Directional Spectra-based Clustering for Visualizing Patterns of Ocean Waves and Winds

Carolina Euán & Ying Sun

To cite this article: Carolina Euán & Ying Sun (2019): Directional Spectra-based Clustering for Visualizing Patterns of Ocean Waves and Winds, Journal of Computational and Graphical Statistics, DOI: [10.1080/10618600.2019.1575745](https://doi.org/10.1080/10618600.2019.1575745)

To link to this article: <https://doi.org/10.1080/10618600.2019.1575745>

View supplementary material [↗](#)

Accepted author version posted online: 15 Feb 2019.

Submit your article to this journal [↗](#)

Article views: 23

View Crossmark data [↗](#)

7Directional Spectra-based Clustering for Visualizing Patterns of Ocean Waves and Winds

Carolina Euán* and Ying Sun

CEMSE Division, King Abdullah University of Science and Technology (KAUST)

*Correspondence to: carolina.euancampos@kaust.edu.sa

February 6, 2019

Abstract

The energy distribution of wind-driven ocean waves is of great interest in marine science. Discovering the generating process of ocean waves is often challenging and the direction is the key for a better understanding. Typically, wave records are transformed into a directional spectrum which provides information about the wave energy distribution across different frequencies and directions. Here, we propose a new time series clustering method for a series of directional spectra in order to extract the spectral features of ocean waves and develop informative visualization tools to summarize identified wave clusters. We treat directional distributions as functional data of directions, and construct a directional functional boxplot to display the main directional distribution of the wave energy within a cluster. We also trace back when these spectra were observed, and we present color-coded clusters on a calendar plot to show their temporal variability. For each identified wave cluster, we analyze wind speed and wind direction hourly to investigate the link between wind data and wave directional spectra. The performance of the proposed clustering method is evaluated by simulations and illustrated by a real-world dataset from the red sea.

Keywords: Directional energy distribution, Functional boxplot, Functional data analysis, Spectral analysis, Time Series.

1 Introduction

Sea waves are one of the most important environmental factors affecting the design and construction of maritime structures (Goda, 2000). To ensure they meet their designated performance, a full understanding of wave patterns is needed. In general, wave modeling is challenging because of its randomness. Naive methods that represent sea waves with a constant height and period ignore the randomness of wave behavior and produce a poor description of sea waves. The randomness of this behavior is explained by the generation process of sea waves, its interaction with local wind, and its propagation. Wind-generated waves start in deep water and become swell when they move out of the generating area. Another transformation of the waves is caused by diffraction from islands or headlands. These physical phenomena produce small and large wave patterns/systems moving in several directions. One of the challenges is to estimate the random generating process of a sea wave, using the recording of its observation. When swell mixes with local wind-generated waves, it cannot be easily identified using the wave record; however, it can be clearly identified using the wave spectrum (Ochi, 1998).

In marine engineering, wave records are typically transformed into a directional spectrum $S(\omega, \theta)$, which provides information about the wave energy distribution across different frequencies, ω , and directions, θ . The wave spectra without considering directions $f(\omega)$ can be viewed as the integration of $S(\omega, \theta)$ over all directions. Swell can be identified using wave spectra if the local wind-generated waves are developed in higher frequencies. However, if this is not the case, ignoring the direction might fail to separate the two systems. In the analysis of complex wave data, it is necessary to consider the directional aspects of the spectrum, which motivates us to develop new clustering methodologies based on two-dimensional (2D) directional wave spectra. Our proposed methods aim to identify different wave patterns along time. We also propose a set of visualization tools for 2D clustering results.

In oceanography, the most commonly used method for identifying wave systems is the partitioning method (PM) proposed by Gerling (1992). The PM consists of two steps: 1) each directional spectrum is segmented into significant partitions that contain unimodal spectra; and 2) for each partition, P_i , compute the peak frequency, $\omega_p = \operatorname{argmax}_{\omega \in P_i} S(\omega, \theta)$, peak direction, $\theta_p = \operatorname{argmax}_{\theta \in P_i} S(\omega, \theta)$, and significant wave height $H_s = 4 \sqrt{\int_{P_i} f(\omega) d\omega}$. Partitions with similar (ω_p, θ_p) are assumed to belong to the same system. Obviously, the PM has several disadvantages, such as there are many tuning parameters need to be selected in the procedure, the definition of the significant partition is subjective and it tends to overestimate the number of wave systems. New methodologies have been proposed to overcome these limitations (see Kerbiriou et al., 2007; Portilla et al., 2009; Ailliot et al., 2013), but they are not entirely satisfactory. Our method is different from the PM. Instead of partitioning the spectrum with summary statistics, we treat the spectrum as a 2D function. By clustering these directional spectra along time, we analyze the seasonal wave patterns. Specifically, we propose the Hierarchical Directional Spectra-based (HDS) clustering method to identify directional spectra with energy concentrated in similar directions and frequencies.

When considering both frequencies and directions of the spectrum, visualizing the clustering results becomes much more challenging. We propose a set of tools for clustering visualization: 1) a functional boxplot of the wave spectra to visualize variability in frequency within a cluster, 2) a directional functional boxplot to visualize the most common directional distribution of the energy in each cluster and its variability, 3) a calendar plot colored by cluster membership to visualize the clustering patterns across various time periods during an entire year.

The visualization tools together with the clustering methodologies are valuable tools to analyze complex and large wave data sets. We demonstrate our

methods on the wave data collected from the Red sea. The Red sea is a stretched basin that covers 2250 km from northwest to southwest. Waves are mainly generated by local wind, and the swell causes only a small proportion of the waves. Section 2 describes the wave and wind data and the challenges when analyzing this data set. The rest of the paper is organized as follows. Section 3 introduces the directional spectrum and stochastic models for sea waves. Then, the proposed clustering methodology based on the directional spectra is described in Section 4. Section 5 illustrates the directional fbplot for visualizing the clustering results. Section 6 presents simulated examples and comparisons with PM. Finally, Section 7 shows the data analysis of wind and waves on the Red sea.

2 Data Description

Before 2009, there were no wind or wave measurements on the Red sea, and research relied then on physical models. In 2009, the first buoy was installed nearby Thuwal, Longitude 38.5 and Latitude 22.16, see Figure 1 (Farrar et al., 2009). Buoy data have since become a valuable resource in calibrating physical models used to reproduce wind and waves in the entire red sea area (Ralston et al., 2013; Langodan et al., 2014, 2016). Wave data are recorded at 2 Hz during the first 17 minutes of each hour. Based on these recordings the directional periodograms are computed, $S(\omega, \theta)$ with $\omega \in (0, 0.5)$ and $\theta \in [0, 360)$. The buoy also measures the wind direction and wind speed.

Figure 2 shows 12 directional spectra of the wave data recorded hourly, along with the wind speed and directions for January 16, 2010. We observe one image per hour that corresponds to the directional periodogram, arrows indicate wind direction and its length represents wind speed. Figure 2, clearly shows that the directional spectrum can have high variability, even within a day. This might be partially caused by high noise levels in the buoy data. If we consider four direction quadrants, NE, NW, SW and SE, we can see that the wind direction and

the highest mode of the wave spectra lie in the same direction quadrant when the wind speed is high. We can then conclude that wave patterns are dominated by wind-driven waves. If the directional spectrum shows that the energy is concentrated in two different directions, it is reasonable to consider the presence of swell. Usually, the energy at higher frequency agrees with the wind direction and the other direction is assigned to the swell direction. In the Red sea data, the presence of swell does not last for a long period of time and the wind-driven waves can be produced in low frequencies that are close to the swell frequency.

We analyze buoy data over the period Nov/14/2009 - Jul/27/2010 for a total of 6087 directional spectra, after removing 44 hours of data with missing values. Data simulated from physical models, developed by Langodan et al. (2014), are also available at this location. To identify possible wave patterns and facilitate the subsequent association analysis using available wind data, we apply the HDS clustering method. Also, we compare clustering results from buoy data and physical model output data to analyze whether the physical model can reproduce wave and wind patterns from the observed buoy data. This information can be useful for the calibration of the physical model. Data simulated from physical models are less noisy than those from different locations on the Red sea.

3 Statistical Models for Sea Waves

We first introduce stochastic models for sea waves. These models, which have been used since the 1950s (Pierson, 1955; Longuet-Higgins, 1957), represent sea waves as a linear superposition of infinite elementary waves. For a fixed location s , assuming that the spectral density exists, the height (vertical displacement) of the sea surface at time t , $X_s(t)$ is modeled by

$$X_s(t) = \int e^{it\omega} \sqrt{f_s(\omega)} dW(\omega), \text{ where } dW \text{ is complex white noise, and } f_s(\omega)$$

represents wave energy in frequency ω at the location s .

Let's consider the JONSWAP (Joint North-Sea Wave Project) spectral family as an example (Hasselmann et al., 1973). This is a parametric family of spectral densities frequently used in oceanography with

$$f_s(\omega) = \frac{g^2}{\omega^5} \exp(-5\omega_p^4 / 4\omega^4) \gamma^{\exp(-(\omega - \omega_p)^2 / 2 \omega_p^2 r^2)} \quad (1)$$

where g is the acceleration of gravity, $r = 0.07$ if $\omega \leq \omega_p$ and $r = 0.09$ otherwise; $\omega_p = \pi / T_p$ and γ is a parameter. Other parameters for the model are the significant wave height H_s and the spectral peak period T_p , both of which might depend on location s . With the JONSWAP model, the swell spectrum exhibits a peak much sharper than the wind-driven waves spectrum (Goda, 2000). However, this model does not consider directions.

Let $S_s(\omega, \theta)$ be the directional spectrum of $X_s(t)$. Then, for a given spectral density $f_s(\omega)$, $S_s(\omega, \theta)$ is commonly considered as

$$S_s(\omega, \theta) = f_s(\omega) D_s(\omega, \theta), \quad (2)$$

where $D_s(\omega, \theta)$ is called the spreading function or directional distribution of energy at frequency ω (Borgman, 1969) and satisfies $\int D_s(\omega, \theta) d\theta = 1$. In this case, $X_s(t)$ is modeled by

$$X_s(t) = \int_{\omega} \int_{-\pi}^{\pi} e^{i(t\omega + \mathbf{k} \cdot \mathbf{r})} \sqrt{S_s(\omega, \theta)} dW(\omega, \theta),$$

where $\mathbf{k} \cdot \mathbf{r} = k_1 \sin(\theta) + k_2 \cos(\theta)$. For simplicity, we denote $f(\omega)$, $D(\omega, \theta)$, and $S(\omega, \theta)$ without the sub-index s unless it is necessary to distinguish between two different locations. Longuet-Higgins et al. (1963) and Cartwright (1961) proposed the first models for spreading functions. Longuet-Higgins et al. (1963) considered $D(\omega, \theta)$ as a power of cosine function centered in a peak direction θ_0 , i.e.,

$$D(\omega, \theta) = G_0 \cos^{2m} \left(\frac{\theta - \theta_0}{2} \right), \quad (3)$$

where θ_0 is the peak direction, G_0 is a normalized constant and $m = m_{max} (\omega / \omega_p)^5$ if $\omega < \omega_p$ and $m = m_{max} (\omega / \omega_p)^{-5/2}$ in all other cases, and m_{max} is a parameter. Statistical models for the directional spectrum can be found in Liu et al. (2003).

Throughout this paper, we use the JONSWAP family (1) and the spreading function in (3) to simulate the directional spectrum $S(\omega, \theta)$. $S(\omega, \theta)$ can be estimated either parametrically via maximum likelihood methods, entropy methods, and Bayesian methods, or nonparametrically via the directional periodogram. Parametric methods rely on wave model assumptions which are normally unknown in practice. Therefore, we choose to use a nonparametric smoothing method to estimate $S(\omega, \theta)$ from the noisy directional periodograms.

4 Directional Spectra-based Clustering Methods

When oscillations are of interest, time series clustering methods based on spectra can be applied (Xu and Wunsch, 2005; Caiado et al., 2015). A similarity in spectra means analogous behavior in the oscillatory wave patterns. Alvarez-Esteban et al. (2016) developed a hierarchical time series method for determining stationary intervals of random sea waves. Their method has been used to identify stationary intervals and their statistical properties in different sea conditions (Euán et al., 2014; Alvarez-Esteban et al., 2016). Another methodology using a time series clustering approach is the hierarchical spectral merger algorithm (HSM) developed by Euán et al. (2018). In the HSM method, each potential cluster is estimated with a representative spectrum, for each step of the clustering algorithm. These methods are developed for spectral density functions of frequencies without considering directional information. We propose the **Hierarchical Directional Spectra-based (HDS)** method to cluster directional wave spectra with energy distributed in similar directions and frequency, which

can be viewed as a generalization of the HSM method for directional wave spectra.

HDS Algorithm.

Let $\{S_i(\omega, \theta)\}$ for $i = 1, \dots, n$ be a set of n directional spectra.

Step 1.

Compute the normalized directional spectra, $\{S_i^N(\omega, \theta)\}$, where

$$S_i^N(\omega, \theta) = \frac{S_i(\omega, \theta)}{\int S_i(\omega, \theta) d\theta d\omega}, i = 1, \dots, n$$

. Start with n clusters, $\{\mathcal{C}_i\}_{i=1, \dots, n}$, each of them containing one of the directional spectra.

Step 2.

Find the most similar clusters as follows,

1. Compute the dissimilarity measure between the directional spectra of each cluster using the total variation (TV) distance, i.e,

$$D(S_i^N, S_j^N) = \frac{1}{2} \int |S_i^N(\omega, \theta) - S_j^N(\omega, \theta)| d\theta d\omega.$$

2. Find the two clusters that have the smallest TV distance, $\mathcal{C}_{i_1}$ and $\mathcal{C}_{i_2}$, and merge the directional spectra into a new cluster, $\mathcal{C}_{New} = \mathcal{C}_{i_1} \cup \mathcal{C}_{i_2}$. Remove $\mathcal{C}_{i_1}$ and $\mathcal{C}_{i_2}$ from the cluster set and update.

Step 3.

The representative directional spectrum in $\mathcal{C}_{New}$ is computed by averaging

$$S_{New}^N(\omega, \theta) = \frac{1}{\#\mathcal{C}_{New}} \sum_{S_j^N \in \mathcal{C}_{New}} S_j^N(\omega, \theta)$$

all directional spectra in the cluster, i.e.,

Step 4.

Repeat Steps 2 and 3 until there is only one cluster left.

To choose the number of clusters we consider the scree plot criteria proposed in Euán et al. (2018). In general, other distances might be chosen, but we prefer the total variation distance due to useful properties such as bounded between $[0,1]$, detection of small differences between functions and the geometrical interpretation. The HDS clustering method can be very useful when both direction and frequency are of interest. However, the high noise level in real data sets may produce a large number of clusters. We illustrate this with an example from the Red Sea data set.

Example 1. Consider a small subsample of the KAUST buoy data set that contains four consecutive hours, 16:00 to 19:00, of Jan 03, 2010. Figure 3(a) shows the observed directional spectra. These spectra are very dispersed in direction and the presence of noise produces differences between the spectra even in consecutive recordings. After normalizing all these functions (i.e., dividing them by their norms), we compute the total variation (TV) distance between the $S(\omega, \theta)$ functions. The computed values between consecutive directional spectra are 0.252, 0.276 and 0.288. These results can be interpreted as less than 75% of the total energy comes from the same frequency and direction. Then, these spectra could be assigned to different clusters.

For applications where the cluster in either the frequency or the direction is of primary interest, we also propose an algorithm to perform the clustering marginally. Assume the spreading model (2) for the directional spectrum holds. We define the two marginal densities as follows:

$$f(\omega) = \int S(\omega, \theta) d\theta \quad \text{and} \quad g(\theta) = \int S(\omega, \theta) d\omega.$$

Note that the marginal density on frequency $f(\omega)$ corresponds to the frequency spectra. Marginal clustering with respect to frequency is equivalent to the HSM clustering. Then, $g(\theta)$ is the directional distribution and represents the total energy in direction θ . We also consider the normalized directional distribution

$g^N(\theta) = \frac{g(\theta)}{\int_{\theta} g(\theta) d\theta}$ which represents the proportion of the energy in direction θ . In

Example 1, we compute both the corresponding $g^N(\theta)$ and the TV distance values between these functions. Figure 3(b) plots the directional distributions. The computed TV values are 0.075, 0.052 and 0.090 which are much smaller differences and reflects better the similarity in direction of the directional spectra.

The **direction-only** version of the HDS clustering method, HDSd, is based on the normalized directional distributions, $\{g_i^N(\theta)\}, i = 1, \dots, n$. The only difference is that in Step 2 of the HDS algorithm, instead of 2D TV distance, we compute the one dimensional TV distance, i.e.,

$$D(g_i^N, g_j^N) = \frac{1}{2} \int |g_i^N(\theta) - g_j^N(\theta)| d\theta.$$

If the HDS algorithm assigns $\{S_i(\omega, \theta)\}_{i=1, \dots, l}$ from l consecutive time segment, i.e., $[t_1, t_2), [t_2, t_3), \dots, [t_l, t_{l+1})$, to be in the same cluster, it implies that within the time period $[t_1, t_{l+1})$, the waves show similar directions and frequencies. In contrast, when the HDSd method is applied it does not imply similarity in frequency spectra. The HDSd method is useful in situations where the range of frequencies is almost constant and when there is major interest in directional features of the spectrum. In **Example 1**, the HDSd assigns three consecutive spectra to the same cluster, which is more interpretable for this data. In general, we recommend to perform the HDS method both jointly and marginally, and use the visualization tools we proposed to decide which algorithm is more appropriate for a given dataset. We will illustrate this point in Section 7 when analyzing the Red sea data.

5 Directional Functional Boxplots

For functional data clustering, it is also challenging to summarize or visualize the identified clusters which contain a set of functions. In this section, we present a

methodology to visualize clustering results of functional data. The main tool for visualization of functional data is based on the concept of functional data ranking. A review on the most commonly used ranking methods for functional data can be found in Lin and Zhou (2017). Although we focus on the clustering of directional spectra, our methodology can be used to summarize the identified clusters from any clustering algorithm with functional objects. The visualization tools that we developed use R (R Core Team, 2018) packages such as fda (Ramsay et al., 2017), ggplot2 (Wickham, 2009) and shiny app (Chang et al., 2018).

Let $\{S_1(\omega, \theta), \dots, S_{n_k}(\omega, \theta)\}$ be a set of directional spectra that belong to a cluster $\mathcal{C}$. Then, we compute the set of frequency spectra and directional distributions, $\{f_1(\omega), \dots, f_{n_k}(\omega)\}_{\omega \in [0, .5)}$ and $\{g_1(\theta), \dots, g_{n_k}(\theta)\}_{\theta \in [0, 360)}$. We visualize cluster information with the functional boxplots (fbplots) of the frequency spectra and the directional fbplots of the directional distributions.

Functional boxplots. Sun and Genton (2011) proposed the fbplots to visualize functional data based on functional data ranking. Let $\{y_1(t), \dots, y_n(t)\}_{t \in \mathcal{I}}$ be a set of functional data and let $\{y_{[1]}(t), \dots, y_{[n]}(t)\}_{t \in \mathcal{I}}$ be the corresponding ordered functions according to the decreasing values of the modified band depth introduced by López-Pintado and Romo (2009). In the fbplot, the border of the box is computed by considering the band that represents the 50% of the deepest curves, i.e.,

$$C_{0.5} = \left\{ (t, y(t)) : \min_{r=1, \dots, \lceil n/2 \rceil} y_{[r]}(t) \leq y(t) \leq \max_{r=1, \dots, \lceil n/2 \rceil} y_{[r]}(t) \right\}.$$

The box gives an indication of the spread of the central 50% of the curves. The curve inside the box is the median curve $y_{[1]}(t)$, which corresponds to the deepest curve. The fbplot detects possible outliers as those curves outside the fences, which are obtained by inflating the envelope of the 50% central region by a factor 1.5.

Directional functional boxplots. In the directional fbplot, the functional boxplot is plotted using polar coordinates. Since directional distributions are defined in $[0, 360)$ degrees, it is natural to consider a circular plot for these functions. We consider the use of polar coordinates to plot the directional distributions. With a polar coordinate, the direction θ becomes an angle and the proportion of the energy in a fixed direction becomes a radius. Then, along the directions which the curve is pulled away from the center are those driving major energy. One advantage of the directional fbplot over the fbplot is the fact that the directional fbplot allows us to naturally match the angle (in degrees) with the directions and that there is no boundary effect.

Example 2. Consider two different directional spectra, $S_1(\omega, \theta)$ and $S_2(\omega, \theta)$ as in (2) with peak at $(.3, \pi / 4)$ and $(.2, 4\pi / 5)$, respectively. Also, consider $S_3(\omega, \theta)$ as $a_1 S_1(\omega, \theta) + a_2 S_2(\omega, \theta)$ where $a_i = \int S_i(\omega, \theta) d\theta d\omega$. Figure 4(a) plots these directional spectra. We simulate 50 noisy observations of each directional spectrum and consider them as a cluster. The features of the 3 clusters is summarized in Figure 4(b) and (c) via the fbplots of frequency spectra and directional fbplots of directional distributions. For Clusters 1 and 2, the fbplots clearly show the differences in frequency and the directional fbplot shows that energy is distributed in different directions, NE and NW. It also shows that the variability is higher in Cluster 1 than in Cluster 2. The fbplot of Cluster 3 suggests the presence of two dominant frequencies and the directional fbplots indicates the presence of two major directions. This can be explained by S_3 which has two modes from S_1 and S_2 respectively.

6 Simulation Study

Here, we present a simulation study to test the HDS clustering method. We examine its accuracy when used to identify clusters with specific features of the directional spectrum. Let's consider $S(\omega, \theta)$ as a directional spectrum defined by (2), where $f(\omega)$ belongs to the JONSWAP parametric family and $D(\omega, \theta)$ is a

cosine type spreading function as in (3). We then assume that the observed directional spectrum of a wave recording follows the model below

$$S_{obs}(\omega, \theta) = S(\omega, \theta) + \varepsilon(\omega, \theta), \quad (4)$$

where $\varepsilon(\omega, \theta)$ is a Gaussian field as in Ailliot et al. (2013) with

$$\Sigma_{i,j} = \text{Cov}(\varepsilon(\omega_i, \theta_i), \varepsilon(\omega_j, \theta_j)) = \sigma^2 \exp \left(-\lambda_1 \left| \frac{1}{\omega_i} - \frac{1}{\omega_j} \right| - \lambda_2 (1 - \cos(\theta_i - \theta_j)) \right),$$

$\lambda_1 = 0.5$, $\lambda_2 = 10$ and σ varies from 0.001 to 0.05.

We consider two experiments with three different directional spectra. Table 1 shows the corresponding parameters for the spectral density and spreading functions. Figure 5(a) shows the directional spectra corresponding to the parameters in Experiment 1. We can see that the peaks of $S_1(\omega, \theta)$, $S_2(\omega, \theta)$ and $S_3(\omega, \theta)$ slightly differ in directions and peak frequency. Figure 5(c) shows the directional spectra corresponding to the parameters in Experiment 2, where $S_3(\omega, \theta) \approx S_3^1(\omega, \theta) + S_2(\omega, \theta) + S_1(\omega, \theta)$. In this second experiment, two of the directional spectra are unimodal with peaks at different directions; the third mimics the effect of wind and swell together with a bimodal directional spectrum. For each experiment, we simulate 50 replicates of each of the three directional spectra, i.e., $C_k = \{S_k^1(\omega, \theta), S_k^2(\omega, \theta), \dots, S_k^{50}(\omega, \theta)\}$ for $k = 1, 2, 3$, where $S_k^j(\omega, \theta)$ is simulated from model (4). Then, we apply the HDS and HDSd to classify these 150 directional spectra into three clusters. We also compare the HDS clustering method to the PM.

For the PM, we apply the watershed algorithm to each directional spectra implemented in the EBImage package in R (Pau et al., 2010). We select the *significant* partitions as the ones that have at least 10% of the total energy. Then, we compute the peak frequency and direction for each partition and perform hierarchical clustering based on these quantities. Note that if a directional

spectrum has two partitions, they might be assigned to different clusters. We let two directional spectra belong to the same cluster if at least one of their partitions is assigned to the same cluster. Thus, it might be possible that one spectrum belongs to two clusters.

To evaluate the clustering performance, we consider the similarity index commonly used when the true clustering structure is known (Alvarez-Esteban et al., 2016). Let $\{G_1, \dots, G_g\}$ and $\{C_1, \dots, C_k\}$, be the set of the g real groups and a

$$\text{Sim}(C, G) = \frac{1}{g} \sum_{i=1}^g \max_{1 \leq j \leq k} \text{Sim}(C_j, G_i), \quad \text{where}$$

$$\text{Sim}(C_j, G_i) = \frac{2 |C_j \cap G_i|}{|C_j| + |G_i|}.$$

We consider $M = 1000$ replicates per experiment and compute the similarity index for each of the three methods PM, HDS, and HDSd. Figure 5(b) shows the boxplots of the similarity index for each method when increasing the noise level. Both HDS and HDSd identified the three clusters when the noise level is low. However, the HDSd faces difficulties when increasing the noise, which is expected since the peak directions of $S_2(\omega, \theta)$ and $S_3(\omega, \theta)$ are very close. Figure 5(d) shows the boxplots for Experiment 2., where both HDS and HDSd performed very well. In contrast, PM performed poorly in both experiments. It is mainly because PM tends to detect at least two partitions even when the directional spectra is unimodal. Although it is difficult to compare the HDS clustering to the PM as they are derived from different principles, the HDS shows higher accuracy in identifying the right clustering structure for our purpose. However, we also noticed that although the HDS is more robust than PM, when the noise level increases, the performance of the HDS clustering gets worse. The robustness might be improved by considering more sophisticated smoothing methods although it is often not trivial and requires further research.

7 Red Sea Data Analysis

To study the waves and wind patterns using the Red sea dataset, we apply the HDS clustering method to both buoy data and data from the physical model obtained over the time period Nov/14/2009 - Jul/27/2010. The Red sea is a narrow basin, and the waves do not exhibit a high variability in frequency, due to the geographical location. We found that the peak frequency of the wave spectra is concentrated on the interval $(.15,.22)$ and does not show much variability. In fact, a clustering based on frequency spectra only suggests low number of clusters (see scree plot on Figure 6(b)). The variability of the directional spectral is mostly caused by the variability in direction. Figure 6(a) shows the hourly directional distribution functions of the buoy data for this period. We observe that the dominant wave direction is NE. Although this plot provides a nice summary of the directional properties of the wave data, it is hard to visualize interesting features for small time scale, due to a large amount of data. We applied both the HDS and the HDSd to summarize the data on a hourly scale. The identified clustering structure match well in the leading clusters, although the scree plot of the HDS clustering in 2D suggests a higher number of clusters (see Figure 6(b)). Here, we only present the HDS clustering results and the clustering results for HDSd are in the supplementary material.

7.1 Clustering Results of Buoy Data and Physical Model

We apply the HDS method to the buoy data and to data from the model separately to visualize whether the physical model output can reproduce the observed features of the buoy data. To do this, we choose 30 clusters based on the scree plot criteria (Figure 6(b)) although the scree plot suggests even higher number of clusters. Since clusters with a small number of members are not very informative, we only present the first 9 leading clusters. Then, we match the labeling of the clusters for the model out data with those identified for the buoy data by considering similarities between the spectral features (peak direction and peak frequency).

Figure 7 presents the clustering visualizations based on fbplots of frequency spectra and directional fbplots of directional distributions. For both the buoy and physical model output, the larger cluster (Cluster 1) appears more than 50% of the time, indicating that the wave energy is mainly propagated in NE direction for this particular location. Other potentially influential directions are represented by clusters 2-9, but they occur less than 10% of the time. Clusters 2 and 3 are of particular interest as their peaks along two directions may indicate the presence of swell.

To visualize the temporal distribution of the identified clusters, we show the color-coded clusters on a calendar plot (Figure 8). We can clearly see, from Figure 8, that physical model output matches well the buoy data over the time period Nov/14/2009 - Jul/27/2010. An example of this is the energy from the SW direction for Clusters 4 and 9 (pink and yellow). The two longest time periods during which these clusters appear are approximately 13-19 Jan 2010 and 10-17 March 2010.

After comparing visually the clustering structures obtained from buoy data with those from the physical model output, we compute the Rand statistics (Rand, 1971) to measure the consistency between the two clustering structures. We obtain an index value between 0 and 1, calculated by the co-membership between two clustering results (Tibshirani and Walther, 2005). We compute the Rand statistics by using the *clusteval* package in R (Ramey, 2012). We find the computed value to be 0.644, which is relatively high given high noise levels in the buoy data. This value suggests that the clusters from the buoy data and the physical model output are similar.

7.2 Wind and Waves Patterns

For KAUST buoy data, we investigate the distribution of wind data for each of the identified wave clusters, first by comparing the histogram of the hourly wind directions. Figure 9 presents two examples for Clusters 1 and 2. Cluster 1 is

observed mainly from Nov 14 to Jul 17 in the buoy location with one major direction of the wave energy from NE, as shown in Figure 9(a). Figure 9(b) shows the histogram of the hourly wind directions for the same time periods as Cluster 1. We can see that the major wind direction is also from the NE, suggesting that waves over the time period Nov 14 - Jul 17 are driven by the wind. Cluster 2 corresponds to a wider spread of energy, i.e., the peaks of the energy distribution functions appear along multiple directions. For this cluster, the distribution of the wind directions is also more spread-out, although NE is still the dominant direction.

We then analyze wind speed, and find that, except for clusters 3 and 5, the mean wind directions match well with the wave energy direction. Interestingly, the means of the wind speed for clusters 3 and 5 are among the smallest, slower than 3 meters per second, which may not be strong enough to drive the waves. Cluster 7 is another case with a slow wind, but with two wave energy directions one, one of them suggesting the possible presence of swell. Cluster 7 only appears less than 2% of the time. The complete results of our wind analysis are presented in Table 2.

8 Discussion and Conclusions

The ocean wave and wind data analyzed in our study are very large and complex and contain long time periods of hourly measurements of wave directional spectra, wind speed and direction. We treat the directional spectra as functional data, which better represent the wave systems. In this paper, we propose a clustering method for directional distributions of wave energy over time. The clusters we obtain represent the different wave patterns observed in the Red sea; they help us to associate wind and wave data. We then analyze wind direction and speed for each wave cluster, and cluster information is shown via the fbplot and the directional fbplot.

We also develop a set of visualization tools to compare multiple complex datasets. For example, we compare clustering results of buoy data with simulated data from a physical model output. In general, both data sets show similar results that reflect the fact that the physical model can reproduce accurately the wave systems in the Red Sea. This methodology can be easily applied to datasets with a similar structure, for other locations or time periods. We can also apply our comparative study of buoy data and physical model data for model calibration. With a calibrated physical model, simulations of wave and wind data from the Red sea can be performed at higher temporal and spatial resolutions.

For space-time data, we often need to select subsets for a set of locations or time periods. Therefore, we developed a Shiny App for interactively visualizing wave clusters and wind data,

https://carolinaeuan.shinyapps.io/wind_and_waves_visualization/. In this app, the user can select different time periods (months) from the buoy and physical model data. For clustering results, the user can decide between different numbers of clusters to test their stability, or choose a number of clusters based on the screeplot being displayed. A color-coded calendar plot appears when the user clicks on a specific cluster the fbplot or directional fbplot, and wind information is displayed above the plot.

It is also worth pointing out that although we presented our method and data analysis for a special class of functions, the directional spectra, the proposed HDS clustering method with the total variation distance as the dissimilarity measure is generally suitable for clustering any 2D functions. However, similar numerical studies as presented in this paper should be performed to investigate its performance.

9 Acknowledgements

We thank Dr. Ibrahim Hoteit from *Earth Fluid Modeling and Prediction* group for sharing the Red Sea data set that was used in this paper.

References

Ailliot, P., V. Monbet, and C. Maisondieu (2013). Dynamical partitioning of directional ocean wave spectra. *Probabilistic Engineering Mechanics* 33, 95–102.

Alvarez-Esteban, P. C., C. Euán, and J. Ortega (2016). Statistical analysis of stationary intervals for random waves. In *Proceedings of the 26th International Offshore and Polar Engineering Conference*, pp. 305–311.

Alvarez-Esteban, P. C., C. Euán, and J. Ortega (2016). Time series clustering using the total variation distance with applications in oceanography. *Environmetrics* 27 (6), 355–369. env.2398.

Borgman, L. E. (1969). Directional spectra models for design use. Offshore Technology Conference.

Caiado, J., E. A. Maharaj, and P. D'Urso (2015). *Handbook of Cluster Analysis*, Chapter Time Series Clustering, pp. 241–263. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis.

Cartwright, D. (1961). The use of directional spectra in studying the output of a wave recorder on a moving ship. In *Proceedings of the Conference Ocean Wave Spectra*, Easton, Maryland, pp. 203–218.

Chang, W., J. Cheng, J. Allaire, Y. Xie, and J. McPherson (2018). *shiny: Web Application Framework for R*. R package version 1.1.0.

Euán, C., H. Ombao, and J. Ortega (2018, Apr). The hierarchical spectral merger algorithm: A new time series clustering procedure. *Journal of Classification* 35(1), 71–99.

Euán, C., J. Ortega, and P. C. Alvarez-Esteban (2014). Detecting stationary intervals for random waves using time series clustering. In *Proceedings of the 33rd. International Conference on Ocean and Arctic Engineering*, pp. 1–7. ASME.

Farrar, J. T., S. J. Lentz, J. H. Churchill, P. R. Bouchard, J. C. Smith, J. N. Kemp, J. Lord, G. P. Allsup, and D. S. Hosom (2009). King abdullah university of science and technology (KAUST) mooring deployment cruise and fieldwork. Technical Report DOI: 10.1575/1912/3012, King Abdullah University of Science and Technology.

Garnier, S. (2018). *viridis: Default Color Maps from 'matplotlib'*. R package version 0.5.1.

Gerling, T. W. (1992). Partitioning sequences and arrays of directional ocean wave spectra into component wave systems. *Journal of Atmospheric and Oceanic Technology* 9(4), 444–458.

Goda, Y. (2000, 01). *Random Seas and Design of Maritime Structures*, Volume 33. World Scientific.

Hasselmann, K., T. Barnett, E. Bouws, H. Carlson, D. Cartwright, K. Enke, J. Ewing, H. Gienapp, D. Hasselmann, P. Kruseman, A. Meerburg, P. Mller, D. Olbers, K. Richter, W. Sell, and H. Walden (1973). Measurements of wind-wave growth and swell decay during the joint north sea wave project (JONSWAP). *Deutschen Hydrographischen Zeitschrift* 12, Deutsches Hydrographisches Institut Hamburg.

Kerbiriou, M.-A., M. Prevosto, C. Maisondieu, A. Babarit, and A. Clément (2007). Influence of an improved sea-state description on a wave energy converter production. Number 42711, pp. 463–473.

Langodan, S., L. Cavaleri, Y. Viswanadhapalli, and I. Hoteit (2014). The red sea: A natural laboratory for wind and wave modeling. *Journal of Physical Oceanography* 44 (12), 3139–3159.

Langodan, S., Y. Viswanadhapalli, H. P. Dasari, O. Knio, and I. Hoteit (2016). A high-resolution assessment of wind and wave energy potentials in the red sea. *Applied Energy* 181, 244–255.

Lin, Z. and Y. Zhou (2017, Dec). Ranking of functional data in application to worldwide PM₁₀ data analysis. *Environmental and Ecological Statistics* 24 (4), 469–484.

Liu, Y., X.-H. Yan, and M.-Y. Su (2003). Directional spectrum of wind waves: Part I - model and derivation. *Journal of Ocean University of Qingdao* 2 (1), 1–12.

Longuet-Higgins, M. S. (1957). The statistical analysis of a random, moving surface. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 249 (966), 321–387.

Longuet-Higgins, M. S., D. E. Cartwright, and N. D. Smith (1963). Observations of the directional spectrum of sea waves using the motions of a floating buoy. In *Ocean Wave Spectra, proceedings of a conference, Easton, Maryland*, pp. 111–136. National Academy of Sciences: Prentice-Hall.

López-Pintado, S. and J. Romo (2009). On the concept of depth for functional data. *Journal of the American Statistical Association* 104 (486), 718–734.

Ochi, M. K. (1998). *Ocean waves: the stochastic approach*. Cambridge, U.K.; New York: Cambridge University Press.

Pau, G., F. Fuchs, O. Sklyar, M. Boutros, and W. Huber (2010). EBIImage an R package for image processing with applications to cellular phenotypes. *Bioinformatics* 26 (7), 979–981.

Pierson, W. J. (1955). Wind generated gravity waves. Volume 2 of *Advances in Geophysics*, pp. 93–178. Elsevier.

Portilla, J., F. J. Ocampo-Torres, and J. Monbaliu (2009). Spectral partitioning and identification of wind sea and swell. *Journal of Atmospheric and Oceanic Technology* 26 (1), 107–122.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ralston, D. K., H. Jiang, and J. T. Farrar (2013). Waves in the red sea: Response to monsoonal and mountain gap winds. *Continental Shelf Research* 65, 1–13.

Ramey, J. A. (2012). *clusteval: Evaluation of Clustering Algorithms* (R package version 0.1 ed.). <https://CRAN.R-project.org/package=clusteval>.

Ramsay, J. O., H. Wickham, S. Graves, and G. Hooker (2017). *fda: Functional Data Analysis*. R package version 2.4.7.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66 (336), 846–850.

Sun, Y. and M. G. Genton (2011). Functional boxplots. *Journal of Computational and Graphical Statistics* 20 (2), 316–334.

Tibshirani, R. and G. Walther (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics* 14 (3), 511–528.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Xu, R. and D. Wunsch (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks* 16 (3), 645–678.

Fig. 1 KAUST buoy location. Red icon points the exact location of the KAUST buoy which is located nearby Thuwal.

Fig. 2 Directional spectra of the Buoy data in polar coordinates for January 16, 2010. Radius corresponds to frequency ω , angle corresponds to directions θ , and color scale (Garnier, 2018) is the value of $S(\omega, \theta)$. Arrows indicate wind direction, and its length represents wind speed. When the wind speed is high, the wind direction and the highest mode of the wave spectra lie in the same direction quadrant.

Fig. 3 A subsample of buoy data observed from 16:00 to 19:00 hours of Jan 03, 2010. (a) Normalized directional spectra in polar coordinates. Radius corresponds to frequency ω , angle corresponds to directions θ , and color scale is the value of $S(\omega, \theta)$. (b) The corresponding normalized directional distributions. The four spectra concentrate in a narrow frequency band but widely direction range; the main direction is more straightforward to identify with the directional distributions.

Fig. 4 Clustering visualization plots of **Example 2**. For the unimodal spectra, S_1 and S_2 , the corresponding directional fbplots and fbplots are unimodal concentrate in the main direction and frequency respectively. For the bimodal spectrum, S_3 , the directional fbplot is pulled away from the center in two different directions while the fbplot of the frequency spectra suggests a unimodal spectrum.

Fig. 5 Simulation study. (a)-(c) Normalized directional spectral. Radius corresponds to frequency ω , angle corresponds to directions θ , and color scale is the value of $S(\omega, \theta)$. (b) - (d) Box plots of similarity index with different noise level and clustering methods, PM, HDS and HDSd (from left to right). Compared to the PM, the HDS shows higher accuracy in identifying the right clustering structure.

Fig. 6 (a) Directional distribution $g(\theta)$ in buoy location for the period Nov/14/2009 - Jul/27/2010. Wave energy is concentrated in the NE direction most of the time. (b) Scree plot of the clustering methods: HDS, HDSd and HSM. Dissimilarity values for HDS are higher than the HDSd.

(a) KAUST buoy

(b) Model output

Fig. 7 Clustering Visualization. Right: directional fbplot using polar coordinates. Cluster IDs and colors between buoy data and physical model clusters match when considering similarities between the spectral features (peak direction and peak frequency). Left: fbplot of the frequency spectra per cluster. Subtitles indicate the percentage of the total period of hours that belong to the corresponding clusters and the percentage of detected outliers (curves outside

the fences). Clustering when considering the model output data is highly similar to clustering when considering the buoy data.

Fig. 8 Color-coded calendar plot for buoy data and physical model data. Dates are mapped by row and hours are mapped by column. Color represents cluster IDs. If a cluster has a similar temporal location between the two different data sets, we say that the model output data reproduces the wave energy distribution accurately during this period.

Fig. 9 Buoy data for Clusters 1 and 2. (a) fbplots and directional fbplots; (b) histogram in Euclidean and polar coordinates of wind direction of the corresponding hours in each cluster. For wind data, the y-axis and radial axis represent the density value for each bin.

Table 1 Spectral parameters for Experiment 1 and Experiment 2.

	Experiment 1			Experiment 2		
	$S_1(\omega, \theta)$	$S_2(\omega, \theta)$	$S_3(\omega, \theta)$	$S_1(\omega, \theta)$	$S_2(\omega, \theta)$	$S_3^1(\omega, \theta)$
T_p	1 / .22	1 / .22	1 / .18	1 / .3	1 / .25	1 / .2
H_s	1	1	1	1	2	1
γ	1	1	1	2	2	2
s_{max}	3	5	5	5	5	5
θ_0	$\pi / 2$	$\pi / 4$	$\pi / 8$	$\pi / 4$	$\pi / 2$	$3\pi / 2$

Table 2 Wind analysis showing direction and speed per cluster. Clusters 3, 5, and 7 (in bold font), have the smallest wind speed.

Cluster Direction (degree)			Speed (<i>m / s</i>)			
	Median	Mean	Sd	Median	Mean	Sd
1	56.8	76.16	68.2	6	5.88	2.29
2	68.95	98.15	84.12	4.7	4.7	1.98
3	65.9	111.57	96.47	3	3.13	1.56
4	210.25	195.6	88.95	3.7	4.14	2.27
5	145.9	153.82	96.03	2.9	2.95	1.33
6	53.4	81.52	84.49	8.3	7.71	2.44
7	52.15	85.45	87.88	3.3	3.34	1.54
8	41	52.39	59.44	3.1	3.12	1.98
9	203.1	178.47	106.6	3.7	3.78	1.47