High frequency radar error classification and prediction based on K-means methods

Zhaoyi Wang, Marie Drevillon, Pierre de Mey-Frémaux, Elisabeth Remy, Nadia K Ayoub, Dakui Wang, Bruno Levier

frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 04 November 2024
DOI 10.3389/fmars.2024.1448427

OPEN ACCESS

EDITED BY
Chunyan Li,
Louisiana State University, United States

REVIEWED BY
Fulvio Capodici,
University of Palermo, Italy
Gui Gao,
Southwest Jiaotong University, China

*CORRESPONDENCE
Marie Drevillon
mdrevillon@mercator-ocean.fr
Dakui Wang
dakui.nmefc@gmail.com

RECEIVED 14 June 2024
ACCEPTED 16 October 2024
PUBLISHED 04 November 2024

CITATION
Wang Z, Drevillon M, De Mey-Frémaux P, Remy E, Ayoub N, Wang D and Levier B (2024) High frequency radar error classification and prediction based on K-means methods.
Front. Mar. Sci. 11:1448427.
doi: 10.3389/fmars.2024.1448427

High frequency radar error classification and prediction based on K-means methods

Zhaoyi Wang[1,2], Marie Drevillon[3]*, Pierre De Mey-Frémaux[4], Elisabeth Remy[3], Nadia Ayoub[4], Dakui Wang[2]* and Bruno Levier[3]

[1]Tianjin Key Laboratory for Marine Environmental Research and Service, School of Marine Science and Technology, Tianjin University, Tianjin, China, [2]Key Laboratory of Research on Marine Hazards Forecasting, National Marine Environmental Forecasting Center, Beijing, China, [3]Mercator Ocean International, Toulouse, France, [4]Université de Toulouse, LEGOS, Laboratory of Space Geophysical and Oceanographic Studies, Toulouse, France

This study aims to characterize the high frequency radar and numerically simulated low-frequency filtered currents in the south-eastern Bay of Biscay (study area) using a K-means classification algorithm based on an improved Euclidean Distance calculation method that does not take missing values. The errors between observations and simulations was estimated and predicted based on this classification method. Results indicate that predominantly eastward (northward) currents over the Spanish (French) continental shelf/slope in winter and more variable currents in the west and south-west in summer. The model classification results for circulation characteristics are in relatively good agreement with HF radar results, especially for currents on the Spanish (French) shelf/slope. In addition, the probabilistic relationship between observed and modeled currents was explored, obtaining the probability of occurrence of modeled current groups when each group of observed currents occurs. Finally, predictions of model and observed current errors were made based on the classification results, and it was found that the predictions based on the classification of all data had the smallest errors, with a 17% improvement over the unclassified control experiment. This study provides a foundation for subsequent model error testing, forecast product improvement and data assimilation.

KEYWORDS

high frequency radar, ocean current, the bay of biscay, K-means, error estimate

1 Introduction

The South-eastern Bay of Biscay (SE-BoB) is distinguished by the presence of canyons (e.g. Capbreton Canyon), abrupt changes in coastal orientation and narrow shelves and slopes. The current is barotropic and quite weak throughout the year, and the vertical gradient of horizontal currents is higher in summer than in winter due to stronger

stratification of the water column (Rubio et al., 2013a). During the winter, the surface currents in the SE-BoB are mainly associated with the Iberian Poleward Current (IPC), which affects the upper 300m of the water column (Cann and Serpette, 2009).The IPC flows up the slope, moving warm surface water along the Spanish coast to the east and along the French coast to the north (Cann and Serpette, 2009; Charria et al., 2013). In summer, this flow is reversed and three times weaker than in winter (Solabarrieta et al., 2014). Wind-induced flow is the main driver of surface ocean circulation in the region (Fontán and Cornuelle, 2015; Fontán et al., 2013; Kersalé et al., 2016; Solabarrieta et al., 2015). In autumn and winter, southwesterly winds dominate, producing northward and eastward drift over the shelf. In spring, the winds change to northeast causing the currents to shift west-southwest along the Spanish coast. In summer, the situation is similar to that of spring, but the total drift direction changes more often due to weaker winds, making the currents more variable (Solabarrieta et al., 2015).

High frequency (HF) radar is a land-based remote sensing technology that has proven to be a cost-effective tool for monitoring coastal areas at ranges of up to 200 km. Paduan and Washburn (2013) provide a detailed description of HF radar technology. Röhrs et al. (2015) used HF radar measurements of currents in comparison with *in-situ* observations of currents to find that wave-induced Stokes drift was not included in the HF radar currents. Oceanographic HF radar is mainly used to measure the ocean surface current field for various applications such as search and rescue, oil spill monitoring, marine traffic information or improvement and data assimilation for numerical current models (Paduan and Rosenfeld, 1996; Gurgel et al., 2001; Roarty et al., 2019, 2019; Rubio et al., 2018; Gurgel et al., 2001).

The HF radar coastal system in the Basque Country has been operational since the beginning of 2009. It consists of two radar stations, one at Cape Higer (1.78°W,43.38°N) and the other at Cape Matxitxako (2.75°W,43.45°N), as shown in Figure 1A, emitting at a bandwidth of 40 KHz, a frequency of 4.86 MHz and an average radiated power of 40W. The system provides long-range hourly, with running averages of 3 hours, surface current data with a spatial resolution of 5.12 km (width of the ranging unit) and 2~3 meters depth. Detailed validation of currents observed by Basque HF radar has been carried out by Solabarrieta et al. (2016; 2015; 2014) and Rubio et al. (2018; 2020; 2019; 2011). Arzoo and Rathod (2017) and Caballero et al. (2020) used radar observations with tidal and near-inertial oscillations removed in combination with satellite data to significantly improve the regional mean dynamic topography (MDT). Solabarrieta et al. (2015) used a linear auto-regressive model with empirical orthogonal function (EOF) decomposition of the HF radar historical data series to predict currents with good results.

In recent years, artificial intelligence methods have been increasingly used in marine data processing and ocean forecasting, and a number of results have been achieved (Cao et al., 2024; Gao et al., 2024). Clustering is a method of grouping data on the basis of their attributes, and the attributes of all elements in each group must be similar (Rehioui et al., 2016). There are various types of clustering, such as data mining algorithmic clustering, dimensionality reduction clustering, and parallel clustering (Zerhari et al., 2015). Among available tools, partition clustering as a type of data mining algorithmic clustering is integrated with different algorithms such as K-means, K-modes, K-medoids, PAM, etc (Sajana et al., 2016). Many authors have

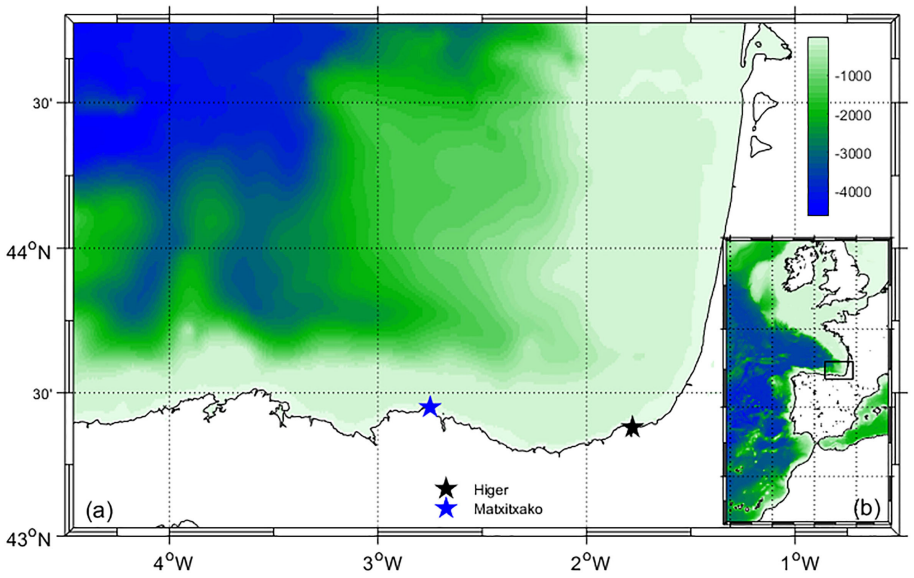

FIGURE 1
The domain and topography of study area **(A)** and IBI model **(B)**, the blue and black star in **(A)** show the location of the Matxitxako and Cape Higer HF radar stations, respectively.

applied clustering methods to study sea current characteristics, Solabarrieta et al. (2015) used a non-linear K-means classification algorithm to obtain a comprehensive description of wind and HF radar-derived currents in the southeastern Bay of Biscay. The K-means clustering algorithm (Wu et al., 2019) is one of the widely used clustering implementations and its use is very common due to its best performance for large data sets (Jothi et al., 2019; Wu et al., 2019) and is more appropriate for data space exploration as it considers data close to the edges of the data space (Camus et al., 2011). In the standard K-means algorithm, K points are first selected as the initial prime, where each prime represents a cluster. All objects in the dataset are then assigned to the prime with the smallest distance. After all data items have been assigned, the prime is recalculated until no further objects change their clusters (Velmurugan and Santhanam, 2011). In calculating the center of mass, distance calculation is required, and many scholars have also studied and improved the mathematical measures of distance calculation (Arzoo and Rathod, 2017), including Euclidean Distance (Saqib et al., 2021), Manhattan Distance (Bora and Gupta, 2014), Minkowski Distance (Charulatha et al., 2013), etc. Among them, the Euclidean Distance is a familiar and direct line between two elements or the minimum distance between two objects (Saqib et al., 2021), which is a more common method of distance calculation. Solabarrieta et al. (2015) used the k-means clustering technique to characterize the main ocean surface circulation patterns in the study area, at scales from several days to inter-annual.

There is usually a discrepancy between the ocean currents observed by HF radar and the simulated ocean currents. This is firstly due to the inaccuracies of the model, especially in coastal areas, where the parameters such as water depth, topography, and bottom friction are often inaccurate, resulting in the inability of the numerical model to accurately represent the characteristics of ocean currents (Hirose et al., 2001; Wang et al., 2023). At the same time, due to the poor signal-to-noise ratio at the long range by HF radar, there are also errors in the measurement of HF radar at the edge of the observation area. In addition, during the grid processing, geometric dilution in the statistical process caused by the positions of grid cells with regard to the locations of the HF radar stations may also cause some errors (Cann and Serpette, 2009; Rubio et al., 2013b). Based on the data assimilation method, it is possible to better utilize the HF radar observations of ocean currents to optimize the numerical simulation of ocean current errors (Shulman and Paduan, 2009). In this paper, the clustering method and longtime series of surface current data from HF radar and modeling are used to cluster the errors of observation and simulation, and to research the errors character and errors prediction between the simulation and observation. By studying the background error covariance of the numerical simulation, a foundation is laid for the assimilation of HF radar surface current observations. This paper first describes the HF radar data used, the numerical model data and the cluster analysis method, the third part presents the results of the cluster analysis of the observed and modeled errors, the fourth part presents the prediction method for the errors between simulation and observation and finally the summary.

2 Data and methods

2.1 Modeling data

The modeling data, based on NEMOv3.6 and covering the northeast Atlantic waters as shown in Figure 1B, was obtained from the Copernicus Marine Service (CMEMS). The IBI product (Sotillo et al., 2015) is defined on a regular grid of 1/36° (~2~3km) with latitudes ranging from 26°N to 56°N and longitudes ranging from 19°W to 5°E, with a mean step size of 0.027788°, resulting in a horizontal grid extending to 1081 × 865 grid points. Atmospheric forcing is provided by hourly ECMWF fields, including the 10-m wind, surface pressure, 2-m air temperature, 2-m specific humidity, precipitation rate, short-wave and long-wave radiation fluxes. Open boundary marine environmental information is sourced from a global model (Lellouche et al., 2018) with1/12° resolution. Tidal forcing is supplied by the 11 tidal harmonics (M2, S2, N2, K1, O1, Q1, M4, K2, P1, Mf and Mm) from the FES2014 solution (Lyard et al., 2021). The reduced-order Kalman filter assimilation scheme (Tranchant et al., 2019) was used to assimilate altimeter data, *in-situ* temperature and salinity vertical profiles and satellite sea surface temperature. The data from the IBI operational system (hind-cast only) of the Bay of Biscay surface currents (U and V components) cover the period 2017-2021. The model outputs have been duly validated, including an assessment of high frequency variability of surface fields, with the results available in a product quality document (Levier et al., 2023).

2.2 HF radar data

Data from Euskoos, encompassing reprocessed and near real-time data, spans the period of 2009-2021 and is available from CMEMS. This delayed mode product, designed for reanalysis purposes, integrates the best available version of *in situ* data for ocean surface currents. It includes the TODAL delay time dataset and is specifically designed for near-surface current measurements from HF radars. The observational coverage in this study aligns with the numerical model, spanning a total of 5 years from 2017 to 2021. Initially, we extract the good observed data points identified by a flag of 1 in the CMEMS dataset.

The seasonal averages of the surface current from the HF radar observation are presented in Figure 2, with winter representing December to February, spring representing March to May, summer representing June to August and autumn representing September to November. The HF radar data from 2017 to 2021 were further processed using a 24-Hour (24H) low-pass filter, which separates the HF radar observations into low-frequency current components

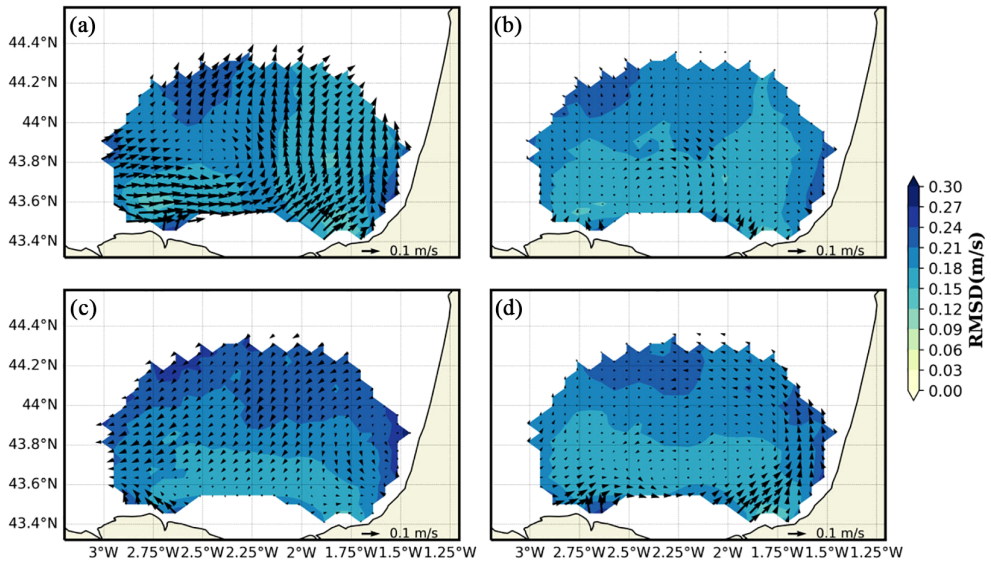

FIGURE 2
Seasonal HF radar observed surface currents (arrow) and the RMSD of HF radar currents and IBI modeling currents (colors). **(A)** Winter, **(B)** Spring, **(C)** Summer and **(D)** Autumn.

and high-frequency current components. Figure 3 shows the percentage contribution of these two components to the total ocean currents. On average, the >24H low-frequency current components account for 57.6% throughout the year, reaching ~70% in the Spanish shelf/slope region, with a maximum of 72.0% in winter and a minimum of 45.8% in summer. A Butterworth low-pass filter with a cut-off period of 24 hours is applied to the hourly surface currents to filter out the <24H high-frequency currents and >24H low-frequency currents, mirroring the process applied to the hourly modeling data. Through above analysis, the possibility of the >24H currents is relatively high. In addition, the low-frequency currents are less variable and more

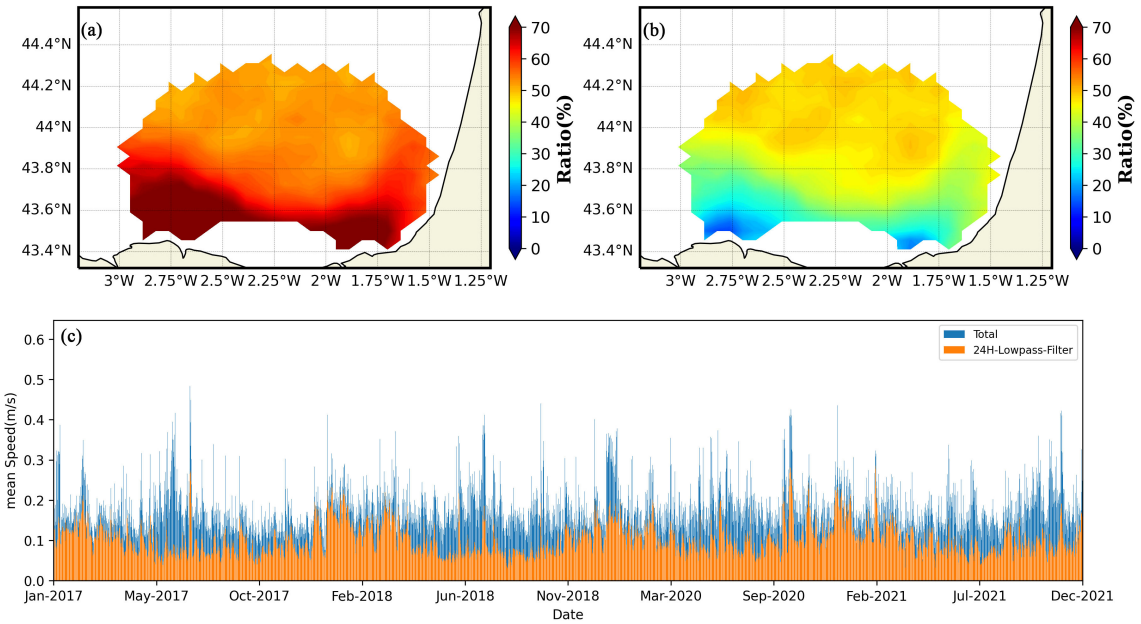

FIGURE 3
Percentage of >24H **(A)** and <24H **(B)** radar surface current to the total current. **(C)** Temporal of HF radar surface mean current, the blue bar is total current and the orange bar is <24H-lowpass filter current.

stable, with environmental factors such as wind and pressure being the main influencing factors. Therefore, the subsequent analysis is based on the >24H low-frequency currents.

2.3 K-means clustering algorithm

The primary objective of this study is to classify HF radar observations and predict the errors between observations and simulations, so it is important to cover all of the data space. The KMA clustering algorithm partitions the high-dimensional data space into clusters or groups, each defined by a prototype and comprising data for which the prototype is the most similar. However, a significant limitation of this technique is the inability to accommodate missing values in the time series analyses. To address this issue of missing values, a new distance calculation function is used in KMA, which computes the average distance between each grid point in the data after remove missing value and the corresponding grid point in the prototype. The points with missing values are automatically excluded when calculating the distance classification in this function, ensuring the maximal utilization of information from all observations.

The HF radar observations and the IBI simulation results consist of N-dimensional vectors $X = (x_1, x_2,\ldots, x_N)$, where N represents the total amount of data. After data pre-processing, the analysis covers a span of 5 years, encompassing N=36400 hours. To optimize the utilization of available data for training, the data from 2019, which notably has a relatively small amount of data at only 5915 hours of observations, was designated as the evaluation data for predictions. Meanwhile, the data from other years (2017, 2018, 2020, 2021) were utilized as the training data. Each vector $x_k = (x_{1k}, x_{2k},\ldots, x_{nk})$ of sea current speed, derived from the HF radar data and the model data, consists of n=410 nodes. Guanche et al. (2013) have previously discussed the criteria for selecting the appropriate number of groups. The number of groups chosen for the spatial classification, obtained using KMA in our experiments, was determined by calculating the variation of the distance of each group from the centroids with the number of groups. We found an elbow-like inflection point with a significant decrease in distance from6 groups and more (not shown) for the change in this distance at M = 6. Based on the above approach, KMA was applied to the classify the 24H-lowpass-filteredradar-observed current dataset, resulting in the defined values of the centroids $v_k = (v_{1k}, v_{2k},\ldots, v_{nk})$ from the same n-dimensional M = 6 groups in the original data.

3 Classification results

The traditional classification methods (such as classification by month or season, as shown in Figure 2) have limitations when dealing with complex and varied datasets. These methods often rely on preset rules or experience, lacking the flexibility and objectivity driven by data, and thus struggle to fully capture and utilize the intrinsic characteristics and detailed information within the data. For complex and volatile datasets like ocean currents, these limitations are particularly evident. In contrast, cluster analysis, as a modern data analysis method, possesses greater objectivity, flexibility, depth, and adaptability, making it better suited to meet the needs of data analysis. The approach in this paper involves using KMA analysis to classify the currents in both the observed and simulated datasets. This allows for examining the morphological characteristics and variability of the currents, as well as the relationship between the observed and simulated datasets. Subsequently, the root mean square deviation (RMSD) between observations and simulations is computed for each group, based on the classification of the observed datasets.

3.1 HF radar currents character

The results of the KMA analysis of hourly HF radar currents are depicted in the 3x2 lattice of Figure 4. The six groups identified represent most of the known surface circulation features in the study area. OG (Observation Group) -01 and OG-03 are groups characterized by weak westward flowing currents, with mean velocities of 6.9cm/s and 10.1cm/s respectively. These groups have the highest probability of occurrence, which is calculated by dividing the number of times each groups by the total number, at 41.2% and 20.3% respectively. OG-02 and OG-04 consist of currents predominantly flowing northward, constituting 10.5% and 14.8% of the surface mean currents in the area. They exhibit strong northward currents, with mean velocities of 15.9 cm/s and 11.4 cm/s, respectively. OG-02 specifically represents northeastward currents on the Spanish shelf/slope. OG-05 and OG-06 are dominated by the east-northward flow, with mean velocities of 20.6cm/s and 15.9cm/s respectively, on the Spanish (French) shelf/slope. Their structures are notably similar, particularly on the Spanish and French continental shelves/slopes, where the currents are mainly eastward and northward. OG-06 is slightly weaker than OG-05, and a cyclonic vortex is observed offshore north of the Spanish shelf.

3.2 Simulation currents character

Figure 5 shows a 3x2 lattice of the results of the KMA analysis of the simulated currents. The MG (Modeling Group) -01 current group exhibits a northwestward (northeastward) current on the French (Spanish) shelf/slope, with a probability of occurrence of 9.3%. Its circulation structure somewhat similar to that of OG02, both has a northward mean flow direction. The MG-02 current group shows a westward mean flow direction, similar to that of OG-01. The MG-03 current group has a very strong eastward (northward) current on the Spanish (French) outer shelf/slope, with a probability of occurrence of 8.1%.Meanwhile, the MG-04 current group has a northwesterly mean current direction, with a probability of occurrence and a more similar circulation structure to

10.3389/fmars.2024.1448427

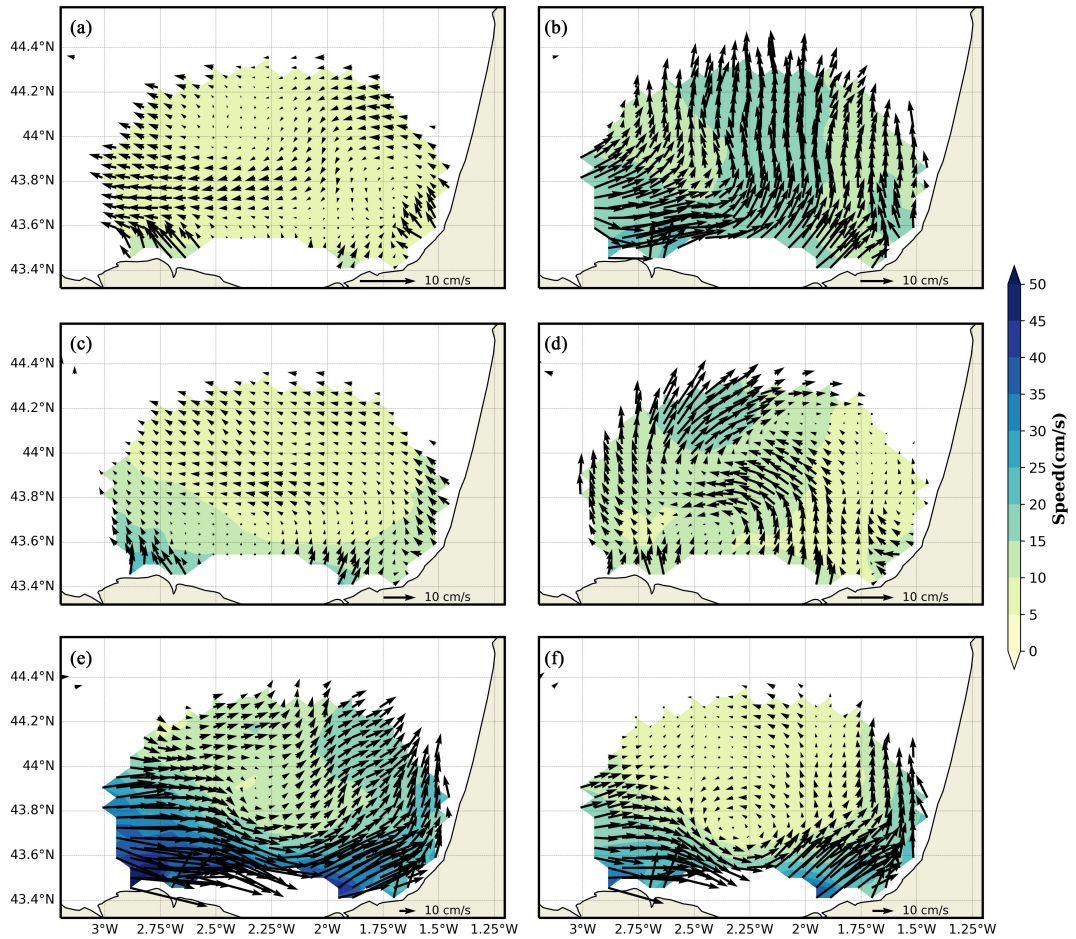

FIGURE 4
3x2 lattice of the K-means analysis applied to the HF radar surface current data, wherein The color represents the average absolute velocity, and the arrow represents the average flow direction calculated from the average meridional flow and average latitudinal flow. **(A)** OG01, **(B)** OG02, **(C)** OG03, **(D)** OG04, **(E)** OG05, **(F)** OG06.

OG-03. The MG-05 current group has a very strong eastward (northward) current on the Spanish (French) shelf/slope, with a circulation structure very similar to OG-05 and a probability of occurrence of 3.6%. The MG-06 current group on the Spanish (French) shelf/slope has a strong eastward (northward) current and a cyclonic eddy structure in the middle of the observed area, which is the same as the OG-06 current structure. The various subgroups of the simulated currents can be found to correspond to their respective subgroups in the observed currents.

3.3 Relationship between OGs and MGs

Figure 6 shows the monthly probability of occurrence and number of occurrences for each group, with a significant decrease in the number of observations occurring in July-August 2020 and June 2021. As shown in Figure 6, OG-01 has a very high probability and number of occurrences around the summer of 2017-2018, with a significant decrease in the number of occurrences into 2020. The frequency of OG-03 had a significant increase in 2020. OG-04 occurs mainly in the winter of 2017, 2020 and 2021. The total number of occurrences of OG-05 is relatively low, with a total of only 1209 occurrences (Table 1), and mainly occurring in 2020.

Table 1 provides the probability of occurrence of each group in different years and seasons. There is a clear seasonal variation in OG-02-04-05, with an occurrence rate of over 40% in winter, while the rest of the groups are less seasonal, with OG01 occurring less than 10% in winter and over 30% in all other seasons. In terms of year of occurrence, OG-05 has a 46% probability of occurring in 2020, while OG-02 and OG-05 have less than 15% probability of occurring in 2017 and 2021, respectively. The probability of occurrence for the other groups is more evenly distributed across the years, largely ranging from 15% to 35%.

The probabilistic linkage between the 6 OGs and 6 MGs is given in Figure 7. For each OGi, Pij denotes the is the probability of occurrence for each MGj, which is computed by dividing the

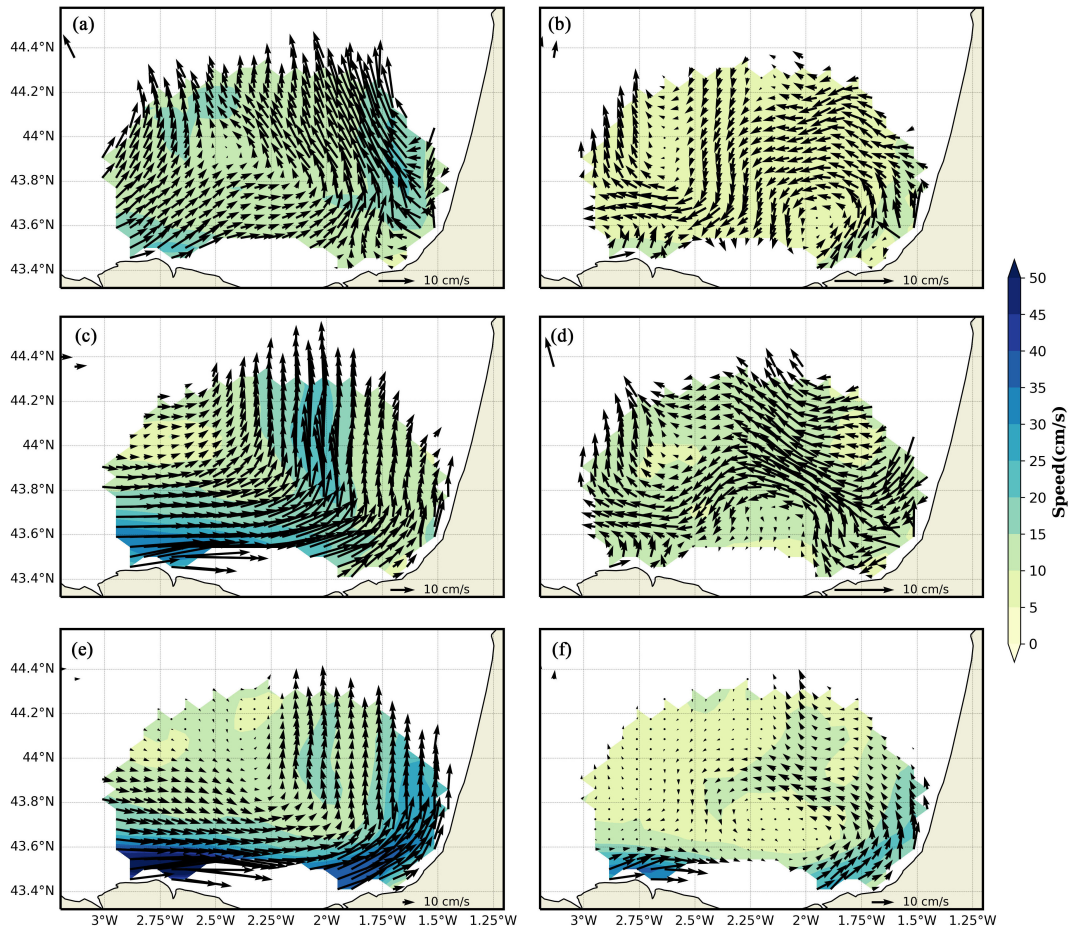

FIGURE 5
3x2 lattice of the k-means analysis applied to the modeling surface current data. **(A)** MG01, **(B)** MG02, **(C)** MG03, **(D)** MG04, **(E)** MG05, **(F)** MG06.

number of occurrences of each MGj during the OGi period by the total number of OGi, yielding a multigrid as shown in Figure 7. Each sub-grid indicates the probability of each MG occurrence at each OG (e.g. if OG-01 occurs, there is a P = 56.6% probability that MG-02 will occur).

The probability of MG-02 occurring in OG-01 is 56.6%, with low flow velocities. There is a cyclonic vortex over the French continental shelf, and a westward current in the northern waters of Spain. The probability of MG-04 occurring is high in the OG-04 group, characterized by mainly northward current, with northwestward offshore current on the French continental shelf leading to net transport away from the coast. In OG-06, dominated by stronger coastal current, a 59.7% probability of MG-06 occurrence is observed, indicating good simulation of the robust east (north) oriented currents on the Spanish (French) shelf/slope. The joint occurrence probability of MG-01 and MG-06 exceeds 50% for both OG-02 and OG-03. The currents in these groups are marked by prevailing west-northwest currents in the central part of the studied region, aligning with the observed patterns. OG-05 has a 53% probability of strong currents occurrence on the Spanish (French) shelf/slope, with a combined probability exceeding 85% with MG-03, showcasing strong northward currents in the central section of the study area.

3.4 RMSD distribution of OGs

Based on the analysis provided, there is a significant resemblance between the classification of observed currents and simulated currents. This indicates the model effectively captures the low-frequency components of the current simulation results. On the basis of clustering analysis of HF radar observation, the RMSD was utilized to assess the current error between the observations and simulations. Figure 8 shows the 3x2 lattice of the surface current RMSD centroids, where the arrows are the mean error of U-component and V-component.

For OG-01, the RMSD is low at the center of the observation area but high along the edges, showing southward current deviations on both sides and eastward deviations in the north. The average RMSD for OG-01 is the smallest among the 6 groups at

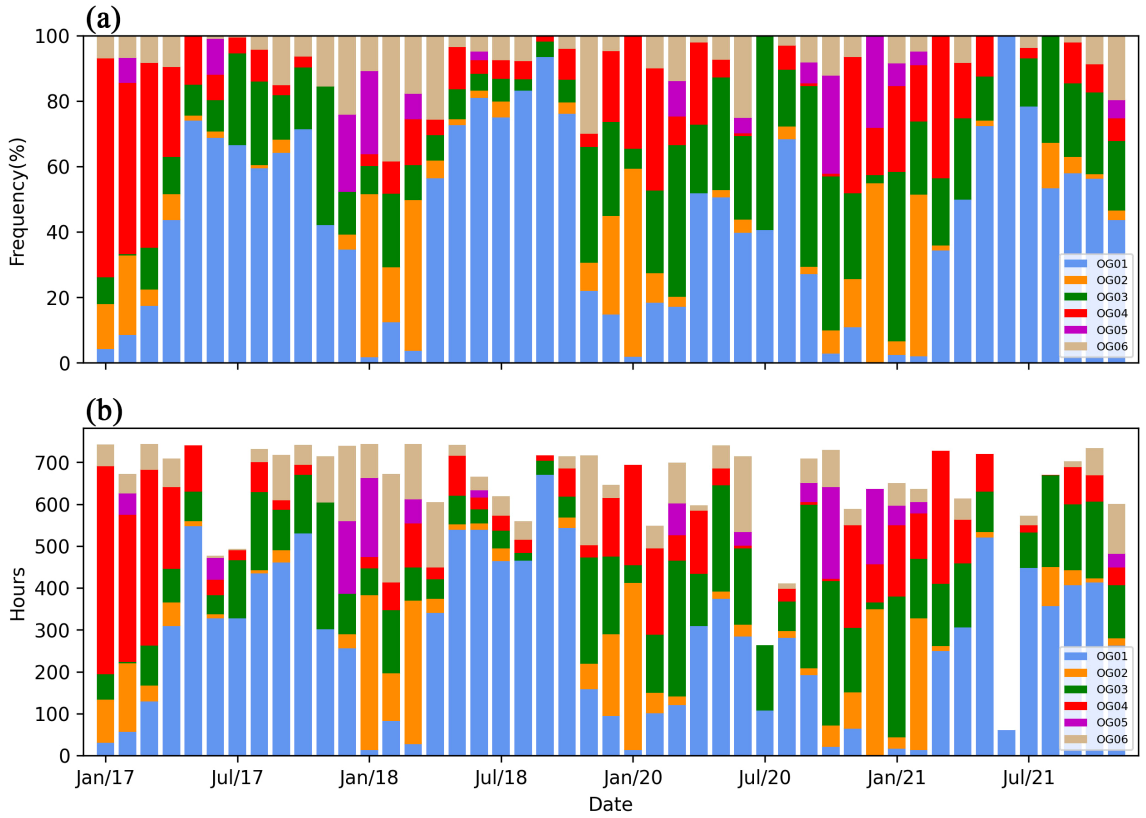

FIGURE 6
Temporal distribution of each RMSD groups during 2017-2021. **(A)** the proportion of the different groups, **(B)** the number of observation hours for the different groups.

11.5 cm/s. The RMSD for OG-02 exhibits relatively high, averaging 14.3 cm/s, with larger errors in the domain's middle and near the French shelf/slope. OG-03 demonstrates elevated RMSD, particularly in the western section of the observation area, averaging 13.4 cm/s, with westward current deviations. The RMSD for OG-04 is notably higher in the central-northern region, showing northwestward current deviation but lower errors near the French shelf/slope. The RMSD for OG-05 is the highest, averaging at 16.6 cm/s, particularly on the western and southeastern sides, with eastward current deviation.OG-06's RMSD is higher on the Spanish shelf and lower in the central region, with an average RMSD of 14.7 cm/s. The most significant errors are found in the southern and eastern parts of the area, where the current is strong along the Spanish and French coasts. Directional errors are detected near the stations and in the northernmost region, far from both stations.

TABLE 1 These values represent the variability of each total current RMSD group, during the study period.

Groups	Total	Proportion							
		2017	2018	2020	2021	Winter	Spring	Summer	Autumn
OG1	12570	0.30	0.31	0.15	0.24	0.05	0.30	0.33	0.32
OG2	3201	0.14	0.37	0.32	0.16	0.66	0.17	0.06	0.10
OG3	6194	0.21	0.17	0.36	0.27	0.20	0.25	0.19	0.36
OG4	4502	0.39	0.15	0.24	0.22	0.42	0.38	0.06	0.13
OG5	1209	0.23	0.22	0.46	0.10	0.56	0.11	0.08	0.25
OG6	2809	0.25	0.38	0.21	0.16	0.31	0.24	0.13	0.32

The values are the % of occurrence of each group during each year, winter, spring, summer or autumn periods, considering winter months, December to February, spring months, March to May, summer months, June to August, and autumn months, September to November.

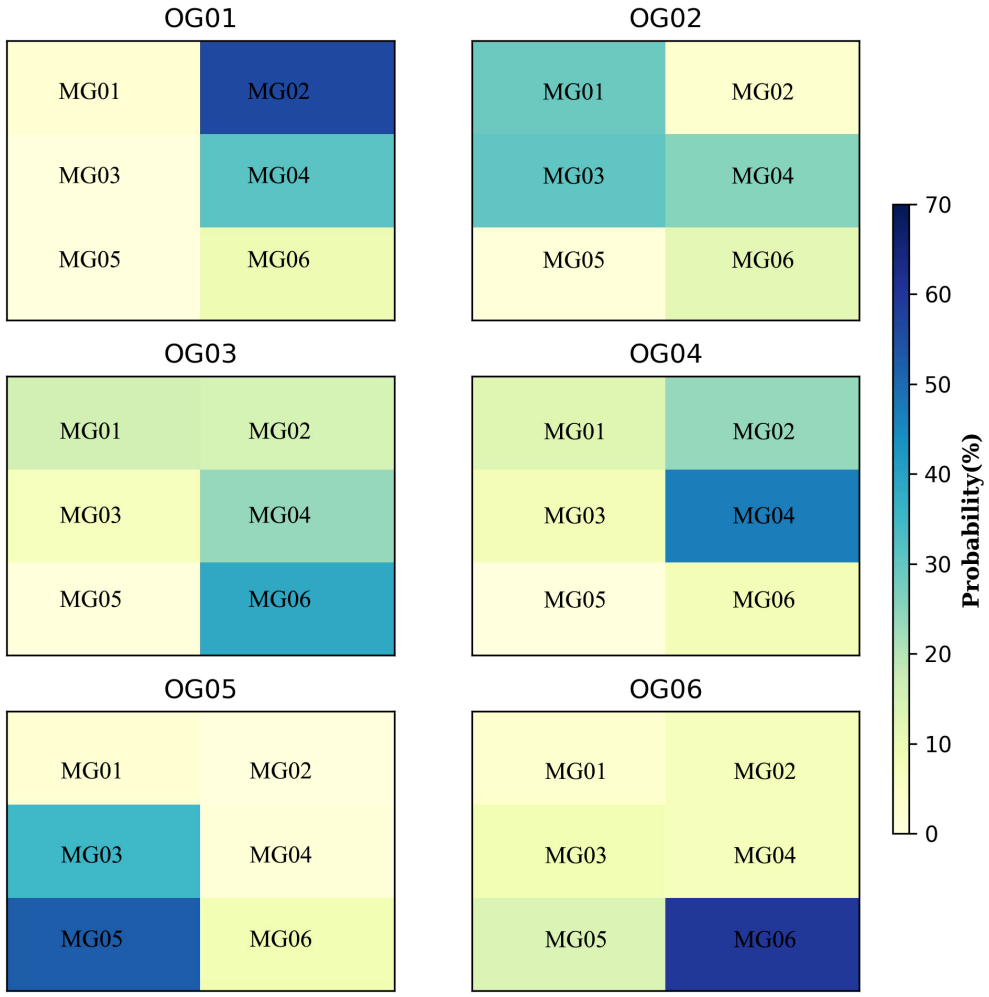

FIGURE 7
Occurrence probabilities of the 6MG projected into the HF radar current lattice.

4 Error estimate

The preceding analysis indicates that the distribution of RMSD between observations and simulations exhibits a significant correlation with the locations of the observation sites and the magnitude of the observed flow velocity. Since the HF radar sites are roughly located at the same latitude, the location of the observation sites can be simplified to latitude. The distribution of RMSD along latitude and longitude was converted to a scatter plot along latitude and speed, and then mapped onto a latitude × speed grid. While increasing with the increasing flow velocity, the RMSD decreases and then increases with latitude, due to the fact that at low latitudes, the angles between the radials collected by the two HF radar stations decreases as well leading to a higher error in the vector sum of the radials. It is noteworthy that the maximum RMSD value of OG-05-06 was observed at around 43.5°N and for speeds between 30 and 40 cm/s.

With the above processing, the distribution of RMSD (Lat, Speed) for each group is obtained. Next, the error between observation and simulation is predicted based on the classification and current observations of HF radar in 2019. Using the latitude and speed information from each observation point, the RMSD between observations and simulations is predicted by interpolation in the RMSD (Lat, Speed) distributions. This allows prediction of the RMSD using only HF radar observations, without the need for simulations. The predictions are then compared to the actual model minus observed RMSD for verification. The average error prediction results for winter and summer are shown in Figures 9A, B. In addition, the actual error results are shown in Figures 9C, D. In winter, the true errors (Figure 9C) exhibit a pronounced west-east pattern, with higher values in the west and lower values in the east. The predicted values (Figure 9A) also display a similar east-west dipole, albeit with under estimated high values in the west and overestimated low values in the east. In the summer, the

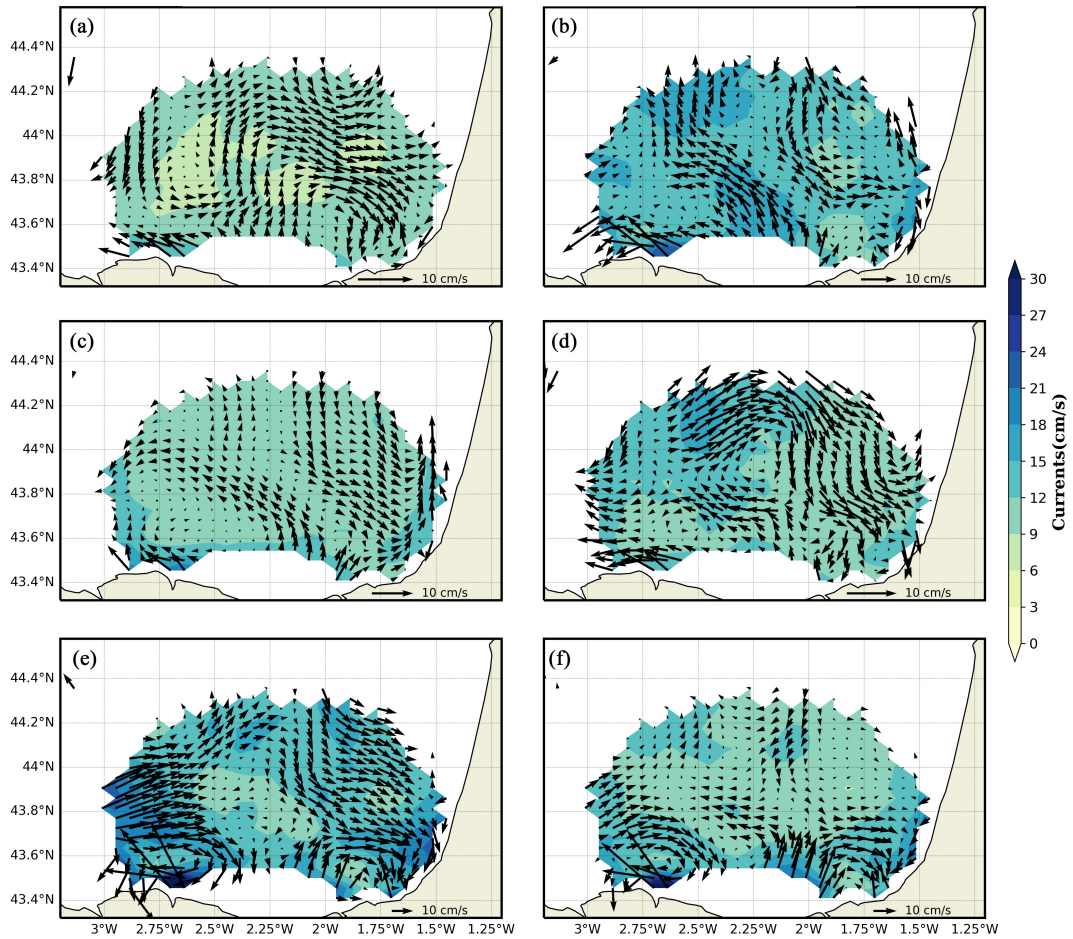

FIGURE 8
3x2 lattice of the surface current RMSD between observation and simulation, wherein the arrows stand for the direction of the surface current bias. **(A)** Group 01, **(B)** Group 02, **(C)** Group 03, **(D)** Group 04, **(E)** Group 05 and **(F)** Group 06.

discrepancy between the simulation and observation (Figure 9D) is minimal in the south-central region and more pronounced at the periphery of the area. The prediction (Figure 9B) effectively captures the large-scale characteristics of the error.

Building on the error prediction based on the classification of all data (2017, 2018, 2020, 2021), this study also attempts to manually categorize the data by seasons and months before conducting error prediction using the KMA classification. Figure 9E shows the temporal distribution of mean RMSD. The variation of the error for each group of experiments is basically consistent with the true value, with the exception of the experiment result based on monthly data classification, which exhibits the largest error with an RMSE of 2.46 cm/s (Table 2). This discrepancy may be attributed to the smaller amount of data available for classification each month after the data is divided by month. As shown in Table 2, the experiment with the smallest prediction error based on a full year of data, with an RMSE of 1.98 cm/s, was not significantly different from the error result of classifying data by season (2.01 cm/s). This represents a 16.8% improvement compared to the control experiment that utilized all observation and simulation errors for statistical analysis without classification.

5 Conclusions

The data obtained from the HF radar system of the SE-BOB and the modeling system of IBI provide a large amount of invaluable information for the study of the regional surface water circulation and the controlling physical processes. K-means cluster analysis methods, as unsupervised learning artificial intelligence techniques, have developed rapidly in recent years. The aim of this study is to classify known circulation features based on the cluster analysis method and use the classification results to predict the errors between observations and simulations at different time scales. First, the HF radar observed and model simulated U and V were processed using a 24H-lowpassfilter, to extract >24H currents. Then, the KMA method was used for the classification training of ocean current speed by 4-year observation data, and the similarity between observation and simulation was analyzed. Finally, based on the RMSD results from the observation classification, the RMSD was predicted for2019 to obtain the simulation error, which can provide a foundation for data assimilation. In the study area, the surface currents show significant temporal and spatial variability, the model simulations of low-frequency currents are in good

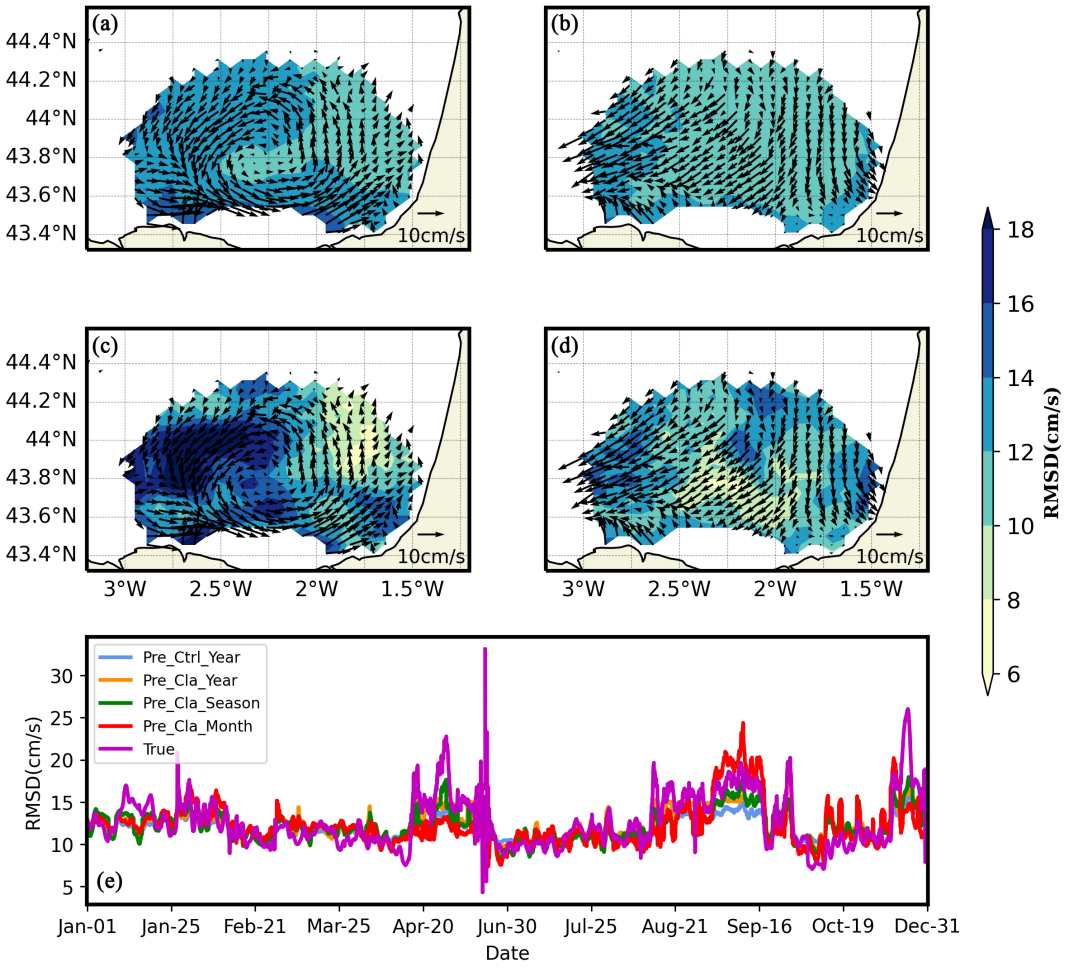

FIGURE 9
Horizontal distribution of predicted **(A, B)** and observed **(C, D)** RMSD values in winter **(A, C)** and summer **(B, D)**, wherein the arrows are HF radar surface currents, and **(E)** temporal distribution of mean RMSD of predict and true: blue line is the RMSD for unclassified control experiments; orange line is the RMSD for experiments with year-round data; green is the RMSD for experiments with quarterly data; red is the RMSD for experiments with monthly data; purple line is for real RMSD between observation and simulation.

TABLE 2 Error statistics for each group of experiments (in cm).

	ME	AE	RMSD
Year	-0.46	1.43	1.98
Season	-0.43	1.49	2.01
Month	-0.58	1.74	2.46
CTRL	-0.69	1.77	2.38

agreement with the observations, and the classification-based experiments can also give good error predictions. This study will provide a basis for subsequent model error testing, forecast product improvement and data assimilation.

Some known features of surface currents in the region (Cann and Serpette, 2009; Rubio et al., 2013a; Solabarrieta et al., 2014) are described by 6 current groups separated by KMA analysis. The main current is eastward (northward) on the Spanish (French) shelf/slope in winter, and more variable currents to the west and southwest in summer. The circulation characteristics of the model's classification results are in good agreement with the HF radar results, especially for currents on the Spanish (French) shelf/slope.

Clusters of westward currents have lower mean velocities but occur more than 60% frequently. About 25.3% of the average surface currents in the region are dominated by northward currents, especially in the central part of the region, with mean velocities ~15 cm/s. The east-northward current on the Spanish (French) shelf/slope is very strong, with mean velocities exceeding 20 cm/s, and a cyclonic vortex exists offshore north of the Spanish continental shelf.

Prediction errors in the experiments based on classification of all data reached ~2.0 cm/s, with an improvement of ~17% over the control experiments without the classification process. However, the error prediction in the classification experiments based on monthly data was not satisfactory and may improve as

the amount of data increases. This study will provide the basis for subsequent model error testing, forecast product improvement and data assimilation.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Author contributions

ZW: Conceptualization, Methodology, Validation, Writing – original draft. MD: Conceptualization, Formal analysis, Methodology, Project administration, Writing – review & editing. PDM-F: Conceptualization, Methodology, Writing – review & editing. ER: Supervision, Writing – review & editing. NA: Methodology, Writing – review & editing. DW: Writing – review & editing. BL: Data curation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by National Key Research and Development Program of China (No.2021YFC3101604; No.2021YFC3101504).

Acknowledgments

Thanks Copernicus Marine Service (CMEMS) for providing the IBI model outputs and the Euskoos HF radar data (reprocessed data and near real-time data).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Arzoo, K., and Rathod, K. R. (2017). K-Means algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8. 2. *Int. Res. J. Eng. Technol.* 4, 2363–2368.

Bora, J. D., and Gupta, A. K. (2014). Effect of different distance measures on the performance of K-means algorithm: an experimental study in Matlab. *Int. J. Comput. Sci. Inf. Technol.* 5 (2), 2501–2506. doi: 10.48550/arXiv.1405.7471

Caballero, A., Mulet, S., Ayoub, N., Manso-Narvarte, I., Davila, X., Boone, C., et al. (2020). Integration of HF radar observations for an enhanced coastal Mean Dynamic Topography. *Original Res.* 7. doi: 10.3389/fmars.2020.588713

Camus, P., Mendez, F. J., Medina, R., and Cofiño, A. S. (2011). Analysis of clustering and selection algorithms for the study of multivariate wave climate. *Coast. Eng.* 58, 453–462. doi: 10.1016/j.coastaleng.2011.02.003

Cann, B. L., and Serpette, A. (2009). Intense warm and saline upper ocean inflow in the southern Bay of Biscay in autumn–winter 2006–2007. *Continental Shelf Res.* 29, 1014–1025. doi: 10.1016/j.csr.2008.11.015

Cao, C., Bao, L., Gao, G., Liu, G., and Zhang, X. (2024). A novel method for ocean wave spectra retrieval using deep learning from sentinel-1 wave mode data. *IEEE Transaction Geosci. Remote Sens.* 62, 1–16. doi: 10.1109/TGRS.2024.3369080

Charria, G., Lazure, P., Cann, L. B., Serpette, A., Reverdin, G., Louazel, S., et al. (2013). Surface layer circulation derived from Lagrangian drifters in the Bay of Biscay. *J. Mar. Syst.* 109, S60–S76. doi: 10.1016/j.jmarsys.2011.09.015

Charulatha, B., Rodrigues, P., Chitralekha, T., and Rajaraman, A. (2013). A Comparative study of different distance metrics that can be used in Fuzzy Clustering Algorithms. *Software Syst. Green Computing Tamil Nadu India* 2013.

Fontán, A., and Cornuelle, B. (2015). Anisotropic response of surface circulation to wind forcing, as inferred from high-frequency radar currents in the southeastern B ay of B iscay. *Ocean Sci.* 120, 2945–2957. doi: 10.1002/2014JC010671

Fontán, A., Esnaola, G., Sáenz, J., and González, M. (2013). Variability in the air–sea interaction patterns and timescales within the south-eastern Bay of Biscay, as observed by HF radar data. *J. Geophys. Res. Oceans* 9, 399–410. doi: 10.5194/OS-9-399-2013

Gao, G., Yao, B., Li, Z., Duan, D., and Zhang, X. (2024). Forecasting of sea surface temperature in eastern tropical pacific by a hybrid multiscale spatial–temporal model combining error correction map. *IEEE Transaction Geosci. Remote Sens.* 62, 1–22. doi: 10.1109/TGRS.2024.3353288

Guanche, Y, Mínguez, R., and Méndez, J. (2013). Autoregressive logistic regression applied to atmospheric circulation patterns. *Climate Dynamics*. 42 (1–2), 537–552. doi: 10.1007/s00382-013-1690-3

Gurgel, K. W., Essen, H. H., and Schlick, T. (2001). "The University of Hamburg WERA HF radar-theory and solutions," in *Proc. of the 1st International Radiowave Oceanography Workshop ROW-2001*. (Oregon, USA: First International Radiowave Oceanography Workshop (ROW2001)), 19–25.

Hirose, N., Fukumori, I., and Zlotnicki, V. (2001). Modeling the high-frequency barotropic response of the ocean to atmospheric disturbances: Sensitivity to forcing, topography, and friction. *J. Geophysical Res.* 106, 30987–30995. doi: 10.1029/2000JC000763

Jothi, R., Mohanty, S. K., and Ojha, A. (2019). DK-means: a deterministic k-means clustering algorithm for gene expression analysis. *Pattern Anal. Appl.* 22, 649–667. doi: 10.1007/s10044-017-0673-0

Kersalé, M., Marie, L., Cann, B. L., Serpette, A., Lathuilière, C., Boyer, A. L., et al. (2016). Poleward along-shore current pulses on the inner shelf of the Bay of Biscay. *Estuar. Coast. Shelf Sci.* 179, 155–171. doi: 10.1016/j.ecss.2015.11.018

Lellouche, J. M., Greiner, E., Le Galloudec, O., Garric, G., Regnier, C., Drevillon, M., et al. (2018). Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1/12° high-resolution system. *Ocean Sci.* 14, 1093–1126. doi: 10.5194/os-14-1093-2018

Levier, B., Guillaume, R., Romain, E., Elodie, G., Stefania, C., Arancha, A., et al. (2023). "Atlantic - European North West shelf - ocean physics analysis and forecast product," (Copernicus Marine Service), 34.

Lyard, F. H., Allain, D. J., Cancet, M., Carrère, L., and Picot, N. (2021). FES2014 global ocean tide atlas: design and performance. *Ocean Sci.* 17, 615–649. doi: 10.5194/os-17-615-2021

Paduan, J. D., and Rosenfeld, L. K. (1996). Remotely sensed surface currents in Monterey Bay from shore-based HF radar (Coastal Ocean Dynamics Application Radar). *J. Geophysical Res.* 101, 20669–20686. doi: 10.1029/96JC01663

Paduan, J. D., and Washburn, L. (2013). High-frequency radar observations of ocean surface currents. *Annu. Rev. Mar. Sci.* 5, 115–136. doi: 10.1146/annurev-marine-121211-172315

Rehioui, H., Idrissi, A., Abourezq, M., and Zegrari, F. (2016). DENCLUE-IM: A new approach for big data clustering. *Proc. Comput. Sci.* 83, 560–567. doi: 10.1016/j.procs.2016.04.265

Roarty, H., Cook, T., Hazard, L., George, D., and Harlan, J. (2019). The global high frequency radar network. *Front. Mar. Sci.* 6, 164. doi: 10.3389/fmars.2019.00164

Röhrs, J., Sperrevik, A. K., Christensen, K. H., Broström, G., and Breivik, Ø. (2015). Comparison of HF radar measurements with Eulerian and Lagrangian surface currents. *Ocean Dynamics* 65, 679–690. doi: 10.1007/s10236-015-0828-8

Rubio, A., Caballero, A., Orfila, A., Hernández-Carrasco, I., Ferrer, L., González, M., et al. (2018). Eddy-induced cross-shelf export of high Chl-a coastal waters in the SE Bay of Biscay. *Remote sens. Environ.* 205, 290–304. doi: 10.1016/j.rse.2017.10.037

Rubio, A., Fontán, A., Lazure, P., González, M., Valencia, V., Ferrer, L., et al. (2013a). Seasonal to tidal variability of currents and temperature in waters of the continental slope, southeastern Bay of Biscay. *J. Mar. Syst.* 109, S121–S133. doi: 10.1016/j.jmarsys.2012.01.004

Rubio, A., Hernández-Carrasco, I., Orfila, A., González, M., Reyes, E., Corgnati, L., et al. (2020). A lagrangian approach to monitor local particle retention conditions in coastal areas. copernicus marine service ocean state report. *J. Oper. Oceanogr.* 13, 570–583.

Rubio, A., Manso-Narvarte, I., Caballero, A., Corgnati, L., Mantovani, C., Reyes, E., et al. (2019). The seasonal intensification of the slope Iberian Poleward Current. *J. Oper. Oceanogr.* 12, 13–18

Rubio, A., Reverdin, G., Fontán, A., González, M., and Mader, J. (2011). Mapping near-inertial variability in the SE Bay of Biscay from HF radar data and two offshore moored buoys. *Geophys. Res. Lett.* 38. doi: 10.1029/2011GL048783

Rubio, A., Solabarrieta, L., Gonzalez, M., Mader, J., Castanedo, S., Medina, R., et al. (2013b). Surface circulation and Lagrangian transport in the SE Bay of Biscay from HF radar data. *Oceans IEEE*. doi: 10.1109/OCEANS-Bergen.2013.6608039

Sajana, T., Rani, C. M. S., and Narayana, K. V. (2016). A survey on clustering techniques for big data mining. *Indian J. Sci. Technol.* 9, 1–12. doi: 10.17485/ijst/2016/v9i3/75971

Saqib, S., Ditta, A., Khan, M. A., Kazmi, S. A. R., and Alquhayz, H. (2021). Intelligent dynamic gesture recognition using CNN empowered by edit distance. *Comput. Materials Continua* 66, 2061–2076. doi: 10.32604/cmc.2020.013905

Shulman, I. G., and Paduan, J. D. (2009). Assimilation of HF radar-derived radials and total currents in the Monterey Bay area. *Deep Sea Res. Part II: Topical Stud. Oceanography* 56, 149–160. doi: 10.1016/j.dsr2.2008.08.004

Solabarrieta, L., Frolov, S., Cook, M., Paduan, J., Rubio, A., González, M., et al. (2016). Skill assessment of hf radar–derived products for lagrangian simulations in the bay of biscay. *J. Atmospheric Oceanic Technol.* 33, 2585–2597. doi: 10.1175/JTECH-D-16-0045.1

Solabarrieta, L., Rubio, A., Cárdenas, M., Castanedo, S., Esnaola, G., Méndez, F. J., et al. (2015). Probabilistic relationships between wind and surface water circulation patterns in the SE Bay of Biscay. *Ocean Dynamics* 65, 1289–1303. doi: 10.1007/s10236-015-0871-5

Solabarrieta, L., Rubio, A., Castanedo, S., Medina, R., Charria, G., and Hernandez, C. (2014). Surface water circulation patterns in the southeastern Bay of Biscay: New evidences from HF radar data. *Continental Shelf Res.* 74, 60–76. doi: 10.1016/j.csr.2013.11.022

Sotillo, M. G., Cailleau, S., Lorente, P., Levier, B., Reffray, G., Amo-Baladrón, A., et al. (2015). The MyOcean IBI Ocean Forecast and Reanalysis Systems: operational products and roadmap to the future Copernicus Service. *J. Oper. Oceanogr.* 8, 63–79. doi: 10.1080/1755876X.2015.1014663

Tranchant, B., Remy, E., Greiner, E., and Legalloudec, O. (2019). Data assimilation of Soil Moisture and Ocean Salinity (SMOS) observations into the Mercator Ocean operational system: focus on the El Niño 2015 event. *Ocean Sci.* 15, 543–563. doi: 10.5194/os-15-543-2019

Velmurugan, T., and Santhanam, T. (2011). A survey of partition based clustering algorithms in data mining: An experimental approach. *Inf. Technol. J.* 10, 478–484. doi: 10.3923/itj.2011.478.484

Wang, Z., Liu, G., Li, W., Wang, H., and Wang, D. (2023). Development of the operational oceanography forecasting snystem in the northwest pacific. *J. Physics: Conf. Ser.* 2486. doi: 10.1088/1742-6596/2486/1/012032

Wu, M., Li, X., Liu, C., Liu, M., Zhao, N., Wang, J., et al. (2019). Robust global motion estimation for video security based on improved k-means clustering. *J. Ambient Intell. Humanized Computing* 10, 439–448. doi: 10.1007/s12652-017-0660-8

Zerhari, B., Lahcen, A. A., and Mouline, S. (2015). Big Data Clustering: Algorithms and Challenges. International Conference on Big Data, Cloud and Applications BDCA'15.2015..