Ocean Sci., 9, 171–182, 2013
www.ocean-sci.net/9/171/2013/
doi:10.5194/os-9-171-2013

Ocean Science
Open Access

A clustering analysis of eddies' spatial distribution in the South China Sea

J. Yi[1], **Y. Du**[1], **X. Wang**[1,2], **Z. He**[3,4], **and C. Zhou**[1]

[1]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Science and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[2]Geomatics College, Shandong University of Science and Technology, Qingdao 266510, China
[3]State Key Laboratory of Tropical Oceanography (LTO), South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, Guangdong, China
[4]College of Ocean and Earth sciences, Xiamen University, Xiamen 361005, China

Correspondence to: Y. Du (duyy@lreis.ac.cn)

Received: 19 September 2012 – Published in Ocean Sci. Discuss.: 7 November 2012
Revised: 23 January 2013 – Accepted: 31 January 2013 – Published: 22 February 2013

Abstract. Spatial variation is important for studying the mesoscale eddies in the South China Sea (SCS). To investigate such spatial variations, this study made a clustering analysis on eddies' distribution using the K-means approach. Results showed that clustering tendency of anticyclonic eddies (AEs) and cyclonic eddies (CEs) were weak but not random, and the number of clusters were proved greater than four. Finer clustering results showed 10 regions where AEs densely populated and 6 regions for CEs in the SCS. Previous studies confirmed these partitions and possible generation mechanisms were related. Comparisons between AEs and CEs revealed that patterns of AE are relatively more aggregated than those of CE, and specific distinctions were summarized: (1) to the southwest of Luzon Island, AEs and CEs are generated spatially apart; AEs are likely located north of 14° N and closer to shore, while CEs are to the south and further offshore. (2) The central SCS and Nansha Trough are mostly dominated by AEs. (3) Along 112° E, clusters of AEs and CEs are located sequentially apart, and the pairs off Vietnam represent the dipole structures. (4) To the southwest of the Dongsha Islands, AEs are concentrated to the east of CEs. Overlaps of AEs and CEs in the northeastern and southern SCS were further examined considering seasonal variations. The northeastern overlap represented near-concentric distributions while the southern one was a mixed effect of seasonal variations, complex circulations and topography influences.

1 Introduction

The South China Sea (SCS) is the largest semi-enclosed marginal sea in northwest Pacific Ocean. It has a large NE–SW oriented abyssal basin with greatest depth of 5567 m. There is an approximate $3.5 \times 10^6\,\mathrm{km}^2$ of total area with mean depth of 1212 m. The SCS is well linked to the adjacent seas through several straits or channels, connecting to the Pacific Ocean through the Luzon Strait, to the Sulu Sea through the Mindoro Strait, to the East China Sea through the Taiwan Strait and to the Java Sea through the Karimata Strait and so on.

Mesoscale eddies in the SCS form a hot research topic in recent years. Much work tries to investigate their dynamic characteristics and formation mechanisms, especially using statistical approaches (Wang et al., 2003; Wu and Chiang, 2007; Xiu et al., 2010; Li et al., 2011). Spatial variation is important for statistical analysis because patterns or rules in different regions are generally distinguished from each other. While some researchers focus their studies on particular regions of SCS (Fang et al., 2002; Wu and Chiang, 2007; Yuan et al., 2007; Li et al., 2003; Yang et al., 2000; Yuan and Li, 2008), some others study the whole maritime area of SCS by geographical divisions (Li et al., 2011; Wang et al., 2003; Lin et al., 2007; Chen et al., 2011). Wang et al. (2003) first proposed a four-zone division for the SCS according to limited knowledge of eddy generation mechanisms, and used it to group eddies for analysis. Chen et

Published by Copernicus Publications on behalf of the European Geosciences Union.

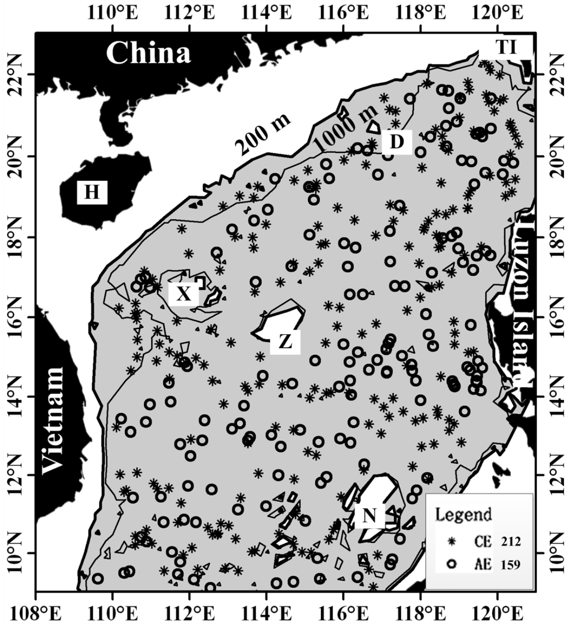

Fig. 1. Map of study area. Birth locations of cyclonic and anticyclonic eddies are symbolized by the asterisk and circle, respectively, and the numbers of them are showed in the legend. Black solid lines indicate 200-m (bold) and 1000-m (thin) isobaths. Some geographic names are represented by initials. H = Hainan Island; TI = Taiwan Island; D = Dongsha Islands; X = Xisha Islands; Z = Zhongsha Islands; N = Nansha Trough.

al. (2011) and Li et al. (2011) also discussed their statistical findings of different geographic zones. However, few studies have explored the spatial characteristics embedded in eddies' generation data or discussed the spatial pattern reflected by their locations. It is interesting to compare this spatial pattern with current geographical divisions and related studies.

So, in this paper, we try clustering analysis to interpret the spatial variations in eddy generation and resolve clusters to known mechanisms or speculations. K-means, a classical centre-based approach, is adopted to perform the clustering analysis. Four questions will be answered in this paper: (a) do eddies in the SCS have clustering features in their generation locations? (b) Are there different characteristics between generation locations of AEs and CEs? (c) How many clusters are evident in these generation locations? (d) What are the spatial patterns of AEs and CEs?

The paper is organized as follows. Section 2 briefly introduces the eddy data and clustering method. Section 3 presents clustering results combined with extensive discussion of the clustering characteristics and spatial patterns. Section 4 summarises the whole paper.

2 Data and method

2.1 Eddy data

The sea surface height (SSH) data used for eddy detection are produced by the operational 1/32° global Naval Research Laboratory's (NRL) Layered Ocean Model (NLOM) nowcast/forecast system (http://www7320.nrlssc.navy.mil/global_nlom32/scs.html). The model assimilates SSH from 3 satellite altimeters (ENVISAT, JASON-1 interleaved and JASON-2). Altimeter track data over a 3-day window are assimilated each day.

The reliability of the model performance has been validated by Du et al. (2011). Comparison between anomaly of SSH (SSHA) derived by averaging NLOM SSH data over 2003–2009 and sea level anomalies (SLA) observations provided by Archiving Validation and Interpretation of Satellite Data in Oceanography (AVISO) showed that deterministic eddies found in SLA were well reflected by the SSHA. Comparison of current speed simulated by NLOM with cruise measurements also confirms the data validation. For more details refer to Du et al. (2011).

The eddy identification algorithm adopted in this paper is an adaptation of SSH-based automated procedure proposed by Chelton et al. (2011). Brief descriptions can be found in Appendix A and Wang et al. (2012). The 1/32° NLOM numerical output started from 1 November 2003. But SSH data were available from 28 April 2005 on NLOM Live Access Server (http://apdrc.soest.hawaii.edu/las/getUI.do). So using the output SSH data from 2005 to 2011, we identified and tracked 371 eddies (159 AEs and 212 CEs). The map of study area and birth locations of these eddies are showed in Fig. 1. To be noted that only the centre of each eddy's birth state, which denotes where it is generated, is used in the clustering analysis.

2.2 Clustering analysis

The goal of clustering analysis is to group data into meaningful subgroups (clusters) (Jain and Dubes, 1988; Kaufman and Rousseeuw, 1990). The more the similarity (homogeneity) within a cluster and the greater the difference between clusters, the better or more distinct the clustering patterns are (Tan et al., 2005). As eddies' distribution has certain concentration trends which is mentioned in (Huang et al., 1992; Wang et al., 2003), we choose the K-means approach, which is simple and capable of detecting center-based patterns. For the purpose of brevity, only basic ideas and important facts of this approach are presented. Details of concrete algorithms can be found in (Jain and Dubes, 1988; Tan et al., 2005).

K-means was first proposed by MacQueen (1967). It defines prototypes in terms of centroids. K is the desired or specified number of clusters. First, K initial centroids are randomly selected. Data points are assigned to the nearest centroid to form initial clusters. Second, re-compute the

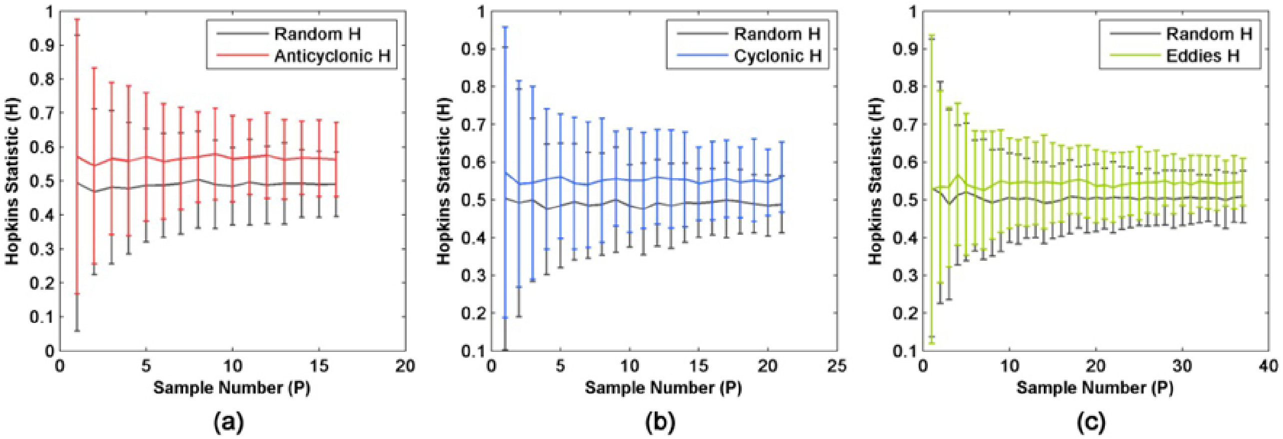

Fig. 2. Error bar plot of Hopkins statistic. Error bar is twice the standard deviation, and curve is mean of H. **(a)** Hopkins statistic of AE and random sample. **(b)** Hopkins statistics of CE and random sample. **(c)** Hopkins statistics of all eddies and random sample.

centroids by averaging the belonging data points and assign them to these new centroids again. Repeat this process until no point changes assignment, then we get the final clusters.

K-means dutifully produces clusters no matter whether the data exhibit an aggregative predisposition or not. To prevent meaningless results, it is essential to evaluate the "clustering tendency" of the eddy data. Hopkins statistic is applied for this evaluation. Specific information about clustering tendency and Hopkins statistic are described in Appendix B.

Another important fact is that K-means is a heuristic approach which produces good but not optimal clusters. Objective functions are required to measure the performance and help find out unbiased and relative "best" clustering results. In this paper, the sum of the squared error (SSE) is used to measure the cohesion of cluster, and the between-groups sum of squares (SSB) is used to measure the separation. Silhouette coefficient (SC), another adopted parameter, combines cohesion and separation in evaluating cluster validity. More details about evaluating the goodness of clustering and these objective functions are described in Appendix C.

The number of clusters is an important input for K-means. In this study, it means how many places are eddies usually generated in. Since there are no exact answers to this question, we need to find a way out that could suggest objective answers. Given the assumption that good performance will be achieved if the data are clustered by right cluster number, optimal values of evaluation parameters like SSE or SC usually hint the answers. In other words, the knees of the SSE curve or the peaks of the SC curve shall indicate the possible right cluster number.

3 Results and discussion

First, the Hopkins statistic results are presented to answer the following questions. (a) Has the eddies' distribution a clustering tendency? (b) What is the correct cluster number? The latter is discussed separately for cyclonic eddy, anticyclonic eddy and the whole. The third part gives an intensive discussion about how AE and CE are different in spatial variations.

3.1 Hopkins statistic

It is suggested that the amount of sample data (P) be less than 10 % of the whole when computing the Hopkins statistic (Ripley, 1977). To facilitate observation, we made a series of experiments whereby increasing the sample data from 1 to 10 % of the total amount. In each calculation of Hopkins value (H) 100 trials are carried out, and the mean values are plotted in Fig. 2. Results of random pattern are presented for comparison.

H value of random patterns fluctuates around 0.5, and becomes stable at 0.5 with accretion of the sample amount as expected. Comparatively, H values of AEs, CEs and all eddies are clearly above 0.5 axis. The H mean for AEs, CEs and all eddies at the largest sample amount are 0.563, 0.560 and 0.547, respectively. But Banerjee and Dave suggested that the H value of a clustered data set should be greater than 0.7. In our opinion, high H values indicate obvious aggregated characteristics, while lower H values, which exceed 0.5, truthfully reflect the weak tendency of eddies' distribution and difficulty for clustering. So, the clustering analysis should be carried out very carefully and objectively to derive a reliable result from plenty of experiments.

To definitely prove that eddies' distribution is not random, we applied the *t*-test to H values calculated at the largest sample amount. The results are showed in Table 1. The values in the last column show that no matter AEs or CEs or the whole, the assumption that the mean H equals to that of random is denied significantly at the 0.05 level. So, we conclude that eddies' distribution in the SCS shows a weak clustering tendency.

Table 1. *T*-test analysis.

	p	Descriptive statistics			Probability of Lilliefors Test[b] for Normal Distribution	Probability of Levene's Test[c] for Equality of Variance	Probability of t-test for Equality of Means
		N	Mean	Standard deviation			
AEs	16	100	0.563	0.055	0.060	0.046	< 0.001[a]
Random		100	0.491	0.047	> 0.200		
CEs	21	100	0.560	0.047	> 0.200	0.044	< 0.001[a]
Random		100	0.488	0.038	> 0.200		
All	37	100	0.547	0.031	> 0.200	0.331	< 0.001
Random		100	0.508	0.034	> 0.200		

[a] Denotes the *t*-test result is computed under the assumption of unequal variance; [b] Lilliefors test is used to test the null hypothesis that data come from a normally distributed population (Lilliefors, 1967); [c] Levene's test is an inferential statistic used to assess the equality of variances (Brown and Forsythe, 1974).

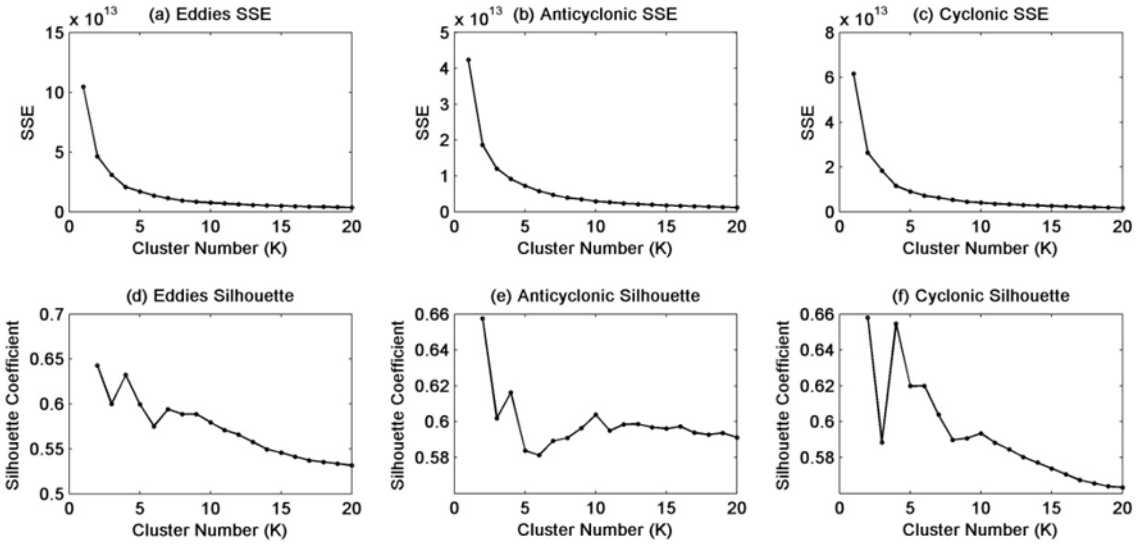

Fig. 3. Plots of sum of squared error and silhouette coefficient.

3.2 Correct number of clusters

To determine the correct number of clusters, we apply K-means by specifying K from 1 to 20 and replicate 1000 times on every K. Figure 3 shows the plots of the median of SSE and silhouette coefficient against variable K.

As mentioned in Sect. 2.2, the knees of the SSE curve or the peaks of the SC curve hint at the possible right cluster number. But as we can see in Fig. 3, it is not easy to decide which is the optimal value, for there are more than one local peak and curve knee. For example, 4, 5, 6 and 7 can all be viewed as the knees of the SSE curve of all eddies (Fig. 3a), while 2, 4, 7 and 9 are reasonable peaks of the SC curve of all eddies (Fig. 3d). It should be known that the spatial pattern of eddy generation may contain a hierarchical clustering structure, and these multiple optima are direct indicators. If the partition goes too coarse, valuable spatial variations will be obscured. If it goes too fine spatial variations would be wrongly torn apart and cause difficulties for interpretation. A practical solution to finding the correct number, not too coarse and not too fine, is to take into consideration both the knees of SSE curve and peaks of SC curve. The lower the SSE, the better the partition. But we do not go to extremes since the curve of SSE is approximating zero axis and will equal 0 when cluster number equals the number of data points. The higher the local peak of SC curve the better it is. So we select as correct the number of clusters for which the SSE is small and the SC is a local maximum and as large as possible.

Based on this criterion, the correct cluster number of all eddies is four. $K = 2$ is discarded for the SSE is the largest. The SC value of $K = 4$ is the second maximum and the SSE value is relatively small. The correct number of AEs and CEs are also four for the same reason. It is interesting that the cluster number equals the number of geographic zones

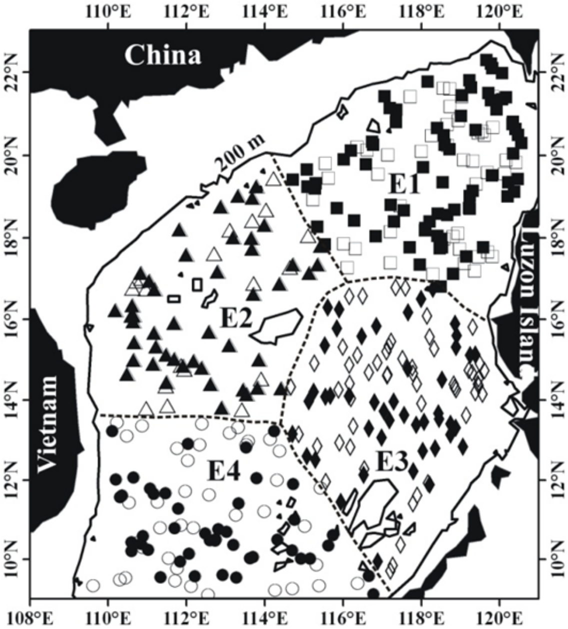

Fig. 4. Four-cluster results. The dashed lines schematically depict the boundary of the four regions. Symbols of anticyclonic eddies are unfilled while those of cyclonic eddies are filled with black. Different groups are distinguished by different geometry shapes.

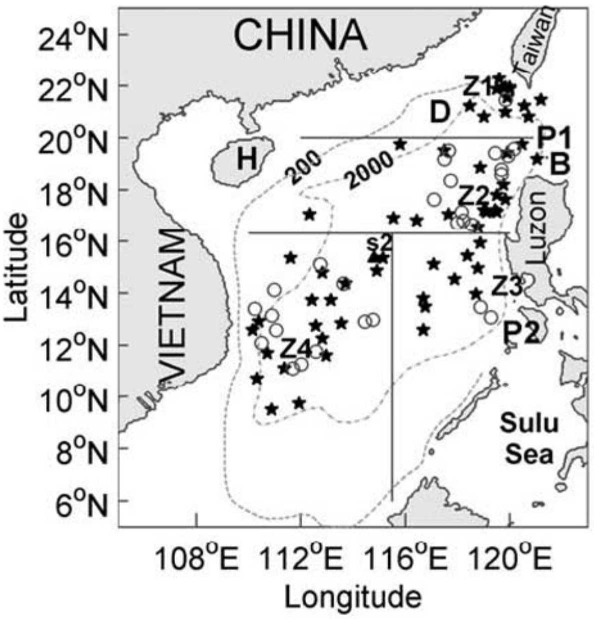

Fig. 5. Four geographic zones of SCS, adopted from Wang et al. (2003). Circles denote cyclonic eddies, and stars anticyclonic eddies.

in Wang et al. (2003). Comparisons of them are detailed in Sect. 3.3.1.

Besides, alternative choices are carefully made to gain finer clustering results which may unveil more details of the spatial variations. The next local maximum after $K = 4$ in each SC curve is selected as the alternative choice. That is, $K = 7$ for all eddies, $K = 10$ for AEs, and $K = 6$ for CEs.

3.3 Clustering results

As a matter of fact, K-means can only guarantee a local minimum by optimizing the objective function to produce relative better and acceptable clustering results. Global optimization achieved by exhaustive searches of all possible choices is unrealistic in applications. So, in this study, the clustering procedure is executed 100 times with K random initial centroids each time. And the objective function we adopted is defined as follows:

$$\mathrm{O}(i) = \mathrm{SSE}'(i) - \mathrm{SSB}'(i) - \mathrm{SC}'(i) \qquad (1)$$

Where i denotes the i^{th} time execution, and SSE$'$, SSB$'$ and SC$'$ denote the Gaussian normalized value of SSE, SSB and SC, respectively. The result that minimizes the objective function is selected for the following analysis and discussion.

3.3.1 Four-cluster result

Here, the four-cluster result for all eddies (Fig. 4) derived from K-means clustering is compared with the four geographic partition of SCS (Fig. 5). In Fig. 4, the four clusters are labeled by E1 (west of Luzon Strait), E2 (around the Xisha Islands), E3 (southwest of Luzon Island), and E4 (southern SCS). E3 and E4 are much overlapped with Z3 and Z4 of Wang et al. (2003). This pair represents eddies generated to the southwest of Luzon Island and southeast of Vietnam. But the northern boundary of E4 is about 13.5° N, whereas Z4 is at 16° N. E2 representing eddies generated around the Xisha Islands has no reasonable matches in the 4 zones. Possible generation mechanisms of eddies in this zone were not discussed in Wang et al. (2003). But previous studies found anticyclonic eddies generated around this area (Huang et al., 1992; Zhong, 1990; Guan, 1997), and they are likely to be caused by frontal instability and modulated by topographic influences (Chai et al., 2001). The region to the west of Luzon Strait, where Kuroshio intrusion happens, is densely populated by eddies. While the clustering method groups these eddies into E1, Z1 and Z2 divide this area according to different generation mechanisms. Eddies generated in Z1 are probably caused by frontal instability at the Kuroshio intrusion (Wang et al., 2000; Su, 2004). And Qu et al. (2000) found intense positive wind stress curl would generate cyclonic eddies in eastern Z2.

Based on these comparisons, we conclude that both spatial partitions confirm regional differences, but eddies generated around Xisha Islands are not well reflected by 4-zone separation, yet 4-cluster separation fails to distinguish Z1 and the nearby eastern Z2 where the mechanisms are supposed to be different.

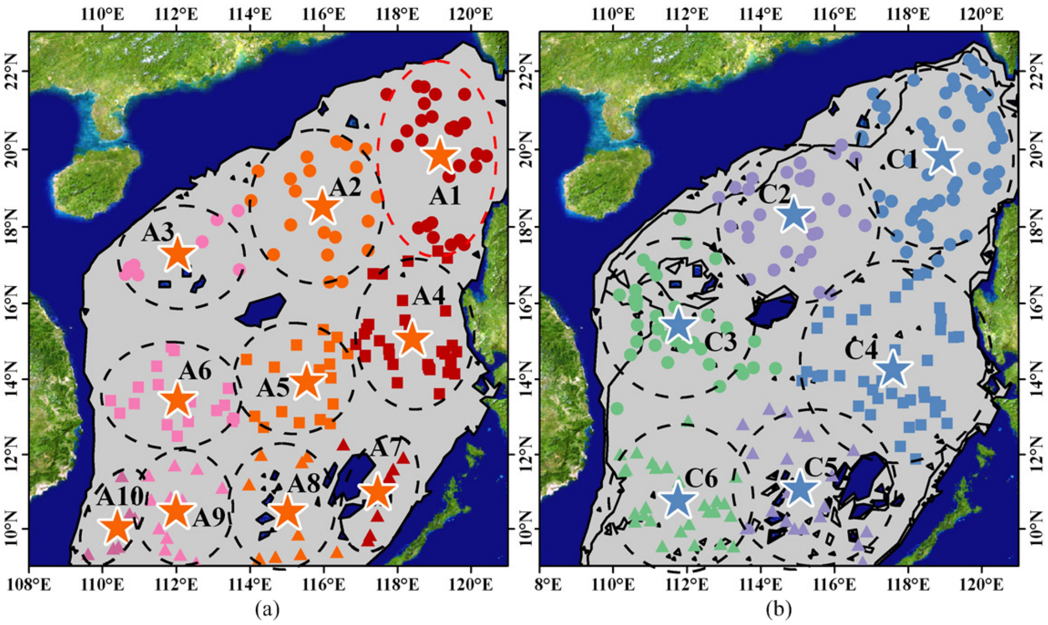

Fig. 6. (a) Ten-cluster result for anticyclonic eddies. Orange stars are cluster centroids. Black dashed lines outline approximate area of clusters. Different groups are distinguished by different colors or symbols. **(b)** Six-cluster result for cyclonic eddies. Blue stars are cluster centroids. Black dashed lines show approximate area of clusters. Different groups are distinguished by different colors or symbols.

Besides, some clear distinctions between AEs and CEs can be recognized from the 4-cluster results. In area E3, AEs concentrate to the north of 14° N, but CEs are mainly to the south. And a few AEs are generated regularly along the Nansha trough. In area E1, some AEs generated near the Luzon Strait are found to the northwest of Luzon Island in Z2 and the rest are mainly focused on the southwest of Taiwan Island in Z1. So, finer clustering results are expected to uncover these distinctive characteristics hidden in the four clusters.

3.3.2 Finer clustering results

We use a larger cluster number to gain the finer clustering results of AEs (Fig. 6a) and CEs (Fig. 6b). Ten clusters of AEs shown in Fig. 6a are divided into three groups: A1, A2, and A3 together represent the northern SCS, A4, A5, and A6 represent the central SCS, and A7, A8, A9, and A10 represent the southern SCS.

First, in northern SCS, A1 represents those AEs generated near the Luzon Strait. This group is situated in the most complex area of the SCS. The Luzon Strait is the deepest channel for water exchange between the SCS and Pacific Ocean. The Kuroshio frequently intrudes into the SCS and influences its circulation, producing AEs southeast of the Dongsha Islands (Li et al., 1998) and southwest of Taiwan Island (Wang and Chen., 1987). Part of the eddies generated northwest of Luzon Island is still improperly included in A1 due to the weakness of K-means in handling data of different density. Concentrations there are supported by Cheng et al. (2005) and Wang et al. (2003). The Ekman transport driven by summer wind stresses is a possible cause for these AEs.

A2 represents those AEs located southwest of the Dongsha Islands. This distribution characteristic is shown in many statistical studies (Wang et al., 2003; Cheng et al., 2005; Lin et al., 2007; Hwang and Chen, 2000). Model results showed that eddies may be caused by the wind stress curl or the Kuroshio intrusions (Cai et al., 2002a, b).

As for A3, Chai et al. (2001) reported three anticyclonic eddies to the east of Hainan Island and frontal instability and topography influences were the causes. Other studies observed eddies moving around. Zhong et al. (1990) found many AEs in this area based on hydrological survey near the northern continental shelf of SCS from 1975–1984. Li et al. (2011) also identified AEs in this area from Lagrangian drifter data from 1979 to present.

Second, in the central SCS, A4 represents the dense cluster southwest of Luzon Island, between 14° N and 16° N. Wang et al. (2008) noted the orographic wind jets associated with the winter monsoon wind and the mountainous topography along the eastern boundary of SCS can spin up AEs and CEs. A5 represents AEs generated in the SCS center; however, while these AEs were also confirmed in Wang et al. (2003) statistical study, generation mechanisms about them were few. A6 represents part of the dipole structure off the Vietnam coast. By principal component analysis of SLA and wind stress curl, Shaw et al. (1999) found that the generation of this dipole structure was related to an eastward jet following the line of zero curl. Wang (2004) and Wang et al. (2006) suggested vorticity transports from the nonlinear effect of the western boundary currents were crucial for the generation of the dipole structure.

Third, in the southern SCS, four clusters are produced. A7 well captures the AEs along the Nansha Trough. A8 groups the eddies generated around the Nansha Islands. A9 and A10 south of 12° N may be viewed as the different patterns of the anticyclonic eddy of the dipole structure off the Vietnam coast. In previous studies, Cai et al. (2002a) suggested that interaction between strong barotropic shelf currents and the local topography can generate anticyclonic eddies over the deep trough (A7) in winter using a coupled single-layer and two-layer hybrid model. Fang et al. (1997) pointed out that seasonal monsoon was a important driving force for the southern gyres and eddies. Guan and Yuan (2006) analyzed eddies generated in southwest of SCS (A9 and A10) using SSHA from TOPEX/ERS and found the seasonal variations were strong, especially between July and December.

Figure 6b illustrates the six clusters of CEs. C1 represents CEs generated near the Luzon Strait. Compared with A1, CE locations are so scattered that no distinguishable patterns are recognized between northwest of Luzon Island and southwest of Taiwan Island. Many studies show that CEs northwest of Luzon Island have strong seasonal variations (Fang et al., 1998; Nitani, 1970; Qu et al., 2000), which explains the dispersed distribution of CEs. Qu et al. (2000) suggested the intense positive wind stress curl offshore northwest of Luzon Island was likely mechanism. But Liu and Su (1992) found the positive vorticity advected westward from the Kuroshio front may produce cyclonic eddies as well.

C2 represents CEs generated southwest of the Dongsha Islands, where CEs were frequently identified by previous studies (Su et al., 1999) and the mechanism is basically the same as in A2. CEs in the C3 are mainly generated north of 14° N, and form the northern part of the dipole structure off the Vietnam coast. CEs in C4 concentrate between 12° N and 14° N to the southwest of Luzon Island, which also has been reported in Wang et al. (2003) and Cheng et al. (2005). The mechanism is the same as in A4 and in-depth discussion can be referred to (Wang et al., 2008). C5 and C6 represent CEs formed in the cyclonic gyres of the southern SCS, south of 12° N. For corresponding mechanism studies refer to Guan and Yuan (2006) and Fang et al. (1997).

3.3.3 Comparisons between AEs and CEs

Figure 7 combines the centroids of AEs and CEs to show how their spatial patterns differ. The spatial dimension of each cluster is represented by a circle of which the radius is defined by the mean distance from these eddies' generation locations to the centroid.

First, more clusters and smaller spatial domains of AEs indicate patterns of AE are relatively more aggregated than those of CE. Second, differences between them can be summarized as follows: (1) to the southwest of Luzon Island, AEs and CEs are basically generated apart; AEs are likely to the north of 14° N and closer to the shore, whereas CEs are to the south and 100 trials are carried out from shore. (2) The central SCS and the Nansha Trough are dominated by AEs. (3) Clusters A3, C3 and A6 are located sequentially apart, along 112° E, and C3-A6 represent the dipole eddies off Vietnam; A9 and A10 may also be the south part, the anticyclonic eddy, of the dipole. (4) Southwest of the Dongsha Islands, AEs are concentrated to the east of CEs.

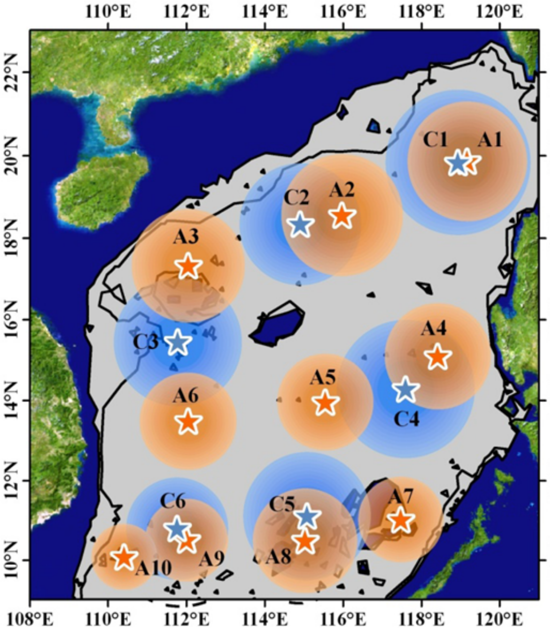

Fig. 7. Centroids and spatial domains of cyclonic and anticyclonic eddies. Centroids of AEs and CEs are shown by stars of different colors. The circles in different colors represent the spatial dimension of each cluster, and the radius is defined by the mean distance from these eddies' generation locations to corresponding centroid.

Beside these distinctions, AEs and CEs are overlapped in northeastern and southern SCS. Do these overlaps simply explain the similarities there, or are there any hidden differences behind the overlaps? To answer these questions, we reexamined the eddies with regard to seasonal monsoons. The winter monsoon is dominated by northeasterly wind while the summer monsoon is dominated by southwesterly wind (Hellerman and Rosenstein, 1983). So we remapped these eddies in northeastern (Fig. 8) and southern SCS (Fig. 9) according to the two monsoon seasons.

In northeastern SCS, eddies generated during different monsoon seasons show similar distribution patterns. Specifically, the center of northeastern SCS is dominated by AEs which are surrounded by CEs generated around Xisha, northwest of Luzon and southwest of Taiwan Island. The similarities indicate that those anticyclonic rings shedding from the loop currents caused by the Kuroshio intrusions may occur during any season of the year, which are consistent with the

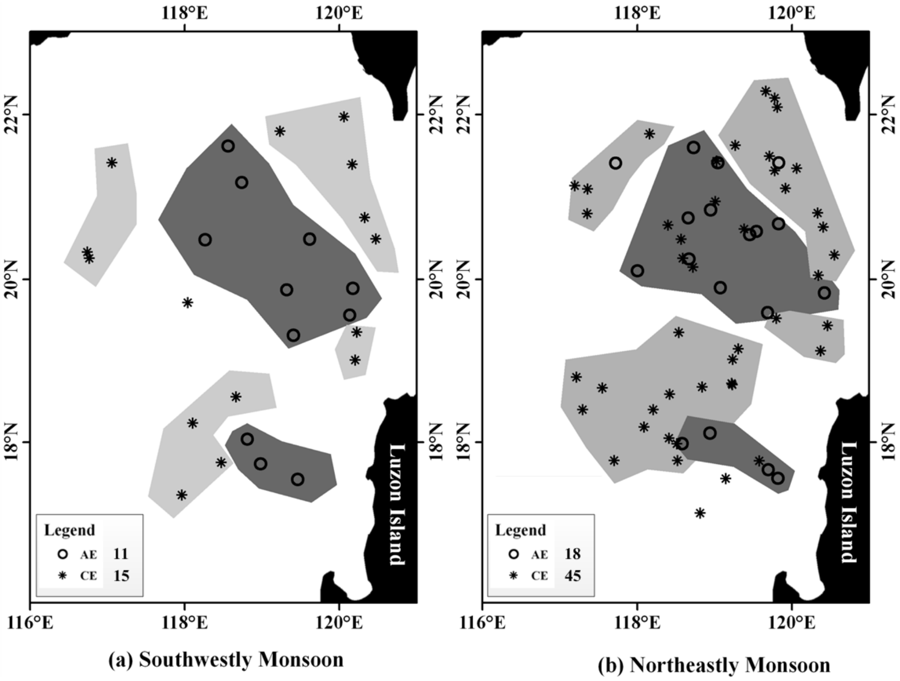

Fig. 8. Eddies of A1 and C1 generated in northeastern SCS during different monsoon seasons. Concentration regions of AEs and CEs are represented by different colors.

study of Yuan et al. (2006), and that the CEs nearby could be induced by the instability of the horizontal shear stress of velocity around the loop currents. On the other hand, during northeasterly monsoon some CEs exist in the center area of northeastern SCS where few CEs are identified during southwesterly monsoon. Such difference could be due to different patterns of the Kuroshio intrusions. Yuan et al. (2006) pointed out that the anticyclonic intrusions of the Kuroshio are more frequent in summer than in winter, but in winter the intrusion path would turn cyclonically toward the southwest at the center of northeastern SCS. As a result, the turns may probably be recognized as CEs which weakened the dominance of AEs in this area in winter. In summary, the spatial patterns of AEs and CEs in northeastern SCS are mainly controlled by the Kuroshio intrusions. The impact seasonal variations make on the pattern is not strong. As for the overlap of the centroids, it is precisely caused by the concentric distribution of AEs and CEs.

In the southern SCS, AEs are preponderant in southwesterly monsoon while CEs are preponderant in northeasterly monsoon. The spatial patterns are quite different in the two monsoon seasons. During southwesterly monsoon, AEs are concentrated to the southeast of Vietnam between 110° E and 112° E, while CEs are mainly located nearby between 112° E and 114° E. During northeasterly monsoon, CEs dominate almost the whole area, AEs are scattered. Distributions of AEs and CEs are not entirely reversed with the monsoon change. Hence, overlaps in southern SCS cannot be explained only by seasonal variations. Fang et al. (1997) thought one eddy observed in the southwestern SCS during southwest monsoon season was not produced directly by wind forcing. Instead, local baroclinic instability, boundary current on the slope and the topography there can be other important causes as well.

4 Conclusions

In this paper, we carried out a clustering analysis using K-means approach on 371 eddies to investigate the spatial variations among their birth locations. We carefully examined the clustering tendency of eddies' distribution by Hopkins statistic. Results showed their aggregation is weak but not random. The correct number of clusters for AEs, CEs and the whole are indicated by optimal values of SSE curve and Silhouette coefficient. $K = 4$ is the common optimal for all and it equals the number of Wang et al. (2003) geographic partitions. On the other hand, since clustering results may contain hierarchical characteristics, we retained alternative choices of cluster number for finer results and in-depth analysis.

Comparisons between 4-cluster results and 4 geographic zones confirmed regional differences in eddies' distribution. But neither of them well characterized the differences in space and between AEs and CEs. Eddies generated around Xisha Islands are not well reflected by 4-zone separation, meanwhile 4-cluster separation fails to distinguish Z1 and the nearby eastern Z2 where the mechanisms are supposed to be different.

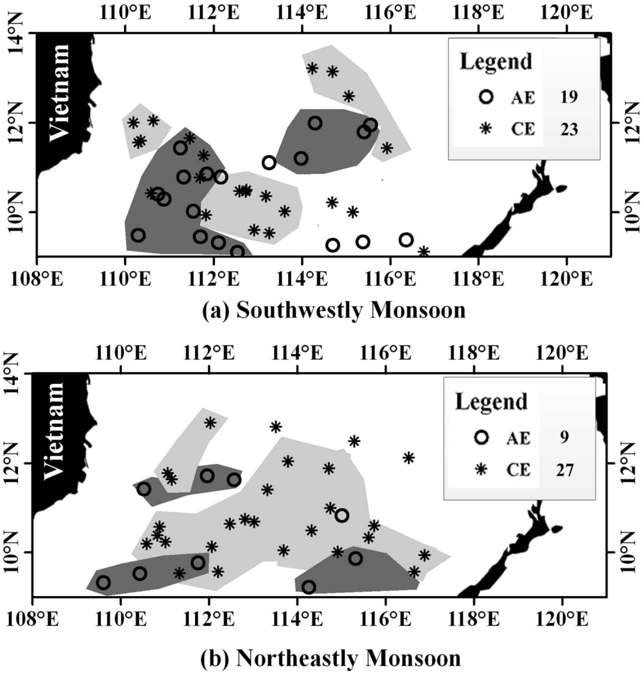

Fig. 9. Eddies of A8, A9, A10, C5 and C6 generated in southern SCS during different monsoon seasons. Concentration regions of AEs and CEs are represented by different colors.

Finer clustering results showed 10 regions where AEs were densely populated and 6 regions for CEs. Previous studies confirmed these partitions and possible generation mechanisms were related. Comparisons between AEs and CEs revealed that patterns of AEs are relatively more aggregated than those of CEs, and distinctions were summarized as follows: (1) southwest of Luzon Island, AEs and CEs are basically generated apart; AEs are likely north of 14° N and closer to the shore, whereas CEs are south and farther from shore. (2) The central SCS and the Nansha Trough are dominated by AEs. (3) Clusters A3, C3 and A6 are located sequentially apart, along 112° E, and C3–A6 represent the dipole eddies off Vietnam. (4) Southwest of the Dongsha Islands, AEs concentrate east of CEs.

Overlaps of AEs and CEs in the northeastern and southern SCS were further examined considering seasonal variations. Spatial patterns of AEs and CEs in northeastern SCS were mainly controlled by the Kuroshio intrusions. The impact of seasonal variations is not significant. Overlap was primarily caused by concentric characteristic of eddies' distribution, and spatial patterns of AEs and CEs in southern SCS were not entirely reversed with the monsoon wind. Overlaps cannot be explained only by seasonal variations; complex circulations and topography there can be other important causes as well.

Finally, the importance of spatial variations in surveying eddies motivates us to investigate patterns as well as distinctions of eddies' distribution. Different generation mechanisms are expected to be reflected by proper geographic partitions, which will also facilitate interpreting eddies' dynamic characteristics in statistical studies. However, such partitioning may not always be the case. Some areas may have different generation mechanisms coexisting. In that case, spatial clustering analysis will fail to distinguish them by geographical partitions. So, further studies are required for better understanding the generation mechanisms of mesoscale eddies in the SCS. Besides, as an eddy is a dynamic phenomenon with evolution processes, the generation is only starting part of the whole evolution. An investigation of the spatiotemporal characteristics of eddies' evolution processes is very necessary in the future.

Appendix A

Eddy identification procedure

The identification procedure applied in this paper is an adaption of that used in Chelton et al. (2011). Eddies' boundaries are delineated by contours of SSH and the centers are represented by local extrema. First, to define the boundary, contours are constructed at 1-cm intervals for instantaneous SSH field. A 1-cm interval yields a good compromise between minimum resolvable eddy amplitude and well-defined and compact eddy interiors (Chelton et al., 2011). Then, those closed contours that satisfy the following criteria are kept for next step:

1. The shape is similar to circle or ellipse. A shape parameter, SI, is introduced to measure the similarity. It is defined as follows:

   $$\mathrm{SI} = \frac{S}{\pi \cdot \left(\frac{D}{2}\right)^2}, \tag{A1}$$

   where S denotes the area of the contour, D denotes the maximum distance between any pairs of points on the contour. Contours whose $\mathrm{SI} > 0.4$ are retained.

2. Diameter of the eddy, which is defined by the diameter of an circle with equivalent area, is limited from 45 to 500 km.

3. The distance between any pair of points along the contour must be less than a specified maximum. The maximum-distance threshold is 400 km for latitudes above 25°, and is increased linearly to 1200 km at the equator (Chelton et al., 2011).

Next, among these selected contours, spatial containment relationship is calculated by program interfaces provided by the commercial ArcGIS Engine Software Develop Kit (http://www.esri.com/software/arcgis/arcgisengine). Those contours that are not contained by any others are defined as the boundary of eddy, and the local extremes within them are identified as centers. Eddies with amplitude less than 2 cm and surviving no more than 2 weeks are excluded from analysis. The amplitude is defined as the absolute difference between SSH value of the center and the boundary.

Appendix B

Clustering tendency

The term "clustering tendency" refers to the problem of deciding whether data exhibit a predisposition to cluster into natural groups, without identifying these groups (Jain and Dubes, 1988). It is recommended that random and regularly spaced patterns not be submitted to clustering algorithms (Jain and Dubes, 1988). Since K-means and other clustering algorithms dutifully cluster data into groups no matter whether they are naturally clustered or purely random, it is essential to assess clustering tendency to prevent meaningless results. Hopkins statistic, which is simple and suitable for two-dimensional data, is used for testing whether eddies' distribution shows an aggregative tendency, or they are just randomly distributed. The Hopkins statistic H is defined as follows:

$$H = \frac{\sum_{i=1}^{p} u_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}. \tag{B1}$$

There are p points randomly generated in the data domain, and p sample points randomly selected from the original data. u_i is the nearest neighbor distance of the i^{th} randomly generated point in the original data set, and w_i is the nearest neighbor distance of the i^{th} sample point in the original data set. The H varies from 0 to 1, and it tends to approximate 0.5 when the sums of u_i and w_i are nearly identical, indicating that the data are similar to random points. Values near 1 and 0 indicate obvious clustering and regularly distributed patterns, respectively.

Appendix C

Goodness of clustering

Cohesion and separation are two measures that evaluate how well the clustering fits the data. In Euclidean space, the SSE can be used to measure the cohesion of both individual clusters (individual SSE) and overall clustering (total SSE). It is defined as follows:

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \text{dist}\,(c_i, x)^2, \tag{C1}$$

where dist is the standard Euclidean distance, K is the cluster number, and c_i denotes the i^{th} cluster and c_i the centroid of c_i.

SSB, the sum of the squared distances of a cluster centroid to the overall centroid of all data points, is a traditional measure of separation. The relationship between SSE and SSB is one falling and the other rising (Tan et al., 2005). The SSB is defined as follows:

$$\text{SSB} = \sum_{i=1}^{K} m_i \,\text{dist}\,(c_i, c)^2, \tag{C2}$$

where c denotes the centroid of all data points, and m_i denotes the number of objects in the i^{th} cluster. In evaluation, the lower SSE is better while the higher SSB is better (Tan et al., 2005).

Another popular method that can be used to measure goodness of clustering is called the silhouette coefficient, which combines cohesion and separation in evaluating cluster validity. The silhouette coefficient is defined as follows:

$$s\,(i) = \frac{b\,(i) - a\,(i)}{\max\{a\,(i)\,, b\,(i)\}}, \tag{C3}$$

where $a(i)$ denotes the average distance from the i^{th} point to all other points within the same cluster; the average distance from the i^{th} point to all points in any cluster not containing the points is calculated, and $b(i)$ is the minimum value with respect to all clusters.

As defined, the silhouette coefficient ranges from -1 to 1. A well-clustered point usually has a small $a(i)$ but a relatively large $b(i)$, and the coefficient is close to 1. On the contrary, a negative value is generated only when $a(i) > b(i)$, indicating that the point is probably wrongly clustered. In contrast to SSE and SSB, the silhouette coefficient is calculated at every data point. Thus, to measure the goodness of individual or overall clustering, we typically use the silhouette mean of the cluster, or of the whole.

Acknowledgements. This is a contribution to Research Project 41071250 of the National Science Foundation of China and 088RA500KA of the Innovation Projects of the State Key Laboratory of Resource and Environment Information System, Chinese Academy of Sciences. We thank the US Naval Research Laboratory for providing the SSH data set and appreciate the comments from reviews and readers.

Edited by: J. M. Huthnance

References

Banerjee, A. and Dave, R. N.: Validating clusters using the hopkins statistic, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2004), Budapest (Hungary), 149–153, 2004.

Brown, M. B. and Forsythe, A. B.: Robust Tests for the Equality of Variances, J. Am. Stat. Assoc., 69, 364–367, doi:10.1080/01621459.1974.10482955, 1974.

Cai, S., Su, J., Gan, Z., and Liu, Q.: The numerical study of the South China Sea upper circulation characteristics and its dynamic mechanism, in winter, Cont. Shelf Res., 22, 2247–2264, doi:10.1016/s0278-4343(02)00073-0, 2002a.

Cai, S., Su, J., Gan, Z., and Liu, Q.: The numerical study of the dynamic mechanism of the South China Sea upper circulation in summer, Acta Oceanol. Sinica, 24, 1–7, 2002b (in Chinese).

Chai, F., Xue, H., and Shi, M.: Hydrographic characteristics and seasonal variation of three anticyclonic eddies on the northern continental shelf of the South China Sea, in: Oceanography in China, 13, edited by: Xue, H., Chai, F., and Xu, J, Ocean Press, Beijing, 2001 (in Chinese).

Chelton, D. B., Schlax, M. G., and Samelson, R. M.: Global observations of nonlinear mesoscale eddies, Prog. Oceanogr., 91, 167–216, doi:10.1016/j.pocean.2011.01.002, 2011.

Chen, G., Hou, Y., and Chu, X.: Mesoscale eddies in the South China Sea: Mean properties, spatiotemporal variability, and impact on thermohaline structure, J. Geophys. Res., 116, C06018, doi:10.1029/2010jc006716, 2011.

Cheng, X., Qi, Y. Q., and Wang, W. Q.: Seasonal and interannual variabilities of mesoscale eddies in the South China Sea, J. Trop. Oceanogr., 24, 51–59, 2005 (in Chinese).

Du, Y., Fan, X., He, Z., Su, F., Zhou, C., Mao, H., and Wang, D.: Extraction of spatial-temporal rules from mesoscale eddies in the South China Sea based on rough set theory, Ocean Sci., 7, 835–849, doi:10.5194/os-7-835-2011, 2011.

Fang, W., Guo, X., and Huang, Y. : Observation and study on the circulation in the southern South China Sea, Chinese Sci. Bull., 42, 2261–2271, 1997 (in Chinese).

Fang, G., Fang, W., Fang, Y., and Wang, K.: A survey of studies on the South China Sea upper circulation, Acta Oceanogr. Taiwanica, 37, 1–16, 1998.

Fang, W., Fang, G., Shi, P., Huang, Q., and Xie, Q.: Seasonal structures of upper layer circulation in the southern South China Sea from in situ observations, J. Geophys. Res., 107, 3202, doi:10.1029/2002jc001343, 2002.

Guan, B.: Warm eddy in the open sea east of Hainan Island, J. Oceanogr. Huanghai Bohai S., 4, 1–7, 1997 (in Chinese).

Guan, B. and Yuan, Y. C.: Overview of studies on some eddies in the China seas and their adjacent seas, the South China Sea and the region east of Taiwan, Acta Oceanol. Sinica, 28, 1–16, 2006 (in Chinese).

Hellerman, S. and Rosenstein, M.: Normal Monthly Wind Stress Over the World Ocean with Error Estimates, J. Phys. Oceanogr., 13, 1093–1104, doi:10.1175/1520-0485(1983)013¡1093:nmwsot¿2.0.co;2, 1983.

Huang, Q., Wang, W. Z., Li,Y. X., Li, Z. W. and Mao, M.: General situations of the current and eddy in the South China Sea, Adv. Earth Sci., 7, 1–9, 1992 (in Chinese).

Hwang, C. and Chen, S.-A.: Circulations and eddies over the South China Sea derived from TOPEX/Poseidon altimetry, J. Geophys. Res., 105, 23943–23965, doi:10.1029/2000jc900092, 2000.

Jain, A. K. and Dubes, R. C.: Algorithms for clustering data, Prentice-Hall Inc., New Jersey, USA, 320 pp., 1988.

Kaufman, L. and Rousseeuw, P. J.: Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley, 1990.

Li, L., Nowlin Jr., W. D., and Jilan, S.: Anticyclonic rings from the Kuroshio in the South China Sea, Deep-Sea Res. Part 1, 45, 1469–1482, doi:10.1016/s0967-0637(98)00026-0, 1998.

Li, Y., Cai, W., Li, L. and Xu, D.: Seasonal and interannual variabilities of mesoscale eddies in northeastern South China Sea, J. Tropical Oceanogr., 22, 61–70, 2003 (in Chinese).

Li, J. X., Zhang, R., and Jin, B. G.: Eddy characteristics in the northern South China Sea as inferred from Lagrangian drifter data, Ocean Sci., 7, 661–669, doi:10.5194/os-7-661-2011, 2011.

Lilliefors, H. W.: On the Kolmogorov-Smirnov test for normality with mean and variance unknown, J. Am. Statist. Assoc., 62, 399–402, 1967.

Lin, P., Wang, F., Chen, Y., and Tang, X.: Temporal and spatial variation characteristics on eddies in the South China Sea, Statistical analyses, Acta Oceanol. Sinica, 29, 14–22, 2007 (in Chinese).

Liu, X. and Su. J.: A reduced gravity model of the circulation in the South China Sea, Oceanologia Et Limnologia Sinica, 2, 167–174, 1992 (in Chinese).

Macqueen, J.: Some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symp. Mathematical Statist, Probability, 281–297, 1967.

Nitani, H.: Oceanographic conditions in the sea east of Philippines and Luzon Strait Summer of 1965 and 1966, The Kuroshio – A Syposium on Japan Current, Honolulu, Hawaii, 213–232, 1970.

Qu, T.: Upper-Layer Circulation in the South China Sea, J. Phys. Oceanogr., 30, 1450–1460, doi:10.1175/1520-0485(2000)030¡1450:ulcits¿2.0.co;2, 2000.

Ripley, B. D.: Modelling Spatial Patterns, J. Roy. Statis. Soc. B, 39, 172–212, 1977.

Shaw, P.-T., Chao, S.-Y., and Fu, L.-L.: Sea surface height variations in the South China Sea from satellite altimetry, Oceanol. Acta, 22, 1–17, doi:10.1016/s0399-1784(99)80028-0, 1999.

Su, J.: Overview of the South China Sea circulation and its influence on the coastal physical oceanography outside the Pearl River Estuary, Cont. Shelf Res., 24, 1745–1760, doi:10.1016/j.csr.2004.06.005, 2004.

Su, J., Xu, J., and Cai, S.: Gyres and eddies in the South China Sea, in: Onset and evolution of the South China Sea monsoon and its interaction with the ocean, edited by: Ding, Y., and Li, C., China meteorological Press, Beijing, 272–279, 1999.

Tan, P.-N., Steinbach, M., and Kumar, V.: Introduction to Data Mining, Addison Wesley, 2005.

Wang, G. H.: Discussions on the movement of mesoscale eddies in the South China Sea (in Chinese), Ph.D., dissertation, Sch. of Phys. Oceanogr., Ocean Univ. of China, Qingdao, China, 2004.

Wang, G. H., Su, J. L., and Chu, P. C.: Mesoscale eddies in the South China Sea observed with altimeter data, Geophys. Res. Lett., 30, 2121, doi:10.1029/2003gl018532, 2003.

Wang, G. H., Chen, D. K., and Su, J. L.: Generation and life cycle of the dipole in the South China Sea summer circulation, J. Geophys. Res., 111, C06002, doi:10.1029/2005jc003314, 2006.

Wang, G. H., Chen, D. K., and Su, J. L.: Winter eddy genesis in the eastern South China Sea due to orographic wind jets, J. Phys. Oceanogr., 38, 726–732, 2008.

Wang, L., Koblinsky, C. J., and Howden, S.: Mesoscale variability in the South China Sea from the TOPEX/Poseidon altimetry data, Deep-Sea Res. Part 1, 47, 681–708, doi:10.1016/s0967-0637(99)00068-0, 2000.

Wang, X., Du, Y., Fan, X., and Yi, J.: Research on automatic identification method of ocean mesoscale eddies, J. Trop. Oceanogr., in review, 2012.

Wang, Z. and Chen, Q. : The warm eddy in northern South China Sea, Journal of Oceanography in Taiwan University, 18, 92–103, 1987 (in Chinese).

Wu, C.-R. and Chiang, T.-L.: Mesoscale eddies in the northern South China Sea, Deep-Sea Res. Part 2, 54, 1575–1588, doi:10.1016/j.dsr2.2007.05.008, 2007.

Xiu, P., Chai, F., Shi, L., Xue, H., and Chao, Y.: A census of eddy activities in the South China Sea during 1993–2007, J. Geophys. Res., 115, C03012, doi:10.1029/2009jc005657, 2010.

Yang, K., Shi, P., Wang, D., You, X. and Li, R.: Numerical study about the mesoscale multi-eddy system in the northern South China Sea in winter, Acta Oceanol. Sinica, 1, 27–34, 2000 (in Chinese).

Yuan, D., Han, W., and Hu, D.: Surface Kuroshio path in the Luzon Strait area derived from satellite remote sensing data, J. Geophys. Res., 111, C11007, doi:10.1029/2005jc003412, 2006.

Yuan, D., Han, W., and Hu, D.: Anti-cyclonic eddies northwest of Luzon in summer-fall observed by satellite altimeters, Geophys. Res. Lett., 34, L13610, doi:10.1029/2007gl029401, 2007.

Yuan, D. and Li, R.: Dynamics of eddy-induced Kuroshio variability in Luzon Strait, J. Trop. Oceanogr., 27, 1–9, 2008 (in Chinese).

Zhong, H.: Stuctures of the density circulation, in: Report of Decadal Hydrographic Series Survey of the Shelf and Adjacent Waters of the Northern South China Sea, edited by: Ma, Y., Ocean Press, Beijing, 1990 (in Chinese).