Effective forecast of Northeast Pacific sea surface temperature based on a complementary ensemble empirical mode decomposition–support vector machine method

LI Qi-Jie, ZHAO Ying, LIAO Hong-Lin and LI Jia-Kang

Institute of Sciences, University of Science and Technology of the Chinese People’s Liberation Army, Nanjing, China

ABSTRACT

The sea surface temperature (SST) has substantial impacts on the climate; however, due to its highly nonlinear nature, evidently non-periodic and strongly stochastic properties, it is rather difficult to predict SST. Here, the authors combine the complementary ensemble empirical mode decomposition (CEEMD) and support vector machine (SVM) methods to predict SST. Extensive tests from several different aspects are presented to validate the effectiveness of the CEEMD-SVM method. The results suggest that the new method works well in forecasting Northeast Pacific SST at a 12-month lead time, with an average absolute error of approximately 0.3 °C and a correlation coefficient of 0.85. Moreover, no spring predictability barrier is observed in our experiments.

1. Introduction

The oscillation of sea surface temperature (SST) has substantial impacts on the climate. Anomalously high SST near the equator (between 5°S and 5°N and the Peruvian coast) causes the El Niño phenomenon, while low SST in this area brings about the La Niña phenomenon, both of which impose considerable influence on temperature, precipitation and wind globally (Luo 2000; Yuan, Yang, and Li 2012; Lian 2014). Owing to the considerable effect that these issues have on humans, many efforts (Feng et al. 2007; Shang et al. 2013; Meng, Lin, and Tang 2015) have been made to analyze SST. In doing so, it has become clear that accurate observations and effective predictions of SST are very important. However, owing to its seriously nonlinear nature, non-stationary and strongly stochastic properties, it is rather difficult to analyze and predict SST. To the best of our knowledge, there is no single method that has been verified to predict SST accurately.

Empirical mode decomposition (EMD) is a signal processing method originally proposed by Huang et al. (1998). This method decomposes original signals into several intrinsic mode functions (IMFs) with different frequencies. These IMFs can be regarded as an orthogonal product of the original signal; nonetheless, the non-stationarity of the resulting IMFs has been improved greatly. Actually, the EMD method has been applied widely in engineering, meteorology, and some other fields (Chen 2012; Yu et al. 2014; Liang et al. 2015). However, the EMD algorithm has its own defects in practical applications; for instance, the endpoint effect and mode confusion. The endpoint effect has been removed by several approaches, including the extension technique (Xu 2009; Shi et al. 2014). Meanwhile, in order to resolve the problem of mode confusion, Wu and Huang (2009) proposed the ensemble empirical mode decomposition (EEMD) algorithm, which improves the EMD method by adding some random white noise.
noise to balance the random error in the original signal. In their study, several EMDs were run for the same signal by adding some random white noise, allowing them to obtain the final IMFs by averaging the resulting IMFs in each run. However, the accumulation of man-made white noise always introduces large reconstruction error in the EEMD method. This problem, though, can be solved to a large extent via complementary ensemble empirical mode decomposition (CEEMD), as suggested by Yeh and Shieh (2010). In this paper, we apply the CEEMD algorithm to decompose the SST data with small reconstruction error for our prediction.

There are many existing approaches, such as curve fitting, mean generating function periodic extrapolation, neural networks, and the support vector machine (SVM) method, for predicting nonlinear time signals. Here, we are interested in the SVM method, which is a machine learning algorithm proposed by Vapnik (1995). Based on Vapnik–Chervonenkis dimension theory and the principle of structural risk minimization, the SVM method has its own advantages in the prediction of nonlinear time signals. Examples of its application (Chen 2012; Cai, Zhang, and Yang 2014) can be found in several fields, such as computer engineering, control and decision-making, and meteorology.

We combine the CEEMD and SVM methods in this paper to construct a new procedure for SST forecasting. Extensive numerical tests are presented to verify the feasibility of our method.

2. SST data and the CEEMD-SVM method

2.1. SST data

Observations show that the SST in the Northeast (NE) Pacific Ocean has been higher than usual in the past few years. In particular, from 2014 to the first half of 2016, the SST in this region was 2 °C higher than normal, and has had obvious impacts on both coastal and inland areas (Bond et al. 2015). In this paper, we consider time series of 408 months (1982–2015) of mean values of SST anomaly (SSTA) data in the region (40°–50°N, 150°–135°W). As shown in Figure 1, the observed SSTA series is obviously nonlinear, evidently non-periodic, and strongly stochastic.

2.2. CEEMD-SVM method

The CEEMD-SVM method for forecasting the SSTA sequence proceeds as follows:

Step 1. The CEEMD method is applied to decompose the historical SSTA data into several modes (IMFs and the remainder mode, R) with different frequencies.

Step 2. For each available mode, we divide the time sequence into two parts, which serve the learning and control groups, respectively (the same division is always used for each mode). Then, the SVM is trained and validated using the learning and control groups separately.

Step 3. For each available mode, the trained SVM is used to generate the predictive data.

Step 4. The predictive result of the SSTA is reconstructed by the forecasting data of all modes.

3. Data decomposition by CEEMD

Using the CEEMD method, the original SSTA sequence is decomposed into eight modes (Figure 2), including seven IMFs and the remainder mode, R. It is observed that the frequencies of seven IMFs decrease successively and the corresponding periods increase. Also, the remainder mode, R, is monotonically increasing, in accordance with the upward trend of the original SSTA data.

Obviously, the modes IMF_4–IMF_7 are intrinsically regular in time since they have their own characteristic frequencies, while the characteristic frequencies of the first few IMFs, especially IMF_1, are indistinct. In other words, the non-periodic and strongly stochastic properties of the original SSTA data are mainly inherited in the first several

Figure 1. The temporal variation of the observed SSTA data of the Northeast Pacific (40°–50°N, 150°–135°W).
IMFs, while the regular components in the SSTA data are separated within the other modes, such as IMF₄–IMF₇ and the remainder mode, R. This characteristic is important for our prediction of the SSTA sequence because, compared with a direct prediction of the observed SSTA data, the prediction accuracy of regular modes could be significantly increased.

Figure 2. Eight CEEMD modes of original SSTA data (units: °C).
Notes: The modes IMF₄–IMF₇ are intrinsically regular in time since they have their own characteristic frequencies, while the characteristic frequencies of the first few IMFs, are indistinct.

Figure 3. The reconstruction error of CEEMD for the total 408 months.

Figure 4. The average absolute error of forecasting results in 2006–2015 between the simulated SSTA and the observed data.
improved. Naturally, it should be expected that our CEEMD-SVM method generates a better prediction than the SVM method if the reconstruction error of CEEMD is negligible.

The reconstruction error of CEEMD is defined by the absolute value of difference between the original data and the summation of all decomposed modes. Figure 3 depicts the reconstruction error of our decomposition for the SSTA data. It is apparent that the reconstruction error is around $10^{-16}$ °C for most of the total 408 months, and only for 5 months is the error located between $2.5 \times 10^{-16}$ °C and $5 \times 10^{-16}$ °C. That is to say, the eight modes in Figure 2 are effective and acceptable.

4. Forecast experiments via the SVM method

The effectiveness of SVM prediction, including the prediction accuracy, the predictable time length, and the practical effect from different starting months of prediction, will be examined here before we apply it to generate the predictive data of each mode.

Table 1. Forecasting errors in 2006–2015 (units: °C) and the correlation coefficients between the simulated and observed data.

| Year | Average absolute error | Maximum absolute error | Minimum absolute error | Correlation coefficient |
|------|-------------------------|------------------------|------------------------|-------------------------|
| 2006 | 0.1242                  | 0.2815                 | 0.0241                 | 0.9006                  |
| 2007 | 0.1607                  | 0.4130                 | 0.0024                 | 0.8324                  |
| 2008 | 0.1570                  | 0.4650                 | 0.0074                 | 0.7258                  |
| 2009 | 0.2691                  | 0.8156                 | 0.0395                 | 0.8940                  |
| 2010 | 0.3158                  | 0.7192                 | 0.0436                 | 0.9403                  |
| 2011 | 0.2672                  | 0.6678                 | 0.0072                 | 0.5988                  |
| 2012 | 0.1694                  | 0.4130                 | 0.0278                 | 0.9225                  |
| 2013 | 0.3052                  | 0.7980                 | 0.0459                 | 0.9112                  |
| 2014 | 0.3008                  | 0.7452                 | 0.0346                 | 0.7737                  |
| 2015 | 0.3030                  | 0.5791                 | 0.0910                 | 0.8661                  |

4.1. Twelve-month forecasting experiments

We design a class of 12-month forecasting experiments with the initial time in January to test the accuracy of SVM prediction by running it 10 times. The SSTA data of 2006–2015 are forecasted year by year. The main trend of the predictive series is the same as the real data in the 12-month prediction experiments. One can see that, for each month in 2006–2015, the average absolute error between the simulated SSTA and the observed data is less than 0.4436 °C (Figure 4). The corresponding forecasting errors in three different measures are also reported in Table 1. We find that the forecasting errors are generally small and acceptable. Actually, the average error is approximately 0.3 °C, the maximum error is less than 0.8 °C, and the average correlation coefficient between the simulated data and the observed data reaches 0.85. Hence, the 12-month SVM prediction is effective.

Note that the final prediction error is accumulated by the forecasting errors of all modes generated by the CEEMD method. We find that the average predicted error of the first mode, IMF1, reaches 0.2 °C and provides the largest contribution to the overall forecasting error; however, this is not unexpected. Actually, due to the evidently non-periodic and strongly stochastic properties (see Figure 2), it is rather difficult to forecast IMF1 in general.

4.2. Thirty-six–month forecasting experiments

To examine the predictable time length of the SSTA forecast via the CEEMD-SVM method, we design 36-month forecasting experiments. Similar to the experiments reported in Section 4.1, in this group of experiments the SSTA data are forecasted from 2013 to 2015.

Figure 5. The temporal variation of the simulated data of thirty-six–month forecasting and the observed data in 2013–2015 (units: °C).
problem (Zhang, Yu, and Duan 2012). Specifically, the prediction of ENSO events possesses seasonal dependence and the prediction error always increases fastest in spring (April, May), as compared with other seasons. Thus, starting the prediction from different months may affect the forecasting results. To examine this, the 12-month forecasting experiments are run five times, from 2010 to 2014, using twelve starting months (from January to December).

The forecasting results based on the different starting months are plotted in Figure 6, and the corresponding prediction error and correlation coefficients between the forecasting results and the observed data are reported in Table 3. We find that the forecasting results based on the 12 different starting months, from January to December, are acceptable and effective. Actually, the average absolute errors are approximately 0.3 °C and the correlation coefficients are no less than 0.73. Additionally, compared with other starting months, the forecasting starting in January, February, or March introduces relatively smaller prediction error. Forecasting errors become increased when the prediction starts in April or May, but the forecasting starting from June to December successfully avoids the negative influence produced by starting in April or May.

As a result, for the present SSTA data and the CEEMD-SVM method, there is no spring predictability barrier. The reason for the disappearance of this phenomenon is not clear, so we cannot rule out its reappearance when predicting SSTA sequences in other regions using our CEEMD-SVM method.

4.3. Sensitivity of the SSTA forecast to the starting month

According to previous experience, climate forecasting may encounter the so-called spring predictability barrier.

Table 2. Thirty-six–month forecasting errors (units: °C) and the correlation coefficients between the simulated and observed data.

| Year | Average absolute error | Maximum absolute error | Minimum absolute error | Correlation coefficient |
|------|------------------------|------------------------|------------------------|------------------------|
| 2013 | 0.3815                 | 1.021                  | 0.0339                 | 0.9112                 |
| 2014 | 0.6295                 | 1.361                  | 0.0577                 | 0.3067                 |
| 2015 | 0.7052                 | 1.521                  | 0.2479                 | 0.1814                 |

As a result, for the present SSTA data and the CEEMD-SVM method, there is no spring predictability barrier. The reason for the disappearance of this phenomenon is not clear, so we cannot rule out its reappearance when predicting SSTA sequences in other regions using our CEEMD-SVM method.
Table 3. The absolute forecasting errors from different starting months (units: °C) and the correlation coefficients between the simulated and observed data.

| Forecast starting month | Average absolute error | Maximum absolute error | Minimum absolute error | Root-mean-square error | Correlation coefficient |
|-------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| Jan.                    | 0.2580                 | 0.7812                 | 0.0102                 | 0.3160                 | 0.8172                 |
| Feb.                    | 0.2961                 | 0.8753                 | 0.0026                 | 0.3554                 | 0.8582                 |
| Mar.                    | 0.2939                 | 0.8732                 | 0.0035                 | 0.3596                 | 0.8112                 |
| Apr.                    | 0.3144                 | 1.018                  | 0.0323                 | 0.3882                 | 0.7463                 |
| May.                    | 0.3198                 | 1.062                  | 0.0113                 | 0.4026                 | 0.7519                 |
| Jun.                    | 0.3129                 | 1.189                  | 0.0024                 | 0.3924                 | 0.7623                 |
| Jul.                    | 0.3372                 | 1.149                  | 0.0219                 | 0.4276                 | 0.8033                 |
| Aug.                    | 0.3059                 | 1.139                  | 0.0055                 | 0.3886                 | 0.7981                 |
| Sep.                    | 0.3083                 | 0.9718                 | 0.0012                 | 0.3899                 | 0.7763                 |
| Oct.                    | 0.3270                 | 0.7201                 | 0.0086                 | 0.3892                 | 0.7951                 |
| Nov.                    | 0.3027                 | 0.9164                 | 0.0083                 | 0.3679                 | 0.7347                 |
| Dec.                    | 0.3268                 | 1.063                  | 0.0077                 | 0.4114                 | 0.8032                 |

5. Summary

In this paper, the CEEMD and SVM methods are combined to predict the monthly SSTA sequence in the NE Pacific Ocean, which is characterized by seriously nonlinear and evidently non-periodic and strongly stochastic properties.

1. The observed SSTA is decomposed using CEEMD, and several modes with different periods are obtained. Compared with the original sequence, the decomposed modes stabilize the non-stationarity and improve the periodicity, which lays the foundation for the forecasting.

2. Owing to the evidently non-periodic and strongly stochastic properties of the first mode, IMF1, it contributes the largest prediction error to the overall forecasting error, with an averaged error of around 0.2 °C.

3. For the present SSTA data and the CEEMD-SVM method, no spring predictability barrier is observed. However, we cannot explain why this is the case, meaning it might reappear when predicting SSTA sequences in other areas.

4. From the experimental results, the CEEMD-SVM method works well when forecasting the SSTA data with a 12-month lead time, and the forecasting result becomes worse when the prediction time is prolonged. Actually, the suggested method is effective for a 12-month SSTA forecast.

Finally, it is important to highlight that this paper considers SSTA data from 1982 to 2015 only, and only for the region of (50°–40°N, 150°–135°W) in the NE Pacific. It therefore serves as a reference for studying SST prediction in other regions.

Acknowledgements

The authors wish to extend their sincere thanks to Dr. LIU Juanjuan and Dr. LIN Pengfei, State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics, Institute of Atmospheric Physics, Chinese Academy of Sciences, who assisted in providing the SST data, and also put forward valuable comments. In addition, the authors wish to extend particular thanks to the anonymous reviewers, whose comments were very helpful.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was supported in part by the Major Research Plan of the National Natural Science Foundation of China [grant number 91530204]; the State Key Program of the National Natural Science Foundation of China [grant number 41430426].

References

Bond, N. A., M. F. Cronin, H. Freelend, and N. Mantua. 2015. “Causes and Impacts of the 2014 Warm Anomaly in the NE Pacific.” Geophysical Research Letters 42 (9): 3414–3420.

Cai, J. H., S. J. Zhang, and L. Yang. 2014. “Nanjing Monthly Average Temperature Prediction Base on Empirical Mode Decomposition and Support Vector Machine.” Mathematics in Practice and Theory 44 (22): 103–111.

Chen, R. 2012. “The Study on Short Time Forecast of Wind Speed Based on EMD and SVM.” Science & Technology Vision 26: 90–92.

Feng, X. B., K. Wang, Z. H. Zhou, and J. L. Wang. 2007. “Characterizing Rainfall Time Series in the Loess Plateau and Periodical Links with Sea Surface Temperature.” Journal of Irrigation and Drainage 26 (2): 56–59.

Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. Liu. 1998. “The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Nonstationary Time Series Analysis.” Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 454 (1971): 903–995.

Lian, T. 2014. “Fundamental Types and Genesis of El Niño-Southern Oscillation.” PhD diss., ZheJiang University.

Liang, Y. J., C. Ren, L. L. Liu, G. F. Pang, and X. Y. Yang. 2015. “Dam Deformation Prediction Based on EMD and GA-BP Neural Network.” Journal of Guilin University of Technology 35 (1): 111–116.
Luo, G. Y. 2000. “A General Survey of the Studies on El Niño and La Niña.” *Scientia Geographica Sinica* 20 (3): 264–269.
Meng, Q. J., P. F. Lin, and X. H. Tang. 2015. “Interannual Variability of Sea Surface Temperature in the Tropical Oceans in the 20th Century.” *Marine Sciences* 39 (3): 59–65.
Shang, K., J. H. He, Z. W. Zhu, and F. X. Zhan. 2013. “Comparison Between Correlations of Heat Content and Sea Surface Temperature over Western Pacific Warm Pool with Spring Persistent Rains.” *Geographical Science* 33 (8): 986–992.
Shi, P. M., J. S. Jiang, B. Liu, and J. Wang. 2014. “Improved Method of EMD and Its Applications Based on Boundary Characteristic Scale Matching Extension Method.” *Chinese Mechanical Engineering* 25 (12): 1616–1623.
Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.
Wu, Z., and N. E. Huang. 2009. “Ensemble Empirical Mode Decomposition: A Noise-assisted Data Analysis Method.” *Advances in Adaptive Data Analysis* 1 (01): 1–41.
Xu, S. Y. 2009. “Improved Algorithm of Empirical Mode Decomposition Based on Extension Technique.” *Journal of Jilin University* 27 (4): 359–365.
Yeh, J. R., and J. S. Shieh. 2010. “Complementary Ensemble Empirical Mode Decomposition: A Noise Enhanced Data Analysis Method.” *Advances in Adaptive Data Analysis* 2 (02): 135–156.
Yu, Y., D. C. Chi, T. T. Chen, M. Sun, and C. Lu. 2014. “Drought Forecast Application of BP Prediction Models Based on EMD in Ling River Basin.” *Journal of Shenyang Agricultural University* 45 (1): 68–72.
Yuan, Y., H. Yang, and C. Y. Li. 2012. “Study of El Niño Events of Different Types and Their Potential Impact on the Following-summer Precipitation in China.” *Acta Meteorologica Sinica* 70 (3): 467–478.
Zhang, Y. L., Y. Q. Yu, and W. S. Duan. 2012. “The Spring Prediction Barrier of ENSO in Retrospective Prediction Experiments as Shown by the Four Coupled Ocean-atmosphere Models.” *Acta Meteorologica Sinica* 70 (3): 506–519.