Multivariate Wave Climate Using Self-Organizing Maps

PAULA CAMUS

Environmental Hydraulics Institute IH Cantabria, Universidad de Cantabria, Cantabria, Spain

ANTONIO S. COFIÑO

Santander Meteorology Group, Department of Applied Mathematics and Computer Sciences, Universidad de Cantabria, Cantabria, Spain

FERNANDO J. MENDEZ AND RAUL MEDINA

Environmental Hydraulics Institute IH Cantabria, Universidad de Cantabria, Cantabria, Spain

(Manuscript received 7 February 2011, in final form 16 May 2011)

ABSTRACT

The visual description of wave climate is usually limited to two-dimensional conditional histograms. In this work, self-organizing maps (SOMs), because of their visualization properties, are used to characterize multivariate wave climate. The SOMs are applied to time series of sea-state parameters at a particular location provided by ocean reanalysis databases. Trivariate (significant wave height, mean period, and mean direction), pentavariate (the previous wave parameters and wind velocity and direction), and hexavariate (three wave parameters of the sea and swell components; or the wave, wind, and storm surge) classifications are explored. This clustering technique is also applied to wave and wind data at several locations to analyze their spatial relationship. Several processes are established in order to improve the results, the most relevant being a preselection of data by means a maximum dissimilarity algorithm (MDA). Results show that the SOM identifies the relevant multivariate sea-state types at a particular location spanning the historical variability, and provides an outstanding analysis of the dependency between the different parameters by visual inspection. In the case of wave climate characterizations for several locations the SOM is able to extract the qualitative spatial sea-state patterns, allowing the analysis of the spatial variability and the relationship between different locations. Moreover, the distribution of sea states over the reanalysis period defines a probability density function on the lattice, providing a visual interpretation of the seasonality and interannuality of the multivariate wave climate.

1. Introduction

The wave climate at a particular location is usually defined by empirical statistics (mean, standard deviation) or by the empirical or analytic univariate probability density function (PDF) of one of the following parameters: significant wave height H_s, mean period T_m, and mean wave direction θ_m. The combination of two parameters is analyzed by similarly bivariate PDFs and empirical statistics. In the case of three parameters, the analysis is usually limited to the empirical probability function $p(H_s, T_m, \theta_m)$, sorting the values in classes and visualizing the results using two-dimensional histograms [e.g., the H_s–T_m for a given directional sector $\Delta\theta$ (Holthuijsen 2007)].

In the last decade, long-term reanalysis databases have been developed (see, e.g., Pilar et al. 2008; Ratsimandresey et al. 2008; Weisse et al. 2002; Dodet et al. 2010; Sebastião et al. 2008). In addition to high spatial and temporal data resolution, the number of wave parameters to define each sea state has also increased considerably. Apart from long-term hourly time series of H_s, T_m, and θ_m, the ocean reanalysis databases provide other parameters such as wind velocity W_{10}, wind direction β_W, storm surge S_s, swell significant wave height H_{s2}, and even the directional spectra. However, the description of wave

Corresponding author address: Fernando J. Mendez, Environmental Hydraulics Institute, IH Cantabria, Universidad de Cantabria, E.T.S.I. Caminos Canales y Puertos, Avda de los Castros s/n, 39005, Santander, Spain.
E-mail: mendezf@unican.es

DOI: 10.1175/JTECH-D-11-00027.1

climate combining the different parameters available is still reduced to two-dimensional histograms of just two variables.

Several statistical methods have been developed in the field of data mining to efficiently deal with huge amounts of information [see Cofiño et al. (2003) for some applications in meteorology]. These techniques extract features from the data, providing a more compact and manageable representation of some of the important properties contained in the data. Standard methods in data mining include clustering techniques, which obtain a set of reference vectors representing the data; one of the most powerful of these techniques is self-organizing maps (SOMs). The SOM algorithm computes a set of M prototypes or centroids, with each of them characterizing a group of data, preserving the topology of the data in the original space in a low-dimensional lattice. The cluster centroids are forced with a neighborhood adaptation mechanism to a space with a smaller dimension (usually a two-dimensional regular lattice), which is spatially organized. SOMs have been applied to different geophysical parameters across several disciplines; in meteorology, for example, they classify atmospheric patterns and derive relations with local precipitation in order to downscale to local stations (Cavazos 1999; Gutiérrez et al. 2005). In hydrology they have been used to identify homogeneous regions for regional frequency analysis (Lin and Chen 2005); and in oceanography, they relate satellite-derived sea level with the sardine recruitment (Hardman-Mountford et al. 2003) to identify sea surface temperature and wind patterns from satellite data (Richardson et al. 2003), to extract spatial patterns of ocean current variations from moored velocity data (Liu and Weisberg 2005), or to analyze the biogeochemistry dynamic in the Adriatic Sea by a combination with a k-means technique (Solidoro et al. 2007). A recent completed review of SOM application in meteorology and oceanography can be found in Liu and Weisberg (2011).

In this work, the SOM technique is applied to wave reanalysis data in order to graphically analyze the combination of three or more wave parameters and extract the ''sea-state types'' defined by the ensemble-considered parameters. In the simple case of considering three parameters H_s, T_m, and θ_m, for example, the objective could be the identification of the range of significant wave heights and the range of directions associated with the sea states with the largest mean periods. Another more complex example could be the determination of the proportion of sea component or swell component of the most frequent sea states at a particular location.

Additionally, the spatial information provided by wave reanalysis databases could be classified by SOM to analyze the combination of several variables at different locations at the same time and to extract sea-state spatial patterns.

Therefore, in this work, the self-organizing maps are applied to hourly time series of meteorology–ocean parameters to characterize the multivariate wave climate. To explore the ability of the SOM to analyze wave climate, different examples for three-, five-, and six-dimensional data are shown for a particular location, and a spatial wave climate characterization is also presented.

This paper is organized as follows. In section 2, the data used to characterize wave climate at different locations and with different parameters is described. In section 3, the proposed methodology to define multidimensional wave climate at a specific location is explained; and trivariate, pentavariate, and hexavariate applications are presented. The spatial multivariate wave climate characterization is presented in section 4, including the steps of the methodology and the results of a particular application. Finally, conclusions are given in section 5.

2. Data

Several locations around the Spanish coast (Fig. 1) have been considered to carry out different multivariate wave climate characterizations by means of the SOM. The wave data used is extracted from the SIMAR-44 database, developed by Puertos del Estado (Spain) using the wave model (WAM) and forced by 10-m winds from the Regional-Scale Model (REMO; Jacob and Podzun 1997). The temporal coverage spans 44 yr (1958–2001) with an hourly resolution and a spatial resolution of $1/8°$–$1/12°$.

Figure 2 shows the empirical bivariate distribution of the hourly time series of H_s and θ_m of almost 400 000 sea states at the locations considered around the Spanish coast. This directional distribution provides information about the direction of the most frequent sea states as well as the largest significant wave heights. These examples clearly show the different typology of wave climate along Spanish coast. The most energetic sea states come from the northwest; the range of wave directions is narrower at Santander, coming from the west to the northeast; while at Villano, the range spans from the southwest to the northeast. At Gran Canaria Island, the most energetic sea states mainly come from the northwest to northeast. At Cadiz, sea states come from a wider range of directions, the most frequent ones being west and southeast. Wave climate at Almeria has a clear bimodality around the east-northeast and west-southwest directions. In Tarragona, sea states are from all possible directions, with the most energetic and the most frequent sea states from the east-northeast and south-southwest.

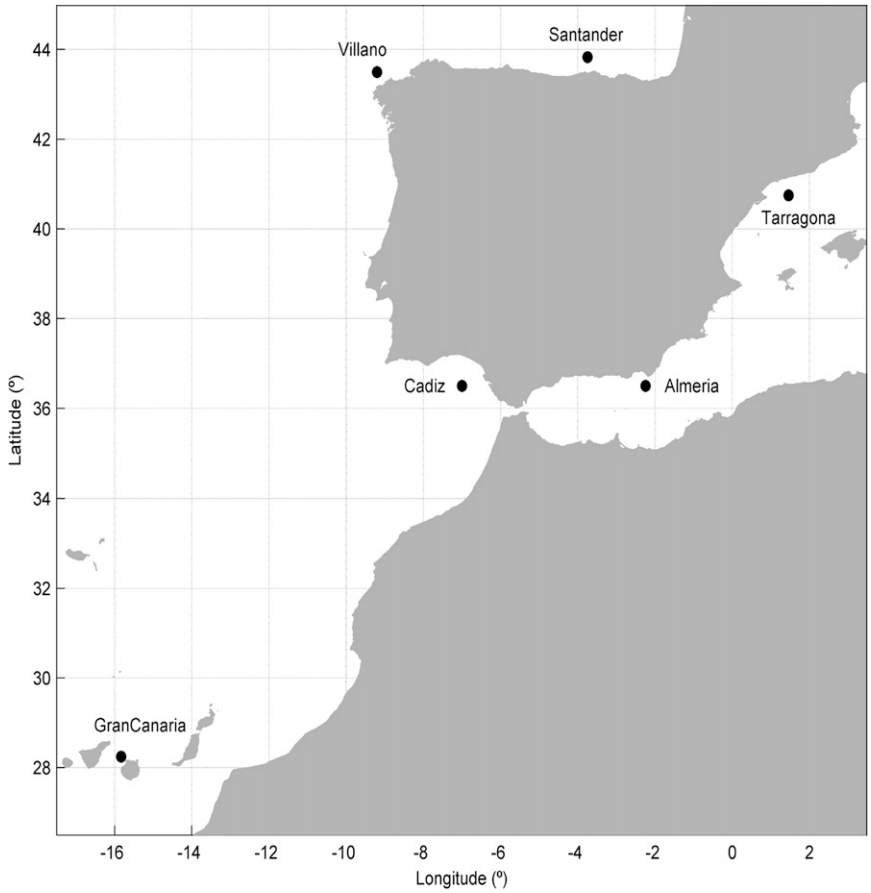

FIG. 1. Locations along the Spanish coast for wave climate characterization.

3. Multivariate wave climate at a particular location

In this section, the SOM algorithm is applied to different combinations of sea-state parameters: significant wave height H_s, mean period T_m, mean wave direction θ_m, wind velocity W_{10}, wind direction β_W, sea significant wave height H_{s1}, sea mean period T_{m1}, sea mean direction θ_{m1}, swell significant wave height H_{s2}, swell mean period T_{m2}, swell mean direction θ_{m2}, and storm surge S_s, from a grid node of the SIMAR-44 database to characterize the multivariate wave climate at a particular location.

a. Methodology

The wave reanalysis data are defined by scalar and directional variables of different magnitudes, which require several processes and some modifications of the SOM algorithm. Therefore, a methodology has been developed in order to obtain proper wave climate classifications by means of the SOM. The methodology has been divided into the following several steps: (a) normalization of the sea-state parameters, (b) preselection of the input data using the maximum dissimilarity algorithm (MDA), (c) application of SOM with a Euclidian–circular (EC) distance, and (d) denormalization of obtained clusters. A sketch of the methodology for a trivariate sea-state definition $\{H_s, T_m, \theta_m\}$ is shown in Fig. 3, and the steps of the methodology are explained below. The Matlab SOM Toolbox (Vesanto et al. 2000) is used based on the sequential algorithm with a linearly initialization and on the other default tuneable parameters.

1) STEP A: NORMALIZATION

The normalization step is required to give a similarly weight when applying the selection and classification algorithms. The multivariate database is defined as $\{H_{s,i}, T_{m,i}, \theta_{m,i}\}$; $i = 1, \ldots, N$, where N is almost 400 000 sea states (44 years of hourly data). The scalar variables are normalized by scaling the variables values between [0, 1] with a simple linear transformation, which requires two parameters—the minimum and maximum value of the two scalar variables,

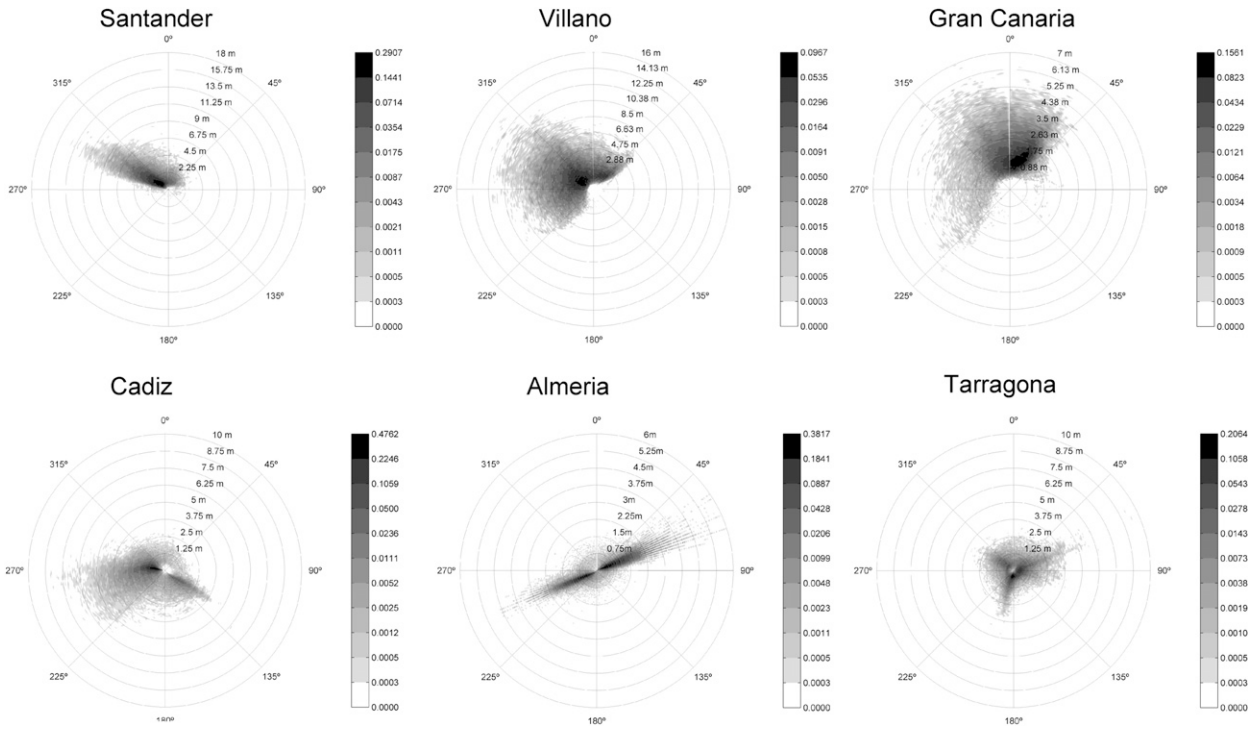

FIG. 2. Empirical joint distribution of H_s and θ_m at the different locations considered along the Spanish coast.

$$H_s^{\min} = \min(H_s); \quad H_s^{\max} = \max(H_s)$$
$$T_m^{\min} = \min(T_m); \quad T_m^{\max} = \max(T_m).$$

For the circular variables (defined in radians or in sexagesimal degrees using the scaling factor $\pi/180$), taking into account that the maximum difference between two directions over the circle is equal to π and the minimum difference is equal to 0, this variable has been normalized by dividing the direction values between π, therefore rescaling the circular distance between [0, 1].

2) STEP B: PRESELECTION

After these transformations, the dimensionless input data $\mathbf{X}_i = \{H_i, T_i, \theta_i\}; i = 1, \ldots, N$ are defined as

$$H_i = \frac{H_{s,i} - H_s^{\min}}{H_s^{\max} - H_s^{\min}}; \quad T_i = \frac{T_{m,i} - T_m^{\min}}{T_m^{\max} - T_m^{\min}}; \quad \theta_i = \frac{\theta_{m,i}}{\pi}.$$

The preselection removes the redundancy of wave reanalysis data, avoids most of the clusters from being representative of these conditions, and enables a wide variety of possible sea-state types to be obtained. The MDA selects a representative subset of size P from a database of size N (Camus et al. 2010). Therefore, in this case, given a data sample $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$, consisting of N three-dimensional vectors, a subset of P vectors $\{\mathbf{X}_1^D, \ldots, \mathbf{X}_P^D\}$ representing the diversity of the data is obtained by applying this algorithm. The selection starts initializing the subset by transferring one vector from the data sample $\{\mathbf{X}_1^D\}$. The rest of the $M - 1$ elements are selected iteratively, calculating the dissimilarity between each remaining data point in the database and the elements of the subset, and transferring the most dissimilar one to the subset (Kennard and Stone 1969). Many variants, depending on the precise implementation of the initialization and the definition of the most dissimilar vector, are available (Willet 1996). In this work, the initial datum of the subset is the sea state with the largest value of significant wave height. The dissimilarity between each remaining vector in the database and each vector in the subset is calculated, and a unique dissimilarity between each vector in the database and the subset is established as the minimum one. For example, if the subset is formed by R ($R \leq P$) vectors, the dissimilarity between the vector i of the data sample $N - R$ and the j vectors belonging to the R subset is calculated as

$$d_{ij} = \|\mathbf{X}_i - \mathbf{X}_j^D\|; \quad i = 1, \ldots, N - R; \quad j = 1, \ldots, R,$$

where $\| \ \|$ is defined as the Euclidean–circular distance

$$\|\mathbf{X}_i - \mathbf{X}_j^D\| = \{(H_i - H_j^D)^2 + (T_i - T_j^D)^2 + [\min(|\theta_i - \theta_j^D|, 2 - |\theta_i - \theta_j^D|)]^2\}^{1/2}.$$

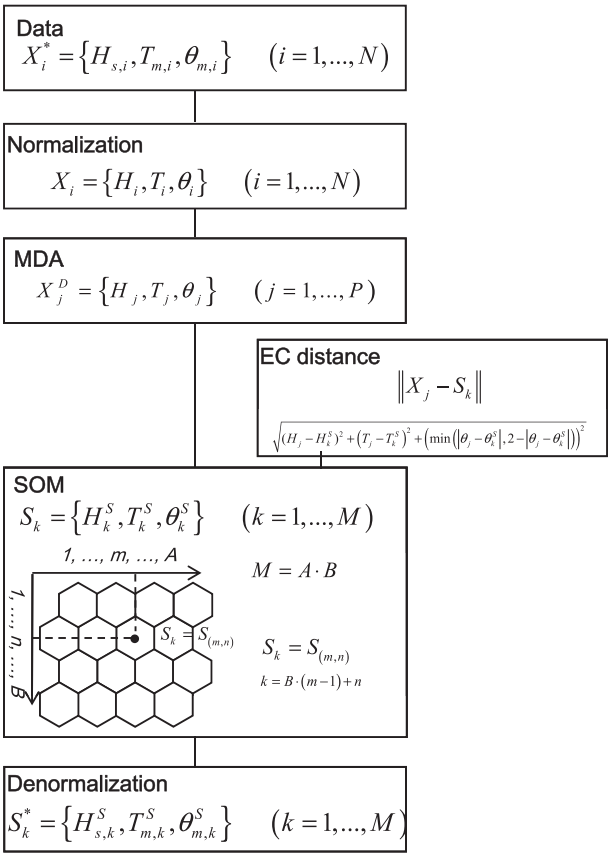

FIG. 3. Methodology to characterize wave climate by SOM at a specific location.

Subsequently, the dissimilarity $d_{i,\text{subset}}$ between the vector i and the subset R, is calculated as

$$d_{\mathbf{i},\text{subset}} = \min\{\|\mathbf{X}_i - \mathbf{X}_j^D\|\}; \quad i = 1, \ldots, N - R; \quad j = 1, \ldots, R.$$

Once the $N - R$ dissimilarities are calculated, the next selected element that is with the largest value of $d_{i,\text{subset}}$, the most dissimilar one. The efficient algorithm developed by Polinsky et al. (1996) has been considered to reduce the computation time.

Several tests have been carried out to analyze the influence of the preselected data (P) in the trivariate wave climate SOM classifications at several locations around the Spanish coast. SOMs of different sizes, considering different quantities of preselected data, have been trained. The representativeness of wave climate has been analyzed by means of the error between the real value of the mean significant wave height and the 99th percentile of the significant wave height, calculated by the complete wave reanalysis time series, and the estimated value, calculated using the centroids and its corresponding probability. The preselected data influence the results for SOM sizes lower than 600 clusters. The error in the mean significant wave height decreases when P increases, while the error in the 99th percentile of the significant wave height increases when P increases. However, the magnitude of the error in the 99th percentile is higher than in the mean value and the increase is significant for $P > 10\,000$. Therefore, we recommend $P = 10\,000$.

3) STEP C: SOM

Once the preselection of P elements has been done $\{\mathbf{X}_1^D, \ldots, \mathbf{X}_P^D\}$, the SOM is applied. SOM automatically extracts patterns or clusters of high-dimensional data and projects them into a bidimensional organized space, allowing an intuitive visualization of the classification and the transformation of the distributions from the high-dimensional space into PDFs on the lattice (Kohonen 2000).

Given the database of three-dimensional vectors $\mathbf{X}^D = \{\mathbf{X}_1^D, \ldots, \mathbf{X}_P^D\}$, SOM is applied to obtain M groups defined by a prototype or centroid $\mathbf{S}_k = \{H_k^S, T_k^S, \theta_k^S\}$, with $k = 1, \ldots, M$. The classification procedure starts with a linear initialization of the centroids $\{\mathbf{S}_1^0, \mathbf{S}_2^0, \ldots, \mathbf{S}_M^0\}$. This means that the centroids are initialized along the mdim greatest eigenvectors of the given data, where mdim is the dimension of the map grid. The algorithm adjusts the prototypes iteratively to the data trying to minimize an overall within-cluster distance from the data vectors $\mathbf{S}_k$ to the corresponding centroid vector $\mathbf{X}_j^D$ for each cluster j.

The SOM runs in cycles; during each training cycle, each of the data vectors $\mathbf{X}_j^D$ is considered, and the ''winning'' centroid vector $\mathbf{S}_{w(j)}$ is found to be the one closest to the data vector,

$$\|\mathbf{S}_{w(j)} - \mathbf{X}_j^D\| = \min_k\{\|\mathbf{S}_k - \mathbf{X}_j^D\|, \quad k = 1, \ldots, M\},$$

where $1 \le w(j) \le M$ is the index of the winning reference vector.

The training procedure includes a neighborhood adaptation mechanism in the lattice of projection; thus, not only does the winning centroid move toward the data vector, but the neighboring centroids in the lattice are also adapted to the sample vector,

$$\mathbf{S}_k = \mathbf{S}_k + \alpha h[w(j), k](\mathbf{X}_j^D - \mathbf{S}_k), \quad k = 1, \ldots, M,$$

where $0 \le \alpha \le 1$ is the learning rate and controls the velocity of the adaptation process. The function $h[w(j), k]$ is a neighborhood kernel on the SOM lattice, which determines the rate of change around the winning centroid and projects the topological relationships in the data space

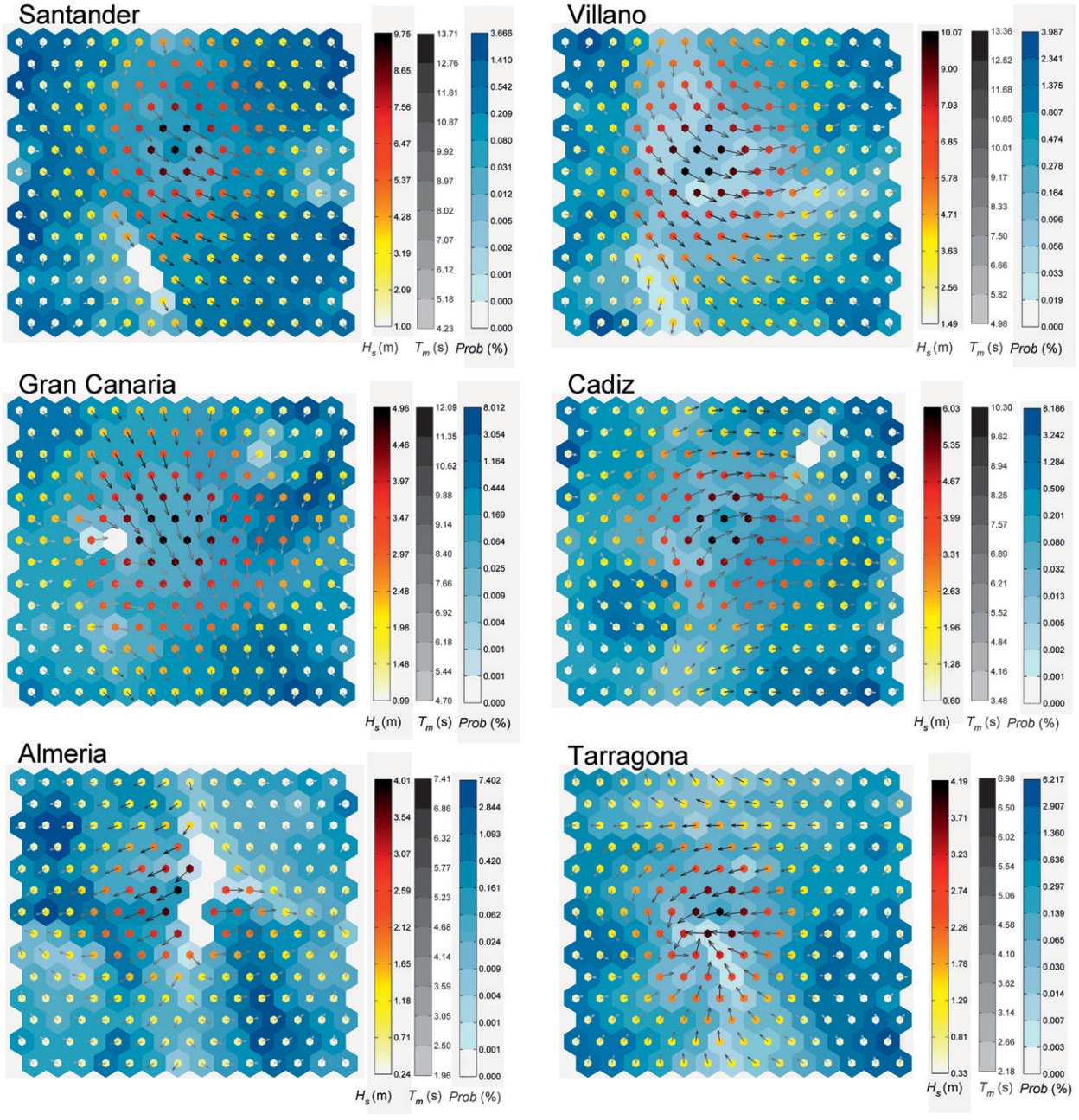

FIG. 4. Trivariate wave climate characterization by means of SOM at different locations along the Spanish coast; H_s, T_m, and θ_m are represented by the size, grayscale, and direction of the arrow.

onto the lattice. This means that similar clusters in the multidimensional space are located together in the lattice of projection. Each cluster of a SOM is defined by two vectors—one in the data space $\boldsymbol{v}_k$ (prototype) and the other one (m_k, n_k) describing the position on the lattice (Fig. 3; SOM). For a given SOM of size $M = AB$, the kth index of a cluster is related with the lattice dimensions and its position in the lattice by the expression $k = B(m - 1) + n$. In this algorithm, the EC distance is also applied as in the MDA,

$$\|\mathbf{X}_j^D - \mathbf{S}_k\| = \{(H_j^D - H_k^S)^2 + (T_j^D - T_k^S)^2 + [\min(|\theta_j^D - \theta_k^S|, 2 - |\theta_j^D - \theta_k^S|)]^2\}^{1/2}.$$

4) STEP D: DENORMALIZATION

Finally, the last step is the denormalization of clusters, applying the opposite transformation of the normalization step:

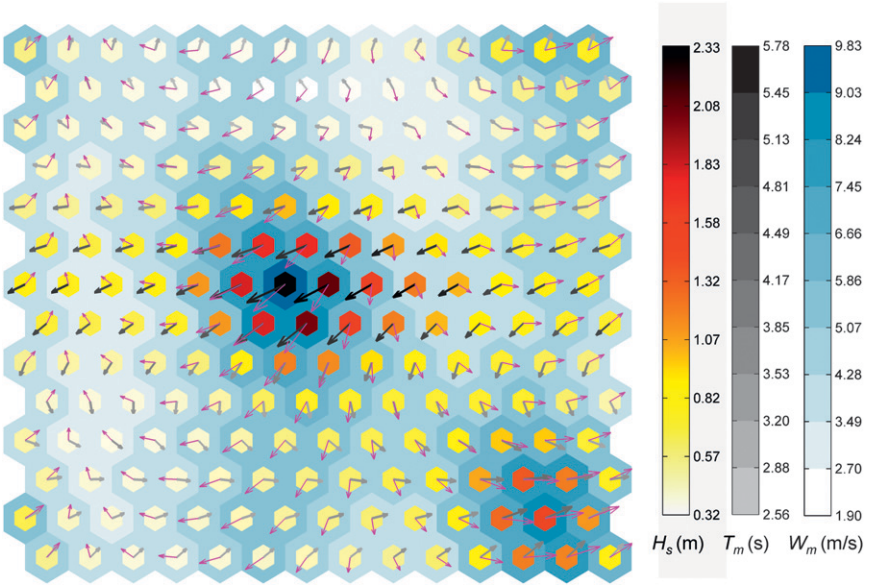

FIG. 5. Five-dimensional wave climate characterization by means of SOM at Almeria; H_s, T_m, and θ_m of each cluster are represented by the size, grayscale, and direction of the gray arrow, and W_{10} and β_W of each cluster are represented by the size and direction of the magenta arrow.

$$H^S_{s,k} = H^S_k(H^{\max}_s - H^{\min}_s) + H^{\min}_s;$$
$$T^S_{m,k} = T^S_k(T^{\max}_m - T^{\min}_m) + T^{\min}_m; \quad \theta^S_{m,k} = \theta^S_k\pi.$$

Although the methodology is explained for only three parameters, the approach is applicable to different combinations of multivariate meteorology–ocean parameters. The main requisite is the identification of the scalar and directional variables in order to establish the normalization and define the Euclidian or circular distance in the similarity criterion of the SOM algorithm.

b. *Trivariate wave climate characterization along the Spanish coast*

The hexagonal self-organizing maps of 14×14 size of trivariate wave climate at the selected locations along the Spanish coast are shown in Fig. 4. Each cell of the SOM represents a cluster defined by $\{H_s, T_m, \theta_m\}$ parameters. The significant wave height H_s, the mean wave period T_m, and the mean wave direction θ_m are represented by the size, intensity of the gray color, and direction of the arrow, respectively. The smaller hexagon, in a light yellow–dark red scale, defines the H_s magnitude. The larger hexagon in a blue scale shows the relative frequency. The input data have been projected into a toroidal lattice, which means that the centroids located on the upper and lower and in the lateral sides of the sheet are joined in the toroidal projection, being similar in the data space. As we can see, similar clusters are located together on the lattice; the magnitudes of the centroid parameters vary smoothly from one cell to another because the SOM algorithm projects the topological relationships of the tridimensional data space in the map.

This technique is able to detect all the possible sea-state types. In the SOM at Santander and Villano, it can be observed that the directional diversity of the clusters matches the corresponding bidimensional distributions (H_s, θ_m). The directional range at Villano varies from the southwest to the northeast direction, being wider than at Santander, which is between the west and northeast directions. The higher diversity of wave directions at Villano can also be observed in the most energetic sea states. Additionally, this trivariate wave climate SOM classification adds the information about the mean period in the analysis of sea-state types. For example, the sea states with the maximum periods at Santander and Villano are from the west-northwest and have medium wave energy (long-period swells).

The SOM at Gran Canaria informs us that wave directions vary between the southwest and northeast directions. The sea-state types with the largest significant wave heights and the ones with the largest periods come from the north-northwest. The most frequent sea-state types are low energy and come from the north-northeast.

The SOM at Cadiz shows that sea states may come from all possible directions. The largest significant wave heights are from the west-southwest. The most frequent

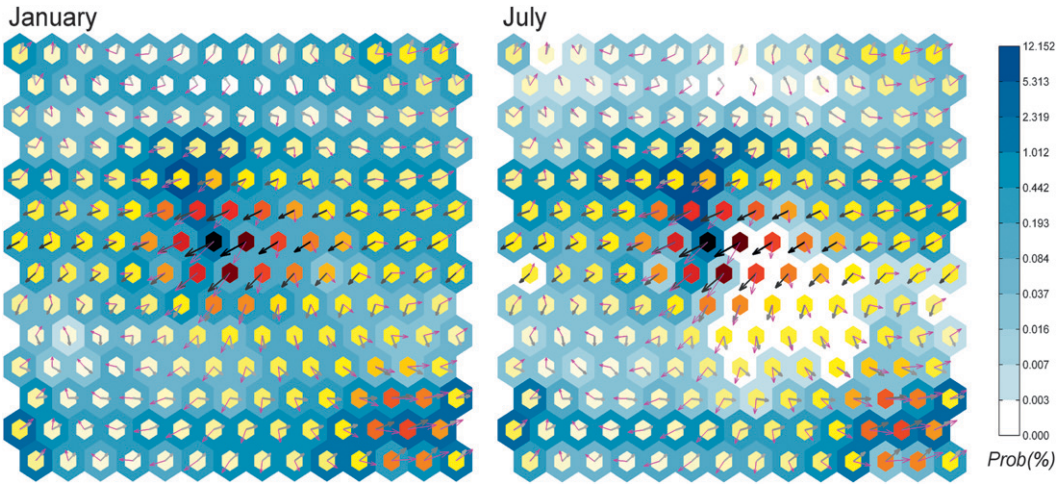

FIG. 6. Seasonality of 5D wave climate characterization by means of SOM at Almeria. (See Fig. 5 caption for an explanation of the map).

wave types and the ones with the highest period are from the west.

The wave climate bimodality at Almeria is reflected in the SOM classification: the sea-state types are mainly from the east-northeast and west-southwest. The sea-state clusters from the east-northeast direction have a higher significant wave height and wave period and are more frequent than the ones from the west-southwest direction.

The SOM at Tarragona detects the diversity of the wave climate with sea-state types coming from all possible directions. The most energetic sea states are from the east-northeast and south-southwest, while the sea-state types with the highest periods are from the east.

c. *Pentavariate wave climate*

In this section, the SOM is applied to five-dimensional wave data at Almeria. Each hourly sea state is defined by the following five parameters: H_s, T_m, θ_m, W_{10}, and β_W.

Figure 5 shows a SOM of $M = 14 \times 14$ size, with each cluster being defined by five parameters, $S_k = \{H_s, T_m, \theta_m, W_{10}, \beta_w\}$, $k = 1, \ldots, M$. In this map, the gray arrows represent the wave characteristics; the size, color intensity, and direction of the arrows are indicated by H_s, T_m, and θ_m, respectively. The size and the direction of the magenta arrows represent W_{10} and β_w. In addition, the smaller hexagon in a scale of light yellow–dark red defines the H_s magnitude and the hexagon in a blue scale defines the wind velocity. As in the previous example, the classification has been projected onto a toroidal lattice.

The bimodality of wave climate at Almeria, detected in the trivariate characterization, is also represented in this pentavariate SOM. We can see two families of energetic sea states from the east-northeast and south-southwest directions, respectively, with associated winds from the same directions. The significant wave height and the wind velocity of the sea-state types from the east-northeast direction are of higher magnitude than the other family of sea-state types. Many clusters represent low energetic sea states from all of the possible directions associated with gentle winds. Most of them are sea states from the first and third quadrant, which are the most frequent sea states at this location, as we have seen in the tridimensional SOM characterization. Additionally, the clusters with the maximum wave periods (Tm >= 5 s) are from the east-northeast, located at the right side of the clusters with the extreme events on the lattice, with associated winds from the northwest to northeast directions.

The seasonality of wave climate at Almeria has been analyzed by calculating the probability of the clusters at each month. The SOM probability density functions in January and July are shown in Fig. 6, as represented by the blue-scale hexagons. All of the sea-state types that are identified occur during the winter months, while during the summer the probability of the sea types defined by waves from the north-northwest to east-northeast directions with associated winds from north-northwest to north-northeast directions is practically null. These blue maps represent very useful quantitative multivariate histograms of wave climate.

d. *Hexavariate wave climate (sea + swell)*

In this application, we have considered the sea and the swell sea-state components at Villano. Each hourly data point is defined by six parameters: the significant wave height, the mean period, and the mean direction of the sea component (H_{s1}, T_{m1}, θ_{m1}), and the significant wave height, the mean period, and the mean direction of the swell component (H_{s2}, T_{m2}, θ_{m2}). Figure 7 shows a SOM

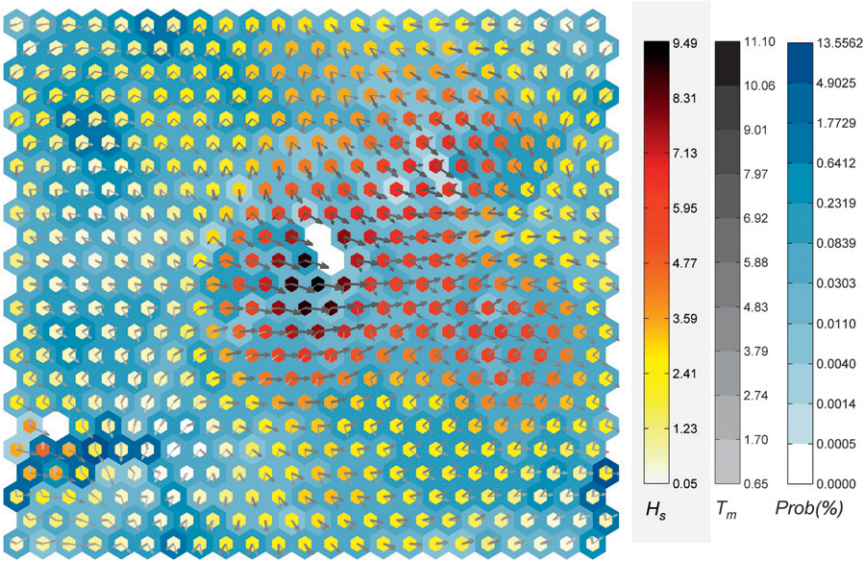

FIG. 7. Six-dimensional wave climate characterization by means of SOM at Villano (sea + swell component); H_{s1}, T_{m1}, and θ_{m1} (sea parameters) of each cluster are represented by the size, grayscale, and direction of the thicker arrow, and H_{s2}, T_{m2}, and θ_{m2} (swell parameters) are represented by the size, grayscale, and direction of the thinner arrow.

of 23 × 23 size, where each cell is defined by these six parameters, with the sea component being represented by the thicker arrow and the swell component by the thinner arrow, with the same symbolism used in the trivariate characterization. In this case, the little hexagon in a light yellow–dark red scale represents the significant wave height (H_s) defined as $\sqrt{H_{s1}^2 + H_{s2}^2}$.

The clusters are defined by a combination of sea and swell components from the southwest to the northeast range of directions and different quantity of energy. The sea-state types with the largest significant wave height are mainly defined by a sea component from the west-northwest to northwest direction with a swell component with very little energy. There are other wave types with an important significant wave height consisting of a sea and swell component, both from the northwest and with an equal amount of energy (located in the upper-right side of the extreme events on the SOM). Other clusters are defined by a sea component from the west and a swell component from the northwest with a similar significant wave height (on the right side of the most energetic clusters). Sea-state types with a sea component from the southwest and more energetic swell components from the northwest are also detected (located in the lower right-hand side of the most energetic clusters). The sea-state types with a well-defined swell component are from a northwest direction (located in the lower left side of the SOM). The clusters with the highest periods are located in the middle–upper right-hand side of the SOM and are defined by a sea component from the northwest and a low energetic swell component from the southwest. Most of the clusters represent low energetic sea states from a wide variety of combinations of directions.

e. *Hexavariate wave climate (wave + wind + storm surge)*

In this wave climate characterization at Tarragona by means of a SOM, each sea state is defined by the following six parameters: significant wave height H_s, mean period T_m, mean direction θ_m, wind velocity W_{10}, wind direction β_W, and storm surge S_s. Figure 8 shows a SOM of 18 × 18 size, where H_s, T_m, and θ_m are represented by the size, the gray color intensity, and the direction of the thicker arrow; W_{10} and β_W are represented by the size and the direction of the thinner arrow. The background of each hexagon represents the storm surge (red scale for positive values and blue scale for negative).

The sea-state types obtained by the SOM are from all possible directions. As in the trivariate classification, the most energetic sea states are from the northeast to the east-southeast directions associated with the strongest winds mainly from the northeast in that location. Other clusters represent considerable energetic sea states from the southwest, with winds from the northwest and with waves and winds from the north-northwest. Most of the SOM clusters correspond to calm situations (very small significant wave heights and gentle winds from different directions). Regarding the storm surge, the positive situations are associated with southern waves. This aspect

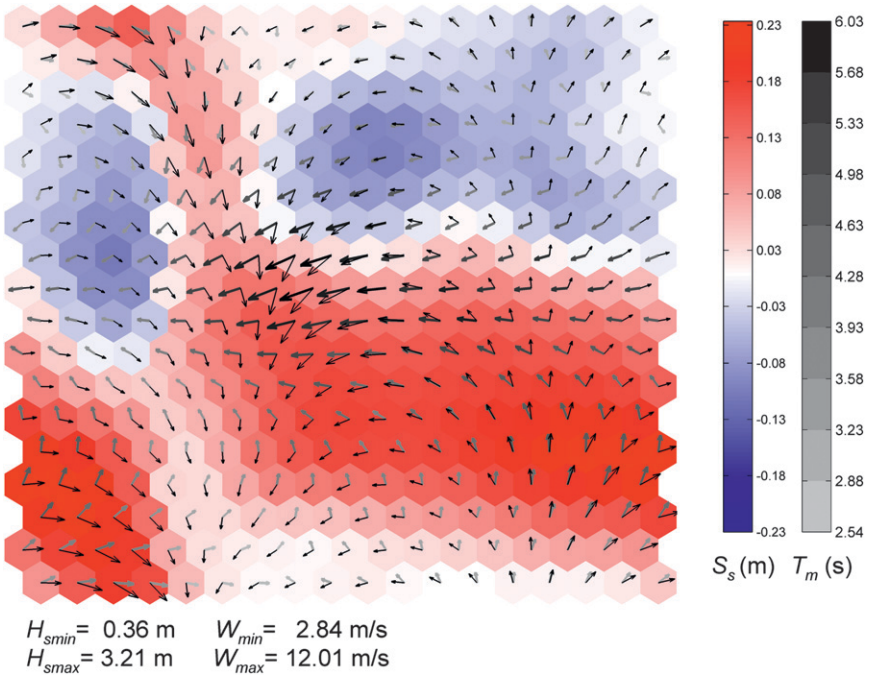

FIG. 8. Six-dimensional wave climate characterization by means of SOM at Tarragona (wave + wind + storm surge); H_s, T_m, and θ_m of each cluster are represented by the size, grayscale, and direction of the thicker arrow, and W_{10} and β_W of each cluster are represented by the size and direction of the thinner arrow.

is very relevant because the SOM is able to reveal the different combinations of wave height, wave direction, and storm surge that occur at a particular site.

4. Spatial variability of multivariate wave climate

The combination of coastal orientation and wave direction produces significant wave climate variations at certain areas. In these situations, a spatial multivariate wave climate characterization defined by offshore data at several locations is necessary. In this section, SOM is used to identify characteristic spatial patterns from the hourly time series of wave and wind fields around the coastal area of interest. The high dimensionality of spatial fields slows down and even spoils the training process of clustering algorithms. Therefore, we have previously considered the problem of dimensionality reduction using principal component analysis (PCA) to extract as much correlation as possible from spatial fields, but keeping the diversity of climate situations. To avoid problems resulting from different scales, all the variables are previously standardized for each grid point, following the procedure that is described later on. As in the characterization of wave climate at a particular location, the amount of similar wave climate situations representing the mean wave conditions requires a preselection for a better SOM classification result. These processes establish a methodology for spatial wave climate characterization, which is described below.

Methodology

The methodology has been divided into the following several steps: (i) standardization of the spatial wave and wind fields, (ii) reduction of data dimensionality by PCA, (iii) preselection of data in the reduced space by MDA algorithm, and (iv) application of SOM to dimensionally reduced preselected data and identification of the closest real data to each centroid. Figure 9 shows a sketch of the methodology for a spatial wave characterization. Each step is described in detail.

1) STEP A: STANDARDIZATION

Each hourly situation is defined by the wave and wind fields around the area of interest, $\mathbf{X}_i^* = \{H_{s,1}, T_{m,1}, \theta_{m,1}, \ldots, H_{s,n1}, T_{m,n1}, \theta_{m,n1}, \ldots, W_{10,1}, \beta_{10,1}, \ldots, W_{10,n2}, \beta_{10,n2}\}_i$; $i = 1, \ldots N$, where $n1$ is the number of wave data locations, $n2$ is the number of wind data locations, and N is the total amount of hourly situations. The wave and wind directions are transformed to x and y components and then standardized (with a zero mean and a standard deviation of one). After these transformations, the dimensionless input data are defined as $\mathbf{X}_i = \{H_1, T_1,$

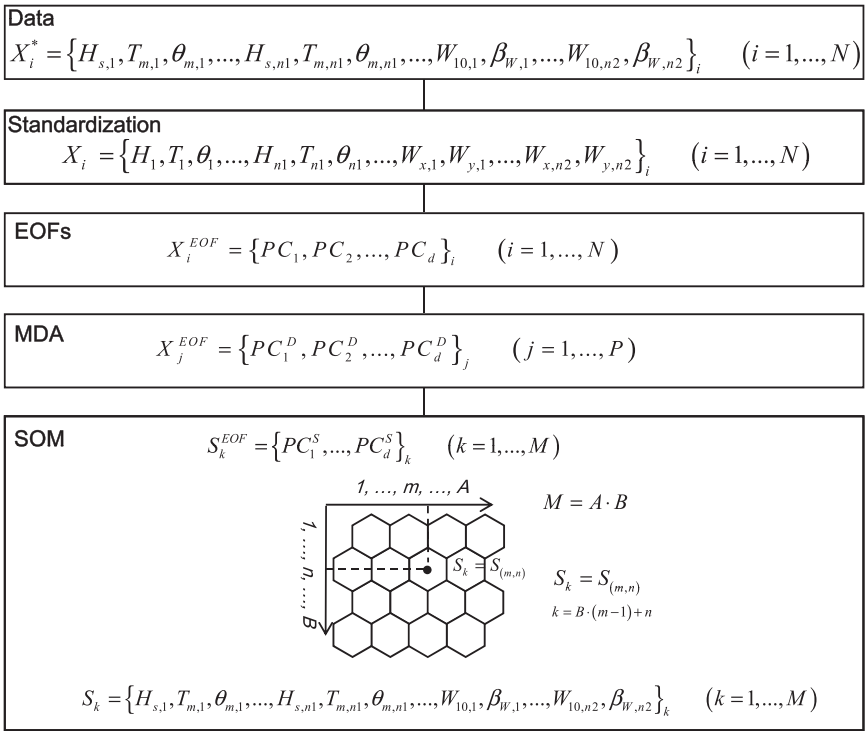

FIG. 9. Methodology for spatial wave climate characterization.

$\theta_1, \ldots, H_{n1}, T_{n1}, \theta_{n1}, \ldots, W_{x,1}, W_{y,1}, \ldots, W_{x,n2}, W_{y,n2}\}_i$; $i = 1, \ldots, N$.

2) STEP B: PCA

The PCA reduces the dimension of the data by means of a projection in a lower dimensional space that preserves the maximum variance of the sample data. The new vectors are formed by the ones where the projected data have the higher variance. Given the spatiotemporal variable $\mathbf{X}_i(x, t_i)$, where x is the spatial data position of dimension $3n1 + 2n2$ and t_i is time, we apply PCA to obtain a new d-dimensional space. The eigenvectors [empirical orthogonal functions (EOFs)] of the covariance matrix of the data define the vectors of the new space. The idea of PCA is to find the minimum d linearly EOFs, so that the transformed components of the original data [principal components (PCs)] explain the maximum variance necessary in the problem at hand. The original data can be expressed as a linear combination of EOFs and PCs,

$$\mathbf{X}(x, t_i) = \mathrm{EOF}_1(x) \times \mathrm{PC}_1(t_i) + \mathrm{EOF}_2(x) \times \mathrm{PC}_2(t_i) + \cdots + \mathrm{EOF}_d(x) \times \mathrm{PC}_d(t_i).$$

Once we apply PCA, our data are defined by the principal components $\mathbf{X}_i^{\mathrm{EOF}} = \{\mathrm{PC}_1, \mathrm{PC}_2, \ldots, \mathrm{PC}_d\}_i$: $i = 1, \ldots, P$.

3) STEP C: MDA

The next step consists of selecting a representative subset of size P using MDA $\mathbf{X}_j^{\mathrm{EOF}} = \{\mathrm{PC}_1, \mathrm{PC}_2, \ldots, \mathrm{PC}_d\}_j$: $j = 1, \ldots, P$. This algorithm has been explained in section 3a. In this case, it is not necessary to implement the EC distance. The first element selected is the one with the largest significant wave height, identified in the original space.

4) STEP D: SOM

SOM is applied to this selected sample in the EOF space to the obtained M clusters, $\mathbf{S}_k^{\mathrm{EOF}} = \{\mathrm{PC}_1, \mathrm{PC}_2, \ldots, \mathrm{PC}_d\}_k$, where $k = 1, \ldots, M$. To avoid the reconstruction of the centroids when projected back to the original space, we have considered the closest data to each centroid and have identified them in the original space. Finally, each cluster is defined by $\mathbf{S}_k^* = \{H_{s,1}, T_{m,1}, \theta_{m,1}, \ldots, H_{s,n1}, T_{m,n1}, \theta_{m,n1}, \ldots, W_{10,1}, \beta_{10,1}, \ldots, W_{10,n2}, \beta_{10,n2}\}_k$; $k = 1, \ldots, M$.

We have applied this methodology in the characterization of wave climate around the south coast of Gran Canaria. In this particular application, we have considered $n1 = 16$ and $n2 = 16$ grid points to define wave and wind fields. To select a convenient dimension reduction of the data, we have computed the reconstruction root-mean-square error (rmse) of variables that define the

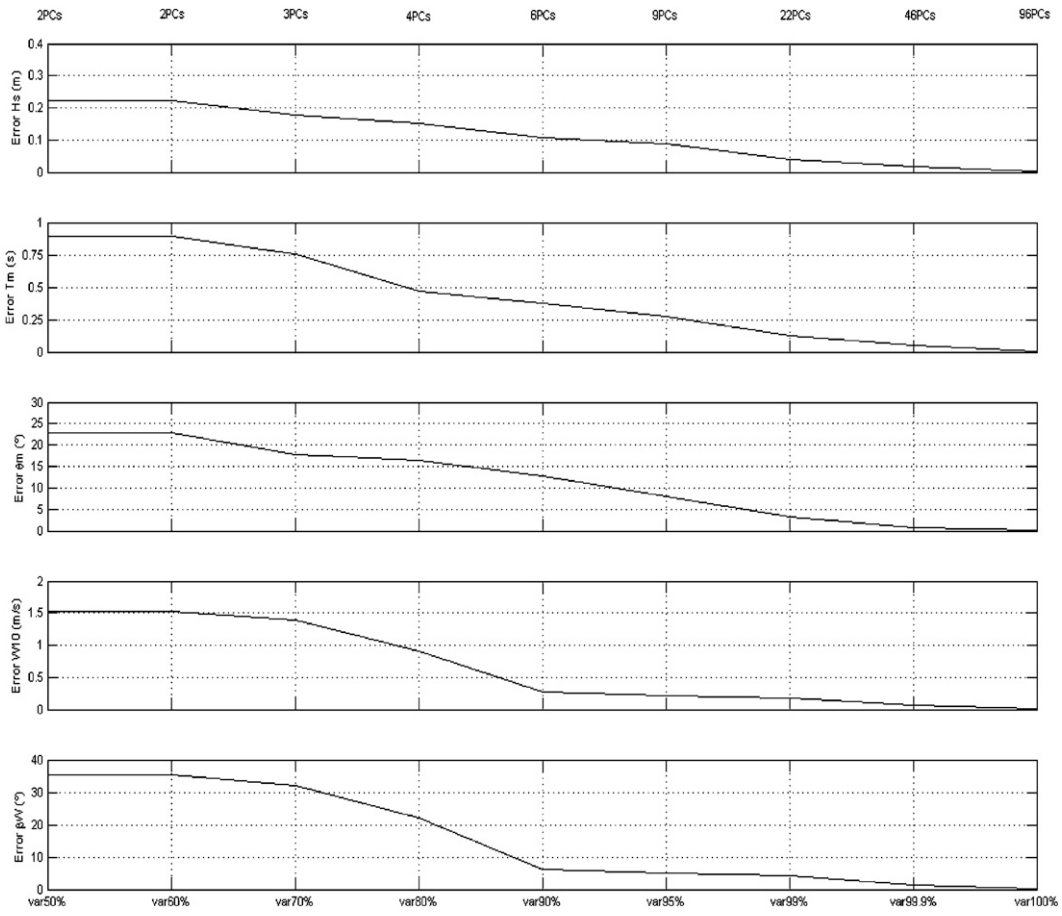

FIG. 10. Reconstruction rmse for each of the five variables in the wave and wind fields used in the spatial wave climate characterization by SOM at Gran Canaria.

wave and wind fields (H_s, T_m, θ_m, W_{10}, and β_W) for an increasing proportion of explained variance of data. Figure 10 shows the mean errors of the five variables for different explained variance and the corresponding number of principal components. We can see that for an explained variance of 95%, the rmse of H_s, T_m, θ_m, W_{10}, and β_W is around 0.09 m, 0.28 s, 8.0°, 0.22 m s^{-1}, and 5°, respectively. Therefore, we can reduce the dimension of the wave and wind fields from 80 to 9 with no significant loss of information and a significant decrease of computational effort.

In this study, a SOM with $M = 49$ patterns in a 7×7 array is constructed. Figure 11 shows the 49 wave climate patterns identified by the SOM around the south coast of Gran Canaria. Each cluster is defined by a wave and wind field at the considered grid nodes. In this case, the input data have been projected into a sheet lattice, which means that the centroids located on the corners of the map are completely different. The fields change gradually across the SOM lattice, from patterns with generally energetic waves and strong winds on the left, to patterns with generally low waves and weak winds on the right. The clusters on the upper left side of the SOM define the most energetic southwesterly wave fields in this area, with wind fields from the same direction. In the opposite lower left corner, the clusters define the highest northeasterly wave and wind fields, which are predominant in this area of study. The wave directions on the right side of the island are coming from the north-northeast, while on the left side of the island the direction is from the north. This pattern, defined by waves from the north to the north-northeast and winds from the north-northeast, persists throughout the lower row and in the middle rows of the array, although the trend is for waves and winds to decrease progressively toward the right side.

Once the characteristic patterns are identified, the frequency of occurrence of each pattern is determined finding out which is the most similar one to each hourly wave and wind field. Monthly PDFs on the SOM lattice are shown in Fig. 12. Every month, the most frequent patterns correspond to the clusters located in the lower–middle

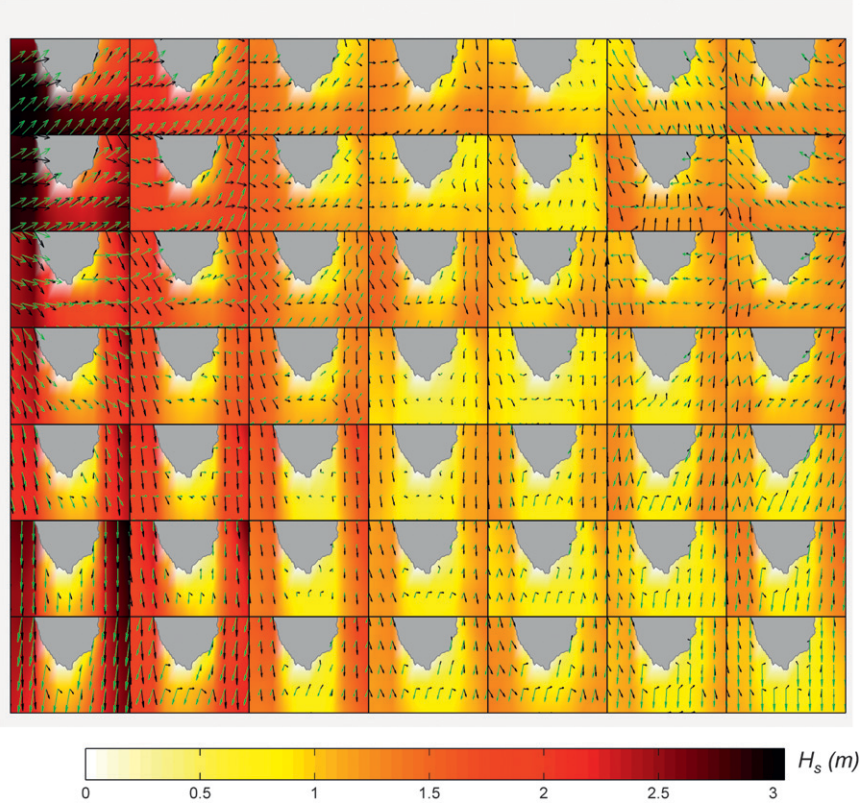

FIG. 11. Spatial wave climate patterns at Gran Canaria, SOM 7 × 7.

right side of the lattice, defined by strong winds from the north-northeast to the northeast and waves from the same directions. Although similar dominant patterns are seen for all months, a monthly variability in the patterns located on the upper right corner of the lattice is highlighted in the frequency maps. These patterns represent the most energetic waves from the southwest direction and the probability of occurrence in summer months is practically null.

5. Conclusions

The SOM is a powerful classification technique to extract patterns from huge amounts of information providing a visual interpretation of high-dimensional data. In this work, the SOM is applied to wave databases to characterize multivariate wave climate at a particular location. A complete methodology is developed, including preselection of data using MDA and new Euclidian–circular distance implemented in the SOM algorithm to work with directional parameters. We define each sea state by three wave parameters $\{H_s, T_m, \theta_m\}$, five $\{H_s, T_m, \theta_m, W_{10}, \beta_W\}$, six $\{H_{s1}, T_{m1}, \theta_{m1}, H_{s2}, T_{m2}, \theta_{m2}\}$, and $\{H_s, T_m, \theta_m, W_{10}, \beta_W, S_s\}$, which are widely used in wave climate characterization at different locations along the Spanish coast. It is shown that the SOM is able to identify multivariate sea-state patterns defining the interaction between wave and wind; sea and swell; and wave, wind, and storm surge. In addition, the prominent visualization properties of the SOM allow for the evaluation of the correlation and dependency between these sea-state parameters by means of visual inspection.

The SOM is also used to identify spatial sea-state patterns. In this case, each situation is defined by the wave field (H_s, T_m, and θ_m in several grid nodes of wave reanalysis database around the area of interest) and the wind field (W_{10} and β_W in the same grid nodes). A different methodology is established, which includes a reduction of the data dimensionality by a principal component analysis (PCA) and preselection by an MDA algorithm, as in the methodology for wave climate characterization at a particular location. From the example presented, we can conclude that the SOM characterization is able to extract the dominant spatial wave patterns in the area of study. This spatial characterization of the wave climate allows the analysis of the spatial relationships between waves at different locations and the associated wind field. It provides a more global description of wave climate, detecting the spatial variability

FIG. 12. Seasonality of spatial wave climate at Gran Canaria.

of wave climate at a certain area of study and the request of wave data at more than a single location to know where waves comes from when they reach a specific coastal stretch.

In addition, the projection of the results in a bidimensional lattice transforms the multivariate distribution into a probability density function on the SOM lattice because of the topology preservation, providing an easy analysis of the probability of sea-state types at different time scales, such as monthly, seasonal, or interannual.

It is finally concluded that the SOM allows an intuitive characterization of the multidimensional wave climate resulting from the visually appealing properties of this clustering technique, improving the knowledge of the wave climate provided by the valuable information of the reanalysis databases. More complex characterizations are possible by means of this useful technique for large datasets.

Acknowledgments. The work was partially funded by projects GRACCIE (CSD2007-00067, CONSOLIDER-INGENIO 2010) from the Spanish Ministry of Science and Technology, MARUCA(200800050084091) from the Spanish Ministry of Public Works, and C3E(E17/08) from the Spanish Ministry of Environment, Rural and Marine Environs. The authors thank Puertos del Estado (Spanish Ministry of Public Works) for providing the reanalysis database.

REFERENCES

Camus, P., F. J. Mendez, R. Medina, and A. S. Cofiño, 2010: Analysis of clustering and selection algorithms for the study of multivariate wave climate. *Coastal Eng.,* **58,** 453–462, doi:10.1016/j.coastaleng.2011.02.003.

Cavazos, T., 1999: Large-scale circulation anomalies conductive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and south-eastern Texas. *J. Climate,* **12,** 1506–1523.

Cofiño, A. S., J. M. Gutierrez, B. Jakubiak, and M. Melonek, 2003: Implementation of data mining techniques for meteorological applications. *Realizing Teracomputing,* W. Zwieflhofer and N. Kreitz Eds., World Scientific, 215–240.

Dodet, G., X. Bertin, and R. Taborda, 2010: Wave climate variability in the North-East Atlantic Ocean over the last six decades. *Ocean Modell.,* **31,** 120–131.

Gutiérrez, J. M., R. Cano, A. S. Cofiño, and C. Sordo, 2005: Analysis and downscaling multi-model seasonal forecasts in Peru using self-organizing maps. *Tellus,* **57A,** 435–447.

Hardman-Mountford, N. J., A. J. Richardson, D. C. Boyer, A. Kreiner, and H. J. Boyer, 2003: Relating sardine recruitment in the Northern Benguela to satellite-derived sea surface height using neural network pattern recognition approach. *Prog. Oceanogr.,* **59,** 241–255.

Holthuijsen, L. H., 2007: *Waves in Oceanic and Coastal Waters*. Cambridge University Press, 387 pp.

Jacob, D., and R. Podzun, 1997: Sensitivity studies with the regional climate model REMO. *Meteor. Atmos. Phys.,* **63,** 119–129.

Kennard, R. W., and L. A. Stone, 1969: Computer aided design experiments. *Technometrics,* **11,** 137–148.

Kohonen, T., 2000: *Self-Organizing Maps*. 3rd ed. Springer-Verlag, 551 pp.

Lin, G.-F., and L.-H. Chen, 2005: Identification of homogeneous regions for regional frequency analysis using the self-organizing map. *J. Hydrol.,* **324,** 1–9.

Liu, Y., and R. H. Weisberg, 2005: Patterns of ocean current variability on the West Florida Shelf using the self-organizing map. *J. Geophys. Res.,* **110,** C06003, doi:10.1029/2004JC002786.

——, and ——, 2011: A review of self-organizing map applications in meteorology and oceanography. *Self-Organizing Maps: Applications and Novel Algorithm Design,* J. I. Mwasiagi, Ed., InTech, 253–272.

Pilar, P., C. Guedes Soares, and J. C. Carretero, 2008: 44-year wave hindcast for the North East Atlantic European coast. *Coastal Eng.,* **55,** 861–871.

Polinsky, A., R. D. Feinstein, S. Shi, and A. Kuki, 1996: Librain: Software for automated design of exploratory and targeted combinatorial libraries. *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery,* I. M. Chaiken and K. D. Janda, Eds., American Chemical Society, 219–232.

Ratsimandresey, A. W., M. G. Sotillo, J. C. Carretero Albiach, E. Álvarez Fanjul, and H. Hajji, 2008: A 44-year high-resolution ocean and atmospheric hindcast for the Mediterranean Basin developed within the HIPOCAS Project. *Coastal Eng.,* **55,** 827–842.

Richardson, A. J., C. Risien, and F. A. Shillington, 2003: Using self-organizing maps to identify patterns in satellite imagery. *Prog. Oceanogr.,* **59,** 223–239.

Sebastião, P., C. Guedes Soares, and E. Alvarez, 2008: 44 years hindcast of sea level in the Atlantic Coast of Europe. *Coastal Eng.,* **55,** 843–848.

Solidoro, C., V. Bandelj, P. Barbieri, G. Cossarini, and S. F. Umani, 2007: Understanding dynamic of biogeochemical properties in the northern Adriatic Sea by using self-organizing maps and K-means clustering. *J. Geophys. Res.,* **112,** C07S90, doi:10.1029/2006JC003553.

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas, 2000: SOM Toolbox for Matlab 5. Helsinki University of Technology Rep. A57, 59 pp.

Weisse, R., F. Feser, and H. Günther, 2002: A 40-year high-resolution wind and wave hindcast for the Southern North Sea. *Proc. Seventh Int. Workshop on Wave Hindcasting and Forecasting,* Banff, AB, Canada, Environment Canada, 97–104.

Willet, P., 1996: Molecular diversity techniques for chemical databases. *Inf. Res.,* **2** (3), Paper 19.