Application of the nonlinear time series prediction method of genetic algorithm for forecasting surface wind of point station in the South China Sea with scatterometer observations*

Jian Zhong(钟剑)†, Gang Dong(董钢), Yimei Sun(孙一妹), Zhaoyang Zhang(张钊扬), and Yuqin Wu(吴玉琴)

China Satellite Maritime Tracking and Control Department, Jiangyin 214431, *China*

(Received 24 March 2016; revised manuscript received 19 July 2016; published online 30 September 2016)

The present work reports the development of nonlinear time series prediction method of genetic algorithm (GA) with singular spectrum analysis (SSA) for forecasting the surface wind of a point station in the South China Sea (SCS) with scatterometer observations. Before the nonlinear technique GA is used for forecasting the time series of surface wind, the SSA is applied to reduce the noise. The surface wind speed and surface wind components from scatterometer observations at three locations in the SCS have been used to develop and test the technique. The predictions have been compared with persistence forecasts in terms of root mean square error. The predicted surface wind with GA and SSA made up to four days (longer for some point station) in advance have been found to be significantly superior to those made by persistence model. This method can serve as a cost-effective alternate prediction technique for forecasting surface wind of a point station in the SCS basin.

Keywords: nonlinear time series prediction, genetic algorithm, surface wind prediction, singular spectrum analysis

PACS: 05.45.Tp, 92.60.Wc **DOI:** 10.1088/1674-1056/25/11/110502

1. Introduction

Surface wind is an important ocean parameter. Forecasting ocean surface wind involves two aspects: the wind speed and wind direction, and the zonal and meridional wind components. Prediction of surface wind speed is very important in many applications, such as planning, construction, and operation-related works in the oceanic areas, and so on. Meanwhile, prediction of the surface wind components constitutes an important component of ocean state prediction using numerical ocean and atmospheric models.[1]

Predictability of the ocean surface wind including wind speed and wind components has achieved considerable success using numerical models in many studies.[1,2] However, the numerical prediction models suffer from various drawbacks, such as incomplete physics, incorrect initial conditions, etc. In particular, when point forecasts at specific locations are required, the numerical prediction models are disadvantageous because there are always extremely complex and are highly computer intensive due to the huge amount of input information, such as the vertical profiles of humidity, temperature, and so on.[3,4]

It is thus interesting to explore the possibility of predicting the surface wind using only past observations without the need for sophisticated numerical models. Over the years, various such data-adaptive approaches, such as linear regression, support vector machines,[5] artificial neural networks (ANNs),[6] genetic algorithm (GA),[7] and so on,[8,9] have been proposed for prediction of the nonlinear data series. The GA algorithm, which is based on Darwin's evolutionary theory,[7,10,11] is a modern and powerful nonlinear data fitting algorithm. One of the advantages of GA algorithm is that it does not require very long time series of wind observations. Another advantage is that it provides an explicit analytical forecast equation.[12]

The predictive skill of GA has been demonstrated in the cases of sea surface temperature (SST) in the Alboran Sea,[13] summer rainfall over India,[14] SST and sea level anomaly in the Ligurian Sea,[15] wave heights in the north Indian Ocean (NIO),[3,4] the tidal currents in the Arabian Sea,[16] and so on. Meanwhile, ocean surface wind prediction with in situ and scatterometer observations using GA have been tested in the NIO and the results show that predictions with GA made up to three days have been found to be quite encouraging.[3] Due to the specific characteristic of the GA algorithm, the prediction models for different point stations and different basins are different.[3,4]

In the present study, the GA technique has been used to forecast the surface wind in the South China Sea (SCS), which is an important ocean due to the high level of scientific and economic interests in the area. In a previous diagnostic study, Basu *et al.*[3] explored the ability of predicting the surface wind in the NIO. Apart from the different basin, another novel feature of the present study is the use of singular spec-

*Project supported by the National Natural Science Foundation of China (Grant Nos. 41230421 and 41605075) and the National Basic Research Program of China (Grant No. 2013CB430101).

†Corresponding author. E-mail: zhj_zhongjian@163.com

© 2016 Chinese Physical Society and IOP Publishing Ltd

trum analysis (SSA) for noise reduction, which is not tested in previous study. It is known that the GA algorithm, in principle, can predict a strictly deterministic, albeit chaotic, and time series. However, the inevitable presence of noise in any physical system and observations introduces spurious features, not amenable to prediction.[16] Hence, a pre-filtering of the data series using a noise reduction technique becomes absolutely necessary. The SSA is arguably the best known data adaptive approach for noise reduction.[1]

Descriptions of the GA algorithm for prediction and SSA for noise reduction are provided in Sections 2 and 3 separately. Section 4 gives the detail of the used scatterometer observations. The results of the GA algorithm compared with the persistence model for forecasting surface wind speed and surface wind components in the SCS are given in Section 5. Section 6 is the conclusion and discussion.

2. Genetic algorithm

The GA algorithm which is based on the Taken's theorem has been described in detail by earlier works.[3,10]

Briefly, given a deterministic time series $\{x(t_i)\}$, $i = 1,\ldots,N$, there exists a smooth map $\beta : \boldsymbol{R}^m \mapsto \boldsymbol{R}$ satisfying

$$x(t) = \beta(x(t-\mu\tau), x(t-(\mu+1)\tau), \ldots, x(t-(\mu+m-1)\tau)), \quad (1)$$

where m is the embedding dimension and τ is a time-lag unit. This is the prediction model for carrying out m-steps ahead predictions. When the mapping β is derived from the observed time series, the evolution of the system can be predicted in time.

The GA algorithm approximates the mapping β using a technique borrowed from evolutionary biology. The algorithm starts with an initial population of N "equation strings", and includes the following major steps: (1) initialization, (2) computing the fitness, (3) ranking the agents, (4) choosing the mates, (5) reproduction and crossover, and (6) mutation. The detailed description of the algorithm can be found in Ref. [3]. The fitness for the equation string g_j which is necessary for step (2) can be computed as

$$\Delta_j^2 = \sum_{t=L+1}^{T} (x(t) - \gamma_j(x(t-\mu\tau), x(t-(\mu+1)\tau), \ldots, x(t-(\mu+m-1)\tau)))^2, \quad (2)$$

where $L = mt$ and T is the total length of the training set and m is known as the embedding dimension. Then, the strength index for each individual can be defined in the following fashion:

$$R_j^2 = 1 - \frac{\Delta_j^2}{\delta^2}, \quad (3)$$

where

$$\delta^2 = \sum_{t=L+1}^{T} (x(t) - \langle x \rangle)^2, \quad (4)$$

where $\langle x \rangle$ denotes the mean value of the training data. R_j^2 is the square of the coefficient of correlation between forecasts and observations, and can be interpreted as the percentage of the training set's total variance explained by the j-th equation string. It is clear that the closer the value of R_j^2 toward unity, the stronger the j-th individual equation string. These steps are run and rerun for a certain number of generations, or until some stopping criterion is satisfied; e.g., when strength index no longer increases anymore. When the best equation string is chosen, the corresponding strength index R_j^2 is expressed as R^2.

3. Singular spectrum analysis

Because of the measurement errors, the scatterometer observations are inevitably a mixture of deterministic part as well as random part. Hence, it is absolutely necessary to reduce the noise before carrying out a GA forecast. The SSA is arguably the best known data adaptive approach for noise reduction, and hence has been adopted in the present study for reducing noise. The SSA was described in detail and used in Ref. [10]. This happened because Takens's theorem, which is the theoretical basis of GA, can only be applied in the absence of noise and an obvious presence of noise in any physical measurement will induce inappropriate reconstructions.

Briefly, for the time series of observations denoted by A, one has to form a trajectory matrix $\boldsymbol{X}$, the rows of which contain w-dimensional vectors of the form $(A(n), A(n-1), \ldots, A(n-w), A(n-1), A(n-2), \ldots, A(n-1-w))$, etc., where n is the total number of points in the given time series, and w is the window of the filter. Then, the covariance matrix $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ is computed and diagonalized to obtain a set of eigenvectors which are the orthogonal singular vectors. A new filtered time series can then be reconstructed considering limited eigenvectors so that the corresponding eigenvalues are above the noise level. A value of $w = 370$ is set in the study to include the annual wind cycle occurring in the SCS, and the filtered time series was constructed by using the first 80 dominant eigenvectors.

4. Data used

In this paper, the time series of daily averaged wind speed (calculated from wind components) and wind components (including zonal and meridional) measured by the scatterometer onboard QuikSCAT satellite at three selected locations in the SCS are used. A scatterometer, which can provide all-day and large-scale wind field information, has become a main instrument to obtain surface wind field.[17,18] The scatterometer

wind used in this paper is distributed by the French Research Institute for Exploitation of the Sea (IFREMER). The wind has been compared with daily, weekly, and monthly averaged forecasts of the European Centre for Medium range Weather Forecast (ECMWF) model, which shows that the average features are well captured over the global oceans. In addition, comparison with the daily averaged Tropical Atmosphere/Ocean (TAO) and the National Data Buoy Center (NDBC) buoy data at Pacific and Atlantic Oceans show small root-mean-square difference.[3] The wind speed prediction is particularly important for marine works, such as ship planning, construction, and so on, whereas the zonal and meridional wind components are necessary for forecasting the sea state with numerical models. Hence, the wind speed and wind components measured by scatterometer have been chosen in order to show that the algorithm performs equally well for both wind-type predictions.

Three locations in the SCS of QuikSCAT scatterometer observations (Station P1 at (20.25° N, 115.75° E), P2 at (12.25° N, 112.25° E), and the location (7.25° N, 107.25° E) for P3) have been used in order to demonstrate the performance of the algorithm in any particular basin of the SCS. The location details and observation duration of the winds are provided in Table 1, and the locations of the three scatterometer observations in the map are shown in Fig. 1. Ideally, to establish any prediction formula, it should have a long time series of observations such that all possible kinds of variability are taken into account. Hence, 3753 observations of scatterometer data for both wind speed and wind components at the SCS are used. The observations used for training and validation consist of ten years, which encompass all possible types of wind variability.

Table 1. Location details and observation duration of remotely sensed scatterometer winds.

Measuring station	Latitude	Longitude	Observation duration
P1	20.25° N	115.75° E	20 July 1999–21 November 2009
P2	12.25° N	112.25° E	20 July 1999–21 November 2009
P3	7.25° N	107.25° E	20 July 1999–21 November 2009

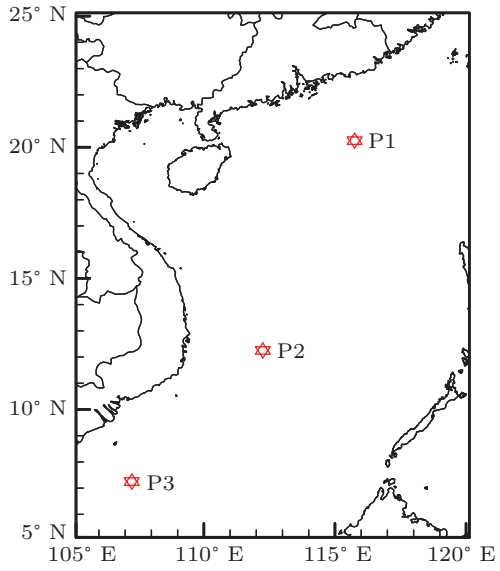


Fig. 1. (color online) Locations of the scatterometer observations.

5. Results

First, the SSA was applied to the WTS (WT stands for the time series of wind speed or wind components) of different point stations using a window size of 370. Then, the GA was applied to predict the filtered time series of the WTS and the fluctuation part (deviation of the filtered WT from the actual WT). The forecast winds were reconstructed with the predicted WTS (including the predicted filtered WTS and the predicted fluctuation part) and compared with the actual winds for evaluating the quality of forecast.

The following parameters were used to train the GA algorithm: the number of equation strings was 60, the total number of arguments and operators allowed was 20 in all the cases, the embedding dimension varied from 4 to as high as 25, and the best embedding dimension m was estimated by trial and error in the present study. The mutation rate was chosen to be 0.01. Finally, the number of iterations required to achieve maximum strength index also varied from case to case and the maximum number was 5000.

Due to the unknown real wind in the previous research,[3] it is instructive to compare the GA forecast with the forecast carried out by a persistence model. For a given WT, the persistence model is defined by the following equation:

$$\mathrm{WT}(t+n) = \mathrm{WT}(t), \tag{5}$$

where n is the lead time of the forecast days. In our case, n varies from 1 to 4. At every grid point station of the present work, the forecasts by both the GA and persistence model methods were carried out. Then, the root mean square errors (RMSE) of the wind speed and wind components forecast for both methods in the time domain were computed.

The first 2960 points of the wind time series were used to train the algorithm and the remaining points were used for validation.

The scatter plots of the wind speed predicted by the GA versus the actual wind speed at P2 point location, and some

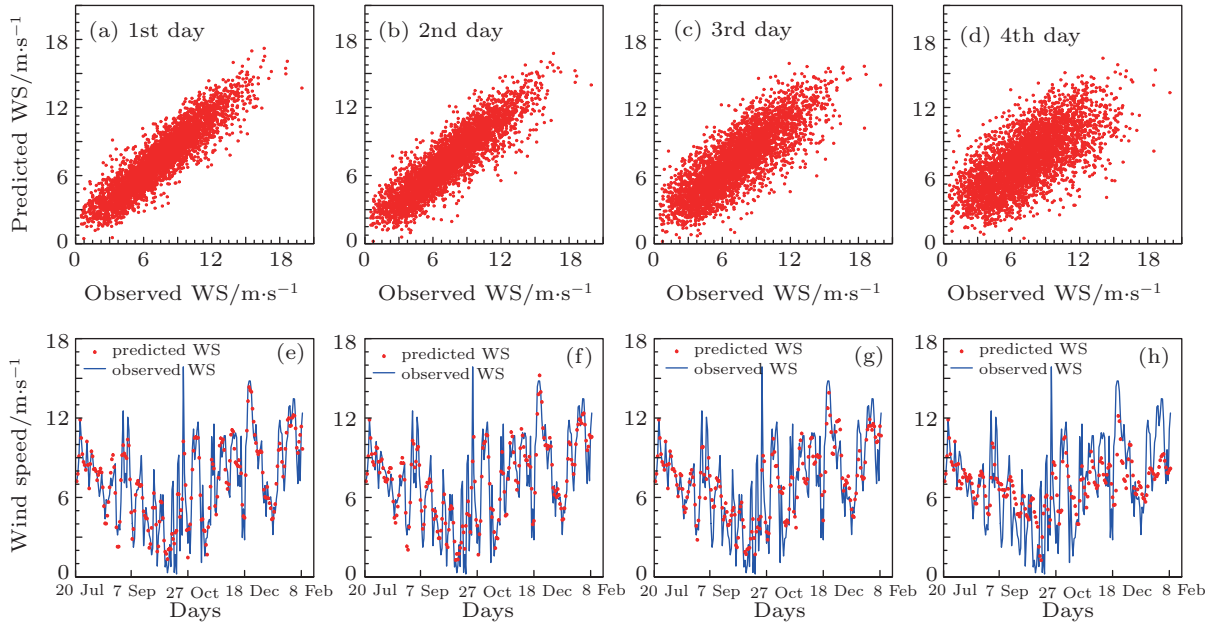


Fig. 2. (color online) The scatter plots of actual and GA predicted wind speed at P2 location from 20 July 1999 to 21 November 2009 (a)–(d). The plots of the time series for actual and GA predicted wind speed at the same location from 20 July 1999 to 27 August 2000 (e)–(h).

part of the corresponding time series at the same location are shown in Fig. 2. The plots are shown for forecasts with leading times from 1 to 4 days. The corresponding figures for P1 and P3 are similar to the case of P2 and are not shown here. It can be easily found that the GA algorithm shows reasonably good performance for predicting wind speed up to 4 days with scatterometer observations. Meanwhile, as the leading time increases, the scatter points diverse more and the prediction errors increase, which means that the performance of the wind speed forecast deteriorates as the leading time increases, which signifies the chaotic nature of the wind speed time series.

Meanwhile, for the 1-day to 4-day forecast, the GA prediction is much better than the persistence model forecast, which is true for the wind speed prediction of all the three points, as shown in Table 2, which shows the comparison of the performance for GA prediction and persistence model prediction. The RMSEs of persistence model prediction are always larger by about 1.0 to 2.0 m/s than GA prediction for a 1-day to 4-day forecast, which are significant differences and improvements.

Although the coefficient of determination R^2 (square of the coefficient of correlation) between the observations and predictions is very similar for the three points, the RMSE of wind speed for P1 point is worse than P2 and P3 points (Table 2) and the difference is about 0.3–0.4 m/s for 1-day to 4-day forecast. This happens because of the lower predictability of wind speed time series for P1 point. The P1 point is closer to the Taiwan strait, where the wind speed always increases when the wind spreads through the strait due to the terrain, and the high wind speed will spread to the P1 which will increase the fluctuant or noise of the whole wind time series and weaken the predictability. Thus, it is not entirely surprising that the performance of GA algorithm is worse for the P1 point than the P2 and P3 points.

The equations for 1-day forecast of wind speed at P2 station are provided in Appendix A.

Table 2. Comparison of predicted wind speed by GA and persistence model at the three locations. The RMSE means root mean square error of forecast in meters/second. R^2 is shown only for the case of GA.

			P1	P2	P3
Forecast (days in advance)	1 day	RMSE (pers.)	2.77	2.25	2.48
		RMSE (GA)	1.58	1.28	1.40
		R^2(GA)	0.82	0.83	0.79
	2 days	RMSE (pers.)	3.77	2.91	3.00
		RMSE (GA)	1.71	1.43	1.48
		R^2(GA)	0.79	0.79	0.77
	3 days	RMSE (pers.)	4.11	3.32	3.38
		RMSE (GA)	2.24	1.81	1.87
		R^2(GA)	0.64	0.67	0.62
	4 days	RMSE (pers.)	4.27	3.50	3.53
		RMSE (GA)	2.84	2.34	2.47
		R^2(GA)	0.42	0.45	0.34

The GA algorithm and persistence model are also used to forecast the zonal and meridional wind components with scatterometer observations. In Fig. 3, the scatter plots and some part of the time series for zonal component of GA algorithm for P2 station are provided, and for the meridional component are shown in Fig. 4. The plots provide the forecasts with lead times from 1 to 4 days. The corresponding figures for P1 and P3 points are not shown for the sake of economy of space. It can easily be seen that the GA algorithm performs reasonably well for predictions with lead times of 1 day, 2 days, 3 days, and 4 days for both zonal and meridional wind components. In addition, the wind components prediction deteriorates with the lead time increases.

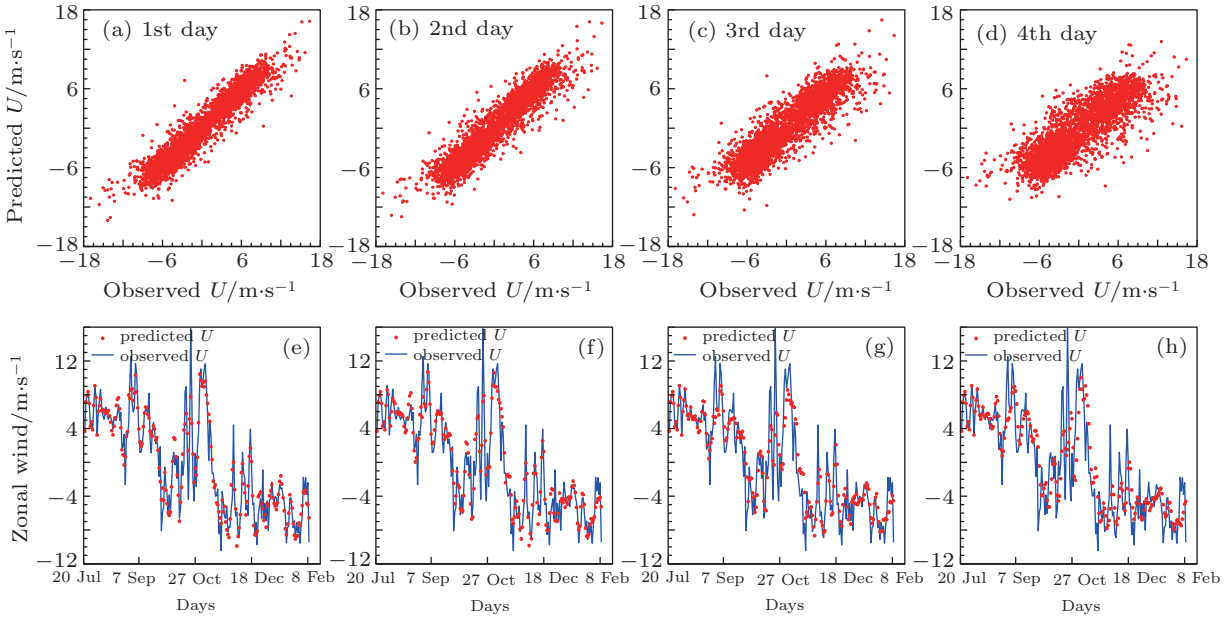


Fig. 3. (color online) The scatter plots of actual and GA predicted zonal wind component at P2 location from 20 July 1999 to 21 November 2009 (a)–(d). The plots of the time series for actual and predicted zonal wind component at the same location from 20 July 1999 to 27 August 2000 (e)–(h).

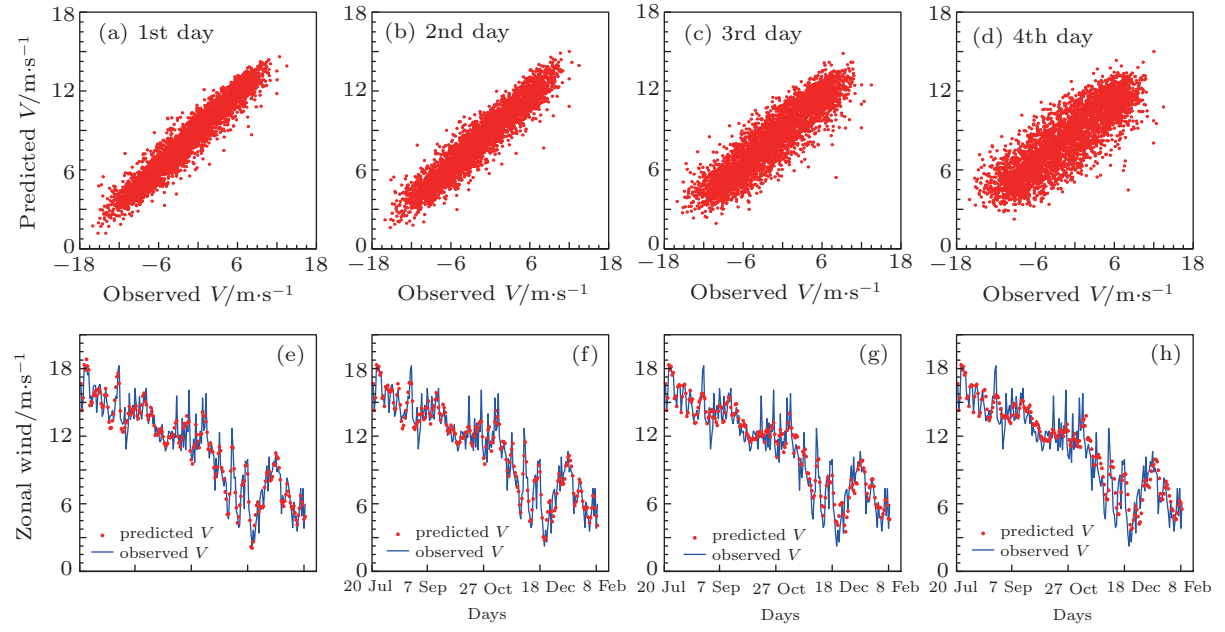


Fig. 4. (color online) Same as Fig. 3, but for the meridional wind.

In Table 3, we show the results of RMSEs for GA forecast and persistence model forecast for all the three points using zonal and meridional components from scatterometer observations. It is clear that the superiority of GA forecast is obvious and always exists for up to 4 days ahead forecast, which is required by many operational applications. The RMSE of persistence model prediction is always larger by about 0.9 to 2.5 m/s than GA prediction for 1-day to 4-day forecast. Also,

for the same day prediction there is a larger RMSE of the persistence model forecast and a larger improvement of the GA forecast. For example, for the 2-day prediction, the RMSEs of zonal and meridional components for persistence model prediction at P1 station are 4.56 and 4.25 m/s, the improvements of the GA prediction are 2.46 and 2.14 m/s, respectively; at the P2 station, the RMSEs for the persistence model prediction are 3.26 and 3.60 m/s, the corresponding improvements of GA prediction are 1.58 and 1.93 m/s, respectively.

Meanwhile, as shown in Table 3, the RMSE of GA algorithm for P1 station is larger than P2 and P3 points, just like the wind speed prediction. The reason is the same as the wind speed prediction due to the specificity of Taiwan strait. From the RMSEs of the 5-day forecast for GA prediction, which are not shown in the Table, we find that the RMSEs of zonal and meridional components for P3 point is 3.28 and 2.39 m/s, respectively, which is still smaller than the RMSEs of the 4-day forecast of P1 point. This means that the predictability of different stations in the SCS basin is different, some stations like P3 point that have higher predictability may be able to forecast more days.

The equations for 1-day forecast of wind zonal and meridional components at P2 station are also given in Appendix A.

Table 3. Comparison of predicted zonal and meridional wind components by GA and persistence model at the three locations.

			P1		P2		P3	
			Zonal	Merid.	Zonal	Merid.	Zonal	Merid.
Forecast (Days in Advance)	1 day	RMSE (pers.)	3.31	3.07	2.67	2.62	2.97	2.25
		RMSE (GA)	1.95	1.84	1.48	1.55	1.62	1.24
		R^2(GA)	0.85	0.88	0.92	0.93	0.92	0.91
	2 days	RMSE (pers.)	4.56	4.25	3.26	3.60	3.49	2.63
		RMSE (GA)	2.10	2.11	1.68	1.77	1.85	1.40
		R^2(GA)	0.82	0.85	0.89	0.91	0.89	0.89
	3 days	RMSE (pers.)	5.03	4.74	3.76	4.13	3.97	2.94
		RMSE (GA)	2.76	2.80	2.16	2.46	2.35	1.68
		R^2(GA)	0.69	0.73	0.82	0.83	0.82	0.84
	4 days	RMSE (pers.)	5.25	4.91	4.00	4.33	4.13	3.02
		RMSE (GA)	3.44	3.51	2.75	3.14	2.95	2.15
		R^2(GA)	0.52	0.57	0.71	0.72	0.72	0.74

Apart from persistence model prediction, one can also use an autoregressive model approach, which is a classical approach for time series modeling. However, the autoregressive approach is a linear approach and has been proven to be unable to improve upon the persistence prediction quite significantly for wind prediction in the Bay of Bengal BOB.[3] Hence, it is not necessary to compare the GA algorithm with the autoregressive approach in the present work.

6. Conclusion and discussion

In the present work, a technique known by the name of GA with SSA has been used to predict surface wind (including wind speed and wind components) in the SCS with scatterometer observations. Since the technique can in principle be used to predict a strictly deterministic time series, whereas the time series of scatterometer wind (WT) inevitable has noise, SSA has been applied to reduce the noise. Then, the GA has been carried out to predict the time series of the WTs (including the filtered WTs and the fluctuation part from SSA).

The GA technique has been used to carry out 1 to 4 days ahead forecast of wind speed and wind components for three points with scatterometer observations in the SCS basin. The forecast has been compared with persistence model forecast and it has been found that GA algorithm is able to significantly improve upon the persistence model forecast with leading times of 1 to 4 days in the SCS basin for both wind speed and wind components prediction. Meanwhile, the predictability for different station in the basin is different, for example, the P1 station that is influenced by Taiwan strait has less predictability, and the improvement of GA algorithm also shows large differences for different station. It can be concluded that the algorithm yields an encouraging performance and has the potential to be used by an operational agency which is interested in the ocean-state forecast of the SCS basin.

One should know that an important advantage of the GA algorithm is that it provides explicit analytical forecast equations for surface wind. Another added advantage is that much less input information is required by this algorithm compared with the sophisticated atmospheric models. In the future, this algorithm could be extended to wind field prediction and it will be tested with the scatterometer wind from the HY-2 satellite that was launched by China.

Appendix A

The analytical equations for 1-day forecasts of wind speed and wind components at location P2 are provided as follows:

$$
\begin{aligned}
\text{trends}(t) &= 2.7\times(\text{trs}(t-1)-\text{trs}(t-2))+\text{trs}(t-3),\\
\text{trendu}(t) &= \text{tru}(t-3)+2.72\times\left(\text{tru}(t-1)-\text{tru}(t-2)-\frac{\text{tru}(t-5)}{16.46\times(\text{tru}(t-1)+31.64)}\right),\\
\text{trendv}(t) &= 2.0\times\text{trv}(t-1)-\text{trv}(t-2)-2.42\left(\frac{\text{trv}(t-2)-\text{trv}(t-4)}{9.27+(\text{trv}(t-2)/(\text{trv}(t-2)-15.61))}\right),\\
\text{fluctv}(t) &= 0.16\times(2.5\times\text{flv}(t-2)+2.0\times\text{flv}(t-3)+\text{flv}(t-4)+\text{flv}(t-5)+\text{flv}(t-6)),
\end{aligned}
$$

$$
\begin{aligned}
\text{WTs}(t) &= \text{trends}(t)+\text{fluctv}(t),\\
\text{WTu}(t) &= \text{trendu}(t)+\text{fluctu}(t),\\
\text{WTv}(t) &= \text{trendv}(t)+\text{fluctv}(t).
\end{aligned}
$$

Here, $\text{WTs}(t)$, $\text{WTu}(t)$, and $\text{WTv}(t)$ are the predicted wind speed, wind zonal component, and wind meridional component at the tth day separately. The $\text{trends}(t)$, $\text{trendu}(t)$, and $\text{trendv}(t)$ are the predictions of the nonlinear trends of wind speed, wind zonal, and meridional components, while $\text{trs}(t-1)$, $\text{tru}(t-1)$, and $\text{trv}(t-1)$ are the corresponding nonlinear trends at day $(t-1)$ separately. These trends have been computed with SSA. The $\text{fluctv}(t)$, $\text{fluctu}(t)$, and $\text{fluctv}(t)$ are the forecasts of the fluctuations of wind speed, wind zonal, and meridional components, and $\text{fls}(t-1)$, $\text{flu}(t-1)$, and $\text{flv}(t-1)$ are the corresponding fluctuations at day $(t-1)$. Similar notation holds good for other days.

Acknowledgment

The authors would like to thank the IFREMER in France for providing the QuikSCAT wind data used in this paper.

References

[1] Sharma R, Sarkar A, Agarwal N, Kumar R and Basu S 2007 *Geophys. Res. Lett.* **34** L03603

[2] Zhong J, Fei J F, Chen X P and Huang X G 2014 *Science China: Earth Sciences* **57** 2690

[3] Basu S, Satheesan K, Sarkar A and Kishtawal C M 2005 *Geophys. Res. Lett.* **32** L24601

[4] Sinha M, Rao A D and Basu S 2013 *J. Oceanogr.* **69** 117

[5] Liu H, Liu D and Deng L F 2006 *Chin. Phys. B* **15** 01196

[6] Ding G, Zhong S S and Li Y 2008 *Chin. Phys. B* **17** 01998

[7] Szpiro G 1997 *Phys. Rev. E* **55** 2557

[8] Yan H, Wei P and Xiao X C 2009 *Chin. Phys. B* **18** 03287

[9] Bilal S, Qureshi M L and Shafqatullah L 2014 *Chin. Phys. B* **23** 030502

[10] Alvarez A 2003 *J. Atmos. Ocean Technol.* **20** 717

[11] Guo J and He Y J 2012 *Chin. J. Oceanol. Limnol.* **30** 336

[12] Sheng Z 2012 *Acta Phys. Sin.* **61** 219401 (in Chinese)

[13] Alvarez A, Lopez C, Riera M, Hernandez G E and Tintore J 2000 *Geophys. Res. Lett.* **27** 2709

[14] Kishtawal C M, Basu S, Patadia F and Thapliyal P K 2003 *Geophys. Res. Lett.* **30** 2203

[15] Alvarez A, Orfila A and Tintore J 2004 *J. Geophy. Res.* **109** C03023

[16] Remya P G, Kumar R and Basu S 2012 *Ocean Engineering* **40** 62

[17] Zhong J, Huang S X, Du H D and Zhang L 2011 *Chin. Phys. B* **20** 034301

[18] Zhong J, Huang S X, Fei J F, Du H D and Zhang L 2011 *Chin. Phys. B* **20** 064301