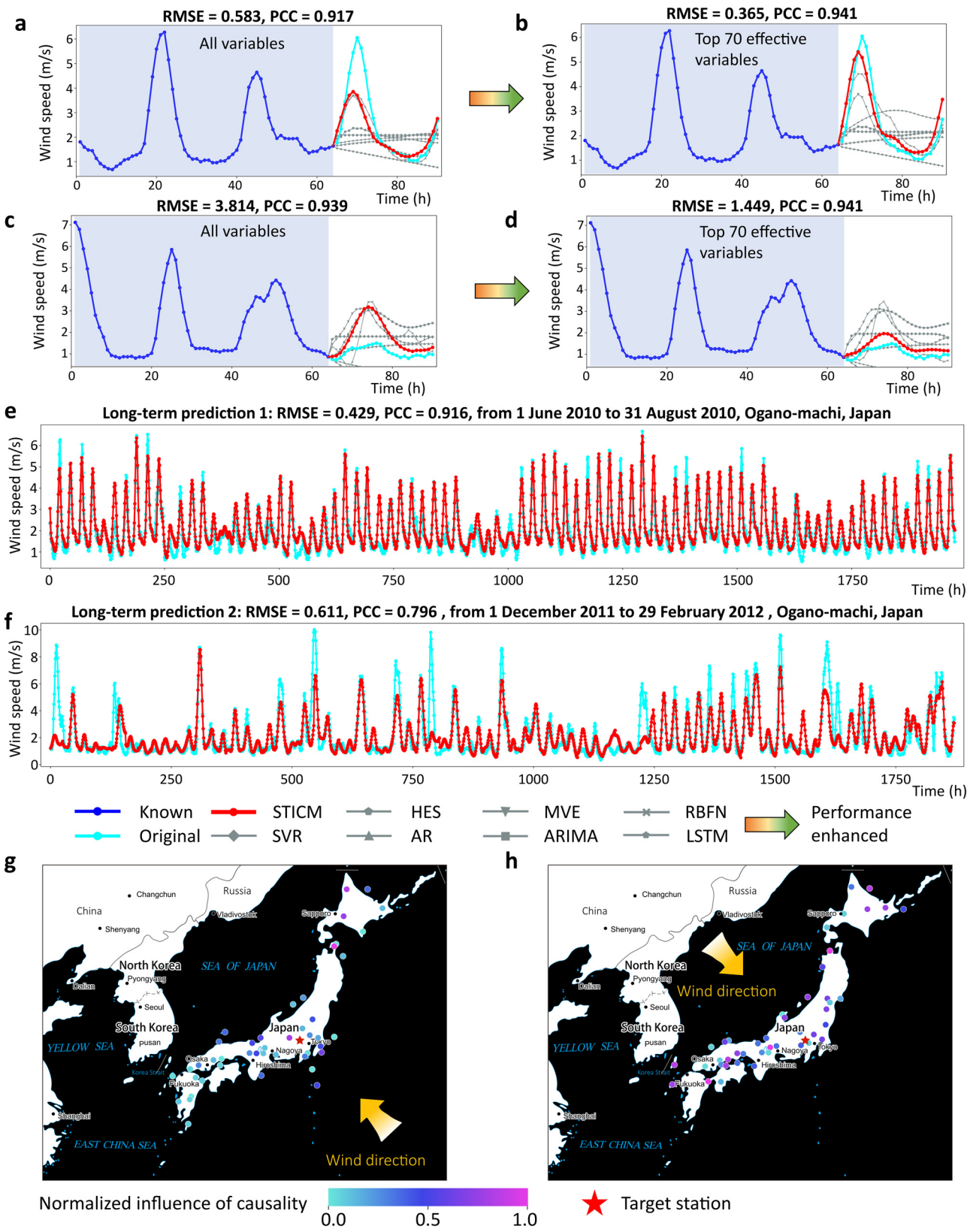

Fig. 5. Wind speed forecasting. The STICM predicts the wind speed of a target station around Tokyo marked by a pink star symbol. Based on the time series from all 155 variables (the wind speed of 155 stations) and from the selected top 70 effective variables, the STICM predicted the future wind speed for two periods ((a) and (c) based on all variables and (b) and (d) based on the top 70 variables) with known length $m = 64$ and forecasting horizon $L - 1 = 26$. The long-term forecastings were performed by the STICM as in (e) and (f), which showed the robustness of the proposed method by predicting the whole season (3 months). The causal relations among the target station and its top 50 effective/causal stations are provided in (g) (for a wet monsoon with wind direction mainly from the southeast) and (h) (for a dry monsoon with wind direction mainly from the northwest) (Based on the standard map with the approval number of GS (2020) 4400 on the standard map service website of the Ministry of Natural Resources of the People's Republic of China, the base map has not been modified.).

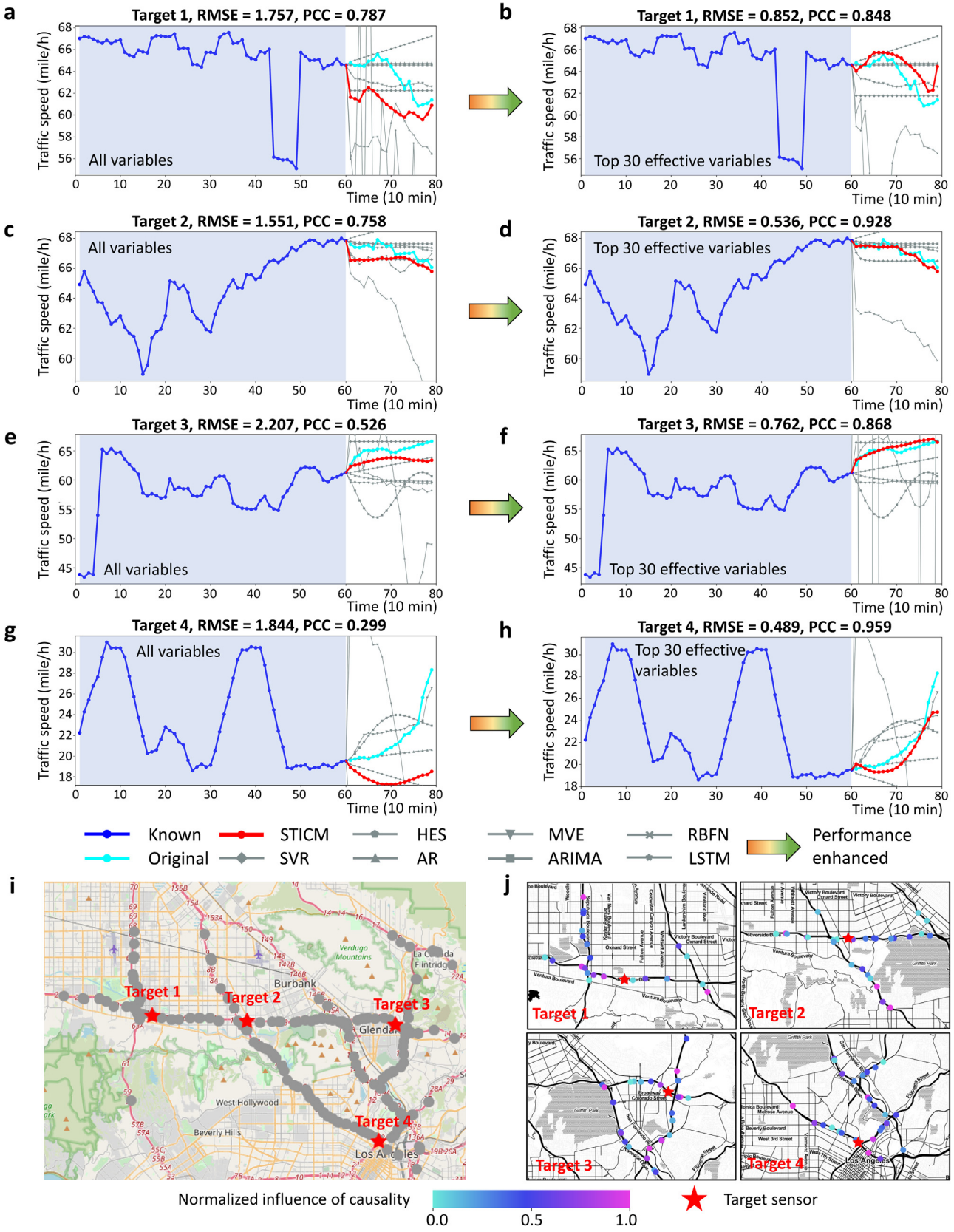

Fig. 6. Traffic speed forecasting. Based on the 207-dimensional traffic speed dataset, the STICM predicted the traffic speed of four target locations/sensors with 60-step known information ($m = 60$) and 19-step forecasting horizon ($L - 1 = 19$), *i.e.*, (a) and (b) for target 1, (c) and (d) for target 2, (e) and (f) for target 3, (g) and (h) for target 4, where the four target locations were marked by red star symbols in (i). By inferring and selecting the top 30 effective variables (*i.e.*, the effective traffic speeds in 30 locations), the forecasting accuracy of the STICM significantly increases and is better than that of the other methods ((b), (d), (f), and (h)). The associations/causal relations among the neighboring locations/sensors are shown in (j).

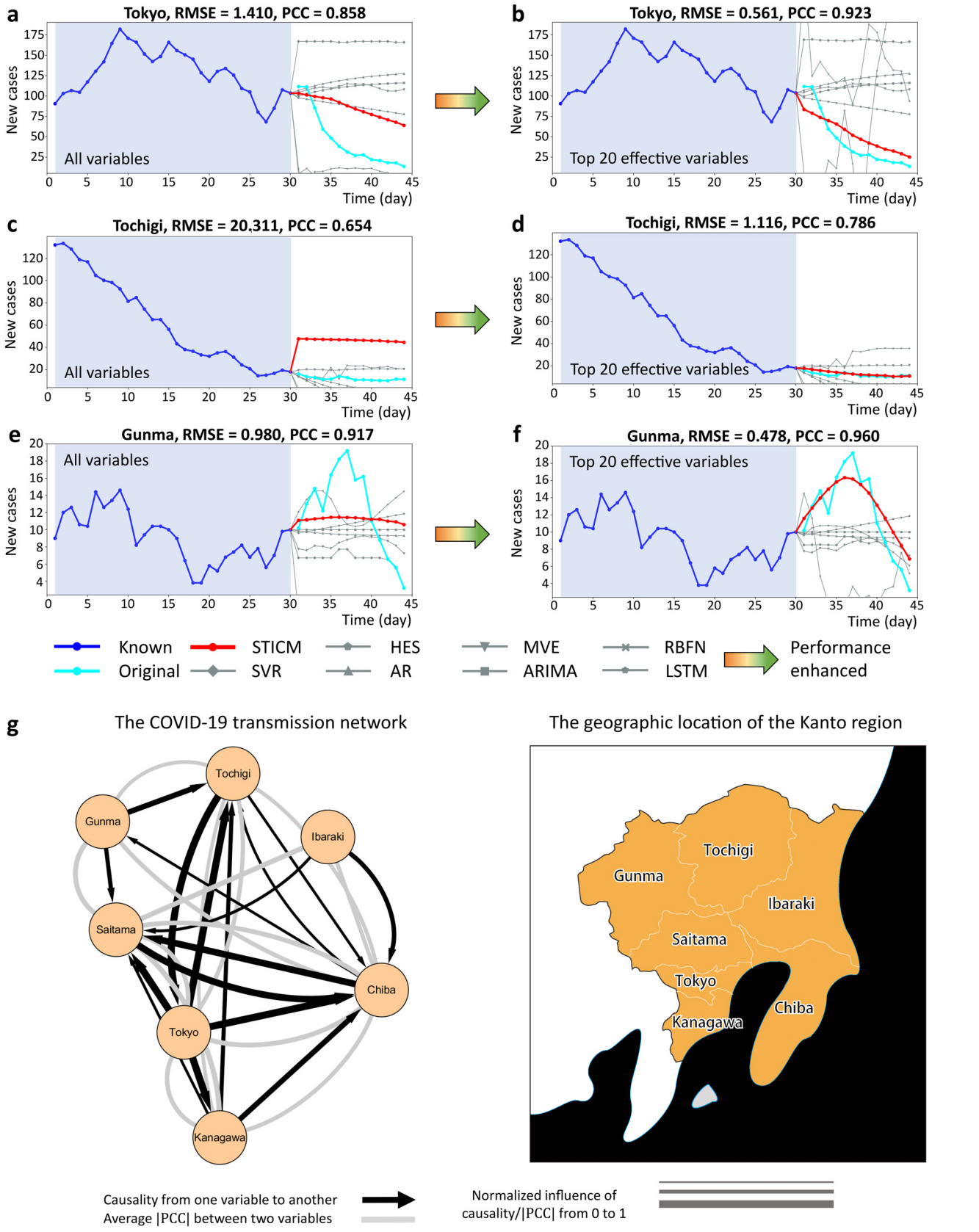

Fig. 7. Predicting the number of COVID-19 patients. Based on the time series of COVID-19 new cases of 47 districts (the left subfigures) or selected top 20 effective districts in each target district (the right subfigures), the STICM predicts the numbers of future new cases, with 30-step known information ($m = 30$) and 14-step forecasting horizon ($L - 1 = 14$), *i.e.*, (a) and (b) for Tokyo, (c) and (d) for Tochigi, (e) and (f) for Gunma. Based on the STICM, the casual network (g) of COVID-19 transmission in the Kanto region, Japan revealed the regulatory relationship in terms of COVID-19 spread among the districts in this region (Based on the standard map with the approval number of GS (2020) 4400 on the standard map service website of the Ministry of Natural Resources of the People's Republic of China, the base map has not been modified.).

In conclusion, the proposed STICM framework has the following advantages compared with traditional forecasting methods. First, the STICM is capable of exploring the time-series data and transforming the spatial information of high-dimensional observables into the temporal information of a target variable. Second, once being trained in a semi-supervised manner, the STICM well solves the primary and conjugate STI equations simultaneously (corresponding to a spatiotemporal convolutional autoencoder), thus making the time-series forecasting robustly even in noise-perturbed cases. Third, in practical applications, the STICM can distinguish the effective/relevant variables, thus unveiling the underlying causal mechanism (in the sense of Granger causality) among massive observables of the dynamical systems. In addition, building on a solid theoretical background of the STI equations and with the TCN causal convolution structure, the STICM opens a new way to explore the spatiotemporal information from high-dimensional time series, and has been validated by the applications to a variety of real-world scenarios.

Author Contributions

Hao Peng, Pei Chen, Rui Liu and Luonan Chen proposed original conceptualization. Hao Peng prepared the data, implemented the software, and conducted the analysis. Hao Peng, Pei Chen and Rui Liu investigated the related works, prepared the figures and wrote the article. Rui Liu and Luonan Chen supervised the manuscript. All authors have read and approved the content of the manuscript.

Data availability

All data needed to validate the conclusions are present in the paper and/or the Supplementary Materials. All data are available at https://github.com/mahp-scut/STICM.

Code availability

The code used in this study is available at https://github.com/mahp-scut/STICM.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

We thank the Japan Meteorological Agency, which provided the datasets of wind speeds used in this study (available via the Japan Meteorological Business Support Center). This work was supported by the National Natural Science Foundation of China (12026608, 62172164, 12271180, 12131020, and 31930022); the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB38040400); Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004); the Special Fund for Science and Technology Innovation Strategy of Guangdong Province (2021B0909050004, 2021B0909060002); the Major Key Project of Peng Cheng Laboratory (PCL2021A12); and JST Moonshot R&D (JPMJMS2021).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fmre.2022.12.009.

References

[1] P.J. Brockwell, R.A. Davis, Time series: theory and methods, Springer Science & Business Media, 2009.
[2] S.-Q. Zhang, Z.-H. Zhou, ARISE: ApeRIodic SEmi-parametric Process for Efficient Markets without Periodogram and Gaussianity Assumptions 2021.
[3] V. Kuznetsov, M. Mohri, Learning theory and algorithms for forecasting non-stationary time series, Adv. Neural Inf. Process. Syst. 28 (2015).
[4] G.E. Box, D.A. Pierce, Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, J. Am. Statist. Assoc. 65 (1970) 1509–1526.
[5] P.J. Rousseeuw, A.M. Leroy, Robust regression and outlier detection, 589, John Wiley & Sons, 2005.
[6] C.C. Holt, Forecasting seasonals and trends by exponentially weighted moving averages, Int. J. Forecast. 20 (2004) 5–10.
[7] R.G. Brown, Exponential smoothing for predicting demand. Operations Research, vol. 5, Inst Operations Research Management Sciences 901 Elkridge Landing Rd, Ste 400, Linthicum HTS, MD 21090-2909; 1957, p. 145–145.
[8] Z. Karevan, J.A. Suykens, Transductive LSTM for time-series prediction: An application to weather forecasting, Neural Netw. (2020).
[9] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[10] J.T. Connor, R.D. Martin, L.E. Atlas, Recurrent neural networks and robust time series prediction, IEEE Trans. Neural Netw. 5 (1994) 240–254.
[11] W.W. Wei, Time series analysis. The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2, 2006.
[12] W.-X. Wang, Y.-C. Lai, C. Grebogi, Data based identification and prediction of nonlinear and complex dynamical systems, Phys. Rep. 644 (2016) 1–76.
[13] A.S. Weigend, Time series prediction: forecasting the future and understanding the past, Routledge, 2018.
[14] P. Chen, R. Liu, K. Aihara, et al., Autoreservoir computing for multistep ahead prediction based on the spatiotemporal information transformation, Nat. Commun. 11 (2020) 4568.
[15] H. Ma, S. Leng, K. Aihara, et al., Randomly distributed embedding making short-term high-dimensional data predictable, Proc. Natl. Acad. Sci. U.S.A. 115 (2018) E9994–10002.
[16] T. Sauer, J.A. Yorke, M. Casdagli, Embedology, J. Stat. Phys. 65 (1991) 579–616.
[17] F. Takens, in: Detecting strange attractors in turbulence. Dynamical systems and turbulence, Springer, Warwick, 1981, pp. 366–381. 1980.
[18] M. Casdagli, Nonlinear prediction of chaotic time series, Physica D 35 (1989) 335–356.
[19] S. Bai, J.Z. Kolter, V. Koltun, An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv:180301271 [Cs] 2018.
[20] J. Gehring, M. Auli, D. Grangier, et al., A convolutional encoder model for neural machine translation. ArXiv Preprint arXiv:161102344 2016.
[21] C. Lea, M.D. Flynn, R. Vidal, et al., Temporal Convolutional Networks for Action Segmentation and Detection. arXiv:161105267 [Cs] 2016.
[22] Y.N. Dauphin, A. Fan, M. Auli, et al., Language modeling with gated convolutional networks, in: International conference on machine learning, PMLR, 2017, pp. 933–941.
[23] K. Cho, B. Van Merriënboer, C. Gulcehre, et al., Learning phrase representations using RNN encoder-decoder for statistical machine translation. ArXiv Preprint arXiv:14061078 2014.
[24] J.H. Curry, A generalized Lorenz system, Commun. Math. Phys. 60 (1978) 193–204.
[25] T.W. Wong, T.S. Lau, T.S. Yu, et al., Air pollution and hospital admissions for respiratory and cardiovascular diseases in Hong Kong, Occup. Environ. Med. 56 (1999) 679–683.
[26] J. Fan, W. Zhang, others, Statistical estimation in varying coefficient models, Ann. Stat. 27 (1999) 1491–1518.
[27] Y. Hirata, K. Aihara, Predicting ramps by integrating different sorts of information, Eur. Phys. J. Spec. Top. 225 (2016) 513–525.
[28] K. Zhang, W. Fan, Forecasting skewed biased stochastic ozone days: analyses, solutions and beyond, Knowl. Inf. Syst. 14 (2008) 299–326.
[29] E. Beninca, J. Huisman, R. Heerkloss, et al., Chaos in a long-term experiment with a plankton community 2008;451:5.
[30] E. Benincà, K.D. Jöhnk, R. Heerkloss, et al., Coupled predator-prey oscillations in a chaotic food web: Coupled predator-prey oscillations, Ecol. Lett. 12 (2009) 1367–1378.
[31] O. Wahltinez, others. COVID-19 Open-Data: curating a fine-grained, global-scale data repository for SARS-CoV-2 2020.
[32] Y. Li, R. Yu, C. Shahabi, et al., Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. arXiv:170701926 [Cs, Stat] 2018.
[33] E.R. Deyle, G. Sugihara, Generalized theorems for nonlinear state space reconstruction, PLoS One 6 (2011).
[34] L.A. Thombs, W.R. Schucany, Bootstrap prediction intervals for autoregression, J. Am. Statist. Assoc. 85 (1990) 486–492.
[35] R.J. Howlett, L.C. Jain, Radial basis function networks 2: new advances in design, Physica 67 (2013).
[36] H. Ye, G. Sugihara, Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality, Science 353 (2016) 922–925.
[37] L. Wang, Support vector machines: theory and applications, 177, Springer Science & Business Media, 2005.
[38] H. Tong, M.K. Ng, Calibration of ϵ- insensitive loss in support vector machines regression, J. Franklin Inst. 356 (2019) 2111–2129.
[39] T. Sofianos, A. Sampieri, L. Franco, et al., Space-Time-Separable Graph Convolutional Network for Pose Forecasting, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, IEEE, 2021, pp. 11189–11198.

[40] Y. Xia, W. Härdle, Semi-parametric estimation of partially linear single-index models, J. Multivariate Anal. 97 (2006) 1162–1184.
[41] R.B. Devlin, K.E. Duncan, M. Jardim, et al., Controlled exposure of healthy young volunteers to ozone causes cardiovascular effects, Circulation 126 (2012) 104–111.
[42] B.-J. Lee, B. Kim, K. Lee, Air pollution exposure and cardiovascular disease, Toxicol. Res. 30 (2014) 71–75.
[43] K. Luo, R. Li, W. Li, et al., Acute effects of nitrogen dioxide on cardiovascular mortality in Beijing: an exploration of spatial heterogeneity and the district-specific predictors, Sci. Rep. 6 (2016) 1–13.
[44] R. Heerkloss, G. Klinkenberg, A long-term series of a planktonic foodweb: a case of chaotic dynamics. Internationale Vereinigung Für Theoretische Und Angewandte Limnologie, Verhandlungen 26 (1998) 1952–1956.
[45] J. Wang, W. Lv, Y. Jiang, et al., A multi-agent based cellular automata model for intersection traffic control simulation, Physica A 584 (2021) 126356.
[46] R. Liu, J. Zhong, R. Hong, et al., Predicting local COVID-19 outbreaks and infectious disease epidemics based on landscape network entropy, Sci. Bull. 66 (2021) 2265–2270.
[47] Z. Lin, J. Feng, Z. Lu, et al., DeepSTN+: Context-Aware Spatial-Temporal Neural Network for Crowd Flow Prediction in Metropolis, AAAI 33 (2019) 1020–1027.
[48] M. Li, Z. Zhu, Spatial-Temporal Fusion Graph Neural Networks for Traffic Flow Forecasting, AAAI 35 (2021) 4189–4196.
[49] R.-G. Cirstea, B. Yang, C. Guo, et al., Towards Spatio- Temporal Aware Traffic Time Series Forecasting, in: 2022 IEEE 38th International Conference on Data Engineering (ICDE), Kuala Lumpur, Malaysia, IEEE, 2022, pp. 2900–2913.
[50] X. Wang, Y. Ma, Y. Wang, et al., Traffic Flow Prediction via Spatial Temporal Graph Neural Network, in: Proceedings of The Web Conference 2020, Taipei Taiwan, ACM, 2020, pp. 1082–1092.
[51] G. Spadon, S. Hong, B. Brandoli, et al., Pay Attention to Evolution: Time Series Forecasting With Deep Graph-Evolution Learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 5368–5384.

Hao Peng received the B.Eng. degree in computer science and technology from South China University of Technology, Guangzhou, China, in 2018. He is currently pursuing a Ph.D. degree in applied mathematics with the School of Mathematics, South China University of Technology, Guangzhou, China. His current research interests include deep neural networks and data mining for complex dynamic systems.

Pei Chen received her B.S. and M.S. degrees from Peking University, in 2007 and in 2014. She received a Ph.D. degree from the School of Computer Science and Engineering, South University of Technology, Guangzhou, China, in 2017. Currently, she is an assistant researcher at the School of Mathematics, South China University of Technology, Guangzhou, China. Her research interest includes deep learning, data mining, and computational biology.

Rui Liu received his B.S. and Ph.D. degrees in applied mathematics from Peking University, in 2005 and 2010. He currently is a full professor at the School of Mathematics, South China University of Technology, Guangzhou, China. His research interests include nonlinear dynamics, modelling, and computational methods.

Luonan Chen received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 1984, and the M.S. and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1988 and 1991, respectively. Since 1997, he has been an associate professor with Osaka Sangyo University, Osaka, Japan, where he became a full professor. Since 2009, he has been a professor and the executive director with the Key Laboratory of Systems Biology, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China. In recent years, he published over 300 journal papers and three monographs in the area of systems biology. His current research interests include systems biology, computational biology, and nonlinear dynamics.