Presented at the 100th AMS Annual Meeting, Boston, MA, January 12-16, 2020

10.4 TRACKING OF WIND-WAVE SYSTEMS USING K-MEANS CLUSTERING

André J. van der Westhuysen[1]

[1]IMSG at NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland, USA

1. INTRODUCTION

Third-generation wind wave models such as SWAN (Booij et al. 1999) and WAVEWATCH III (Tolman et al. 2016) produce directional spectrum output with a very large number of degrees of freedom (100M to 1B per time step on a typical model grid). To reduce this large amount of information, while retaining details of complex wave fields, wave spectrum partitioning algorithms have been developed to identify significant wave components such as swells and wind sea (e.g. Vincent and Soille 1991, see Fig 1). This partitioned model output is increasingly being applied to provide targeted forecasting for specific marine activities, such as favorable wave conditions for recreation, steep wind seas that are hazardous to small craft, long-frequency swell which adversely affect large commercial ships entering ports, and so forth.

Although useful for grouping the wave component data, these partitioning algorithms operate locally in geographical space, independently at each grid point in the model. As such, the coherence of the derived swell and wind sea partitions in geographical space and time is not guaranteed. Hanson and Phillips (2001) developed a nearest-neighbor approach (in wave height-period-direction space) to associate partitions between time steps, thereby establishing the temporal coherence at a given geographical location. Devaliere et al. (2009) extended this approach to establish the spatiotemporal connection between partitions, resulting in coherent wave systems. Their "spiral tracking" method is an agglomerative algorithm, in which individual wave partitions are associated in nearest-neighbor fashion with a set of wave systems, growing in a spiral pattern from the center of the model domain towards the outer boundaries. A weakness of this approach is that it is essentially a serial operation, so that if one nearest-neighbor connection between similar partitions in space or time is missed, a wave system can be erroneously broken off, or mis-associated with another wave system. The result is a flip-flopping pattern between wave systems in time and space, as shown in Fig 2.

In the present study, we aim to remedy this problem by proposing an unsupervised machine learning approach for combining the wave partitions. This task is cast as a clustering problem, which is solved using the well-known k-means algorithm (Lloyd, 1982). As will be shown, the key difference with the previous approaches is that all wave partitions are considered simultaneously in space and time, and grouped on the basis of their common mean characterics. As such, the risk of a missed spatial or temporal connection between individual partitions, as in the previous approaches, is significantly reduced. This paper discusses the development of this cluster-based approach to wave system identification, and its proposed implementation in an operational system for nearshore wind wave forecasting.

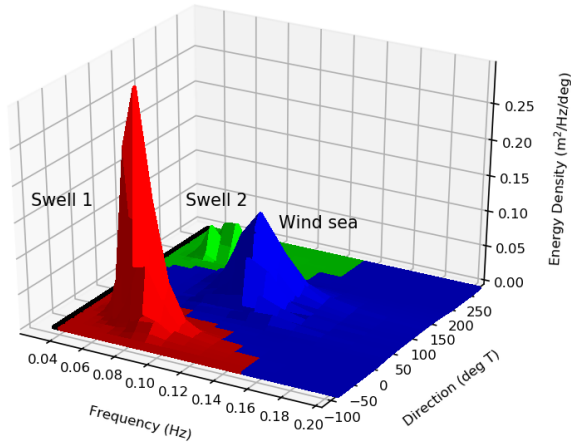

Figure 1: Separation of a directional wave spectrum into three distinct wave partitions (colors).

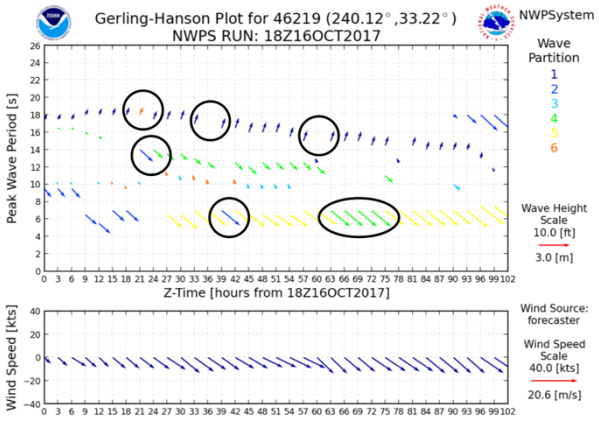

Figure 2: Erroneous "flip-flopping" of partition assignments to wave systems in simulations at NDBC station 46219. Shown are time series of wave systems (colors) in terms of their wave period (vector origin), direction (vector orientation) and height (length). Mis-assignments indicated by circles.

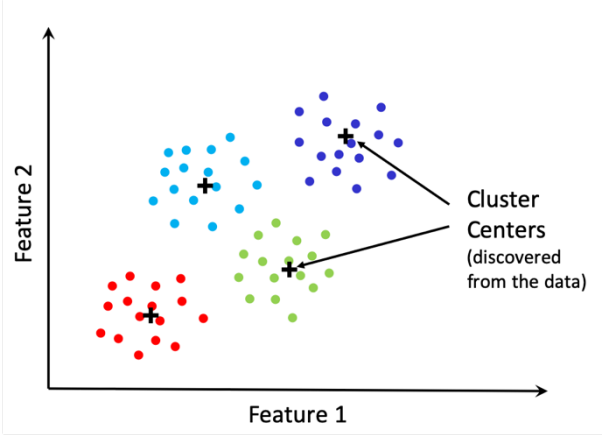

Figure 3: K-means clustering of four clusters over a two-dimensional feature space. Plusses indicate the cluster centers.

2. METHOD

Clustering is an unsupervised machine learning method widely used in science (e.g. Eisen et al. 1998) and business (e.g. Chapman and Feit 2019). One of the most efficient and popular clustering algorithms is k-means (Lloyd, 1982), which attempts to segment data points in compact groups around a fixed number of cluster centers, representing the means across a number of feature dimensions (Fig 3). The basic algorithm has the following steps: (i) initialization of cluster centers at random within the n-dimensional feature space, (ii) assignment of each data point to the nearest cluster center, (iii) recomputing the position of the cluster centers based on the mean of the data points assigned to each, (iv) repeat steps (ii)-(iii) until cluster center positions cease to change. In this study, we used the `KMeans` implementation in Python's machine learning library scikit-learn. The feature space within which the data is clustered differs by application. For the present application, features that distinguish different wave systems are unique combinations of wave period, wave direction, wave height and possibly directional spread, steepness and wave age. Therefore, once the directional wave spectrum has been divided into a set of wave partitions, at each geographical location, these features of each partition can be used as input to the clustering algorithm to identify the wave systems.

An important characteristic of the k-means algorithm is that it does not compute the number of clusters (unlike with e.g. hierarchical clustering) – rather, this has to be specified. Furthermore, in forecasting applications, the number of wave systems is not known a priori, and can vary between different forecasting cycles. The number of clusters therefore becomes a hyperparameter to be tuned. Following Scikit-learn (2019), the k-means analysis is thus repeated with a range of *k* values, and the one yielding the highest silhouette coefficient metric (Rousseeuw, 1987) is selected. The silhouette coefficient measures the normalized average difference between the distances from each data point to all the data points in its nearest neighboring cluster and those in its assigned cluster, on a scale of [-1, 1]. A score of 1 signifies a perfect outcome, with clear separation between clusters. This calculation was done using the `silhouette_score` function in Python's scikit-learn library.

3. DATA

The data is sourced from NOAA's Nearshore Wave Prediction System (Van der Westhuysen et al. 2013). This SWAN-based forecast system is run on computational grids tiled along the United States coastline, having nearshore resolutions of 1.8 km-500 m. It is driven on-demand by wind grids developed by National Weather Service forecasters, and wave boundary conditions from NOAA's operational WAVEWATCH III model. Wave-current interaction is included using surface currents from the Real-Time Ocean Forecast System (RTOFS-Global). Tides and storm surge are accounted for using the Extratropical Surge and Tide Operational Forecast System (ESTOFS).

This model produces output in terms of integral wave parameters, directional wave spectra, and wave partitions. The integral wave parameters from this model can be very complex in regions where many wave systems exist, as seen for the Hawaiian Islands domain shown in Fig 4. The conditions include a NW swell, E trade wind seas, and a S swell from a distant storm, and can be difficult to interpret when viewed in terms of an integral wave parameter field such as shown here.

The input features to the clustering algorithm are the individual wave partitions computed using the Vincent and Soille (1991) algorithm (e.g. Fig 1). This partitioning output includes the component wave period, wave direction, and significant wave height of each computed partition. The data records thus comprise the feature values of each wave partition at each geographical location, at each time step in the wave model simulation.

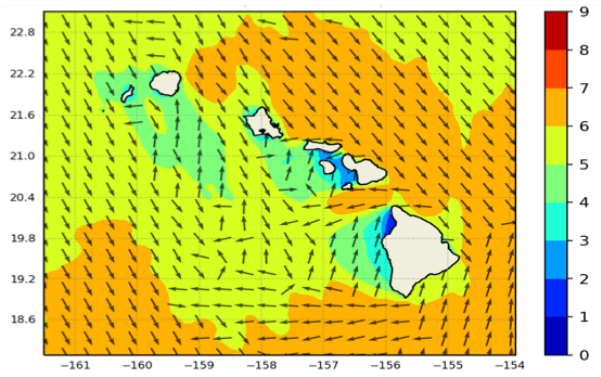

Figure 4: Significant wave height (ft, colors) and mean wave direction (vectors) forecast field over WFO Honolulu on 2019/05/29 at 00Z by the SWAN-based Nearshore Wave Prediction System.

Time	Lat	Lon	HS_PT	TP_PT	DIR_PT	System
t=0	27.0	-79.5	2.1	14.2	60.2	1
t=0	27.1	-79.5	2.2	14.0	59.7	1
t=0	27.2	-79.5	2.0	14.1	58.1	1
t=0	25.0	-82.5	1.2	9.8	295.4	2
t=0	25.1	-82.5	1.4	9.7	293.1	2
t=0	25.2	-82.5	1.1	9.9	294.3	2
t=1	27.0	-79.5	2.2	13.4	55.8	1
t=1	27.1	-79.5	2.4	13.1	55.4	1
t=1	27.2	-79.5	2.1	13.3	54.9	1
t=N	25.2	-82.5	0.5	6.5	275.2	2

Figure 5: Data frame used in the clustering operation. HS_PT, TP_PT and DIR_PT denote the partition wave height, peak period and direction, respectively.

The compilation of this wave partition data into a data frame is shown in Fig 5. The features of each partition appear as columns (e.g. latitude, longitude, partition wave height, period, direction), and the values for each partition appears by row. Note that for a given time level and geographic location, there is typically more than one partition row – this reflects the complex sea state of numerous wave systems. The typical data size is about 6M rows of partitions. The objective of the clustering operation is to assign a wave system label to each partition (last column), so that they can be assembled into spatially and temporally coherent wave system fields.

4. RESULTS

Fig 6 shows the result of the k-means clustering calculation for the wave partitions generated during a 6-day forecast simulation over the WFO Honolulu domain. The clustering is shown in wave parameter space (height-period-direction), such that each partition at each geographical location and time level appears as a point in this parameter space. Each cluster (wave system) assignment is identified with a different color. We can see that there are clear coherent clusters identified in this parameter space, each of which represents a unique wave system, in this case totaling four (hyperparameter set at k = 4). We can thus distinguish systems with combinations of low period and low wave height from the east, higher period and wave height from the northwest, and so forth. Note that time is not plotted as a separate dimension, so that these clusters represent all time levels of the wave systems during the 6-day forecast.

Although helpful in understanding the clustering results, the representation in wave parameter space is difficult to interpret for practical application. We therefore map these results back to a geographical-temporal representation using the timestamp, longitude and latitude features of each labelled wave partition. Fig 7 shows this mapping, from which we can see the four distinct wave systems within the Hawaiian Islands model domain: (i) a northwesterly swell system (first pair of panels); (ii) easterly trade wind seas (second pair of panels); (iii) a southerly swell (third pair of panels), and (iv) a southwesterly swell (fourth pair of panels). We can verify that the produced wave fields are indeed coherent in geographical space, displaying the expected wave dynamics such as shadowing behind the islands, depending on the wave direction and directional spread. Also included in Fig 7 is a radial plot of the simulated directional wave spectrum at the location of NDBC 51003. We can verify that the directional wave spectrum (the primary quantity produced by the spectral wave model) indeed shows the presence of the four wave systems identified by the clustering algorithm.

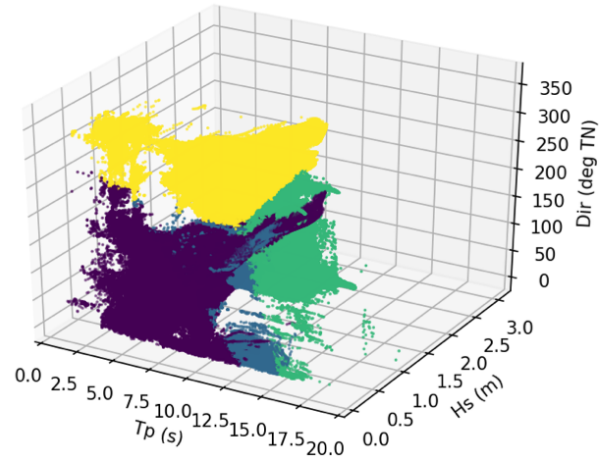

Figure 6: Wave system clustering in three-dimensional wave feature space, for WFO Honolulu. Colors indicate the identified wave systems, for k = 4.

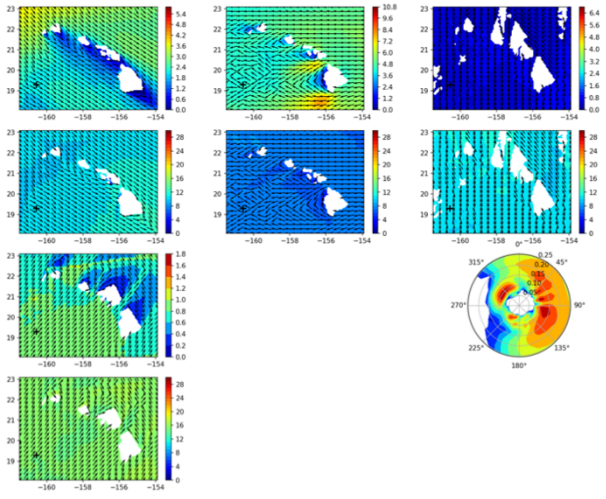

Figure 7: Wave system clustering in geographical space, for WFO Honolulu. Each pair of panels show a distinct wave system, with upper panels showing wave height and direction, and lower panels showing peak period and direction. The plus symbol in the SW corner of the domain indicates the location of the included directional wave spectrum.

4.1 Optimal number of wave systems

As described in Section 2, the k-means clustering algorithm does not yield an estimate of the number of clusters (wave systems). Rather, it is a hyperparameter that must be provided as input to the algorithm. In collaboration with our NOAA weather forecasters, we determined that five is the maximum number of wave systems that can effectively be communicated in forecasts. The clustering algorithm therefore iterates through a number of trial operations with k = 2 to 5, determines the silhouette coefficient for each, and then selects the k that yields the highest value of this performance metric. Figure 8 shows the results for this hyperparameter search for the 2019/05/29 00Z Hawaiian Islands example. For these conditions, the optimal number of clusters (wave groups) was three, and it achieved a quite a high silhouette coefficient of 0.73.

Fig 9 shows the geographical mapping of the clustering result with the optimal value of the hyperparameter k = 3. From the included directional wave spectrum, we can manually verify that for the conditions of this forecast, a total of three clusters (easterly wind seas, northwesterly swell, southwesterly swell) is indeed appropriate. By contrast, if we check the results of the trial runs with k = 5 (Fig 10), we see significantly poorer results. We can see that the third wave system identified here agrees with the third wave system identified above (Fig 9), but that the remaining four wave systems differ. Instead of the coherent and physically realistic wave systems of Fig 9, the k = 5 solution shows incoherent, unphysical wave systems, in particular for systems 1 and 2 (Fig 10, top left and top center). This is an example of a poor clustering result, where wave groups are erroneously pulled apart, because the specified k is greater than the actual number of wave systems that occurred during these conditions. As a result, this trial's silhouette score is lower at 0.59, and is hence not selected as the final clustering result.

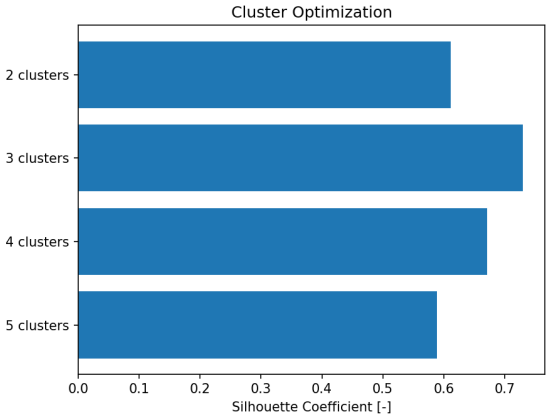

Figure 8: Comparison of silhouette scores for values of the hyperparameter *k* (number of clusters) ranging from 2 to 5.

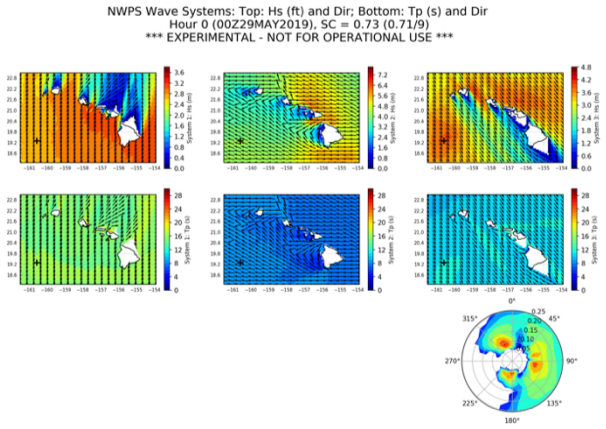

Figure 9: Clustering results over the Hawaiian Islands domain for *k* = 3. Silhouette coefficient = 0.73.

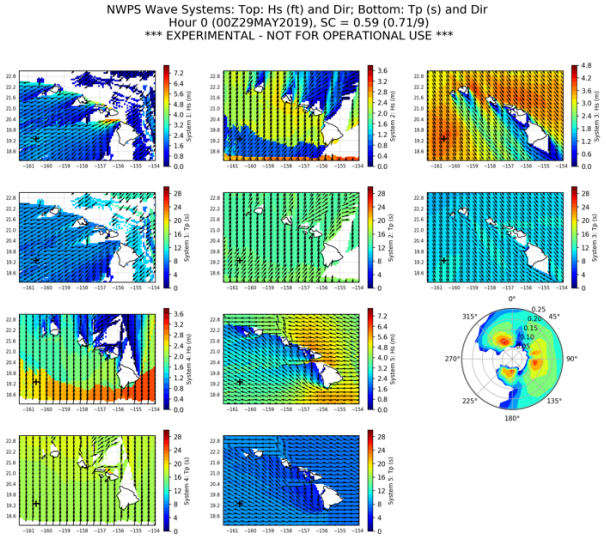

Figure 10: Clustering results over the Hawaiian Islands domain for *k* = 5. Silhouette coefficient = 0.59.

4.2 Temporal consistency

As discussed in Section 1, a concern with existing methods for producing wave systems such as Hanson and Phillips (2001) and Devaliere et al. (2009) is the serial nature of their algorithms. The proposed cluster-based approach solves this problem by segmenting all time levels of the wave partitions simultaneously, making it much less vulnerable to missed temporal connections from one time step to the next.

We can verify this improvement by studying the time series of the above clustering results. Fig 11 and 12 show time examples of the clustered wave model results at the locations of NBDC 51003 and NDBC 51201 in the Hawaiian Islands model domain for the May 29, 2019 00Z forecast cycle. From Fig 9 above we know that there were

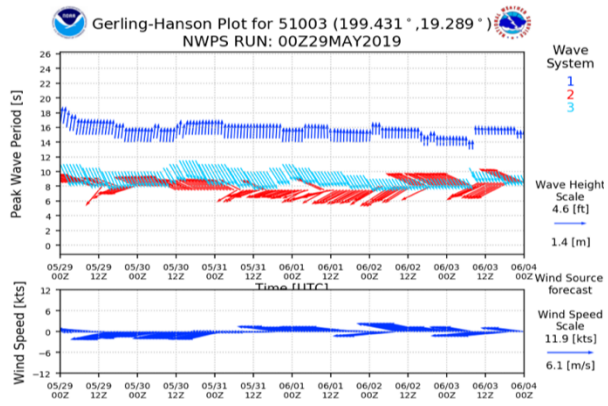

Figure 11: Time series of wave systems (colors) at NDBC 51003 (SW of Hawaiian Islands) for the simulation of 2019/05/29 00Z. Shown are wave period (vector origin), direction (vector orientation) and height (vector length), and the associated wind conditions (bottom panel).

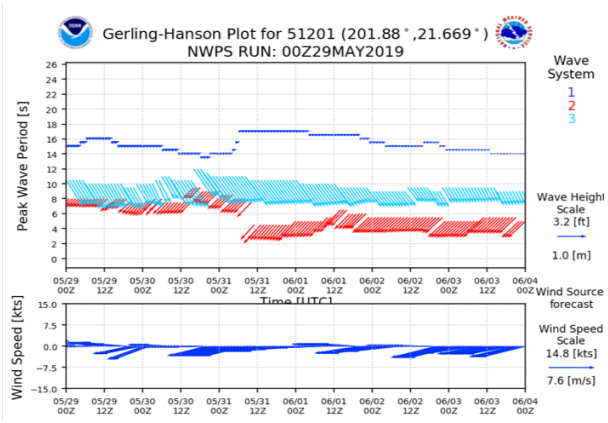

Figure 12: Time series of wave systems (colors) at NDBC 51201 (Oahu North Shore) for the simulation of 2019/05/29 00Z. Shown are wave period (vector origin), direction (vector orientation) and height (vector length), and the associated wind conditions (bottom panel).

three dominant wave systems present during this forecast cycle, namely a northwesterly swell, a long-period southerly swell, and an easterly wind sea. Fig 11 shows the time series of these systems at NDBC 51003, located to the southwest of the Hawaiian Islands. At this open water location, we can clearly see the time series of the three wave systems in different colors. The southerly swell (system 1) has the highest peak period, followed by the lower period of the northwesterly swell (system 3). As expected, the easterly wind sea (system 2) has the lowest peak period. Notice that the color labeling of the time series of all three systems is uniform, without the flip-flopping seen in the example of Fig 2. This indicates that the time series of the wave systems are indeed temporally consistent. One interesting feature in this result is the alternating peak direction in the easterly wind sea (system 2). Inspection of the lower panel, shows that the peak direction of this wind sea generally agrees with the prevailing easterly winds, as expected. However, due to the easterly wind sea condition, there is sheltering on the western side the Big Island of Hawaii (Fig 9, center panels). Since station 51003 is located downwind of this disturbance, it experiences cross-sea conditions within this single wind sea system – waves from one part of this system approach from the north of the Big Island, and those from the other part from the south. The result is alternating peak wave directions from this single wave system.

Further interesting phenomena are seen in Fig 12, showing nearshore results at NDBC 51201 at the North Shore of Oahu. We still see the three wave systems discussed above, but at different magnitudes due to the nearshore location. First, since the location is at the northern face of the island, it is almost completely sheltered from the southerly swell (system 1), which shows up only with a very small wave height. Similarly, the easterly wind sea (system 2) refracts around the island of Oahu, so that it arrives from the northeast at this location (and not as crossing seas, as seen at 51003). Finally, as expected, the northwesterly swell for which this location is famous arrives unobstructed and is seen clearly with the largest wave height as system 3. Notice again that all three these wave systems have been identified coherently in time, without any flip flopping between their assignments.

5. DISCUSSION

In this study, we investigated the application of clustering to the segmentation of wave partitions computed from simulated directional wave spectra into coherent wave systems in space and time. The presented methodology has shown promising results to practical field applications such as the nearshore wave forecasts of the NOAA Weather Forecast Offices, for example those of the Honolulu office shown here.

There are, however, a few remaining challenges in applying clustering for wave system segmentation. First, the number of partitions that each directional wave spectrum is decomposed into does not necessarily match the number of clusters *k* for a given trial. If the number of partitions at a particular location and time step is fewer than *k*, some wave systems are not associated with a partition. This can occur if a wave system is smaller than the full model domain, or is transient and does not last the full analyzed duration. However, the number of partitions can also be greater than *k*, in which case more than one partition can be assigned the same cluster (wave system) label. In this case, the variance of all the assigned partitions is summed to obtain a combined significant wave height for that wave system. The peak period is computed as a weighted average over the assigned partitions on the basis of their variance densities. The peak direction of the combined wave system is the

direction with the greatest variance. This process of averaging over partitions can result in a degree of blending between wave systems, which is not always desirable. However, except if the computed spectrum is very noisy, the optimum number of clusters *k* would generally be equal to, or greater than the number of identified wave partitions, so that this averaging problem would not occur.

A second challenge of the proposed approach is describing conditions that are highly nonstationary. The conditions presented here are approximately stationary in that their wave height-period-direction characteristics do not change significantly during the course of the 6-day forecast. When conditions within a cluster (wave system) change rapidly, such as for example in a hurricane vortex, it is more difficult to identify a unique feature combination with which to associate that cluster. A solution would be to not analyze the entire forecast period simultaneously, but rather to apply a moving window within which the clustering is computed.

A final challenge is in the identification of very young wind seas. Small craft in coastal waters are vulnerable to steep wind seas which can develop within an hour of the start of a local storm. These wind seas manifest themselves as a high-frequency peak in the directional wave spectrum. Hence, they would be identified as a new partition, and subsequently assigned to a wave system cluster. However, in modeled directional wave spectra, we typically do not immediately see the presence of young wind seas if there is ambient swell present from a similar direction. This is because the source term for quadruplet nonlinear interaction used in models such as SWAN tends to smooth disturbances in the high-frequency tail. It therefore takes a few hours for the young wind sea to develop its own spectral peak in the model, and hence its own partition and wave system. Thus, the clustering approach has difficulty in identifying very young wind sea within an hour or two of its origination, even though it is not the root cause of this deficiency. Solutions for this deficiency should therefore be sought in an improved spectral representation in the underlying third-generation spectral wave models.

6. CONCLUSIONS

This study investigated the application of k-means clustering to identify wave systems in numerical wave model forecast runs. Examples of the resulting wave system fields in space and time were shown, as well as the methodology for the selection of the appropriate number of wave systems. From the results of this study, the following can be concluded:

i. K-means clustering is found to be successful in the segmentation of wave partitions into wave systems that are coherent in both space and time. The complex examples over the Honolulu Weather Forecast Office model domain considered here show analyzed wave systems with consistent spatial fields, and time series free of the flip-flopping of system assignments found with the Devaliere et al. (2009) spiral tracking algorithm.

ii. The silhouette coefficient was found to be an effective quality metric for determining the hyperparameter setting *k*, the number of clusters (wave systems) present in the wave condition, which is unknown a priori for a given forecast cycle. Optimal values are sought within the range $k = 2$ to 5, and these agree well with manual inspection of the modeled directional wave spectrum.

iii. Remaining challenges of this cluster-based approach are: (a) Reconciling mismatches between the number of wave partitions identified by partitioning algorithms such as Vincent and Soille (1991) and the optimal wave systems *k* determined using the silhouette coefficient, (b) Computing wave systems for highly nonstationary conditions where wave system characteristics vary rapidly, and (c) Identification of very young wind seas, due to limitations in the underlying modeled directional wave spectra.

7. REFERENCES

Booij, N., R. C. Ris and L. H. Holthuijsen, 1999. A third-generation wave model for coastal regions, Part I, Model description and validation, J. Geophys. Res., 104, C4, 7649-7666.

Chapman, C. and Feit, E. McDonnell, 2019. R for Marketing Research and Analytics, Second Edition, Springer, Switzerland.

Devaliere, E.-M, J. L. Hanson and R. A. Luettich, Jr., 2009. Spatial tracking of numerical wave model output using a spiral tracking search algorithm, Proc. 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, Vol. 2, 404-408.

Eisen, M. B., Spellman, P.T., Brown, P. O. and Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 95(25):14863–8. doi: 10.1073/pnas.95.25.14863.

Hanson, J. L. and O. M. Phillips, 2001. Automated analysis of ocean surface directional wave spectra. J. Atmos. and Ocean Tech., Vol. 18, 277-293.

Lloyd, S. P., 1982. Least squares quantization in PCM. IEEE Transactions on Information Theory. 28 (2): 129-137. doi:10.1109/TIT.1982.1056489.

Rousseeuw, P. J., 1987. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

Scikit-learn, 2019. Selecting the number of clusters with silhouette analysis on KMeans clustering. Retrieved on February 12, 2020, from https://scikit-

learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py.

Tolman, H.L., et al., 2016. The WAVEWATCH III development group (ww3dg), 2016: User Manual and system documentation of WAVEWATCH III version 5.16., College Park, MD, USA. NOAA/NWS/NCEP/MMAB Tech. Note 329:326

Van der Westhuysen, A. J., R. Padilla-Hernandez, P. Santos, A. Gibbs, D. Gaer, T. Nicolini, S. Tjaden, E. M. Devaliere and H. L. Tolman, 2013. Development and validation of the Nearshore Wave Prediction System. Proc. 93rd AMS Annual Meeting, Am. Meteor. Soc., Austin.

Vincent, L. and P. Soille, 1991. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. IEEE Transactions of Pattern Analysis and Machine Intelligence, 13, 583-598.

.