Tseng *et al. Geoscience Letters* (2023) 10:42
https://doi.org/10.1186/s40562-023-00295-6

Geoscience Letters

RESEARCH LETTER

Open Access

Sea surface temperature clustering and prediction in the Pacific Ocean based on isometric feature mapping analysis

John Chien-Han Tseng[1*], Bo-An Tsai[2] and Kaoshen Chung[2]

Abstract

Isometric feature mapping (ISOMAP) is a nonlinear dimensionality reduction method and closely reflects the actual nonlinear distance by the view of tracing along the local linearity in the original nonlinear structure. Thus, the first leading 20 principal components (PCs) of low-dimensional space can reveal the characteristics of real structures and be utilized for clustering. In this study, a *k*-means algorithm was used to diagnose SST clustering based on ISOMAP. Warm and cold El Niño–Southern Oscillation events were subdivided into Central Pacific and Eastern Pacific types, and a two-dimensional cluster map was used to depict the relationship. The leading low-dimensional PCs of ISOMAP were considered as the orthogonal basis, and their trajectories demonstrated meaningful patterns that could be learned by machine learning algorithms. Predictions of SST in the Pacific Ocean were performed using support vector regression (SVR) and feedforward neural network (NN) models based on the low-dimensional PCs of ISOMAP. The forecast skills, the root-mean-square error (RMSE) and anomaly correlation coefficient (ACC), were comparable to those of current numerical models.

Keywords Anomaly correlation coefficient, ISOMAP, *k*-means, Neural network, Principal component analysis, Principal component, Root mean square error, Sea surface temperature, Support vector regression

Introduction

El Niño–Southern Oscillation (ENSO) events are typically divided into two classes: El Niño (warm) and La Niña (cold) events. Regardless of the event class, each event has at least two types of sea surface temperature (SST) anomalies: Central Pacific (CP) anomalies and Eastern Pacific (EP) anomalies (Ashok et al. 2007; Kao and Yu 2009; Kug et al. 2009). This subdivision depends on the geographical locations and evolution of strong anomalies. Typical differences or even slight differences in SST patterns result in various atmospheric circulations and cause chain reactions (James 1994; Trenberth and Hurrell 1994; Trenberth et al. 1998; Alexander et al. 2002), such as subtropical highs with different locations and strengths and tropical cyclones with different locations and trajectories (Tu′uholoaki et al. 2023). Clustering of ENSO SST events enables differentiating certain time events and determining any differences between such events. Therefore, SST clustering should be performed in accordance with not only geographical positions but also quantifiable and objective methods.

EP and CP types of ENSO dynamics or their frequencies could be attributed to global warming (Capotondi et al. 2015). Vecchi and Wittenberg (2010) present numerical climate model findings indicating that global warming leads to weakened Walker's circulations and a decrease in the zonal slope of the thermocline around the equator, potentially resulting in a higher occurrence of CP events (McPhaden 2012). However, Yeh et al. (2012) argue that the thermocline inclination or the number of EP or CP events may be driven by natural variability. The challenge lies in differentiating the

*Correspondence:
John Chien-Han Tseng
jchtsenghome@gmail.com
[1] Central Weather Bureau, 64 GongYuan Road, 100, Taipei, Taiwan
[2] National Central University, Taoyuan, Taiwan

transition processes between CP and EP events, as mixed characteristics are often observed (Hernández et al. 2020), where certain EP events exhibit similarities to CP events for a period before diverging, and vice versa. In essence, no two ENSO events, whether CP or EP, are alike (McPhaden et al. 2006). One potential solution to this issue is the application of clustering techniques, which can help assess similarity and identify transitional processes.

The success of clustering depends on the distribution of data points. For example, data points that are farther apart are less likely to look similar. Isometric feature mapping (ISOMAP) is a separated data point and nonlinear dimensionality reduction method that accurately reflects the actual distance through tracing along the local linearity in the original nonlinear structure (Tenebaum et al. 2000; Balasubramanian et al. 2002). During the last step of ISOMAP, temporal principal components (PCs) and spatial empirical orthogonal functions (EOFs) are generated as the principal component analysis (PCA). Close ISOMAP PC data points have similar patterns, and leading ISOMAP PCs provide proper distributed low-dimensional data points for clustering (Bayá et al. 2008; Tasoulis et al. 2020).

Multiple regression analysis relies on leading PCs to reduce the number of predictor variables in order to achieve improved predictions and more efficient computations (Stock and Watson 2002; Sousa et al. 2007). Sousa et al. (2007) used a feedforward artificial neural network (NN) with PCs to replace a multiple nonlinear regression in order to achieve improved predictions. Brunton et al. (2020) supported this idea and proposed the calculation of computational fluid dynamics (CFD) problems in low-dimensional data points.

Traditional CFD numerical models are based on differential equations with sufficiently small spatial and temporal grids. Fine grid meshes are typically associated with additional data points, a sufficiently small time-integration step to ensure computational stability, and a high computational cost. Data-driven methods, that is, methods in which data are driven by statistics and not by physical laws (e.g., PCA, regression, and NNs), are used to address the limitations of this time-integration step. However, these methods cannot be solely used to solve CFD problems. Instead, the future trend of solving CFD problems is to combine both numerical and statistical methods. The term "data-driven" should have a generalized meaning pertaining to the transplantation of data variations. Brunton and Kutz (2022) used several machine learning algorithms, such as feedforward NNs, autoencoders, and deep learning, to solve CFD problems. However, when data-driven methods are used in the original high-dimensional data space, computational cost represents a substantial burden. These high-dimensional space data points should be counted in terms of low-dimensional data (Brunton et al. 2020; Brunton and Kutz 2022). In other words, if statistical data-driven models do not extract meaningful reduced dimensional data points or reduced features, they will not be able to offer improved forecasts (Wilks 2019).

Machine learning or artificial intelligence (AI) approaches, such as data-driven models or NN have become increasingly popular in the field of numerical weather prediction (NWP) models. These approaches have shown in enhancing various aspects of NWP models including the utilization of observation data, data assimilation, CFD numerical schemes, and post-processing (Bonavita et al. 2021; Dueben et al. 2021). The European Centre for Medium-Range Weather Forecasts (EMMWF) has put forth a 10-year development roadmap that emphasizes the integration of machine learning and NWP models. The developments include using data-driven models or AI models to help physical parameterization calculations (e.g. gravity wave drag, cloud, radiation, etc.), the tangent/adjoint models for data assimilation or ensemble prediction, and downscaling post-processing. Data-driven models offer the ability to learn and simulate the performance of past numerical models without being constrained by complex physical theorems or CFD equations. These approaches reduce the development and maintenance costs associated with creating tangent linear/adjoint models. Researchers no longer need to derive equations or model schemes that precisely correspond to the current nonlinear models. Instead, the calculations are performed through the composition of NN functions, eliminating the need for linearized and transposed operators of the original complex nonlinear models (Hatfield et al. 2021).

In this study, we clustered and predicted Pacific SST by using leading ISOMAP PCs. In terms of clustering, ISOMAP provided well-distributed SST leading PCs. In terms of prediction, the leading SST ISOMAP PC trajectories showed similarities with the Lorenz 63 model variable trajectories. Brunton and Kutz (2022) used a feedforward NN to study Lorenz 63 model variable trajectories. They successfully simulated similar trajectories to those of numerical models. In this study, we hypothesized that the use of appropriate learning algorithms in SST PC prediction and the combination of these PCs with time-fixed spatial SST EOFs would result in improved future predictions. We also hypothesized that spatial SST EOFs would not change in all our prediction experiments and that the learning algorithms would focus on temporal leading SST PC variations. The reason for fixed EOFs is that the leading EOFs are nearly identical especially when calculating to remove the seasonal cycle or monthly climatological mean (Elken et al. 2019).

The forecasting of SST using nonlinear regression and neural networks has been explored in previous studies by Tangang et al. (1997, 1998a, 1998b), Tang et al. (2000), and Wu et al. (2006). These studies utilized leading SST PCs as

predictors in their models, along with sea level pressure and prior SST anomalies. They achieved favorable results in terms of correlation skills, particularly in predicting SST anomalies in the Pacific equator area, by considering the leading 5 PCs. In this work, we introduced some advancements to the existing approach. Instead of using traditional PCs, we adopted ISOMAP PCs, which offered improved distribution for resolving different ENSO events. Furthermore, we extended the hidden layers of the neural network, experiments with different neuron numbers, and incorporated 20 ISOMAP PCs as multi-dimensional space trajectories in our model. These modifications aimed to improve the forecasting accuracy of our NN model.

The remainder of this paper is organized as follows. Section "Data and methods" explains the clustering and learning algorithms used for SST prediction. Section "Clustering" outlines the cluster distributions and actual SST patterns of cluster centroids. It also provides clues for determining the similarity between an SST figure and a specific type of cluster instead of using CP or EP as a qualitative description. Section "Prediction" explains the prediction of SST through support vector regression (SVR) and NNs. Finally, Section "Conclusions and discussion" summarizes the main research findings and potential future improvements in data-driven SST prediction.

Data and methods

SST data were obtained from the fifth version of the Extended Reconstructed Sea Surface Temperature data set of the National Oceanic and Atmospheric Administration (NOAA), National Climatic Data Center, based on data from the Comprehensive Ocean–Atmosphere Data Set, collected from January 1980 to December 2022. El Niño, normal, and La Niña events were differentiated using the Niño 3.4 (170°W–120°W, 5°S–5°N) index of the NOAA Climate Prediction Center. El Niño events were defined in the Niño 3.4 region when the moving 3-month average SST anomaly exceeded 0.5 °C for at least 5 months. In contrast, La Niña or so called anti-El Niño events were defined when the average SST anomaly fell below 0.5 °C for 5 months. Pacific Ocean domain (120°E–60°W, 30°S–30°N) SST was used for clustering and learning algorithm prediction.

In accordance with Tseng (2022), SST ISOMAP was calculated using the distance matrix formed by the covariance of SST anomalies with a nearest neighbor number of 44. The results indicated that the 20 leading SST ISOMAP PCs explained approximately 90% of the variance in the reconstructed geodesic distance data matrix (cf. Figure 7 in Tseng 2022). These 20 reduced low-dimensional components were sufficient for clustering and data-driven model prediction. In addition, the computational cost of clustering and prediction was low, and computational complexity was lower than that in the original high-dimensional space. About ISOMAP and the traditional PCA, we highlighted in the Appendix.

ISOMAP measures the geodesic distance to differentiate the structure of data PC points, so the clustering based on the distance k-means was used in this article. The algorithm k-means for clustering that minimized the total reconstruction function by the proper cluster number. The reconstruction cost function (Theodoridis et al. 2010) was defined as

$$J(\theta, u) = \sum_{i=1}^{N} \sum_{j=1}^{M} u_{ij} \left\| x_i - \theta_j \right\|^2 , \tag{1}$$

where $\theta = (\theta_1^T, ..., \theta_m^T)^T$ represents different cluster center vectors, $\| \cdot \|$ is the Euclidean distance, x_i represents the leading PC vectors, and

$$u_{ij} = \begin{cases} 1, & x_i \text{ is cosest to } \theta_j \\ 0, & \text{otherwise} \end{cases} , \tag{2}$$

where N is the number of leading PC vectors (data points) and M is the number of clusters that should be initially provided. After obtaining a satisfactory distribution of points for the SST ISOMAP PCs, the leading 20 PCs were selected, and k-means clustering was applied using the Euclidean distance calculation.

Three of the leading SST ISOMAP PC temporal trajectories demonstrated spiral circles with similar but not completely identical regular swinging behaviors and with different durations in each spiral cycle (Tseng 2022). These findings prompted a comparison with the 3D trajectories of the Lorenz 63 model, which Brunton and Kutz (2022) employed neural network algorithms to learn and replicate the model's numerical predictions. This comparison sparked the idea of using the leading three PCs, or even more components, as variables for prediction. The approach involved training learning algorithms to predict the PC points at the next time step, multiplying these points by the assumed time-invariant spatial EOFs, and generating predictions for SST. SVR and NNs (Algorithms 1 and 2) were used as learning algorithms for trajectory predictions. The 20 leading ISOMAP PCs were then selected for SVR and NN training, as in the clustering process. After the leading number exceeded 20, the latest 10 year corresponding to the prediction period in El Niño or La Niña event average PCs were used to construct residual PCs (number more than 20). For example, if one wanted to predict SST from May 2019 to Apr. 2020, El Niño to normal year, then the residual PCs would use the time period from Mar. 2016 to Feb. 2017, El Niño to normal year and to La Niña. These residual PCs were selected with spatial EOFs to obtain fine spatial prediction structures. They occupied few variances and did not influence the prediction results. Both SVR and

NN predicted and the residual PCs were multiplied by the climatological spatial EOFs (from January 1980 to December 2018) and used to achieve SST predictions related to actual physical space. The idea of using EOF, leading PCs and residual PCs are shown in Fig. 1. When we calculated ISOMAP EOFs and PCs, the SST climatological mean was removed. We used learning algorithms to predict the leading 20 PCs and filled them with selected residual PCs, then times the climatological spatial EOFs to reconstruct SST anomalies (SSTA). Finally, we could obtain SST predictions by adding the climatological mean SST.

During SVR training, the SST ISOMAP PCs were regarded as independent components and trained individually. The SVR provided the prediction PC values. During the NN training process, the network resembled the feed-forward network algorithm of Brunton and Kutz (2022) and exhibited three hidden layers with ten neurons. One-step time lags in the input PC data were used as the output PC data in the NN. However, given the trade-off between computational efficiency and final prediction accuracy, no particular reason was given for selecting such numbers of hidden layers and neurons. The hidden layers 3–20 and the number of neurons 10–200 had been tested and there were no significant forecast skills improvement.

In NN prediction experiment, we arrange SST data into the training data set and the testing data set as the traditional machine learning did. At the same time, we made the training data set have its own training, validation, and testing data parts. This arrangement was for avoiding overfitting and execution efficiency. First, we kept Jan. 2019–Dec. 2021 period to test NN model. This was the testing data set and was never used in training process. Then, during the NN training process, the period of time from Jan. 1980 to Dec. 2018 was the training data set. There were approximately 60%, 20%, and 20% of the data were randomly chosen for training, validation, and testing, respectively, in training NN model. When the NN model was trained, 60% training data were used to generate the model, 20% training data were used for validation and for tuning the model, the rest 20% training data were used for testing without joining to tune the model. Adding the validation and testing data in training process could effectively generate the model and improve the overfitting. Considering only the efficiency and accuracy of generating NN models, there was still no particular reason to use the 6:2:2 ratio. Consequently, the NN prediction results were not identical in different NN training processed, because the different data (including training, validation, and testing) were used to produce the NN model. To obtain more robust NN results, the NN model training and final prediction test were counted 20 times, and

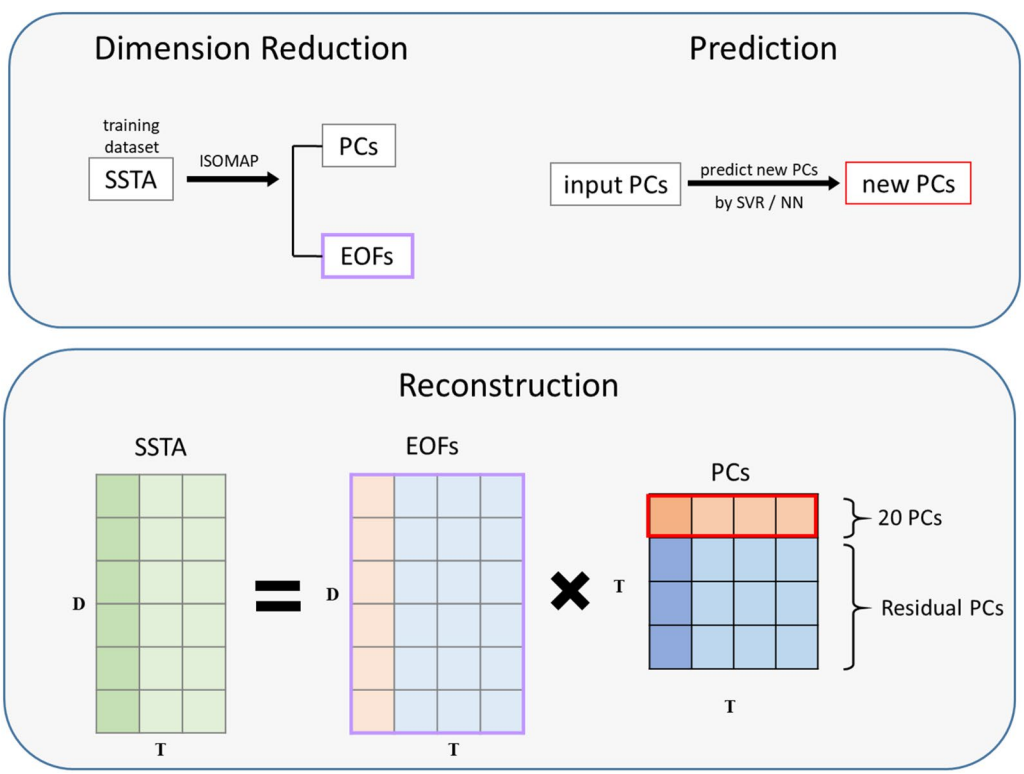

Fig. 1 The idea of using PCs to predict. The leading PCs are used for training algorithms and predicting to next time steps. With residual PCs times climatological EOFs, the predictive data on physical space SST anomalies (SSTA) can be reconstructed

the prediction average was used as the final prediction value. In contrast to NN, the SVR separated the data into training and testing two parts. All training data were used for validation to generate the SVR model. The forecast skills for the final predictions of SVR and NNs were the root-mean-square error (RMSE) and anomaly correlation coefficient (ACC), whose equations and calculation methods are outlined in detail in the appendix.

***Algorithm* 1**: SST prediction by SVR

1. Select the 20 leading ISOMAP PCs and **train** each time evolution PC individually by using SVR
2. Use those different SVRs to predict the testing time leading 20 PCs
3. Combine the 20 predicted leading PCs and previous period residual PCs to obtain a complete PC set.
4. Multiply the complete PC set by the climatological EOFs to recompose the SST prediction pattern.

***Algorithm* 2**: SST prediction by NN

1. Select the leading 20 ISOMAP PCs. Split the PC data into two periods for training and testing
2. Define the input PC time stream as x (1, 2, …, $N-1$) and the output PC time stream as x (2, 3, …, N). Further divide the training data into training (60%), validation (20%), and testing (20%) to generate the **NN model**. This NN model can predict one-step time series prediction
3. Given the starting time $x(T)$ in the testing data

   For $i = T, T+m-1$

   $x(i+1)$ = **NN model**($x(i)$)

   end

   NN model predicts the x from $T+1$ to $T+m$
4. Repeat Step 2-3 20 times to get 20 different NN models. Take the **average forecast values** from the **20** NN predictions.
5. Combine the 20 predicted leading PCs and previous period residual PCs to obtain a complete PC set.
6. Multiply the complete PC set by the climatological EOFs to recompose the SST prediction pattern.

Clustering

Figure 2 depicts the relationship between the cost function defined in Eq. (1) and the number of clusters. No elbow point minimizing the cost function and representing the optimal number of clusters was identified. Therefore, seven clusters were subjectively selected. As shown in Fig. 2, cluster numbers exceeding seven resulted in small values of the cost function. These clusters (greater than seven or more) did not highlight clear differences in the SST distributions.

Figure 3 shows the centroids of seven clusters and the distribution of these clusters. The seven events close to these seven centroids were as follows: two El Niño events (October 1987 and December 2009), two La Niña events (February 2006 and October 2011), and three normal events (September 1984, June 1994, and June 2014). Analysis of these cluster centroids revealed clear El Niño/La Niña CP and EP events. Even normal events were easily divided into three clusters depending on the SST anomalies (warm or cold) in the western coast of South America. Notably, the event observed in September 1984 was a normal SST event, whereas the event observed in October 1984 was a La Niña event. The centroid close to September 1984 lay on the dividing line between normal and La Niña events. In addition, the SST anomaly pattern of cold water in the EP in September 1984 was similar to a La Niña phenomenon. The warm or cold SST anomalies belonged to the CP or EP type could be revealed clearly in this cluster map. Strong ENSO and EP events tended to appear on both sides of the iso 1 axis (x axis), with the first PC approaching their minima or maxima. The first PC had the strong east–west variation pattern around the equator (cf. Figure 6 in Tseng 2022). It corresponded the strong EP events with larger the first PC values. In contrast, CP and weak events were observed in locations with smaller the first leading PC around the value 0 in the iso 1 axis direction. These results are consistent with those of previous studies investigating warm and cold events (Kao and Yu 2009; Yu et al. 2011).

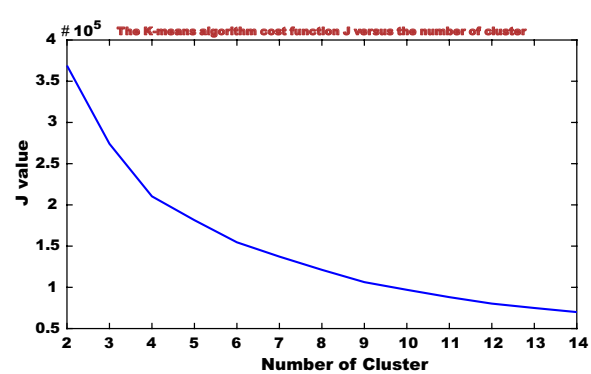

Fig. 2 Clustering cost function and number of clusters in the *k*-means algorithm. A relatively small value of the cost function indicates an appropriate number of clusters

Tracing a single El Niño and La Niña case revealed the unique trajectory or preferable location of this specific case. The EP case did not abruptly skip to a CP pattern. Instead, it evolved over time, as in the case of 1997/98 EP El Niño evolving between 2002/03 CP El Niño and 1998/2000 EP La Niña. The clustering positions and trajectories provided additional clues to analyze the El Niño or La Niña processes instead of analyzing the entire SST anomalies or variations of the Niño 3.4 index. When compositing the similar event cases, this clustering provides the guidance. For the future implementation, the numerical SST model results or their ensemble results can be projected on this clustering map. The model results can be used to examine the differences with these real historical events.

Although our clusters could exhibit CP/EP ENSO events through locations different from the cluster centroids, or provide guidance for composite similar CP/EP cases. These ISOMAP PC clusters marked or anchored centroids or historical points to clearly visualize and examine current or future SST event trajectories. We thought the process of any one event was also important and the dynamics should be restored. When examining the recent La Niña event from August 2020 to February 2023 (Fig. 6a), we could find an evolutionary ISOMAP PC trajectory close to the CP La Niña centroid in February 2002 but ultimately quite different from the 2002 case. We also noticed that this latest La Niña event evolved slowly and the beneath dynamics worth studying in the future. We thought studying this latest case by composite analysis must be very careful.

Prediction

In this section, we compared the predictions of SVR and NN for SST. The testing period was concentrated between 2019 and 2022, preceded by the training period. This was to ensure that the training data set did not contain the testing data used for prediction, and that the learning algorithms had not acquired known answers before making the prediction. To validate the SVR and NN forecasts, 36-month consecutive time periods were selected from Jan. 2019 to Dec. 2021, and their RMSE and ACC values were averaged. Because the SVR could not get good prediction results by using the training data far from the initial forecast month. For example, to predict the Jan. 2020 SST in SVR, the last of training data set time was Dec. 2019. But for NN, the training data set time limit was from Jan. 1980 to Dec. 2018. This arrangement was because the NN prediction results much better than the SVR. Moreover, we found that the last training

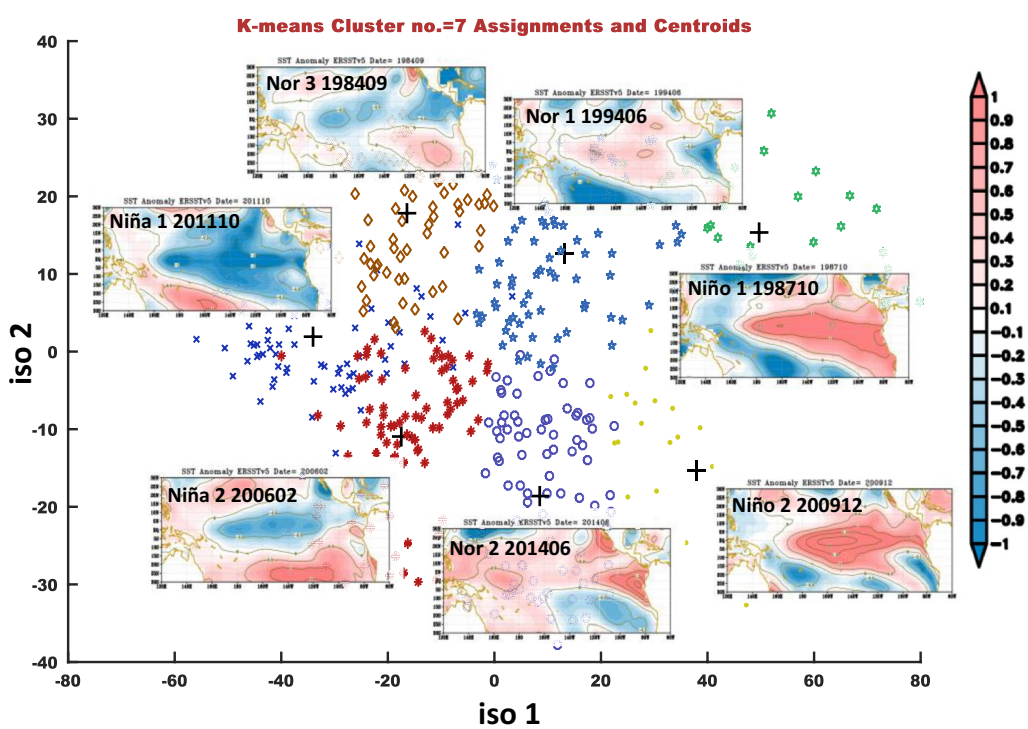

Fig. 3 Two-dimensional projection of seven clusters (different color points) and their centroids (+ sign) with *k*-means clustering based on the 20 leading ISOMAP PCs. The figure depicts SST anomalies close to the centroids

data time in NN prediction did not need to be 1 month before the start prediction month as in SVR.

For the SVR, orthogonal PCs capable of being separated during training were used. Figure 4a depicts the three leading ISOMAP PCs versus time variations for both training and testing scenarios. The blue curves represent the training sets, the green curves represent the testing sets, and the red curves represent the SVR-predicted values. The same training and forecasting procedures were used in the 20 leading PCs. The other 17 ISOMAP PCs training and testing process were not shown in here. As shown in Fig. 3a, the training period spanned from January 1980 to December 2021. In this case, the prediction period spanned from January 2022 to October 2022. Notice that the second PC SVR testing accuracy was worse than the first or the third PCs. Due to the second ISOMAP PC variation differed from the other two PCs, the second ISOMAP PC values exhibited a monotonically decreasing trend over the period of 41 years. SVR always tried to get the predictive values in the final training time and the testing stage back to higher than what actually happened. The second PC monotonically variation was similar to the second PC in Li et al. (2019) monotonically rising findings. Other random time periods exhibited comparable acceptable training results but poor testing results (not shown).

To achieve future predictions, a feedforward NN was used to train the 20 leading PCs within the training period from January 1980 to December 2018. Figure 4b depicts the NN prediction trajectory (blue) and actual observation trajectory (red) for the first three ISOMAP PCs from March 2021 to December 2021. For convenience, only the trajectories composed of the first three PCs are shown in Fig. 4b. Once the prediction trajectory perfectly matched the observation trajectory, the forecast was regarded as the optimal solution. Although we were unable to match the NN trajectory with the observed trajectory, the forecast skills were acceptable and still outperformed the SVR forecasts (Fig. 5). We also noticed that prediction performance should be examined by forecast skills rather than the leading PC trajectories. However, we did not illustrate the NN ISOMAP PCs individually as we did with the SVR, because the NN PCs exhibited similar variations during the training period but unsatisfactory testing results (forecasts). Furthermore, the SVR PC 3D prediction trajectories were not illustrated, because they were inferior to those of the NN.

Figure 5a, b depicts 36-case average forecast RMSE and ACC values. The forecast initial time is from Jan. 2019 to Dec. 2021, with total 36 cases and 48-month predictions. The blue curves belong to the SVR results, and the red curves belong to the NN results. Since SVR forecasts underperform beyond 10 months, 10 is chosen here as

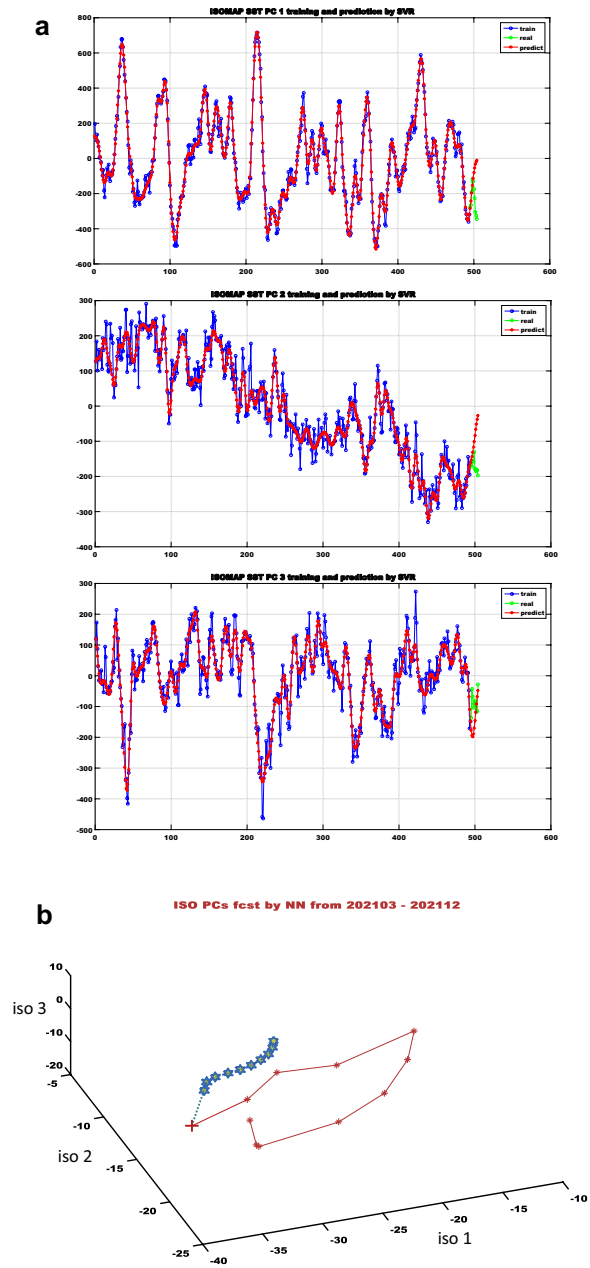

Fig. 4 **a** Three leading PCs from the training data (blue curves), testing data (green curves), and SVR prediction (red curves). The training period lasted from January 1980 to December 2021, and the prediction period lasted from January 2022 to October 2022. **b** Three leading PC trajectories obtained from NNs (blue curve) and true observations (red curve) between March 2021 and December 2021

the number of lead months. The forecasting ability of the NN was superior to that of the SVR, with an RMSE value between 0.3 and 0.4 and an ACC value between 0.68 and 0.8. Some test cases maintained an ACC value of 0.7 for approximately 18–24 months. On the other hand,

for Niño 3.4 area predictions (Fig. 5 red dashed curves), these scores were similar to or even slightly higher than those of SST forecasts in ECMWF seasonal model (Molteni et al. 2011), hybrid coordinate ocean model (Thoppil et al. 2021), and National Centers for Environment Prediction/Environmental Model Center Global Ensemble Forecast System (personal communication with Yuejian Zhu).

An intriguing case occurred between May 2019 and December 2022. SST anomalies revealed El Niño events in the first 2 months, followed by normal events, La Niña events, brief normal events for approximately 2 months, and then La Niña events once again. Figure 6a depicts this evolution of 3D ISOMAP PCs. In Fig. 6c, this actual observation trajectory is represented by a red curve, and

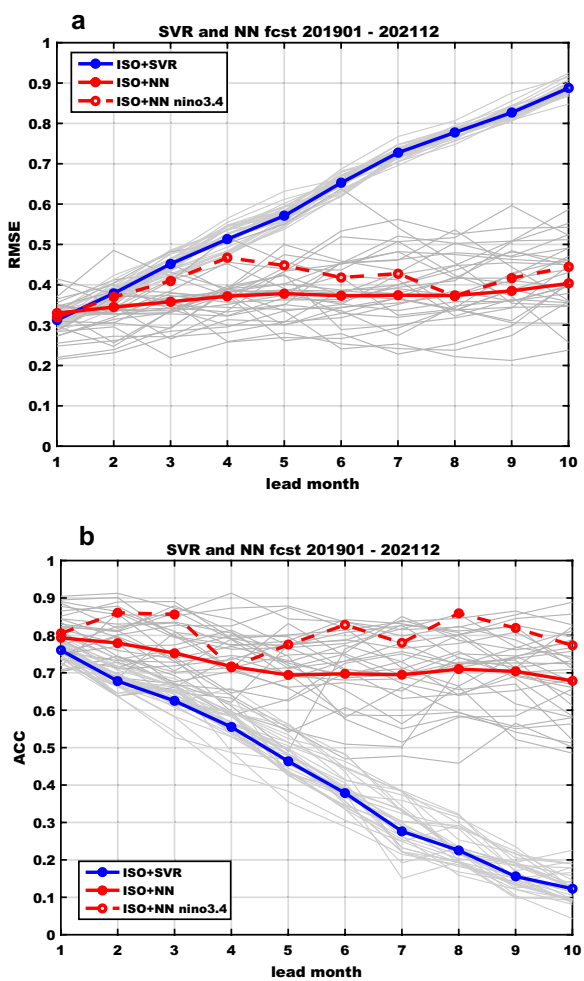

Fig. 5 Forecast initial time from Jan. 2019 to Dec. 2021, 36-case prediction average values of RMSE and ACC versus lead months. The blue curves belong to the SVR, the red curves belong to the NN, and the red dashed curves focus on Niño 3.4 area: **a** RMSE, **b** ACC. The gray curves highlight all 36-case prediction values

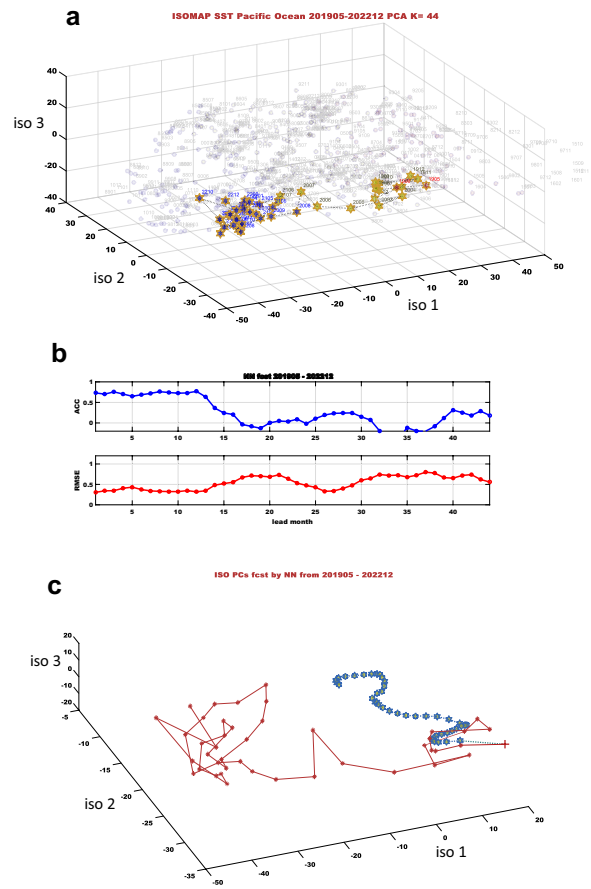

Fig. 6 **a** Three leading SST ISOMAP PC trajectories between May 2019 and December 2022. SST anomalies reveal El Niño events, normal events, La Niña events, normal events again, and then La Niña events. El Niño, normal, and La Niña event are marked by red, yellow, and blue points, respectively. **b** RMSE and ACC values from the NN during this period. **c** NN prediction trajectory (blue curve) and true observation (red curve) during this period

the NN forecast trajectory is represented by a blue curve. Initially, the NN was unable to accurately predict this period but the trend was acceptable. As shown in Fig. 6b, the initial 12-month ACC values remained between 0.6 and 0.7, and the RMSE values remained between 0.4 and 0.7. For the final outputs, using predictive PCs, residual PCs, spatial EOFs, and climatological monthly mean SST, physical space SST predictions could be reconstructed. Figure 7a shows at lead times of the 7–9 months SST predictions, and the validation period is from October to December 2021, the same period as the PC trajectories in Fig. 4b. The ACC values in this period time were 0.6–0.7 and the RMSE values were 0.35–0.4. The right column of the Fig. 7b shows the real observation from ERSSTv5 data.

To summarize, based on Fig. 6, although the ensemble mean forecast trajectory (Fig. 6c blue curve) primarily

followed the second PC variation and not the first PC variation as in the observation period (Fig. 6c red curve), the NN forecasts outperformed the SVR forecasts. At least, the initial 12-month predictions were acceptable. During this period, half of the prediction trajectory samples turned right along the first PC axis positive direction, whereas the other half turned left along the first PC axis negative direction. If the selected ensemble forecasts had large differences in this positive–negative mode, the ensemble average would remove the positive–negative first PC variation and remain the trajectory variation contributed by other PCs that did not differ significantly among the ensemble members. Averaging the results served as a reminder that the data composite analysis would have yielded similar results that would have eliminated the main variation.

Conclusions and discussion

In this study, low-dimensional SST PC points obtained using ISOMAP were used to differentiate clusters and predict variations in time trajectories. By highlighting SST anomaly patterns, these clusters were able to determine the relationships between various ENSO and normal events. With three clusters, the corresponding centroids easily identified El Niño, normal, and La Niña events. The new ENSO index was also used to measure the entire Pacific basin SST anomaly, as opposed to merely the small-area Niño 3.4 index. Overall, the evolution of centroids over time served as an intriguing aspect of ENSO evolution.

By inverting the forecasts with climatological spatial EOFs, we studied the time variations of low-dimensional PC points and used them to predict future SST through SVR and NNs. This method was computationally more efficient than the original high-dimensionality numerical

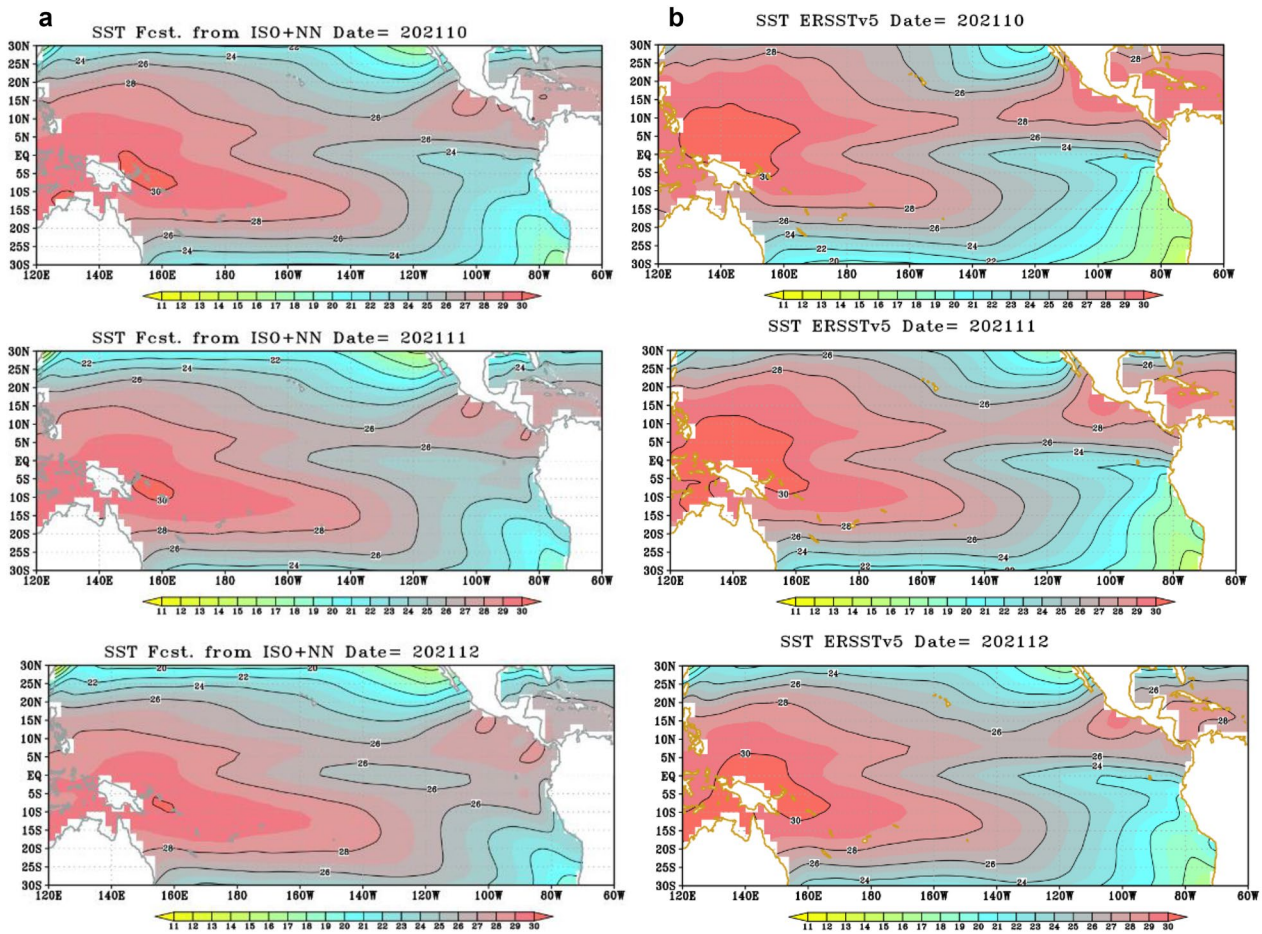

Fig. 7 SST prediction by NN (left column a) and real SST observation from ERSSTv5 (right column b). The prediction initial time is Mar. 2021 in Fig. 3b. The period from Oct. 2021 to Dec. 2021 is the forecast lead month number 7–9. The ACC value is around 0.6–0.7

model, and it produced acceptably accurate long-term forecasts. Although the slow change of the SST model was a possible reason for the forecast accuracy, the forecast based on low-dimensional data points was still one of the potential forecast models to obtain the correct SST in a short time.

Currently, our data-driven models are being developed to predict other atmospheric variables, such as wind vector field data. Because the wind vector fields usually have discontinuity in time variation that are difficult for learning algorithms. Other dimensionality reduction techniques, such as self-organizing maps, which successfully cluster zonal wind in boreal summer (Rousi et al. 2022), probably provide the low-dimensional wind information to do NN forecasts. Moreover, we are employing and evaluating autoencoders and diffusion maps that were used to replace ISOMAP for the prediction of atmosphere–ocean fluids. We also extend this data-driven ISOMAP and NN methods for learning historical and ensemble member predictions in low-dimensional space, which can be used for improving ensemble forecasts.

Despite the need for additional tuning of SVR and NN parameters, such as slack variables and the number of neurons, current models have been worked well. However, the optimal number of hidden layers in NNs remains unclear, and other function compositions may yield superior results. Moreover, because of the limited SST sample size, simple models were used to avoid overfitting. Therefore, future research should investigate low-dimensional changes predicted by numerical weather models as a potential solution.

Appendix

PCA and ISOMAP

Tenebaum et al. (2000) proposed isometric feature mapping (ISOMAP) to solve the classification problem and obtain well-distributed low-dimensional data points. They pointed out that the traditional PCA considers the data under linear framework. For example, traditional PCA taking the data points' time evolution is resolved by the linear view, to measure the data by the Euclidean distance, the line segment. In contrast, the ISOMAP measures the distance between two data points based on the geodesic distance, which more closely reflects the actual distance by the view of tracing along the local linearity in the original nonlinear structure. The geodesic distance is calculated on the framework of the nearest neighbor graph. In brief, ISOMAP constructs the weighted nearest neighbor distance graph first and then solves this weighted distance by traditional PCA (Tenebaum et al. 2000; Tseng 2022). Basically, the key point is to build the reasonable distance matrix. The distance matrix can be regarded as another kind of covariance matrix. If one considers all the data points to be the neighbors, ISOMAP would be degenerated to traditional PCA. Moreover, Izenman (2008) points out when the isotropic kernel function is used in distance matrix, the kernel PCA will be identical to ISOMAP.

Both PCA and ISOMAP adopt the same eigen function solving processes. Now, we take PCA as the example. Given the data matrix.

$\mathbf{Y}_{mn}$,

where m is the spatial dimensions, and n is the temporal dimensions.

Assume that we choose $\mathbf{Y}_{mn}^T \mathbf{Y}_{mn}$ to solve the PCA, then we can get.

$\mathbf{Y}_{mn}^T \mathbf{Y}_{mn}\ \mathbf{P}_n = \mathbf{P}_n\ \Sigma_{nn}$,

where $\mathbf{P}$ is eigen vector matrix and Σ is eigen value matrix. Since $\mathbf{Y}^T\mathbf{Y}$ is symmetric matrix, so $\mathbf{P}^T\mathbf{P} = \mathbf{I}$.

It is easy to get

$$\mathbf{P}^T\mathbf{Y}^T\mathbf{Y}\mathbf{P} = \Sigma$$

and if we assume one matrix $\mathbf{z}$ the relation with $\mathbf{Y}$ and $\mathbf{P}$ as.

$\mathbf{z}_{mn} = \mathbf{Y}_{mn}\ \ \mathbf{P}_{nn}$,

then we could get.

$\mathbf{z}_{mn}^T\ \ \mathbf{z}_{mn} = \Sigma_{nn}$.

That implies

$$\mathbf{Y}_{mn} = \mathbf{z}\mathbf{P}^T = \mathbf{z}/\sqrt{\lambda} \cdot P^T\sqrt{\lambda} = eof \cdot PC \tag{3}$$

If we do not care the magnitude of $\mathbf{z}$ and $\mathbf{P}$, the $\mathbf{z}$ can be EOF and the $\mathbf{P}$ is PC. Or we can rearrange the EOF and PC by square root of eigenvalues λ. In here, notice that the data $\mathbf{Y}$ can be separated into $\mathbf{z}$ spatial modes times $\mathbf{P}$ temporal modes two parts.

Support vector regression

Support Vector Regression (SVR) is a machine learning algorithm that is widely used for regression tasks. It is based on the Support Vector Machines (SVM) algorithm, which is primarily used for classification. SVR uses the concept of support vectors, which are data points that lie closest to the decision boundary. The decision boundary is defined by a hyperplane in a high-dimensional feature space. SVR introduces the concept of ε-insensitive loss function, where errors

within a certain tolerance ε are ignored. The optimization problem in SVR involves finding the hyperplane that maximizes the margin while satisfying the ε-insensitive loss. The SVR formulation involves minimizing the following objective function:

Given kernelized model (Gaussian kernel was used in this article).

$f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b,$

use ε-insensitive loss functionand minimize

$$e_\varepsilon(\mathbf{y}, f(\mathbf{x})) = \begin{cases} 0\ , & if\left|\mathbf{y} - f(\mathbf{x})\right| < \varepsilon \\ \left|\mathbf{y} - f(\mathbf{x})\right| - \varepsilon\ , & otherwise \end{cases}$$

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi^+ + \xi^-) \tag{4}$$

subject to.

$$\mathbf{y} - (\mathbf{w}^T\phi(\mathbf{x}) + b) \le \varepsilon + \xi^+$$
$$(\mathbf{w}^T\phi(\mathbf{x}) + b) - \mathbf{y} \le \varepsilon + \xi^-,$$
$$\xi^+, \xi^- \ge 0$$

where $\mathbf{w}$ represents the weight vector, b is the bias term, $\mathbf{x}$ is the input data, and $\mathbf{y}$ is the corresponding target value. The ξ^+ and ξ^- are slack variables for measuring the distance between data points and the correct hyperplanes. The C parameter controls the trade-off between the margin width and the training error. SVR utilizes a kernel function to map the input data into a high-dimensional feature space, allowing for nonlinear relationships to be captured.

Neural networks

The NN model used in this article consists of an input layer, an output layer, and three hidden layers. Both the input layer and the output layer have 20 neurons, representing 20 principal components (PCs). Each hidden layer contains 10 neurons. The calculation between each layer is defined as follows:

$$\mathbf{y} = h(\sum \mathbf{w}^T\mathbf{x} + b)\ , \tag{5}$$

where $h = \sigma(\mathbf{z}) = \dfrac{2}{1 + e^{-2\mathbf{z}}} - 1.$

The activation function used is a hyperbolic tangent sigmoid function σ followed by a linear combination function that maps the l inputs $\mathbf{x} = (x_1, ..., x_l)^T$ to m hidden neurons $h = (h_1, ..., h_m)^T$. During training process, weights $\mathbf{w}$ and biases b will be adjusted to minimize the loss function, which is the mean square error (MSE) between input PCs and output PCs.

RMSE and ACC

The RMSE is defined as

$$RMSE = \sqrt{\sum_{i=1}^{N} w_i D_i^2} \Big/ \sqrt{\sum_{i=1}^{N} w_i}\ , \tag{6}$$

where D_i is the deviation between the forecast and the verified analysis field and w_i is the weighting function defined by the following cosine latitude:

$w_i = 1\big/\cos\phi_i(\phi$ is latitude),

where N is the number of samples. The ACC is defined as

$$ACC = \frac{\sum_{i=1}^{N} w_i(f_i - \overline{f})(a_i - \overline{a})}{\sqrt{\sum_{i=1}^{N} w_i(f_i - \overline{f})^2 \sum_{i=1}^{N} w_i(a_i - \overline{a})^2}}\ , \tag{7}$$

where $f_i, \overline{f}, a_i, \overline{a}$ are given as follows:

$$f_i = F_i - C_i,\ \overline{f} = \sum_{i=1}^{N} w_i f_i \Big/ \sum_{i=1}^{N} f_i,$$

$$a_i = A_i - C_i,\ \overline{a} = \sum_{i=1}^{N} w_i a_i \Big/ \sum_{i=1}^{N} a_i,$$

where F, A, and C are the forecast, verified analysis field, and climatological value, respectively.

Acknowledgements
The authors thank two anonymous reviewers for their constructive and helpful comments. The author appreciates one of the reviewers for pointing out some errors in the original manuscript. The authors thank Yuejian Zhu in NCEP/EMC for providing forecast skills of NCEP SST predictions.

Author contributions
JC-H T is responsible for setting the direction of the topic, setting the research methods, and selecting the materials. B-AT is responsible for collating data, implementing and drawing related models. KC participates in the discussion of research methods and research results.

Funding
This research is grateful for support from Central Weather Bureau, Taiwan, ROC and the Ministry of Science and Technology, Taiwan, ROC (Grant # MOST 110-2634-F-008-008).

Availability of data and materials
The SST data are from IRI web, and the link is https://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.ERSST/ (accessed on 6 Jan. 2023).

Declarations

Competing interests
Not applicable.

Received: 18 April 2023 Accepted: 24 August 2023
Published online: 07 September 2023

References
Alexander MA, Blade I, Newman M, Lanzante J, Lau NC, Scott JD (2002) The atmospheric bridge: the influence of ENSO teleconnections on air–sea interaction over the global oceans. J Clim 15:2205–2231
Ashok K, Behera SK, Weng H, Yamagata T (2007) El Niño Modoki and its possible teleconnection. J Geophys Res 112:C11007. https://doi.org/10.1029/2006JC003798
Balasubramanian S, Peterson RA, Jarvenpaa SL (2002) Exploring the implications of m-commerce for markets and marketing. J Acad Mark Sci 30:348–361. https://doi.org/10.1177/009207002236910
Bayá AE, Granitto PM (2008) ISOMAP based metrics for clustering. Int Artif 12(37):15–23
Bonavita M, Geer A, Laloyaux P, Massart S, Chrust M (2021) Data assimilation or machine learning. ECMWF Newslett 167:17–22
Brunton SL, Kutz JN (2022) Data-driven science and engineering: machine learning, dynamic systems, and control. Cambridge University Press, Cambridge
Brunton SL, Noack BR, Koumoutsakos P (2020) Machine learning for fluid mechanics. Annual Reviews 52:477–508
Capotondi A, Wittenberg AT, Newman M, Lorenzo ED, Yu JY, Braconnot P, Cole J, Dewitte B, Giese B, Guilyardi E, Jin FF, Kranauskas K, Kirtman B, Lee T, Schneider N, Xue Y, Yeh SW (2015) Understanding ENSO diversity. Bull Amer Meteor Soc 96:921–938. https://doi.org/10.1175/BAMS-D-13-00117.1
Dueben P, Modigliani U, Geer A, Siemen S, Pappenberger F, Bauer P, Brown A, Palkovic M, Raoult B, Wedi N, Baousis V (2021) Machine learning at ECMWF: a roadmap for the next 10 years. ECMWF Techn Mem 878:17
Elken J, Zujev M, She J, Lagemaa P (2019) Reconstruction of large-scale sea surface temperature and salinity fields using sub-regional EOF patterns from models. Front Earth Sci 7:232. https://doi.org/10.3389/feart.2019.00232
Hatfield S, Chantry M, Dueben P, Lopez P, Geer A, Palmer T (2021) Building tangent-linear and adjoint models for data assimilation with neural networks. J Adv Model Earth Syst. https://doi.org/10.1029/2021MS00252
Hernández JDR, Mesa J, Lall U (2020) ENSO dynamics, trends, and prediction using machine learning. WAF 35:2061–2081. https://doi.org/10.1175/WAF-D-20-0031.1
Izenman AJ (2008) Modern multivariate statistical techniques: regression, classification, and manifold learning. Springer Press, New York
James IN (1994) Introduction to circulating atmospheres. Cambridge University Press, Cambridge
Kao HY, Yu JY (2009) Contrasting Eastern-Pacific and Central-Pacific types of ENSO. J Clim 22:615–632
Kug JS, Jin FF, An SI (2009) Two types of El Niño and warm pool El Niño. J Clim 22:1499–1515
Li Y, Chen Q, Liu X, Li J, Xing N, Xie F et al (2019) Long-term trend of the tropical Pacific trade winds under global warming and its causes. J Geophys Res Oceans. https://doi.org/10.1029/2018JC014603
McPhaden MJ, Zhang X, Hendon HH, Wheeler MC (2006) Large scale dynamics and MJO forcing of ENSO variability. Geophys Res Lett 33:L16702. https://doi.org/10.1029/2006GL026786
McPhaden MJ, Zhang X, Hendon HH, Wheeler MC (2012) A 21st century shift in the relationship between ENSO SST and warm water volume anomalies. Geophys Res Lett 38:L15709. https://doi.org/10.1029/2012GL051826
Molteni F, Stockdale T, Balmaseda M, Balsame G, Buizza R, Ferranti L, Magnusson L, Mogensen K, Palmer T, Vitart F (2011) The new ECMWF seasonal forecast system (system 4). ECMWF Techn Memo 656:51
Rousi E, Kornhuber K, Beobide-Arsuaga G, Luo F, Coumou D (2022) Accelerated western European heatwave trends linked to more-persistent double jets over Eurasia. Nat Commun 13:3851. https://doi.org/10.1038/s41467-022-31432-y|www.nature.com/naturecommunications
Sousa SIV, Martins FG, Alvim-Ferraz MCM, Pereira MC (2007) Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ Model Softw 22(1):97–103. https://doi.org/10.1016/j.envsoft.2005.12.002
Stock JH, Watson MW (2002) Forecasting using principal components from a large number of predictors. J Am Stat Assoc 97(460):1167–1179
Tang B, Hsieh WW, Monahan AH, Tangang FT (2000) Skill comparisons between neural networks and canonical correlation analysis in predicting the equatorial Pacific sea surface temperatures. J Clim 13:287–293
Tangang FT, Hseih WW, Tang B (1997) Forecasting the equatorial Pacific sea surface temperatures by neural network models. Clim Dyn 13:135–147
Tangang FT, Hsieh WW, Tang B (1998a) Forecasting the regional sea surface temperatures of the tropical Pacific by neural network models, with wind stress and sea level pressure as predictors. J Geophys Res 103:7511–7522
Tangang FT, Tang B, Monahan AH, Hsieh WW (1998b) Forecasting ENSO events—a neural network-extended EOF approach. J Clim 11:29–41
Tasoulis S, Pavlidis NG, Roos T (2020) Nonlinear dimensionality reduction for clustering. Pattern Recognit 107:107508. https://doi.org/10.1016/j.patcog.2020.107508
Tenebaum JB, Silva VD, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–3232
Theodoridis S, Pikrakis A, Koutroumbas K, Cavouras D (2010) An introduction to pattern recognition: a MATLAB approach. Academic Press, Burlington
Thoppil PG, Frolov S, Rowley CD, Reynolds CA, Jacobs GA, Metzger EJ, Hogan PJ, Barton N, Wallcraft AJ, Smedstad OM, Shriver JF (2021) Ensemble forecasting greatly expands the prediction horizon for ocean mesoscale variability. Commun Earth & Environ 2:89. https://doi.org/10.1038/s43247-021-00151-5
Trenberth KE, Hurrell JW (1994) Decadal atmosphere ocean variations in the Pacific. Climate Dyn 9:303–319
Trenberth KE, Branstator GW, Karoly D, Kumar A, Lau NC, Ropelewski C (1998) Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. J Geophys Res 103:14291–14324
Tseng JCH (2022) An ISOMAP analysis of sea surface temperature for the classification and detection of El Niño & La Niña events. Atmosphere 13:919. https://doi.org/10.3390/atmos13060919
Tu′uholoaki M, Espejo A, Singh A, Damlamian H, Wandres M, Chand S, Mendez FJ, Fa′anunu ′O, (2023) Clustering tropical cyclone genesis on ENSO timescales in the southwest Pacific. Clim Dyn. https://doi.org/10.1007/s00382-022-06497-6
Vecchi GA, Wittenberg AT (2010) El Niño and our future climate: where do we stand? Wiley Interdiscip. Rec.: Climate Change 1:260–270. https://doi.org/10.1002/wcc.33
Wilks DS (2019) Statistical methods in the atmospheric sciences, 4th edn. Elsevier, Amsterdam
Wu A, Hsieh WW, Tang B (2006) Neural network forecasts of the tropical Pacific sea surface temperatures. Neural Netw 19:145–154. https://doi.org/10.1016/j.neunet.2006.01.004
Yeh SW, Ham YG, Lee JY (2012) Changes in the tropical Pacific SST trend from CMIP3 to CMIP5 and its implication of ENSO J Climate25:7764–7771. https://doi.org/10.1175/JCLI-D-12-00304.1
Yu JY, Kao HY, Lee T, Kim ST (2011) Subsurface ocean temperature indices for Central-Pacific and Eastern-Pacific types of El Niño and La Niñaevents. Theor Appl Climatol 103:337–344. https://doi.org/10.1007/s00704-010-0307-6

Publisher's Note