Evaluation of Clustering Techniques for Revealing Dependence Between Wind Speed and Surge Height

Paulina E. Kindermann[1,2(✉)], José A. A. Antolínez[1], and Oswaldo Morales-Nápoles[1]

[1] Delft University of Technology, Stevinweg 1 2628 CN, Delft, Zuid-Holland, The Netherlands
p.e.kindermann@tudelft.nl

[2] HKV Lijn in Water, Informaticalaan 8, 2628 ZD Delft, Zuid-Holland, The Netherlands

Abstract. Extreme storms over the North Sea drive coastal flood risk in the Netherlands, causing high waves and extreme sea levels. Designing flood defenses requires accurate statistical extrapolation of hydraulic load conditions with return periods of 1,000 years or more. This is a challenging task given limited observational data. This study uses a large, simulated dataset (~9,000 years) to explore the statistical dependence between extreme wind speed u and surge height s. Storms were clustered using several techniques. Self-organizing maps (SOM) effectively captured physical relationships, such as the influence of wind direction and tidal offset on storm dynamics, however variability in statistical dependence between u and s for different clusters was better represented using manual clusters. Copula models were fitted to the cluster data, with the BB8 copula outperforming others. This study illustrates the potential of machine learning to identify patterns in large datasets while emphasizing the relevance of manual clustering approaches for revealing nuanced statistical dependencies critical to flood risk assessment.

Keywords: Storm surge modelling · Statistical dependence · Self-organizing maps

1 Introduction

Extreme storms above the North Sea are the main driver of coastal flood risk in the Netherlands, resulting in high waves and extreme sea levels due to storm surge. To mitigate this risk, flood defenses are designed to sustain extreme hydraulic load conditions, defined in terms of water levels and wave heights with return periods of 1,000 years or more [1]. Although observations span approximately 100 years, the derivation of these design conditions relies on statistical extrapolation, introducing uncertainties due to limited data length relative to the target return periods, data inhomogeneities, and missing storm observations [2]. Recent advances in numerical weather prediction have enabled reanalysis and hindcast datasets to emerge as a standard alternative to observational data [3]. However, 1000-year and beyond return period estimates using probabilistic models fitted with approximately 75 years of reanalysis/hindcast might still introduce uncertainty when natural variability is large [3]. Along improved atmospheric reanalysis [4],

C. Coelho et al. (Eds.): CD 2025, COASTALRL 42, pp. 135–141, 2026.
https://doi.org/10.1007/978-3-032-15477-4_22

short, mid, and recently long range forecasting have been made available, such as the seasonal forecast system (SEAS5) by the European Centre for Medium-Range Weather Forecasts (ECMWF)[5]. Recently, the Royal Dutch Meteorological Institute (KNMI) has produced a large dataset of simulated sea levels, using the DCSM-model with wind data from SEAS5 [6]. Effectively, approximately 9,000 years of simulated weather and storm surge are now available, allowing the derivation of higher-precision estimates for the return levels of extreme wind speed and sea levels.

Moreover, this large dataset enables us to study other storm characteristics and mutual correlations in more detail. The aim of this research is to study the statistical dependence between extreme wind speed and surge height, using the large dataset by KNMI and applying copula models. The outcomes are comprehensive dependency structures, conditioned to environmental, causal and consequential aspects of storms, which is achieved by clustering storms, which is important for modelling hydraulic loads in the design of coastal infrastructure.

2 Data and Methods

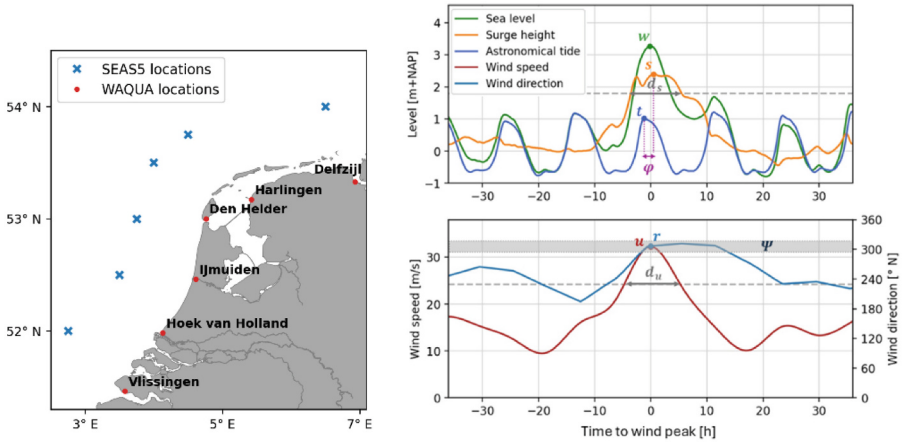
The dataset by KNMI [6] contains simulated sea level time series from the DCSMv5 model [7], forced with wind fields from SEAS5 [5]. The so-called Dutch Continental Shelf Model (DCSM) is a numerical model that solves the two-dimensional shallow-water equations, which is also used for operational forecasting along the Dutch coast. The DCSMv5 model is run with the 6-hourly mean sea level pressure and wind stress from each SEAS5 ensemble member as input, resulting in sea levels in the North Sea with a 10-min resolution [7]. To prevent mutual dependence among ensemble members, the first month of each ensemble is excluded. From a statistical perspective, each ensemble member from SEAS5 represents a possible climate scenario, enabling their combination into a single, extensive dataset equivalent to approximately 9,000 years of data. In past years, KNMI has performed numerous validations of the dataset, using various DCSM-versions and observational data. More details can be found in [6, 8].

2.1 Storm Selection and Parametrization

The first step is to select extreme events from the dataset of simulated wind and sea levels. For this study, six locations along the Dutch coast were selected. Sea level time series were extracted from the closest DCSMv5 output grid cell, and wind data from nearby SEAS5 grid cells, as shown in Fig. 1.a. Since wind speed is one of the main drivers of extreme sea levels, storms are selected using a peaks-over-threshold based on wind speed. The threshold is defined conditional on the wind direction during the peak wind speed, such that 9,000 storms were selected for each of the 16 wind directions, per location. This roughly corresponds to storm events with a return period of one year, given the wind direction.

For each selected storm event, time series of wind speed, wind direction, sea level and astronomical tide are extracted from the dataset, for a period of three days around the storm peak. The residual surge is determined, being the difference between the sea level and astronomical tide. An example of one selected storm event is presented in Fig. 1.b,

including definitions of relevant storm features. The maximum wind speed u in [m/s] is the selection variable and therefore the time index is defined such that u occurs at $t = 0$. Other relevant storm features are the wind direction r during maximum wind speed using the nautical convention [°N], the maximum sea level w in meters [m] with respect to the Dutch reference level [m + NAP], the maximum surge height s [m] and the closest astronomical high tide t [m]. The tidal offset φ being the time difference between s and t in hours [h], which is defined positive if the surge peak precedes the tidal peak, as in the Fig. 1.b. Ψ [h] is a measure for wind rotation, defined as the duration that the wind direction resides within the corresponding 22.5°-sector. Finally, d_u and d_s [h] are the exceedance duration of the 75%-level of maximum wind speed u and surge height s, respectively, during the storm event.



(a) Map of the Netherlands with the selected locations for sea level data (red dots) and wind data (blue crosses).

(b) Temporal evolutions of sea level (green line), residual surge (orange line) and astronomical tide (blue line) in the upper graph, and wind speed (red line) and wind direction (blue line) in lower graph.

Fig. 1. Map of locations **(a)** and example of storm event **(b)**, including definitions of features

2.2 Clustering

The next step is to cluster the selected 9,000 storm events to identify different storm types. Three vector quantization methods, the Self-organizing maps (SOM) [9], the k-means algorithm (KMA) based on storm features and KMA based on the full time evolutions [10], were compared with a 'manual' clustering based on storm features. This manuscript focuses on results from SOM. The SOM algorithm was performed on a hexagonal lattice and the initialization vector is determined using the maximum dissimilarity algorithm (MDA) [11]. The nine storm features, as defined in Fig. 1., were normalized and weighted, based on expert judgment. An evaluation of the inter- and intra-cluster variability, revealed that 196 clusters were a suitable choice. The SOM identifies neurons, each neuron 'holds' adjacent data we define as clusters.

2.3 Statistical Dependence

For each cluster, the statistical dependence between u and s is evaluated. First, the pairs are translated to uniform space using the empirical cumulative distribution function (CDF). Then, the following copula models are fitted to the data: the Gaussian, Student, Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7 and BB8 copula. The goodness of fit is evaluated using the Cramér-Von Mises criterion, S_{CvM} [12].

3 Results

Figure 2 presents the feature values for each node from the trained SOM algorithm for location Vlissingen, with each plot corresponding to one of the nine features defined in Fig. 1.b. The nodes (hexagons) represent clusters of storms, and the colors indicate the feature values for each node. It can be observed that the SOM algorithm effectively captures several physical relationships. For instance, the plots show that southern (180–270°N) wind directions (lower right plot) are generally associated with higher wind speeds (upper left plot) compared to northern wind directions. However, these high wind speeds only result in significant surge heights for western to northern wind directions, as highlighted by the red square. Notably, a few nodes exhibit relatively high sea levels w despite low surge heights s, as indicated by the purple square in the lower center plot. This can be explained by the minimal time offset between the surge peak and astronomical high tide, ϕ, (right center plot), meaning that the maximum surge (almost) coincides with high tide, which results in high sea levels even if the surge is only moderate.

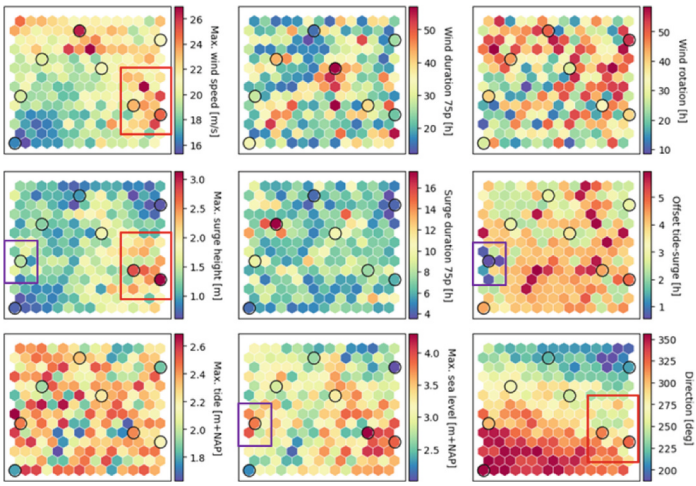


Fig. 2. Resulting feature values for each nodes in a 14x14 grid from the SOM algorithm for Vlissingen. Each plot represents one of the nine features and the color indicates the feature value.

For nodes indicated with black circles in Figs. 2 and 3 shows pair plots of surge height s and wind speed u in the standard normal space, including Pearson's correlation

coefficient (σ) for each quadrant. These nodes were selected to capture variation across cluster's feature values. Some of the feature values corresponding to these eight nodes are indicated in the legend of Fig. 4. Based on the feature values, we would expect a strong correlation between u and s for clusters 158 and 185 (the two black circles in the red square), since these correspond to northern wind directions that exhibit high values of s for high values of u, as discussed previously. However, it can be observed that none of the clusters show a strong upper tail dependence. While clusters 158 and 185 (associated with northern wind directions) indeed show somewhat higher σ-values in the upper joint tail than other clusters, clusters 0 and 5, also linked to northern wind directions, exhibit very low values of σ in the upper corner. Interestingly, cluster 37 shows a somewhat higher σ-value in the upper tail, which may be attributed to its long surge and wind durations. Figure 4 evaluates the performance of various copula models fitted to the data pairs from Fig. 3, with lower S_{CvM}-values indicating a better fit. Among these, the BB8 copula provides the best fit for most clusters.

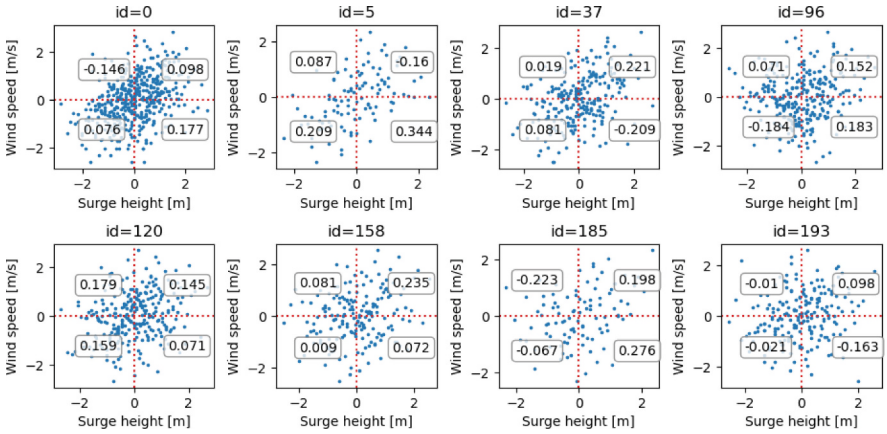


Fig. 3. Pair plots in the standard normal space for eight example nodes from the SOM results for Vlissingen. The Pearson's correlation coefficient is indicated in each corner.

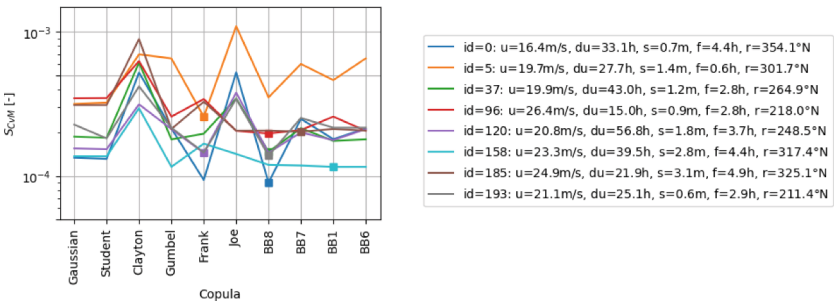


Fig. 4. Performance of various copula models fitted to the data corresponding to eight example nodes, expressed by S_{CvM}. The squares indicate the copula model that best fits the data.

4 Discussion, Conclusions and Future Work

Figures 3 and 4 show no clear link between specific cluster characteristics and the statistical dependence between *u* and *s*. In contrast, manual clustering reveals more variability, with some clusters showing strong upper tail dependence ($\sigma \sim 0.6$) and others weak dependence, alongside clearer differences in copula model performance. However, only eight clusters were evaluated from the SOM algorithm. To fully assess the capability of SOM to capture variations in statistical dependence, the semi-correlations should be evaluated for all 196 clusters, to identify potential patterns and relationships with specific features.

This study demonstrates how machine learning techniques, such as Self-Organizing Maps (SOM), can potentially contribute to effectively identifying patterns in large datasets, revealing key physical relationships in storm dynamics. At the same time, the findings highlight the ability of manual clustering approaches for capturing nuanced statistical dependencies.

References

1. Jonkman SN, Schweckendiek T (2015) Developments in Levee Reliability and Flood Risk Analysis in the Netherlands. Geotechnical safety and risk V, pp. 50–60. IOS press
2. Haigh ID et al (2023) GESLA Version 3: A major update to the global higher-frequency sea-level dataset. Geoscience Data Journal 10(3): 293–314
3. Bauer P, Thorpe A, Brunet G (2015) The quiet revolution of numerical weather prediction. Nature 525:47–55
4. Hersbach H et al (2020) The ERA5 global reanalysis. Q J R Meteorol Soc 146(730):1999–2049
5. ECMWF: SEAS5 User guide (2021)
6. De Valk CF, van den Brink HW (2024) An appraisal of the value of simulated weather data for quantifying coastal flood hazard in the Netherlands. EGUsphere. Preprint
7. Gerritsen H, De Vries H, Philippart M (1995) The Dutch Continental Shelf Model. Coast Estuar Stud 9:425–467
8. Van den Brink HW (2020) Het gebruik van de ECMWF seizoens-verwachtingen voor het berekenen van de klimatologie van extreme waterstanden langs de Nederlandse kust. KNMI
9. Kohonen T (2001) Self-Organizing Maps. Springer Series in Information Sciences, 30
10. Lloyd SP (1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28(2):129–137
11. Camus P, Mendez FJ, Medina R, Cofiño AS (2011) Analysis of clustering and selection algorithms for the study of multivariate wave climate. Coast Eng 58(6):453–462
12. Genest C, Rémillard B, Beaudoin D (2009) Goodness-of-fit tests for copulas: a review and a power study. Insurance Math Econom 44(2):199–213