Statistical modeling of the space-time relation between wind and significant wave height

Said Obakrim[1,1], Pierre Ailliot[2,2], Valerie Monbet[1,1], and Nicolas Raillard[3,3]

[1]University of Rennes 1
[2]University of Brest
[3]IFREMER

November 30, 2022

Abstract

Many marine activities, such as designing ocean structures and planning marine operations, require the characterization of sea state climate. This study investigates the statistical relationship between wind and sea states, considering its spatiotemporal behavior. A transfer function is established between wind fields over the North Atlantic (predictors) and the significant wave height (predictand) in a location in the Bay of Biscay off the French coast. The developed method takes into consideration both wind seas and swells by including local and global predictors. The global predictors' spatiotemporal structure is defined to account for the non-local and non-instantaneous relationship between wind and waves, using a fully data-driven approach. Weather types are constructed using a regression guided-clustering method, and the resulting clusters correspond to different wave systems (swells and wind seas). Then, in each weather type, a penalized linear regression model is fitted between the predictor and the predictand. The validation analysis proves the model's skill in predicting the significant wave height (RMSE = 0.27m); furthermore, the interpretability of the model is discussed.

Statistical modeling of the space-time relation between wind and significant wave height

Said Obakrim [1,2*], Pierre Ailliot [3], Valérie Monbet [1], Nicolas Raillard [2]

[1]Université de Rennes, IRMAR, France.
[2]Laboratoire Comportement des Structures en Mer, IFREMER, Plouzané, France
[3]Laboratoire de Mathématiques de Bretagne Atlantique, UBO, Brest, FRANCE

Key Points:

- A data-driven approach is proposed to determine the wave generation area for any given ocean location
- A weather-type approach that links wind fields and the significant wave height
- Regression-guided clustering improves the prediction of the significant wave height and provides interpretable weather types

*Centre Ifremer Bretagne - ZI de la Pointe du Diable - CS 10070 - 29280 Plouzané

Corresponding author: Said Obakrim, said.obakrim@univ-rennes1.fr

Abstract

Many marine activities, such as designing ocean structures and planning marine operations, require the characterization of sea state climate. This study investigates the statistical relationship between wind and sea states, considering its spatiotemporal behavior. A transfer function is established between wind fields over the North Atlantic (predictors) and the significant wave height (predictand) in a location in the Bay of Biscay off the French coast. The developed method takes into consideration both wind seas and swells by including local and global predictors. The global predictors' spatiotemporal structure is defined to account for the non-local and non-instantaneous relationship between wind and waves, using a fully data-driven approach. Weather types are constructed using a regression guided-clustering method, and the resulting clusters correspond to different wave systems (swells and wind seas). Then, in each weather type, a penalized linear regression model is fitted between the predictor and the predictand. The validation analysis proves the model's skill in predicting the significant wave height (RMSE = 0.27m); furthermore, the interpretability of the model is discussed.

Plain Language Summary

Ocean wave climate has a significant impact on human activities and its understanding is of socio-economic and environmental importance. In this study, we propose a statistical model that predicts the significant wave height, in a location in the Bay of Biscay, using the North Atlantic wind fields. At first, we define the predictors of the model to account for both wind seas and swells. Then, a weather-type based approach is used to construct the link between the predictors and H_s. The proposed method allows to understand the spatiotemporal relationship between wind and waves and predicts well both wind seas and swells.

1 Introduction

A sea state is a statistical description of the sea surface waves generated by wind at a given time and location. The sea state is characterized by a superposition of wind seas and swells (Ardhuin & Orfila, 2018). The local wind generates wind seas, whereas swells are generated in distant areas. Significant wave height (H_s), defined as four times the zeroth moment of the wave power spectrum, is commonly used to describe the sea state. Thus, H_s is an essential measure of wave height and provides information about the wave energy of a given sea state.

Sea state climate characterization has received increasing interest in the past decades. High-quality wave data is essential for many marine applications, such as designing coastal and offshore structures and planning marine operations. Traditional *in situ* measurements obtained from buoys provide the most reliable data for sea state parameters; however, they are only available for the last decades, and they are limited spatially. Numerical models (Hasselmann et al., 1973; Tolman et al., 2009) provide deterministic simulations of spectral wave models from which sea state parameters are extracted. They are a valuable source of data and provide decades of records; although they are computationally expensive and sensitive to the quality of forcing fields (wind, currents, and water levels) (Roland & Ardhuin, 2014).

Statistical models constitute an alternative to numerical models for constructing the wind-waves relationship. These models are not computationally expensive, and once the statistical relationship is estimated, future predictions can be made by assuming that this relationship will stay the same in the future. Models of this kind are known in the literature under the name of statistical downscaling (SD) models (Maraun et al., 2010). General Circulation Models (GCMs) are the primary tools that provide future projections of atmospheric variables. However, on the one hand, these models generally do not

provide ocean sea state parameters. On the other hand, GCMs provide coarse spatial resolution simulations, making them unsuitable for most impact assessment applications. To bridge the gap between what GCMs provide and what industry and policymakers require, SD models establish an empirical relationship between large-scale atmospheric and local mesoscale variables. Besides their computational efficiency, SD models have been compared with numerical models in various studies for ocean wave parameters and other climate variables. Wang et al. (2010) compared these methods in terms of climatological characteristics of the present period using ERA-40 wave data. They found that the statistical models are better at reproducing the observed climate than the dynamical models. Laugel et al. (2014) analyzed these methods for climate projections. The study shows that the statistical downscaling approaches can reproduce both the present climatology and future projections. In addition, they estimate the uncertainties associated with the choice of general circulation models (GCMs) or climate scenarios. However, some challenges remain in modeling the relationship between wind and sea state parameters.

Wind waves are generated by the surface wind. However, it is not only the local wind that defines local waves. Wind from distant regions generates waves that may take days to arrive at the target point; thus, the relationship between wind and waves is neither local nor instantaneous. Therefore, it is necessary to consider a large wave generation area to understand the wave dynamics at a particular location. The ESTELA (Evaluation of Source and Travel-time of wave Energy reaching a Local Area) (Pérez et al., 2014) is a method that defines the wave generation area and wave travel time at any ocean location worldwide. Using its spectral information, the method selects the fraction of energy that travels to the target point from selected source points. The present study uses an entirely data-driven approach to define wave generation area. It is based on estimating waves' travel time from each source to the target point using the maximum correlation between the significant wave height and wind conditions. Therefore, this method is not computationally expensive, and only wind data and H_s at the target point are needed.

Statistical downscaling models have to take into consideration both wind sea and swells, which is challenging in swell-dominated areas (Hemer et al., 2012). Therefore, given the non-instantaneous and non-local relationship between wind and waves, the quality of SD approaches relies on the quality of the definition of atmospheric predictors. Camus, Menéndez, et al. (2014) used a weather types-based SD model to downscale wave parameters in the northwest of Spain. Predictors were defined as the average over three days of the sea level pressure (SLP) and its gradient to account for the superposition of swells. The ESTELA method can help the design of statistical and dynamical downscaling models. For example, Camus, Méndez, et al. (2014) and Hegermiller et al. (2017) used the ESTELA approach to define the spatiotemporal coverage of predictors used in their SD framework.

After defining the predictors, the statistical relationship between predictors and the predictand can be modeled using either multiple linear regression (Casas-Prat et al., 2014), weather types (Camus, Menéndez, et al., 2014), or nonlinear models like neural networks (Baño-Medina et al., 2020). In the case of high-dimensionality and multicollinearity, they may learn well the underlying physical relationships that generate the data; however, they lack the capability of generalizing to new data (overfitting) and may be difficult to interpret. To improve generalization, penalized approaches like Ridge (Hoerl & Kennard, 1970) and LASSO (Tibshirani, 1996), have been proposed. Furthermore, the physical interpretability of statistical models applied to climate is of major interest. Interpretable models provide transparent information about climate and can be trustfully applicable for decision making (Kashinath et al., 2021).

Weather types-based SD approaches consist in finding the leading atmospheric circulation patterns and then fitting a different model between the predictor and predictand in each weather type (Maraun et al., 2010). Weather types can be found by using a clustering algorithm on the predictor (Camus, Menéndez, et al., 2014). This method

constructs clusters without accounting for the local environment. Cannon (2012) proposed a method that accounts for both local and global climate conditions by using a regression-guided clustering algorithm. Instead of using an unsupervised algorithm on atmospheric variables, they used the clustering algorithm on atmospheric data combined with predictions of a regression model that links atmospheric variables with local variables.

This study provides a framework for the wind to waves relationship using an entirely statistical approach. The statistical relationship, henceforth called the transfer function, is a function that links the space-time wind fields over North-Atlantic (predictors) and the significant wave height (predictand) at a single site located in the Bay of Biscay off the French coast. The developed methodology considers wind sea and swells and provides additional information about the spatiotemporal relationship between wind and waves. The main contribution of this work, on one hand, it provides a fully data-driven approach that estimates the travel time of waves from any source point to a target point, which is an essential information for the definition of predictors. On the other hand, it proposes a regression-guided clustering algorithm that account for both global and local climate to construct weather types.

This paper is structured as follows. After describing the data in Section 2, the local predictors are defined in Section 3. Then, in Section 4, the construction of the global predictors is described. Next, in Section 5, the statistical model that combines the local and global predictors is presented. Then, Section 6 presents the results of the SD model. Finally, the study is concluded in Section 7.

2 Data

The atmospheric data used in this work to construct predictors is extracted from the Climate Forecast System Reanalysis (CFSR) (Saha et al., 2010). CFSR is a global reanalysis, developed at the National Centers for Environmental Prediction (NCEP), that covers the period from 1979 to the present with hourly time step and spatial resolution of 0.5°by 0.5°. Extracted data consists of hourly $10m$ zonal and meridional wind components in the north Atlantic (Figure 1).

The historical wave data used in this work is the sea-state hindcast database HOMERE (Boudière et al., 2013) based on the WAVEWATCH III model forced by CFSR wind. The database covers the English Channel and the Bay of Biscay with unstructured computational mesh. It contains 37 parameters and the frequency spectra on high spatial resolution, ranging from 200 m to 10 km, with a one-hour time step.

The point of interest is located in the Bay of Biscay (figure 1) at (25.4°N, 1.6°W). Waves at this point are related to both large-scale conditions in the North Atlantic (swells) and to local conditions (wind seas) (Charles et al., 2012). Swell conditions are generally dominant; however, the highest H_s are generated by strong local storms. To identify different wave systems, energy spectral partitioning methods are used. Homere uses the watershed algorithm (Tracy et al., 2007) to separate wind sea and different swells.

The temporal resolution of both predictors and predictand is upscaled from hourly to 3 hourly resolutions to facilitate the analysis. Both datasets comprise a common period of 23 years, from 1994 to 2016.

3 Local predictor

Wind speed, duration, and the fetch have an important impact on the characteristics of the wind sea (Ardhuin & Orfila, 2018). Hereafter, at time t the variables $U(t)$, $F(t)$, $U(t-1)$, and $F(t-1)$ are considered to construct the local predictors. $U(t)$ is the wind speed at the target point and $F(t)$ is the fetch length at time t, calculated as

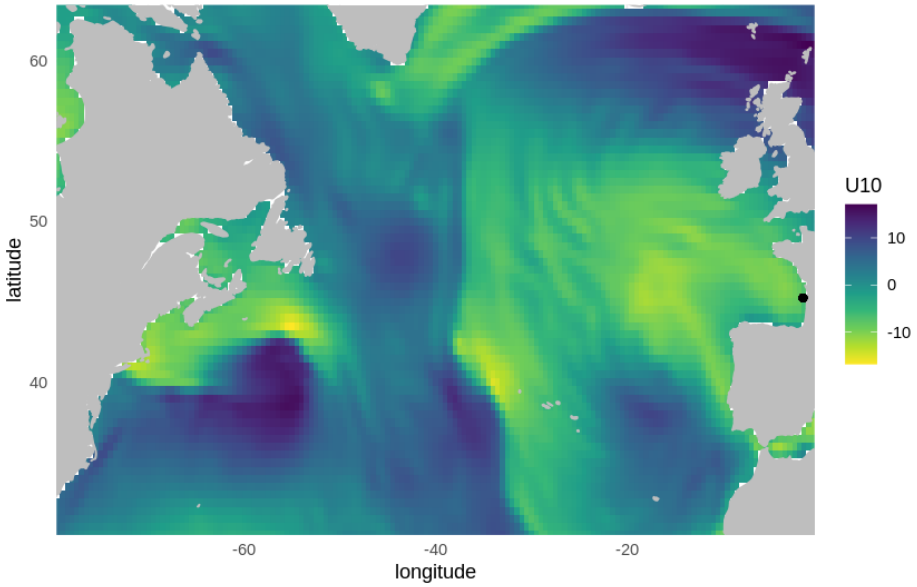

Figure 1. CFSR zonal component in the considered area in 1994-01-01 00h:00. The black point represents the point of interest.

the minimum of the distance from the target point to shore in the direction from which the wind is blowing and $500km$. Lagged wind conditions are considered because they provide information about the temporal variability of the wind and, thus, the duration of wind conditions.

To investigate the capability of local variables to explain H_s, the polynomial regression model

$$H_s(t) = \beta_0^l + X^l(t)\beta^l + \epsilon^l(t) \tag{1}$$

is considered. Where X^l is the local predictor:

$$X^l(t) = \{U(t), U^2(t), U^3(t), U^2(t)F(t), U(t-1), U^2(t-1), U^3(t-1), U^2(t-1)F(t-1)\} \tag{2}$$

β_0^l and β^l are model coefficients, and $\epsilon^l(t)$ is the model error. Model 2 contains polynomial terms and interactions between local variables in order to take into account non-linear relationships between H_s and predictors.

The model is fitted using data from 1994 to 2011 and is assessed in a validation period from 2014 to 2016 using the Pearson correlation r, root mean square error (RMSE), and bias:

$$r = \frac{\sum_{t=1}^{n}(\hat{H}_s(t) - \overline{\hat{H}_s})(H_s(t) - \overline{H_s})}{\sigma_{\hat{H}_s}\sigma_{H_s}} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{H}_s(t) - H_s(t))^2}{n}} \tag{4}$$

$$BIAS = \frac{\sum_{t=1}^{n}(\hat{H}_s(t) - H_s(t))}{n} \tag{5}$$

where $\hat{H}_s(t)$ is the predicted H_s at time t, $\overline{\hat{H}_s}$ and $\overline{H_s}$ are the mean of observed and predicted H_s, respectively; $\sigma_{\hat{H}_s}$ and σ_{H_s} are the standard deviation of predicted and observed H_s, respectively; and n is the number of observations.

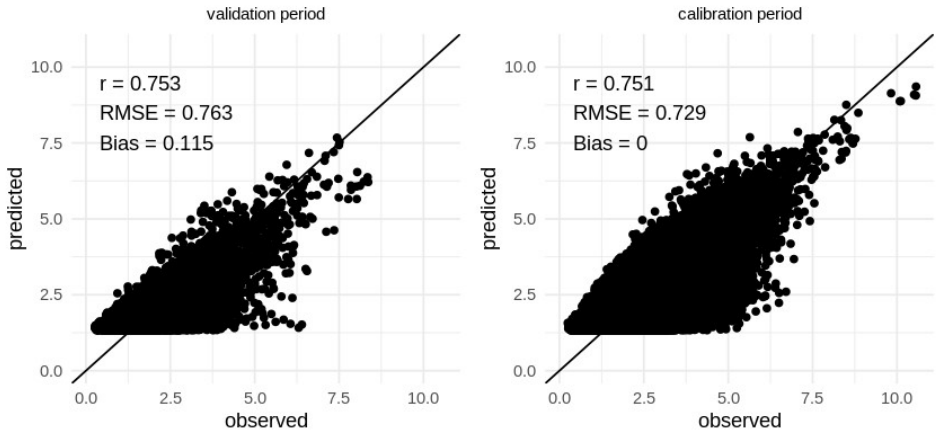

Figure 2. Results of the local model 2 in the validation and calibration period.

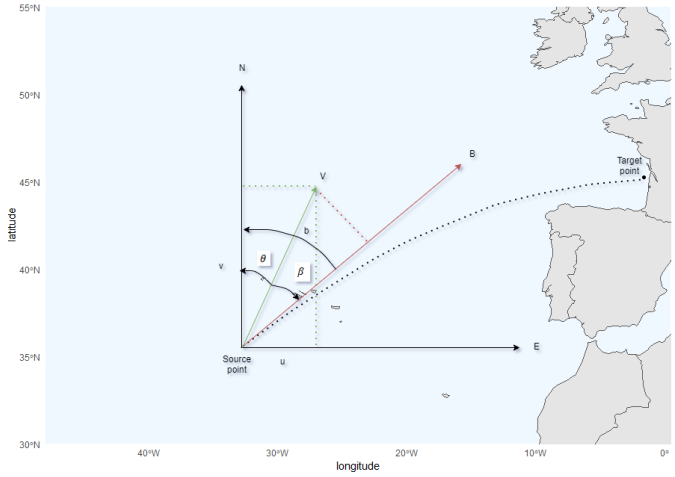

Figure 3. Wind projection representation. The original wind vector V at each source point is projected into the component B defined by the bearing b of the target point from the source point in a great circle path (black dashed line). The great circle is drawn arbitrarily to explain the method and may not be the actual circle path.

Results of the local model 1 are shown in Figure 2. The model poorly predicts small values of H_s, which is expected given that local predictors do not consider swell systems propagated from distant areas. In contrast, the model is better in predicting large values of H_s which can be explained by the fact that extremes are mainly generated by local wind.

4 Global predictor

In order to take swells into account, a global predictor which describes wind condition over the north Atlantic has to be considered. Wind data has two components, the zonal and meridional components. Each of the two components in space and time carries more or less information about the waves observed at the target point at a given date. However, using all of them as inputs to a statistical model is computationally challenging, given the high dimensionality of the data, and may lead to hardly interpretable results due to the strong correlation between wind conditions at closed locations in space

and time. This section defines the global predictor related to the spatio-temporal domain of the wave generation area.

4.1 Spatial coverage

Following Pérez et al. (2014), the spatial coverage of the global predictor is based on the assumption that deep-water waves travel along a great circle path. Therefore, the wave generation area is limited by neglecting grid points whose paths are blocked by land. Furthermore, small islands are not taken into consideration.

4.2 Wind projection

To reduce the dimension of the atmospheric variables and to create a more interpretable model, wind components at each grid point are projected into the bearing of the target point in a great circle path (Figure 3) using the equation:

$$W = U\, cos^{2s}\left(\frac{1}{2}(b-\theta)\right) \tag{6}$$

where W is the projected wind, U is the wind speed, s the spread parameter (Ian R. Young 1999 ("Chapter 5 - Fetch and Duration Limited Growth", 1999)), b the great circle bearing, and θ is the wind direction.

The parameter s controls the amount of wind energy spread in a particular direction; the greater s, the less the wind energy spread is. The spread parameter s should not be too large to avoid losing too much information, especially for grid points near the target point; Hereafter, s is chosen to be equal to 1. Methods to select s for each source point were tested; however, this does not improve numerical results (not shown). Figure 4 illustrates the mean of the projected wind in the four seasons. Strong winds that blow towards the direction to the target point are observed in winter and mostly in the area around 50°N, 40°W.

4.3 Temporal coverage

According to the dispersion relation, the group velocity of waves is expressed as

$$C_g = \frac{gT}{4\pi} \tag{7}$$

where g is the gravitational velocity and T the period. For example, swells whose period is around $15s$ have a group velocity of $11.73m/s$, traveling 50% faster than a $10s$ ocean wave and it takes them about five days to cross the Atlantic from Cape Hatteras to the Bay of Biscay (Ardhuin & Orfila, 2018). Therefore, waves generated at a location j and time t might take time t_j to arrive at the target point.

At each location j and time t, the predictor is defined as the mean of the squared lagged projected wind in a time window, so that

$$X_j^g(t;t_j,\alpha_j) = \frac{1}{2\alpha_j+1}\sum_{i=t-t_j-\alpha_j}^{t-t_j+\alpha_j} W_j^2(i), \quad t_j+\alpha_j+1 \le t \le t_j-\alpha_j+n \tag{8}$$

where α_j controls the length of the time window, t_j is the mean travel time of waves, W_j is the projected wind at location j, and n the total number of observations. Henceforth, the parameter α_j is called the temporal width despite the fact that the length of the temporal wind is equal to $2\alpha_j+1$. Remark that the relationship between the projected wind and H_s seems to be a square relationship (Figure 5) so that in equation (8) the squared projected wind is considered.

The parameters t_j and α_j may be estimated jointly for all locations by minimizing an objective function (least squares, for example); however, such an approach would

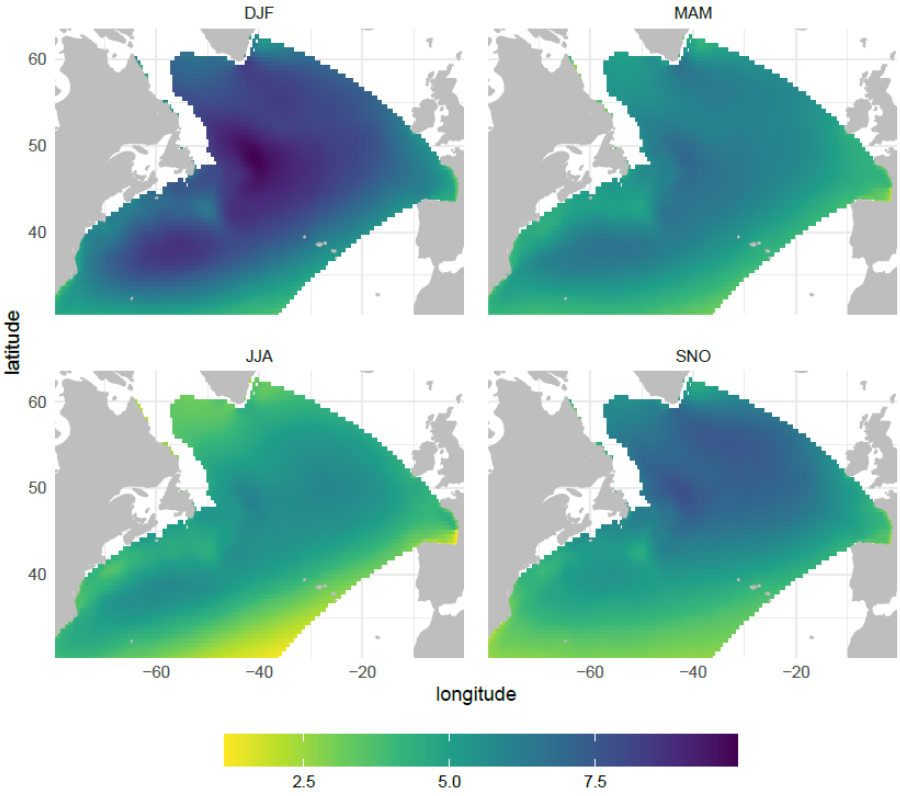

Figure 4. Mean projected wind in the winter, spring, summer, and autumn.

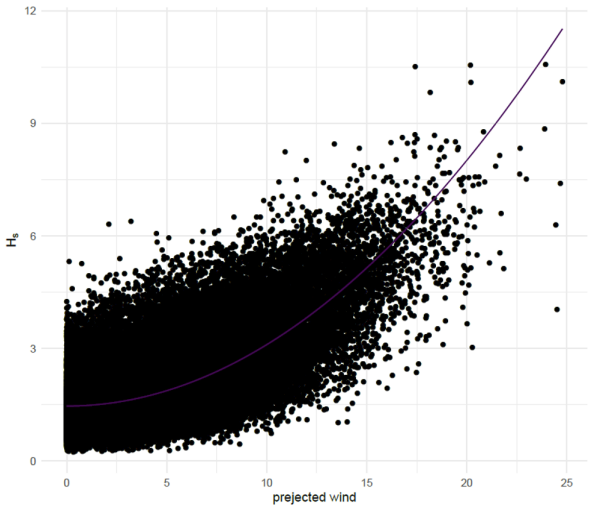

Figure 5. Projected wind at point located in (45.5°N, 3.5°W) versus H_s and the estimated curve line using the model $H_s = aW^2 + b$

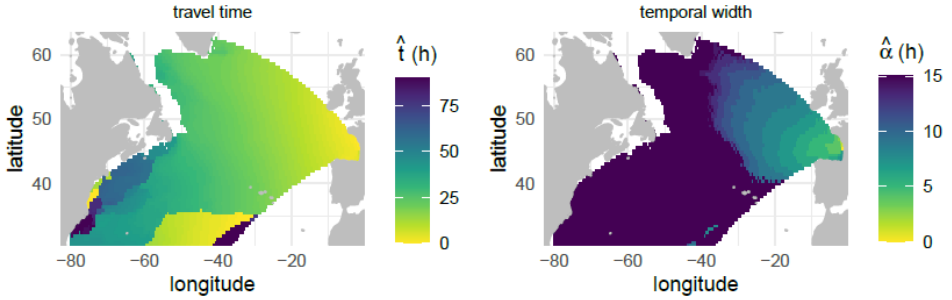

Figure 6. Estimated travel time of waves and the temporal width using equation 9

be non-polynomial and computationally unfeasible due to the combinatorial explosion. Therefore, t_j and α_j are estimated independently for each location using the maximum Pearson correlation between the global predictor and H_s, so that

$$(\hat{t}_j, \hat{\alpha}_j) = arg\max_{t_j,\alpha_j} \big(corr(H_s, X_j^g(t_j, \alpha_j))\big). \tag{9}$$

Figure 6 shows the estimated travel time of waves and the temporal width. Globally, the two parameters are spatially smooth and interpretable. Regions below 35°N seem to have incoherent values of travel time, which may be explained by the fact that waves generated by the wind at these areas have small contributions to the H_s observed at the target location. As expected, the two parameters increase as the distance between the source and target point increases.

Waves generated at a source point situated at (37.5°N, -70.5°W), which is $5642km$ far from the target point, can take on average $180h$ (about 7 and half days) to reach the target point. These waves travel at a velocity of $8.7m/s$; thus, according to the dispersion equation (7), they have an average period of $11.1s$. On one hand, considering $\hat{t}_j + \hat{\alpha}_j$ as the maximum travel time of the waves, at the same source point, waves can also take $225h$ (about 9 days) to reach the target point, with a velocity of $7m/s$ and a period of $9s$. On the other hand, the minimum wave travel time $(\hat{t}_j - \hat{\alpha}_j)$ at the same point is $135h$ (about 5 and a half days) with a velocity of $11.6m/s$ and a period of $14.8s$. Therefore, $t_j - \alpha_j$ and $t_j + \alpha_j$ can be interpreted as the propagation time of long-period waves and short-period waves, respectively.

5 Wind-waves model

5.1 Linear regression model

After defining the predictors, this section presents the statistical downscaling model. Firstly, the linear model that combines the local and the global predictor is considered

$$H_s(t) = X^l(t)\beta^l + X^g(t)\beta^g + \epsilon(t) \tag{10}$$

where β^l and β^g are local coefficients and global coefficients, respectively. Here β^l is not necessarily the same as in equation (1). X_t^l is the local predictor defined in equation (2), X_t^g the global predictor defined in equation (8), and $\epsilon(t)$ is the model error.

5.2 Model fitting

Model (10) can be fitted using least squares method; given by

$$(\hat{\beta}^{ls}) = (X^T X)^{-1} X^T H_s \tag{11}$$

where $X = (X^l, X^g)$ and $\hat{\beta} = (\hat{\beta}^l, \hat{\beta}^g)$.

The least-squares estimates in equation (11) are the best linear unbiased estimates of the parameters. However, since the global predictor is high dimensional (a 67108×5651 matrix), and its variables are highly correlated, the matrix $X^T X$ may be ill-conditioned. Thus, the least-squares estimates become highly sensitive to H_s variations. To address this issue, ridge regression (Hoerl & Kennard, 1970) minimizes the penalized residual sum of squares

$$arg\min_{\beta} \left\| X^l\beta^l + X^g\beta^g - H_s \right\|^2 + \lambda \|\beta^g\|^2 \tag{12}$$

where $\lambda \geq 0$ is the regularization parameter. Remark that the regularization is not applied to the parameters associated to the local predictor. The parameter λ allows to take into consideration the bias-variance trade-off. It can also be viewed as a smoothing parameter, meaning that the greater λ is, the smoother β^g is. A sufficiently smooth β^g may be more interpretable and can help, for example, to identify the source of energy and the contribution of each source point to H_s. However, choosing too large values of λ reduces the prediction performance of the model.

5.3 Regression-guided clustering

Using the global predictor to construct weather types leads to clusters that only account for the global atmospheric circulation and not for the local environment (not shown). This subsection describes a regression-guided clustering method that considers both the global predictor and the predictand.

After estimating the coefficients, the contribution of a source point j at time t to H_s at the target point, is defined as $X_j^g(t)\hat{\beta}_j^g$. The matrix of contributions X_{β^g} is defined as

$$X_{\beta^g}(t, j) = X_j^g(t)\hat{\beta}_j^g. \tag{13}$$

We expect swell systems coming from contributions from distant areas whereas, wind sea will be associated to local contributions. A natural question that arises is whether we can identify these wave systems by using X_{β^g}. Subsequently, the k-means clustering algorithm is used on $X_{\hat{\beta}^g}$ to obtain the weather types (WTs). Finally, the link function can be constructed by fitting the linear regression model (10) at each class. Therefore, Model (10) now becomes

$$H_s(t) = X^l(t)\beta_i^l + X^g(t)\beta_i^g + \epsilon_i(t), \quad \forall t \in I_i \quad i = 1, ..., K \tag{14}$$

where β_i^l and β_i^g are local and global coefficients for the class i. I_i is all time indices that are in class i and K is the total number of WTs.

5.4 The case of two weather types

The hyper-parameters of the model (14) are λ , the number of WTs K, and also the K regularization parameters λ_ks associated to the different weather types (given that, at each weather type a ridge regression is fitted). Given the number of hyper-parameters, it is not computationally feasible to explore all the possible combinations and optimize them simultaneously using for example cross-validation as it usually done in the statistical literature. Instead, we propose the simpler approach described below. At first, we select λ considering only two WTs, then the number of WT for this fixed value of λ, finally λ_ks are fixed for all weather types.

The most usual approach to choose the regularization parameter λ of the ridge regression consists in performing cross-validation and take the value of λ which minimizes a prediction error, typically the RMSE. In the current work, in addition to forecast accuracy, we also intend to obtain a physically interpretable model. Interpretabilty will be

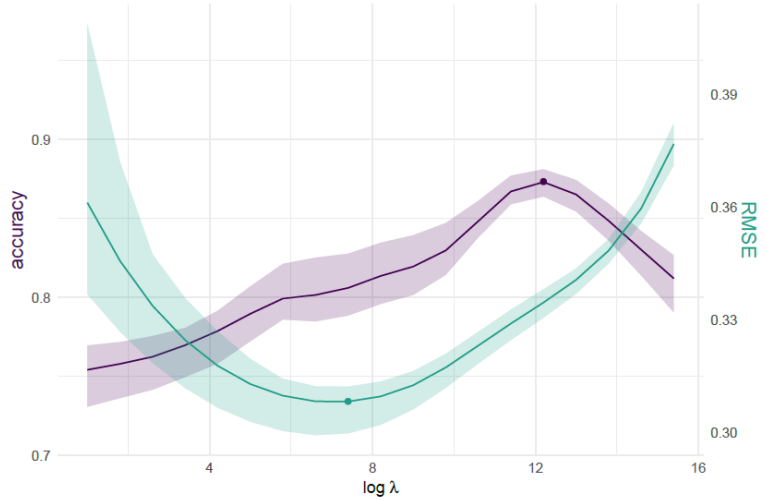

Figure 7. Results of cross-validation: RMSE (green line) and classification accuracy (purple line) versus the logarithm of λ. The red and blue dots correspond to the minimum of RMSE and maximum of accuracy, respectively. The interval for each criterion is defined as the its minimum and maximum.

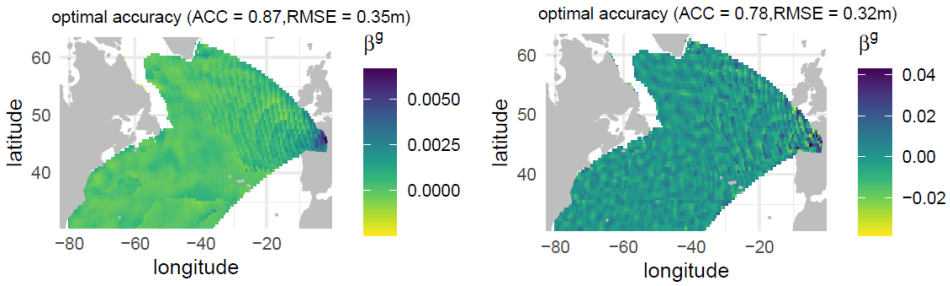

Figure 8. Estimated global coefficients β^g using ridge regression with λ that gives the maximum accuracy (left panel) and minimum RMSE (right panel).

quantified as follows. First, the k-means clustering algorithm is used on the contributions X_{β^g} to identify the leading two clusters. The resulting clusters are then compared with the sea state classification obtained using the energy spectrum partitioning in Homere. The sea states chosen for the comparison are wind sea and swell and the agreement between the two clustering is measured using the classification accuracy

$$\text{accuracy} = \text{correct predictions}/\ \text{sample size} \tag{15}$$

Figure 7 shows that the value of λ that gives the optimal classification accuracy is greater than the one that gives the optimal RMSE. Figure 8 shows the estimated global coefficients β^g using the two different optimal values of the regularization parameter λ. The coefficients obtained using λ that gives the maximum classification accuracy are smoother than the ones obtained when minimizing the RMSE and generally decrease as the distance between the source and target points increases. The optimal λ based on classification is chosen in this study, given that it gives interpretable coefficients, and considering that RMSE does not increase a lot when using λ that gives the maximum accuracy (0.31m to 0.33m).

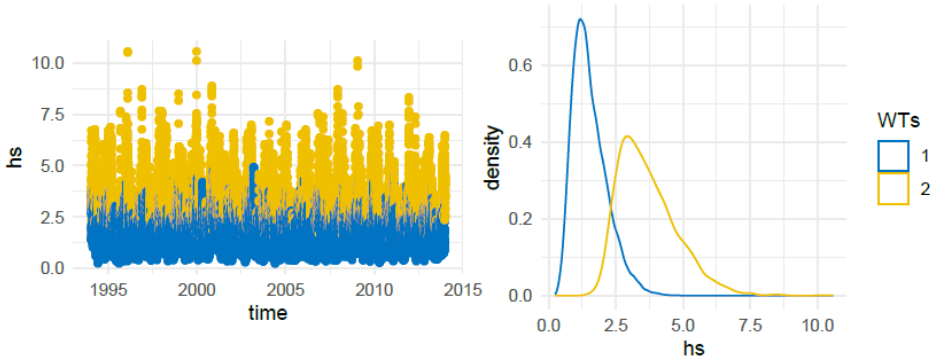

Figure 9. Time series of H_s depending on the clusters (left panel) and empirical density (right panel) in the calibration period.

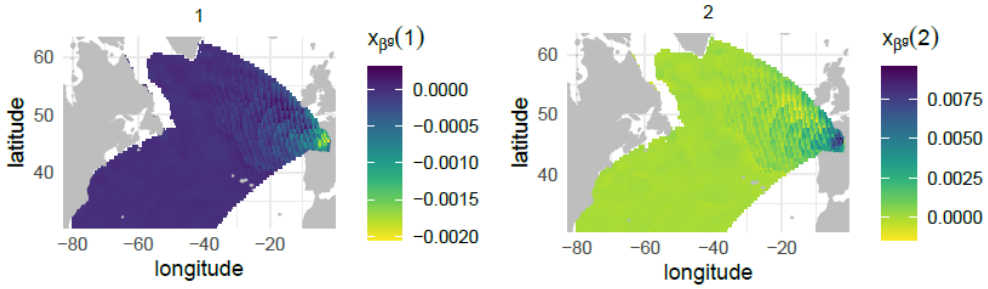

Figure 10. Mean of $X_{\beta g}$ minus the global mean for the cluster 1 (left panel) and cluster 2 (right panel).

classes	1	2
swell	47074	6388
wind sea	974	3904

Table 1. Contingency table of k-means clusters (1 and 2) and Homere sea states classes (swell and sea state) in the calibration period.

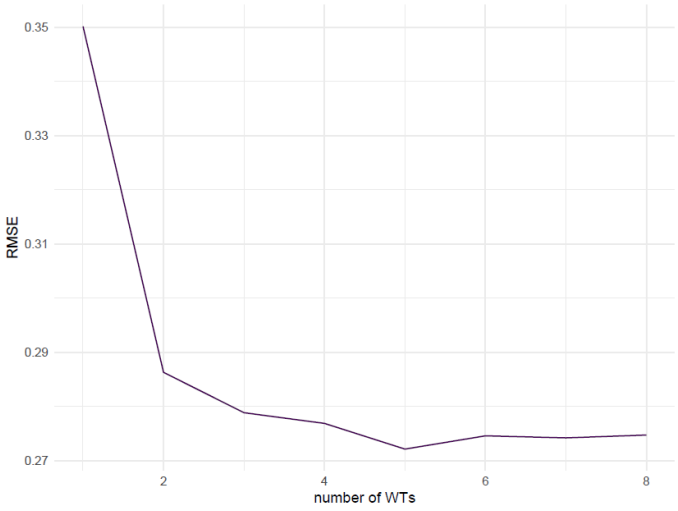

Figure 11. RMSE versus the number of WTs for the validation period.

Figure 9 shows times series of H_s and the corresponding empirical density with respect to the clusters in the calibration period. The most probable cluster is the first one (82%) which corresponds mostly to swells and the second cluster corresponds to wind seas (Table 1). To understand the difference between the two clusters, we define the anomaly of $X_{\beta g}$ in each cluster 1 and 2 as $x_{\beta g}(1)$ and $x_{\beta g}(2)$, respectively

$$\begin{aligned} x_{\beta g}(1) &= \bar{X}_{\beta g}(1) - \bar{X}_{\beta g} \\ x_{\beta g}(2) &= \bar{X}_{\beta g}(2) - \bar{X}_{\beta g} \end{aligned} \tag{16}$$

where $\bar{X}_{\beta g}(1)$ and $\bar{X}_{\beta g}(2)$ are the mean of $X_{\beta g}$ at cluster 1 and 2, respectively and $\bar{X}_{\beta g}$ is the global mean of $X_{\beta g}$. For the first cluster, the local wind around the target point contributes less than the global mean in H_s (Figure 10) and grid points that are far contribute more, as expected when swell systems are dominating. In contrast, in the second cluster, generally associated to wind sea, local wind contributes more than the global mean in H_s.

6 Results

The clusters obtained in the last section seem to be interpretable and correspond to sea state classes of Homere (accuracy = 0.87). However, the number of sea states K may be greater than 2; therefore, a validation analysis is done to select the optimal number of WTs. To do that, for each number of WTs (from 1 to 8), model (14) is fitted using the calibration period and evaluated using the validation period. Figure 11 illustrates the RMSE of H_s as a function of the number of WTs. The optimal number of WTs is 5, and the RMSE seems to decrease significantly from 1 to 5 WTs.

Figure 12 shows the time series of H_s and its empirical density as a function of the five WTs. The resulting WTs depend on the value of H_s; for example, the first WT corresponds to small values of H_s, and the fifth one corresponds to extremes. The other clusters (2 to 4) correspond to intermediate values H_s, in increasing order. The bottom right panel of Figure 12 shows the frequency of occurrence of WTs. The first WT is the most likely, and the fifth one has the smallest probability of occurrence. The transition matrix in the bottom left panel shows that the self-transition probabilities are greater than 0.9 for all WTs, meaning that the WTs are consistent in time. Remark that some transition probabilities are precisely zero; for example, the transition probabilities from the

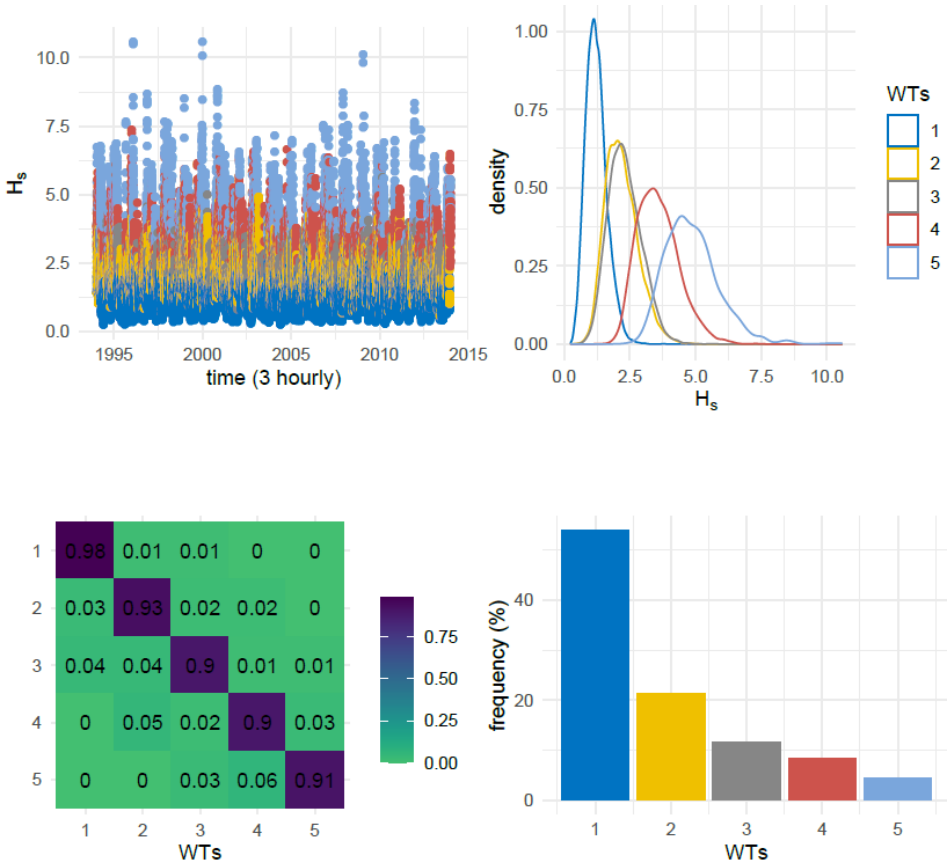

Figure 12. Top left panel: time series of H_s as a function of WTs. Top right: empirical density of H_s as a function of WTs. Bottom left: transition matrix of WTs. Bottom right: Frequency of occurrence of WTs. All figures correspond to the calibration period.

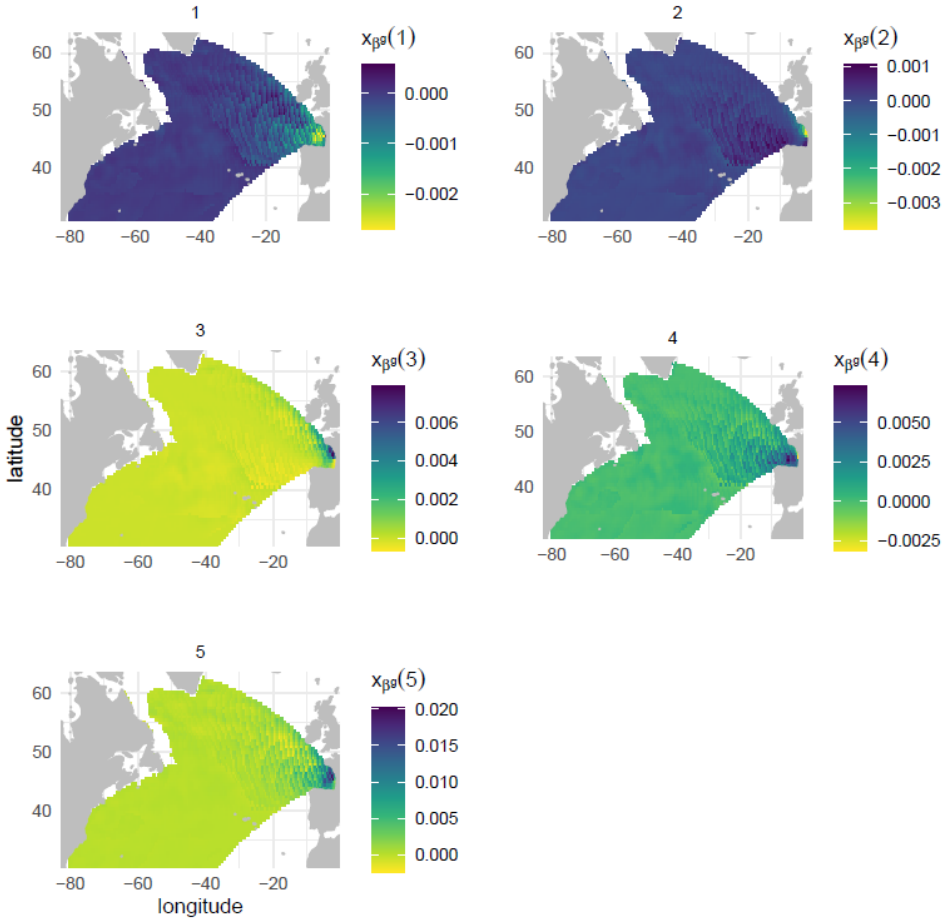

Figure 13. Mean of $X_{\beta g}$ minus the global mean for the five WTs.

1st to the 4th and the 5th WT are equal to zero. This means that the probability to be in extreme sea states after being in the first WT is zero.

Figure 13 shows the mean of $X_{\beta g}$ at each WT where

$$x_{\beta g}(i) = \bar{X}_{\beta g}(i) - \bar{X}_{\beta g},\ i = 1, .., 5 \tag{17}$$

where $\bar{X}_{\beta g}(i)$ is the mean of $X_{\beta g}$ at the ith WT and $\bar{X}_{\beta g}$ is the global mean of $X_{\beta g}$. For the 1st and 2nd WT, contributions of source points far from the target points are greater than the global mean. Therefore, these two classes correspond to swells. In the 3th WT, the local wind contributes more, with moderate winds, in the variance of H_s. The 4th one can be considered a composition of wind sea and swells given that local and far sources points contribute to the variance of H_s. Finally, the 5th WT corresponds to the wind sea where the local source points contribute with the highest intensities of winds creating the highest waves.

The monthly variability of WTs is shown in the left panel of figure 14. As expected, the 5th and 4th WTs occur primarily in winter (December-January-February), and the 1st WT, which corresponds mainly to swells, often occurs during summer. The winter long-term variability of frequency of occurrence of WTs is shown in the right panel of figure 14. The continuous black line corresponds to the mean annual winter of NAO index (Barnston & Livezey, 1987) from 1994 to 2016. The horizontal black line indicates when NAO is greater or less than zero. The long-term variability of weather types seems

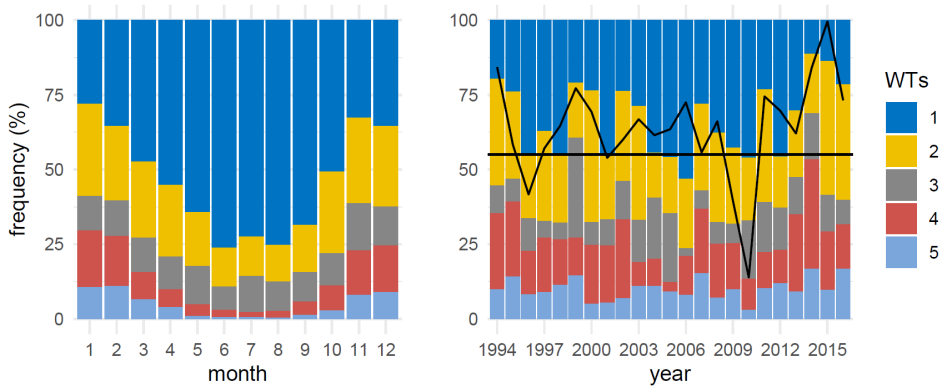

Figure 14. Monthly and annual (in December-january-february) frequency occurrence of WTs in the calibration period. The continuous black line corresponds to the mean annual winter (DJF) time series of NAO (North Atlantic Oscillation) index and the horizontal black line indicates when NAO is less or greater than zero. When the continuous black line is below the horizontal line, the NAO is less than zero.

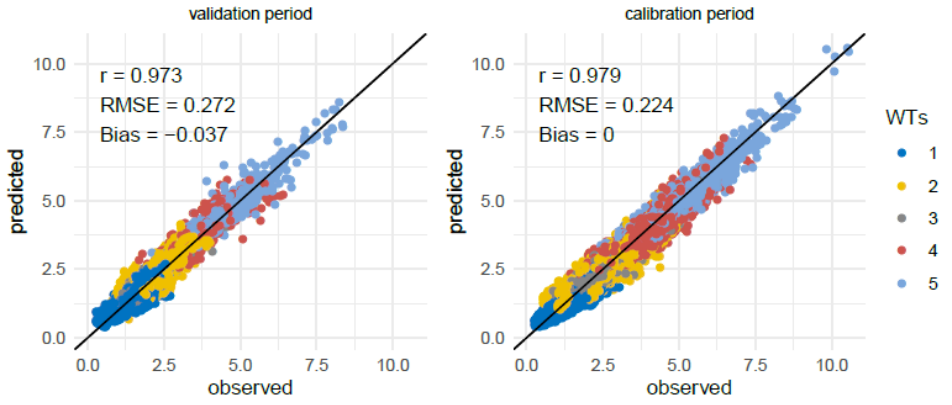

Figure 15. Observed versus predicted values of H_s using the model (14) in the validation and calibration period.

to be related to NAO index. For example, the winter of the year 2010 experienced less extreme waves and the NAO index is less than zero. Whereas, the most extreme sea states are observed in 2014 where the NAO is greater than zero.

Figure 15 and 16 show results of model (14). The model performs well in predicting H_s. The RMSE in the validation period is $0.272m$ for an H_s of mean $1.97m$ and standard deviation of $1.1m$. Comparing these results with those of the local model in Figure 2, it appears that considering the global predictor is important to explain the variability of H_s. Figure 17 illustrates the performance of the downscaling model at each weather type in the validation period. It can be seen that the model in WT 1, 2, and 4 explains less the variability of H_s compared with the model in WT 3 and 5. This can be explained by the fact that in these WTs, the model has to take into consideration sources points that cover the swell generation as seen in Figure 13. In contrast, in WT 4 and 5, the model takes into account mainly local sources points as waves are mainly generated by local wind (Figure 13).

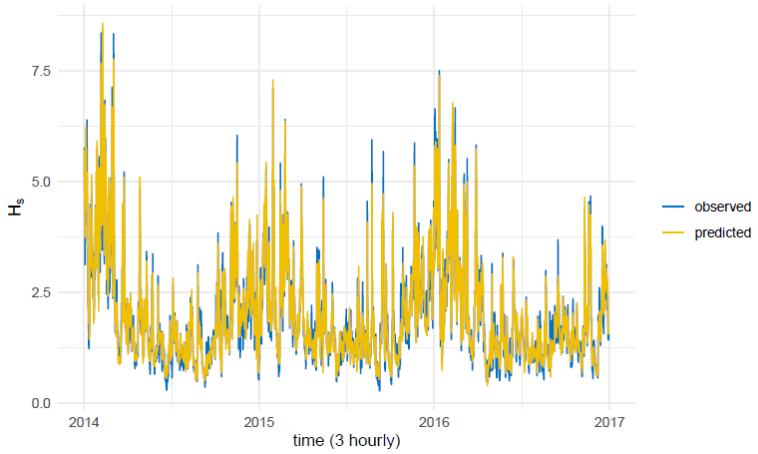

Figure 16. Time series of observed and predicted values of H_s in the validation period.

7 Conclusions

This study proposes a method that describes the spatiotemporal relationship between wind and the significant wave height (H_s). At first, the local model, based on a linear regression between the local wind and H_s, is constructed. However, the model poorly explains the variability of H_s given that the model does not consider the swell generation. Therefore, the global predictor was defined to account for both wind sea and swells. The global predictor is based on the projected wind, which is the wind that goes from source points to the target point in a great circle path. After wind projection, the spatial coverage of the predictor is defined based on the assumption that waves travel along a great circle path. Then its temporal coverage is defined based on two parameters, called the travel time of waves and the temporal width. Both parameters exhibit spatial structure and increase as the distance between the source and target points increases.

The statistical downscaling model combines the local and global predictor to predict H_s using a weather type model. The weather types were constructed using a regression-guided clustering algorithm. The comparison between the Homere sea state classes (wind sea and swell) and two clusters obtained by the clustering algorithm shows a significant resemblance. The predictive model consists of fitting ridge regression between the predictors and the predictand on each WT, and the validation analysis shows that the optimal number of WTs is five. The obtained weather types are interpretable and correspond to different wave systems, and the results of the downscaling model show its skill in predicting H_s. This statistical downscaling method can be extended to other locations. However, for close locations, it will be redundant to define the global predictor and weather types for each location. Therefore, only the local predictor may be adapted to each location.

The methodology presented in this study consists of different steps, from estimating the travel time of waves to finding the weather types. Travel time of waves and weather types can be considered as latent variables and they may be estimated for example using the EM (Expectation-Maximization) algorithm where variables are evaluated based on the prediction of H_s, which can lead to optimal estimations. However, given the complexity of the problem and the high dimensionality of data this solution can be challenging.

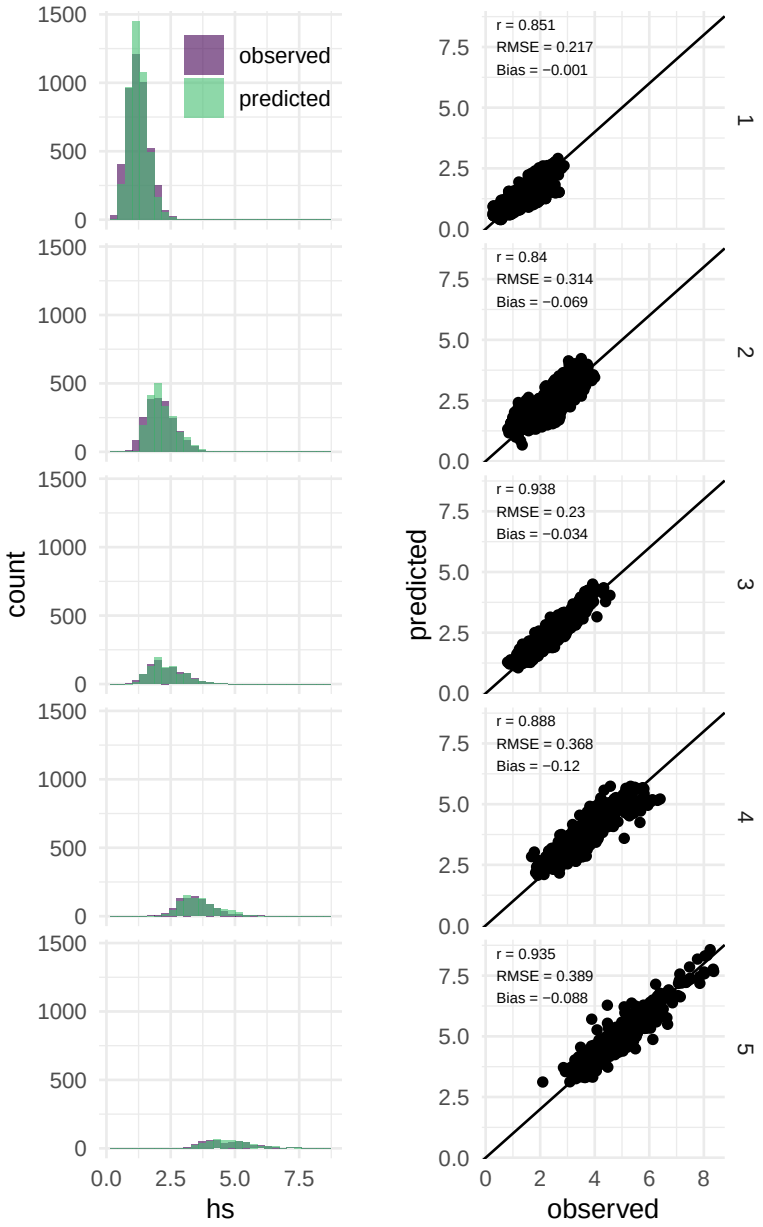

Figure 17. Left panel: histogram of observed versus predicted H_s at each WT. Right panel: scatter plot of observed versus predicted H_s. Both in the validation period.

8 Open Research

The hindcast data Homere is available in their website: `https://marc.ifremer.fr/produits/rejeu_d_etats_de_mer_homere`. The wind data is available from the CFSR website: `https://climatedataguide.ucar.edu/climate-data/climate-forecast-system-reanalysis-cfsr`. Finally, NAO index is obtained from the National Oceanic and Atmospheric Administration website: `https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml`.

The processed data used in this work can be found in: `https://doi.org/10.5281/zenodo.5845423` and the R notebooks are available in: `https://doi.org/10.5281/zenodo.5845250`

References

Ardhuin, F., & Orfila, A. (2018, 08). Wind waves..

Baño-Medina, J., Manzanas, R., & Gutiérrez, J. M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, *13*(4), 2109–2124.

Barnston, A. G., & Livezey, R. E. (1987). Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly weather review*, *115*(6), 1083–1126.

Boudière, E., Maisondieu, C., Ardhuin, F., Accensi, M., Pineau-Guillou, L., & Lepesqueur, J. (2013). A suitable metocean hindcast database for the design of marine energy converters. *International Journal of Marine Energy*, *3*, e40–e52.

Camus, P., Méndez, F. J., Losada, I. J., Menéndez, M., Espejo, A., Pérez, J., ... Guanche, Y. (2014). A method for finding the optimal predictor indices for local wave climate conditions. *Ocean Dynamics*, *64*(7), 1025–1038.

Camus, P., Menéndez, M., Méndez, F. J., Izaguirre, C., Espejo, A., Cánovas, V., ... Medina, R. (2014). A weather-type statistical downscaling framework for ocean wave climate. *Journal of Geophysical Research: Oceans*, *119*(11), 7389–7405.

Cannon, A. J. (2012). Regression-guided clustering: a semisupervised method for circulation-to-environment synoptic classification. *Journal of applied meteorology and climatology*, *51*(2), 185–190.

Casas-Prat, M., Wang, X. L., & Sierra, J. P. (2014). A physical-based statistical method for modeling ocean wave heights. *Ocean Modelling*, *73*, 59–75.

Chapter 5 - fetch and duration limited growth. (1999). In I. R. Young (Ed.), *Wind generated ocean waves* (Vol. 2, p. 83 - 131). Elsevier.

Charles, E., Idier, D., Thiébot, J., Le Cozannet, G., Pedreros, R., Ardhuin, F., & Planton, S. (2012). Present wave climate in the bay of biscay: spatiotemporal variability and trends from 1958 to 2001. *Journal of Climate*, *25*(6), 2020–2039.

Hasselmann, K. F., Barnett, T. P., Bouws, E., Carlson, H., Cartwright, D. E., Eake, K., ... others (1973). Measurements of wind-wave growth and swell decay during the joint north sea wave project (jonswap). *Ergaenzungsheft zur Deutschen Hydrographischen Zeitschrift, Reihe A*.

Hegermiller, C., Antolinez, J. A., Rueda, A., Camus, P., Perez, J., Erikson, L. H., ... Mendez, F. J. (2017). A multimodal wave spectrum–based approach for statistical downscaling of local wave climate. *Journal of Physical Oceanography*, *47*(2), 375–386.

Hemer, M. A., Wang, X. L., Weisse, R., & Swail, V. R. (2012). Advancing wind-waves climate science: The cowclip project. *Bulletin of the American Meteorological Society*, *93*(6), 791–796.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., ... others (2021). Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200093.

Laugel, A., Menendez, M., Benoit, M., Mattarolo, G., & Méndez, F. (2014). Wave climate projections along the french coastline: dynamical versus statistical downscaling methods. *Ocean Modelling*, *84*, 35–50.

Maraun, D., Wetterhall, F., Ireson, A., Chandler, R., Kendon, E., Widmann, M., ... others (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of geophysics*, *48*(3).

Pérez, J., Méndez, F. J., Menéndez, M., & Losada, I. J. (2014). Estela: a method for evaluating the source and travel time of the wave energy reaching a local area. *Ocean Dynamics*, *64*(8), 1181–1191.

Roland, A., & Ardhuin, F. (2014). On the developments of spectral wave models: numerics and parameterizations for the coastal ocean. *Ocean Dynamics*, *64*(6), 833–846.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., ... others (2010). The ncep climate forecast system https://www.overleaf.com/project/616d1ff6df8f1d7e3172abb1reanalysis. *Bulletin of the American Meteorological Society*, *91*(8), 1015–1058.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tolman, H. L., et al. (2009). User manual and system documentation of wavewatch iii tm version 3.14. *Technical note, MMAB Contribution*, *276*, 220.

Tracy, B., Devaliere, E., Hanson, J., Nicolini, T., & Tolman, H. (2007). Wind sea and swell delineation for numerical wave modeling. In *10th international workshop on wave hindcasting and forecasting & coastal hazards symposium, jcomm tech. rep* (Vol. 41, p. 1442).

Wang, X. L., Swail, V. R., & Cox, A. (2010). Dynamical versus statistical downscaling methods for ocean wave heights. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *30*(3), 317–332.