DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Mathématiques*

Par
Said OBAKRIM

Statistical Downscaling and Climate Change in the Coastal Zone

Rapporteurs avant soutenance :

Thomas OPITZ	Chargé de recherches, INRAE, Avignon
Mathieu VRAC	Directeur de recherches CNRS, LSCE, Paris

Composition du Jury :

Président :	Carlo GAETAN	Professeur, Université de Venise
Examinateur·ice·s :	Thomas OPITZ	Chargé de recherches, INRAE, Avignon
	Mathieu VRAC	Directeur de recherches CNRS, LSCE, Paris
	Gwladys TOULEMONDE	Maître de conférences, Université de Montpellier
Dir. de thèse :	Valérie MONBET	Professeur des universités, Université Rennes 1
Co-dir. de thèse :	Pierre AILLIOT	Maître de conférences, Université de Bretagne Occidentale

Acknowledgement

Three years have passed working on this thesis. Three years rich in learning, good experiences and memories. Through the guidance of my supervisors, Nicolas, Pierre, and Valérie, this long journey has been a success. It could only have been achieved with their help and encouragement. A heartfelt thanks to you for your availability whenever I had questions or needed to discuss. I appreciate all your valuable suggestions and constructive feedbacks. Throughout my thesis, you provided me with helpful advice, and I deeply thank you for that.

I am honored that Carlo Gaetan, Mathieu Vrac, Gwladys Toulemonde, and Thomas Opitz reviewed my thesis and judged my presentation. Your remarks, comments, and compliments are appreciated.

A good thesis is only possible with a good atmosphere in the workspace animated by Agathe, Camil, Florian, Mathieu, Nidiana, Sacha, and Simon. Thanks to all of you for maintaining such a great workspace.

Because doing a Ph.D. requires the support of friends, I would like to thank all the friends with whom I spent even a little time during this thesis. Thanks to Jeff, Khalil, Othman, Khaoula, Ayoub, and Hicham for the fun times. Thanks to the climbing team, Florian, Romain, and Simon, for the incredible climbing sessions. Thanks to the ASCI animators Camil, Audrey, Lena, Kevin, and all the rest of the club, for the climbing initiation courses and the outdoor climbing sessions. Thank you all for making my stay in Breizh pleasant and memorable.

Finally, I express my gratitude to my family: my dad and mom, my brothers Ali, Omar, and Yassine, and my sister Fatima. Thank you for supporting me from my first steps and checking on me throughout my thesis despite the distance.

Table of Contents

Résumé

Il existe une forte demande de données de haute qualité sur les vagues océaniques pour plusieurs applications marines (Bitner-Gregersen et al., 2016; Ardhuin et al., 2019). Par exemple, les ingénieurs ont besoin de séries temporelles à long terme de paramètres de vagues, tels que la hauteur significative des vagues (H_s), pour estimer l'occurrence des événements extrêmes, caractériser la climatologie des sites pour les convertisseurs d'énergie marine, et concevoir des structures côtières et offshore ou planifier des opérations maritimes (Kerbiriou et al., 2007). La caractérisation de l'état de la mer est donc nécessaire pour ces nombreuses applications qui nécessitent des séries temporelles étendues avec une résolution spatiale à l'échelle du kilomètre.

Nous distinguons trois méthodes de caractérisation de l'état de la mer: les méthodes d'observation, les modèles numériques et les modèles statistiques. Bien que les méthodes d'observation de l'état de la mer (telles que les mesures in situ) fournissent des données fiables, elles ne fournissent pas une connaissance complète de l'état de la mer dans l'espace et dans le temps (Ardhuin et al., 2019). Les modèles numériques de vagues (Gelci, Cazalé, and Vassal, 1957; Remya et al., 2022; Hemer, Katzfey, and Trenham, 2013) constituent donc une source alternative de données sur les vagues qui permet d'étudier les vagues avec une haute résolution spatiale et temporelle (Boudière et al., 2013). Cependant, comme les modèles numériques nécessitent des calculs intensifs (Laugel, 2013), les méthodes statistiques sont de plus en plus populaires dans la communauté des climatologues en géneral (Wilby et al. (1998); Benestad, Chen, and Hanssen-Bauer (2008); Maraun et al. (2010); Scher (2018); Sungkawa, Rahayu, et al. (2019)) et des océanographes en particulier (Wang and Swail (2006); Wang, Swail, and Cox (2010), Laugel (2013), Camus et al. (2014b)).

Dans cette thèse, nous nous intéressons à la caractérisation des paramètres d'état de mer tel que la hauteur significative des vagues (H_s) en utilisant des méthodes statistiques et d'apprentissage profond. En particulier, nous nous intéressons à la modélisation de la relation entre les conditions du vent de l'Atlantique Nord et les paramètres d'état de la mer à un endroit situé dans le Golfe de Gascogne. Étant donné la multidimensionnalité des données de vent et la relation décalée en temps entre les conditions de vent et les vagues, nous proposons d'abord un cadre général pour sélectionner les covariables pertinentes qui

influencent la hauteur significative des vagues.

Après l'étape de prétraitement, un modèle de régression basé sur les types de temps est proposé pour modéliser la relation entre le vent et les vagues. Les types de temps sont construits à l'aide d'un algorithme de classification puis, pour chaque type de temps, une régression de Ridge est ajustée entre les conditions de vent et la hauteur significative des vagues. Le modèle prédit bien H_s, mais il présente certaines limites, à savoir : (i) la régression de Ridge ne tient pas compte du fait que les covariables ont une structure spatiale ; et (ii) les types de temps sont construits a priori à l'aide d'un algorithme de classification et ils ne sont pas évalués en fonction de la prédiction de H_s. Par conséquent, nous proposons un algorithme d'espérance-maximisation (EM) pour estimer les paramètres de la régression de Ridge généralisée avec des covariables spatiales, puis, pour tenir compte les points (i) et (ii), nous proposons un mélange d'experts de Ridge généralisés estimés à l'aide d'un algorithme EM variationnel. Ce modèle est utilisé comme modèle de régression basé sur les types de temps et ses performances sont supérieures à celles du modèle original.

Les contributions principales de cette thèse sont :

- Proposer un cadre général pour sélectionner les covariables pertinentes qui influencent les états de mer:

  Étant donné que les vagues océaniques sont une combinaison de mer du vent générées localement et de houles générées et propagées à partir des zones éloignées (Ardhuin and Orfila, 2018), les vagues observées à un endroit particulier dépendent des conditions du vent sur une large zone dans une fenêtre temporelle de plusieurs jours (Camus et al., 2014a). Par conséquent, la reproduction de la relation spatio-temporelle entre le vent et les vagues à l'aide de méthodes statistiques et d'apprentissage automatique n'est pas simple (Camus et al., 2014a). Dans cette thèse, nous proposons des étapes de prétraitement pour identifier les covariables pertinentes pour la prédiction des vagues. En particulier, nous proposons deux types de prédicteurs pour les états de mer: un prédicteur local et global. Le prédicteur local est basé sur le vent dans le point d'intéret et le prédicteur global sur les conditions du vents dans l'Atlantique Nord. Les conditions du vent sont caractérisées par deux composantes (zonale et méridionale), qui peuvent être difficiles à prendre en compte directement dans un modèle statistique. Pour résoudre le problème de la multidimensionnalité dans cette étude, nous introduisons la projection

du vent, qui consiste à ne retenir que la fraction du vent soufflant vers le point cible. L'étape de prétraitement proposée permet d'utiliser une seule variable pour chaque point de grille, réduisant ainsi de moitié la dimension du prédicteur global. En outre, Le vent provenant de régions éloignées génère des vagues qui peuvent mettre des jours pour atteindre le point cible. Ainsi, la relation entre le vent et les vagues n'est pas instantanée. Il est donc nécessaire de prendre en compte les conditions du vent décalées pour comprendre la dynamique des vagues à un endroit cible particulier (Pérez et al., 2014). La présente étude utilise une approche entièrement basée sur les données (data-driven) pour définir la zone de génération des vagues. Elle est basée sur l'estimation du temps de voyage des vagues entre chaque source et le point cible en utilisant la corrélation maximale entre la hauteur significative des vagues et les conditions du vent (le vent projeté).

- Étude des méthodes de régularisation adaptées aux problèmes de régression avec covariables spatiales:

  Les données climatiques sont connues pour être de haute dimentionnalité, due au petit nombre d'observations et au grand nombre de covariables, ce qui augmente considérablement le risque de surajustement dans les modèles de regression. De plus, il existe de fortes dépendances spatiales entre les covariables, ce qui crée un problème de multicollinéarité. Par conséquent, les modèles statistiques doivent tenir compte de ces aspects afin d'améliorer la qualité des prédictions et l'interprétation physique du modèle. Les méthodes de régularisation sont largement étudiées dans la littérature et se sont révélées efficaces à cette fin (Hastie et al. (2009)) . La pénalité de Ridge généralisée est un outil puissant pour traiter la multicollinéarité et la haute dimensionnalité dans les problèmes de régression. En outre, la pénalité de Ridge généralisée permet de mettre n'importe quelle structure de covariance sur les coefficients de régression (Wieringen (2015)), ce qui peut être avantageux dans les applications spatiales et climatiques en particulier. Dans cette thèse, nous présentons l'estimateur de Ridge généralisé comme un estimateur a posteriori d'un modèle de variable latente dont les paramètres sont estimés avec l'algorithme Expectation-Maximisation (EM) (Bishop and Nasrabadi (2006)). Une étude de simulation est menée pour évaluer la performance du modèle puis il est appliqué pour estimer la fonction de transfert entre les conditions du vent sur l'Atlantique nord et les vagues dans le Golfe de Gascogne.

- Exploiter les avantages des modèles de mélange dans la modélisation des données hétérogènes pour créer un modèle de régression basé sur les types de temps pour les vagues océaniques:

  La classification du temps en différents systèmes au-delà de la classification printemps, été, automne et hiver s'est avérée avantageuse pour la modélisation des variables climatiques (Camus et al. (2014b), Yarnal and Frakes (1997), Peña-Angulo et al. (2016)). Dans cette thèse, nous allons étudier l'utilisation des modèles de mélange pour les tâches de régression basées sur les types de temps. En particulier, nous introduisons le mélange d'experts de régression pénalisé par Ridge généralisé et un algorithme pour estimer ses paramètres. Nous avons montré que l'utilisation de l'algorithme EM est problématique étant donné que la distribution postérieure de l'étape E n'a pas de forme analytique; nous avons donc proposé une approximation variationnelle (El Assaad et al., 2016) de l'étape E. Une étude de simulation est réalisée pour évaluer la performance du modèle et la méthode est ensuite utilisée comme méthode de régression basée sur les types de temps pour le downscaling de la hauteur significative des vagues.

- Étude de l'utilisation de modèles d'apprentissage profond pour la modélisation de la relation entre le vent et les vagues océaniques:

  Les modèles d'apprentissage profond gagnent en popularité dans la communauté climatique, en raison de leur capacité à construire des représentations hiérarchiques des covariables (Goodfellow, Bengio, and Courville, 2016) et en particulier, les réseaux de neurones convolutionnels (CNN) permettent d'apprendre des caractéristiques spatiales complexes à partir de données spatiales (Gu et al., 2018). Les modèles d'apprentissage profond ont été utilisées dans de nombreuses études pour le downscaling des variables climatiques telles que les précipitations et la température. À notre connaissance, ils n'ont pas encore été utilisées pour le downscaling de l'état de la mer ; par conséquent, dans Michel et al. (2022), nous avons développé un modèle de downscaling pour les paramètres de l'état de la mer en utilisant un modèle de réseau neuronal convolutif. Pour l'instant, les méthodes développées dans cette thèse sont basées sur les prédicteurs définis à l'aide d'une étape de prétraitement, basée sur l'estimation de temps de voyage optimal des vagues en utilisant la corrélation maximale entre H_s et les conditions de vent. Dans cette thèse, nous utilisons la capacité des CNN à extraire des caractéristiques spatiales

et des modèles Long short-term memory (LSTM) à apprendre des dépendances temporelles à long terme afin de construire la fonction de liaison entre H_s et les conditions de vent sans utiliser d'étape de prétraitement.

Abstract

Ocean wave climate has a significant impact on human activities, and its understanding is socioeconomically and environmentally important. In this thesis, we are interested in characterizing sea state parameters such as significant wave height (H_s) using statistical and deep learning methods. In particular, we are interested in modeling the relationship between North Atlantic wind conditions and sea state parameters at a location in the Bay of Biscay. Given the multidimensionality of the wind data and the time-lagged relationship between wind conditions and waves, we first propose a general framework to select the relevant covariates that influence the significant wave height.

After the preprocessing step, a regression model based on weather types is proposed to model the relationship between wind and waves. The weather types are constructed using a clustering algorithm, and then, for each weather type, a Ridge regression is fitted between the wind conditions and the significant wave height. The model predicts H_s well; however, it has some limitations, namely: (i) Ridge regression does not take into account that the covariates have a spatial structure; and (ii) the weather types are constructed a priori using a clustering algorithm, and they are not evaluated based on the prediction of H_s. Therefore, we propose an expectation-maximization (EM) algorithm to estimate the parameters of the generalized Ridge regression with spatial covariates. Then, to account for (i) and (ii), we propose a mixture of generalized Ridge experts estimated using a variational EM algorithm. This model is used as a weather-types-based regression model, and its performance is better than that of the original model.

Finally, the last part of this thesis is devoted to developing deep learning methods for sea state parameters prediction.

Keywords: Downscaling, Sea state, Generalized Ridge, Mixture of experts, EM algorithm, Deep learning

Introduction

Problem statement

There is a strong demand for high-quality ocean wave data for several marine applications (Bitner-Gregersen et al., 2016). For example, engineers need long-term time series of wave parameters, such as significant wave height (H_s), to estimate the occurrence of extreme events, characterize site climatology for marine energy converters, and design coastal and offshore structures or plan marine operations (Kerbiriou et al., 2007). Wave observation methods, numerical models, and statistical methods are three different approaches to this end. Although observations (such as in-situ measurements) provide reliable data, they are limited in space and time (Ardhuin et al., 2019). In contrast, numerical models provide decades of wave data that can cover the entire globe. However, because numerical models are computationally intensive, statistical and data-driven methods are becoming increasingly popular in the climate and meteorology community (Wang and Swail (2006), Laugel (2013), Camus et al. (2014b)).

This thesis aims to develop statistical and machine learning methods adapted to modeling the relationship between wind and ocean wave parameters. In particular, we aim to construct a link function between the wind conditions over the North Atlantic and the significant wave height in the Bay of Biscay.

The main orientations of this thesis are:

- Providing a framework for identifying relevant wind-based covariates for wave parameter prediction.
  Since ocean waves are a combination of locally generated wind waves and swells generated and propagated from distant areas (Ardhuin and Orfila, 2018), the waves

observed at a particular location depend on wind conditions over a large area within a time window of several days (Camus et al., 2014a). Therefore, reproducing the spatio-temporal relationship between wind and waves using statistical and machine learning methods is not straightforward. In this thesis, we propose preprocessing steps for identifying the relevant covariates for wave prediction.

- Studying regularization methods adapted to regression problems with spatial covariates.
  Climate data are known to be multicollinear and high-dimensional, which requires treatment in regression problems. Regularization methods are widely studied in the literature and have been shown to be effective for this purpose (Hastie et al., 2009); but do we need regularization methods suitable for regression in spatial applications?
- Leveraging the benefits of mixture models in heterogeneous data modeling to create a weather-types-based regression model for ocean waves.
  Classifying the weather into different systems beyond the spring, summer, fall, and winter classification has been proven to be advantageous in modeling climate variables (Camus et al. (2014b), Yarnal and Frakes (1997), Peña-Angulo et al. (2016)). In this thesis, we will investigate the use of mixture models for weather-types-based regression tasks.
- Investigating the use of deep learning models for modeling the relationship between wind and ocean waves.
  Deep learning models are gaining popularity in the climate community, given their ability in building hierarchical representations of covariates (Goodfellow, Bengio, and Courville, 2016). In particular, convolutional neural networks (CNNs) allow for learning complex spatial features from spatial data (Gu et al., 2018). In this thesis, we will investigate the use of deep learning methods to predict ocean wave parameters.

Manuscript organization

This thesis is organized as follows:

Chapter 1: The opening chapter is a preliminary introduction to the problem of downscaling and the statistical methods used for this purpose. This chapter also presents

statistical and machine learning methods that are necessary to understand this thesis, such as penalized linear regression, neural networks, weather types, and the Expectation-Maximization (EM) algorithm.

Chapter 2: This chapter recalls the concept of sea state and present the methods used for the characterization of sea states; namely: observation methods, numerical models, and statistical models. Then the wave data used in this thesis is presented.

Chapter 3: In this chapter, we present the framework that will be used in this thesis to identify relevant wind-based covariates for wave parameter prediction. Then, we present a statistical model that links wind conditions over the North Atlantic and waves at a location in the Bay of Biscay. This work contributes to the understanding of the complex relationship between wind and waves using a data-driven approach that can be used for weather and climate studies.

Chapter 4: In this chapter, we presents an expectation-maximization (EM) algorithm for estimating generalized Ridge regression parameters when the covariates have a spatial structure. A simulation study is conducted to assess the performance of the model then it is applied to estimate the transfer function between wind conditions over the north Atlantic and waves at the Bay of Biscay.

Chapter 5: This chapter introduces the mixture of generalized Ridge experts and a variational EM algorithm for estimating its parameters. A simulation study is done to assess the model's performance and the method is then used as a weather types-based regression method for downscaling the significant wave height.

Chapter 6: The developed methodology in the first chapter uses a preprocessing step that defines the predictors of the statistical downscaling model. In this chapter, we propose a deep learning approach that automatically extracts these features from data without a preprocessing step.

Chapter 7: Finally, this chapter summarizes the contributions of this thesis and presents some future research perspective.

Contributions

The articles related to this thesis are listed below:

- Chapter 3 is based on the article *Statistical modeling of the space-time relation between wind and significant wave height*, S.Obakrim, P.Ailliot, V.Monbet, N.Raillard,

2022, https://www.essoar.org/pdfjs/10.1002/essoar.10510147.2

- Chapter 4 is based on the article *EM algorithm for generalized Ridge regression with spatial covariates*, S.Obakrim, P.Ailliot, V.Monbet, N.Raillard, 2022, https://doi.org/10.48550/arXiv.2208.04754
- Chapter 5 is based on the article *Maximum likelihood estimation of a mixture of generalized Ridge regression*, S.Obakrim, P.Ailliot, V.Monbet, N.Raillard, 2022, doi: https://www.essoar.org/pdfjs/10.1002/essoar.10510147.2
- Chapter 6 is based on the article *Deep learning for statistical downscaling of sea states*, M.Michel S.Obakrim, N.Raillard, P.Ailliot, V.Monbet, 2022, doi: https://ascmo.copernicus.org/articles/8/83/2022/. And the article *Learning the spatio-temporal relationship between wind and significant wave height using deep learning*, S.Obakrim, V.Monbet, N.Raillard, P.Ailliot, https://doi.org/10.48550/arXiv.2205.13325

Chapter 1

Statistical and Probabilistic Tools for Downscaling

Contents

Note: This chapter is a preliminary introduction to the statistical and probabilistic tools used for downscaling. In the first section, the problematic of downscaling and the motivations behind it are presented. Then, in the second section, some statistical downscaling methods based on transfer functions are described. Section 3 discusses statistical downscaling approaches based on weather types.

1.1 Statistical downscaling

Anticipating climate change due to greenhouse gas emissions is crucial for impact assessments and policymakers. General circulation models (GCMs) are the primary tools for identifying these changes and climate projection. The main drawback of these models is their coarse spatial resolution (a grid size of about 100-500 km), which makes them unsuitable for most impact assessment applications that require regional and/or local climate projections. For example, hydrological models often require meteorological variables with a resolution of less than 10 km (Boé et al., 2007), and marine energy converters de-

ployment needs wave data at fine-scale (Boudière et al., 2013). To address this problem, downscaling methods derive fine-scale resolution time series of climate variables needed for the impact study.

Downscaling approaches are classified into two approaches, dynamical and statistical. Dynamical downscaling (Xue et al., 2014) is similar to GCM models but at a much higher resolution. Dynamical downscaling models use the results of GCMs and incorporate detailed descriptions of the physical processes that determine the local area to generate realistic climate information at a much finer resolution. However, despite their accuracy, these models require high computational resources and expertise (Hong and Kanamitsu, 2014). An alternative approach to dynamical downscaling is statistical downscaling (SD), which establishes empirical relationships between large-scale atmospheric and local climate variables.

Many statistical downscaling approaches have been proposed in the literature. (Maraun et al., 2010) has classified these methods into perfect prognosis (PP), model output statistics (MOS), and weather generators (WG), based on the type of predictors chosen rather than the type of statistical model. PP is calibrated using observations at large and local scales, and projections are produced using large-scale predictors simulated by GCM. On the other hand, MOS downscaling methods construct a statistical relationship between GCM outputs of the large-scale variables and the observed local variable. MOS models generally use bias correction methods to correct GCM biases (Teutschbein and Seibert, 2012). Finally, weather generators are statistical methods that simulate time series of climate variables that have a distribution close to the observed variables. Future synthetic simulations can be performed by adjusting the parameters of the weather generator to account for future climate conditions (Keller et al., 2017).

Perfect prognosis methods are statistical models that establish empirical relationships between observed large-scale predictors and observed local-scale predictors. These relationships are often inferred using predictors derived from numerical models when the predictors are realistically simulated, hence the name perfect prognosis. PP models fall into three categories: Regression (Hessami et al., 2008), analog (Zorita and Von Storch, 1999), and weather type-based (Camus et al., 2014b) methods. This study focuses on perfect prognosis methods, especially regression and weather type-based methods.

1.2 Downscaling as a regression problem

Suppose that we observe a sample $y = (y_1, ..., y_n)$ of a local-scale variable Y and a matrix X of global-scale variables of size $n \times d$. In the following, we refer to y and X as the predictand and the predictor, respectively. The problem of statistical downscaling can be viewed as a regression problem of the form

$$Y = f(X) + \epsilon \tag{1.1}$$

where ϵ is the model's error. The function f can be estimated either by linear models like linear regression and generalized additive models or by non-linear models like neural networks and other machine learning algorithms. Besides the statistical method, the quality of the downscaling model depends on the quality and the amount of data used for the estimation and the quality of the predictor X (Camus et al., 2014a). This section will discuss methods used to estimate the transfer function f, like linear regression and neural networks.

1.2.1 Linear regression

Assuming that the relationship between the local-scale variable Y and the global-scale variables is linear, linear regression can be used

$$Y = X\beta + \epsilon \tag{1.2}$$

where β is the vector in $\mathbb{R}^d$ of model parameters and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ is the error term with variance σ^2. The model 4.2 can be fitted by minimizing the least squares loss function, which gives the solution

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{1.3}$$

Least squares estimates are the best linear unbiased estimates of the parameters. However, the variables in X might be highly correlated, some variables may be irrelevant, or the problem is high-dimensional ($d \geq n$), and the least-squares estimates have zero bias and large variance. To address this issue, variable selection and shrinkage methods are used. By using these methods, it is possible to increase the bias to reduce the variance of predictions. This section will focus on shrinkage methods, especially Ridge regression. Readers are invited to see (Hastie et al., 2009) for a review of variable selection methods

used for regression.

Shrinkage or penalization methods impose a penalty on the size of regression coefficients by minimizing a penalized residual sum of squares

$$\hat{\beta} = \arg\min_{\beta}\{\|X\beta - y\|_2^2 + \lambda \text{Penalty}(\beta)\} \tag{1.4}$$

where $\lambda \geq 0$ is a hyper-parameter that controls the amount of penalization. The common choices of penalty are:

- Ridge where $\text{Penalty}(\beta) = \|\beta\|_2$ or other L^2 based penalties such as fused and generalized Ridge (see e.g, Wieringen (2015))
- Lasso where $\text{Penalty}(\beta) = \|\beta\|_1$ or other L^1 based penalties (see e.g, Vidaurre, Bielza, and Larranaga (2013) for a review)
- Best subset or L^0 penalty where $\text{Penalty}(\beta) = \|\beta\|_0$ (Huang et al., 2018).

Unlike Ridge regression, which has an analytical solution, Lasso and best subset methods lack closed solutions given that the penalty is not differentiable. Therefore, numerical methods are used to derive solutions. After taking the derivatives over β and equating to zero, the solution for Ridge regression is

$$\hat{\beta} = (X^T X + \lambda I_d)^{-1} X^T y \tag{1.5}$$

where I_d is a $d \times d$ identity matrix with d is the number of variables. Therefore, Ridge regression adds a constant to the diagonal of $X^T X$ before inversion. This solves the problem when the matrix $X^T X$ is singular, often when the covariates are highly correlated.

Large-scale climate variables and GCM outputs are multidimensional and multicollinear, and using them as a predictor in a linear regression statistical downscaling model might be challenging. Therefore, regularization methods can be beneficial (Permatasari, Djuraidah, and Soleh, 2017; Sungkawa, Rahayu, et al., 2019). For example, Hessami et al. (2008) used Ridge regression to downscale precipitation and temperature in eastern Canada and pointed out that Ridge estimates are more robust than ordinary least squares estimates.

1.2.2 Neural networks and deep learning

In the last subsection, we discussed the linear regression model, which is a simple and usually interpretable model. However, in some applications, the assumption that the

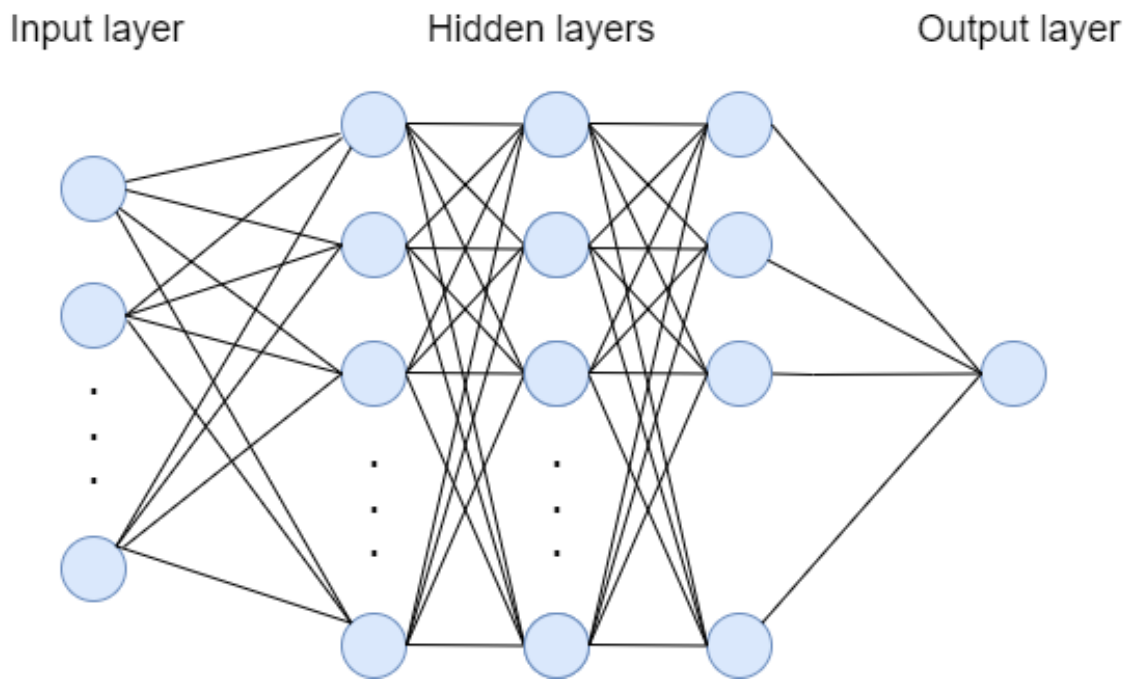

Figure 1.1 – Multi layer perceptron (MLP) architecture.

relationship between the predictor X and the predictand Y is linear might not be accurate. Therefore, non-linear methods like neural networks are gaining popularity in the climate community, for instance, for precipitation (Baño-Medina, Manzanas, and Gutiérrez, 2020), wind (Sailor et al., 2000) and temperature downscaling (Sha et al., 2020).

The multi-layer perceptron (MLP) is a widely used artificial neural networks (ANNs). It consists of a collection of connected neurons which form layers. The first and the last layers are the input and the output layer, respectively, and the layers in the middle are called hidden layers (figure 1.1). Each neuron in the hidden layer computes a real number that corresponds to a non-linear transformation of the linear combination of the previous layer

$$\begin{aligned} Z_m^{(l)} &= \sigma(Z^{(l-1)}\beta_m^l), \;\; m = 1, ..., M_l, \;\; l = 1, ..., L \\ Z^{(l)} &= (Z_1^{(l)}, Z_2^{(l)}, ..., Z_{m_l}^{(l)}), \;\; Z_0 = X \end{aligned} \tag{1.6}$$

where $Z_m^{(l)}$ is the m-th neuron of the hidden layer l, σ is the activation function, β_m^l are coefficients of the neuron m including the intercept, M_l is the number of neurons, and L is the number of hidden layers. Common choices of the activation function are linear, sigmoid, tanh or Rectified Linear Unit (ReLU) (see figure 1.2). For regression, the measure

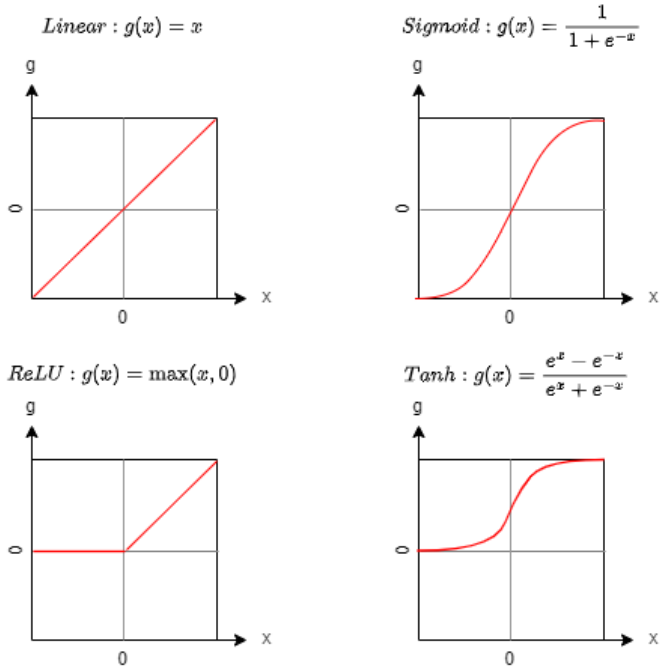

Figure 1.2 – Activation functions used in neural networks.

for goodness of fit is usually the mean squared error (MSE)

$$MSE(\theta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1.7}$$

where θ are model parameters, often called weights. The weights of artificial neural networks are estimated by minimizing the cost function (4.8) and gradient descent is used to find the minimum. Given the structure of the model, gradients can be derived easily using the chain rule. This approach of updating gradients for neural networks is called back-propagation (Hastie et al., 2009).

Using more than the classic three layers (input, hidden, output) in artificial neural networks has given rise to deep learning (LeCun, Bengio, and Hinton, 2015). Recently, deep learning models have grown significantly due to more powerful computers, large data sets, and optimization techniques for training deeper networks. The advantage of deep learning models lies in their ability to build hierarchical representations of inputs, making them able to extract features from complex data structures such as images. Convolutional neural networks (CNNs) are an example of deep learning methods used primarily in computer vision. CNNs use the convolution operator in at least one of their layers.

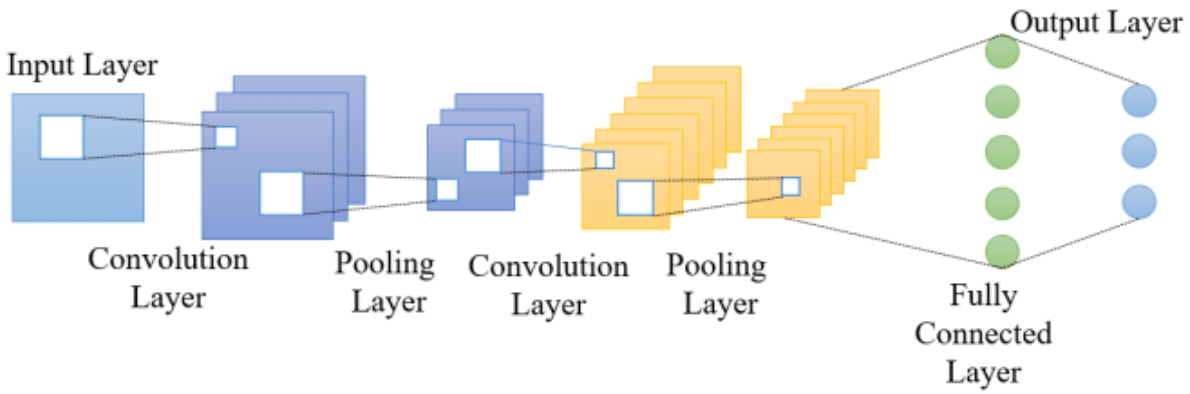

Figure 1.3 – Convolutional neural networks architecture. Illustration source (Gu et al., 2019).

Convolution is a type of linear operator which is applied to inputs in order to extract features. Feature extraction is done by passing a matrix (called filter or kernel) over the inputs and transforming it based on the kernel values.

MLP models are usually called fully connected networks, meaning that each neuron in each layer is connected to all neurons in the next layer. On the other hand, CNNs are known for sharing weights thanks to the convolution operator, as all positions in the image share the same kernels. Weight sharing allows feature extraction to be locally invariant as kernels pass through all image positions. CNNs allow reducing the number of parameters to learn (Yamashita et al., 2018). CNNs contain a series of connected layers (figure 1.4), which are:

- **Input layer**
  The input of CNNs is usually an image represented by a 3D array of dimensions (height, width, depth).
- **Convolutional layer**
  The convolution layer uses filters that apply the convolution operator to input data with respect to its dimensions. Then an activation function (such as ReLU, Sigmoid, or tanh) is applied after the convolution. The resulting output is called the feature map.
- **Pooling layer**
  The pooling layer is a method used to reduce the dimension of features. Common forms of pooling are max pooling and average pooling. It should be noted that there are no learnable parameters in the pooling layer.

- **Fully connected layer**
  After the final convolution or pooling layer, the features are transformed into a one-dimensional vector. Then this vector is passed to fully connected layers (dense layers).

As for MLP, CNNs are trained using back-propagation that optimizes a loss function (which depends on the task: regression or classification), where the parameters to be learned are convolution kernels and weights in the fully connected layer. Several parameters that determine the architecture, such as the filter size, the number of filters, the number of layers, etc., need to be selected.

Artificial neural networks are widely used in downscaling as an alternative for linear regression because of their ability to learn complex and non-linear relationships between large and local scale variables. For example, Cannon and Whitfield (2002) used ANNs for downscaling streamflow conditions over British Columbia and Canada, and their method outperforms stepwise regression. Cawley et al. (2003) used a multi-layer perceptron for downscaling extreme precipitation in the northwest of the United Kingdom. Deep learning approaches are also gaining increasing attention in the climate and meteorology community (Scher, 2018; Rasp, Pritchard, and Gentine, 2018). Many deep learning models have been proposed as statistical downscaling models. For instance, Baño-Medina, Manzanas, and Gutiérrez (2020) proposed a CNN model for downscaling precipitation and temperature over Europe and concluded that their deep learning model outperformed linear and generalized linear models. Sha et al. (2020) used a CNN model to downscale daily minimum and maximum temperature over the western continental United States and showed that their method outperforms simpler downscaling methods.

1.3 Weather types models

Weather typing consists of finding the leading atmospheric circulation patterns that influence mesoscale climate. The classification of weather systems is widely used in meteorology and climatology, and numerous methods have associated weather types to the climatology of precipitation (Yarnal and Frakes, 1997), temperature (Fernández-Montes et al., 2013; Peña-Angulo et al., 2016), and ocean waves (Camus et al., 2014b). Numerous weather typing methods have been proposed in the literature. Ailliot et al. (2015) notes two different methods for constructing weather types; observed or latent weather states. Observed weather types are extracted directly from predictors or other global variables

like sea level pressure. On the other hand, a statistical model estimates latent weather types a posteriori from local or both local and global variables. Weather types-based statistical downscaling approaches find the leading atmospheric circulation patterns and then fit a model between the predictor and predictand in each weather type (Maraun et al., 2010).

1.3.1 Observed weather types

The North Atlantic Oscillation (NAO) is the classical circulation pattern commonly used to define the weather types in the North Atlantic. NAO is defined as the first leading mode from the empirical orthogonal function (EOF) analysis of daily or monthly geopotential height anomalies at the 500 hPa (Hurrell et al., 2003). Weather types can then be defined based on NAO indices such as NAO+, NAO- and blockings.

Observed weather types can also be found using a clustering algorithm on the predictors. The most commonly used clustering algorithm is k-means, which consists of iteratively determining the center of each cluster and assigning the data to the cluster whose center is closest. Boé et al. (2006) proposed a weather types-based SD model for precipitation and temperature. The weather types were constructed using the k-means algorithm, and then a resampling method was used to generate precipitation and temperature conditionally to these weather types. Camus et al. (2014b) applied weather types to downscale ocean wave parameters at two locations on the east coast of the North Atlantic. K-means was used to construct weather regimes, and then the empirical density function of wave parameters was estimated at each regime. Other methods like hierarchical clustering algorithms (Ward Jr, 1963) can be used. For example, Scher (2018) used a hierarchical descending clustering method for constructing weather types using sea level pressure in Australia to statistically downscale rainfall amounts.

1.3.2 Latent weather types

Latent weather types correspond to weather states estimated a posteriori from the data. By considering the weather types as a latent/hidden variable, an optimal clustering that captures the local dynamics can be obtained using a model-based clustering algorithm. For example, hidden Markov models (HMMs) were used for constructing weather regimes for precipitation (Zucchini and Guttorp, 1991; Vrac, Stein, and Hayhoe, 2007). Similarly, mixture models are model-based clustering approaches to infer latent weather

types. (Flecher et al., 2010) used a mixture model to construct weather types for simulating multivariate daily time series of minimum and maximum temperatures, global radiation, wind speed, and precipitation intensity.

Bellone, Hughes, and Guttorp (2000) assumed that precipitation depends on hidden weather types, which were modeled using non-homogeneous hidden Markov models. Depending on the weather types, the rainfall amounts were modeled using the Gamma distribution, and the model parameters were estimated using the Expectation-maximization (EM) algorithm. Ailliot and Monbet (2012) used a weather type-based approach to describe wind speed on the island of Ushant. A non-homogeneous hidden Markov chain modeled the weather types, and the time series of wind speed were simulated with an autoregressive model depending on the weather types. The method is called a Markov-switching autoregressive model, estimated with the EM algorithm. Vrac, Hayhoe, and Stein (2007) compared EM and hierarchical clustering weather types-based methods over North America and found that the EM-based approach is generally more reliable than hierarchical clustering in detecting variability and simulating intraseasonal observed weather patterns. Furthermore, they pointed out that "hierarchical clustering will tend to provide us with strong average information and a sharp distinction between the patterns as they will have significantly different mean values. In contrast, the EM method takes the variance of the data into account to define patterns... That means that a day can belong to more than one pattern at the same time, with different probabilities".

1.4 Expectation-maximization algorithm

One widely used method for inferring latent variables is the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977a). The EM algorithm is an iterative method that maximizes the likelihood when there is missing data. The EM algorithm alternates between the expectation and maximization steps (E-step and M-step, respectively). The E-step calculates the conditional expectation of the complete data log-likelihood given the observations and current parameters. Then in the M-step, the parameters are estimated by maximizing the conditional expectation of the log-likelihood calculated in the E-step.

Suppose we have a set of observed data X and a latent variable Z and suppose that the distribution of X depends on some set of parameters θ. The log-likelihood function is

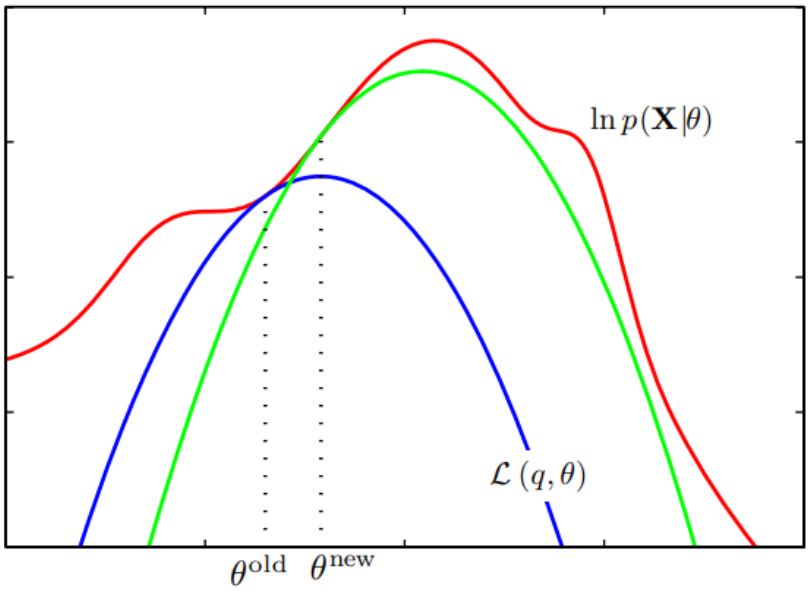

Figure 1.4 – Illustration of the EM algorithm, which involves computing the lower bound and maximizing it with respect to the parameters. Illustration source (Bishop and Nasrabadi, 2006)

given by

$$\ln p(X;\theta) = \ln \sum_z p(X,z;\theta). \tag{1.8}$$

Maximizing 1.8 is problematic given that the sum prevents the logarithm from acting directly on the joint distribution of X and Z (Bishop and Nasrabadi, 2006). Consider $p(X,Z;\theta)$ the complete log-likelihood of the complete data (X,Z), which we suppose is simple to optimize and let q be a distribution over the latent variables Z. We have:

$$\begin{aligned}
\ln p(X;\theta) &= \sum_z q(z) \ln p(X|z;\theta) \\
&= \sum_z q(z)(\ln p(X,z;\theta) - \ln p(z|X;\theta)) \\
&= \sum_z q(z)(\ln \frac{p(X,z;\theta)}{q(z)} - \ln \frac{p(z|X;\theta)}{q(z)}) \\
&= \sum_z q(z) \ln \frac{p(X,z;\theta)}{q(z)} - \sum_z q(z) \ln \frac{p(z|X;\theta)}{q(z)}.
\end{aligned} \tag{1.9}$$

Therefore (Bishop and Nasrabadi, 2006),

$$\ln p(X;\theta) = \mathcal{L}(q,\theta) + KL(q||p) \tag{1.10}$$

where

$$\begin{aligned} \mathcal{L}(q,\theta) &= \sum_z q(z) \ln \frac{p(X,z;\theta)}{q(z)} \\ KL(q||p) &= -\sum_z q(z) \ln \frac{p(z|X;\theta)}{q(z)}. \end{aligned} \tag{1.11}$$

$KL(q||p)$ is the Kullback-Leibler (KL) divergence (Kullback, 1997) between $q(Z)$ and the posterior distribution $p(Z|X;\theta)$, which satisfies $KL(q||p) \geq 0$ and $KL(q||p) = 0$ if and only if $q(Z) = p(Z|X;\theta)$. Therefore, $\mathcal{L}(q,\theta)$ is a lower bound of $\ln p(X;\theta)$ given that $\ln p(X;\theta) \geq \mathcal{L}(q,\theta)$.

Given a current value of the parameters θ^{old}, the E-step maximizes the lower bound $\mathcal{L}(q,\theta)$ with respect to the distribution $q(Z)$ while holding the parameters θ^{old} fixed. The solution to this optimization problem occurs when the Kullback-Leibler divergence equals zero, which corresponds to the case where $q(Z)$ is equal to the posterior distribution $p(Z|X;\theta)$. The lower bound $\mathcal{L}(q,\theta)$ therefore becomes

$$\begin{aligned} \mathcal{L}(q,\theta) &= \sum_z p(z|X;\theta^{old}) \ln \frac{p(z,X;\theta)}{p(z|X;\theta^{old})} \\ &= \sum_z p(z|X;\theta^{old}) \ln p(z,X;\theta) - \sum_z p(z|X;\theta^{old}) \ln p(z|X;\theta^{old}) \\ &= Q(\theta|\theta^{old}) + C \end{aligned} \tag{1.12}$$

where

$$\begin{aligned} Q(\theta|\theta^{old}) &= \sum_z p(z|X;\theta^{old}) \ln p(z,X;\theta) \\ C &= \sum_z p(z|X;\theta^{old}) \ln p(z|X;\theta^{old}). \end{aligned} \tag{1.13}$$

Given that the constant C is independent of θ, the E-step correspond to computing the expectation of the complete-data log-likelihood $Q(\theta|\theta^{old})$. On the other hand, in the M-step, the distribution $q(Z)$ is held fixed, and the lower bound $\mathcal{L}(q,\theta)$ is maximized with respect to the parameters θ. This automatically increases the log-likelihood $\ln p(X;\theta)$ by at least as much as the lower bound does, given that the KL divergence is non-negative. The M-step, therefore, corresponds to maximizing the expectation of the complete-data log-likelihood $Q(\theta|\theta^{old})$.

Algorithm 1: EM algorithm

Input: Observed variable X
Initialization: Initialize the parameters θ
repeat
  E-step
  calculate the quantity $Q(\theta|\theta^{old})$
  M-step
  maximize $Q(\theta|\theta^{old})$ with respect to θ
  $\theta^{old} = \theta^{new}$
  $\theta^{new} = \arg\max_\theta Q(\theta|\theta^{old})$
until *convergence*;

1.5 Conclusions

In this chapter, we recalled the problem of downscaling, especially statistical downscaling. Then, we presented some statistical and data-driven approaches used for SD. Some methods, such as linear regression, neural networks, deep learning, or weather types, were used in the literature for downscaling precipitation and other climate variables. These methods are presented in this chapter primarily for understating this thesis. For example, the EM algorithm was presented in this chapter, given that it is used in the literature on latent weather types and will be widely used in this thesis (in chapters 4 and 5).

Chapter 2

Sea State Characterization

Contents

Note: This second chapter aims to recall the definitions and terminologies related to ocean waves. The first section explains wind-wave generation. Then, the methods used to observe the sea state are explained in the second section. The numerical wave models are recalled in the third section, and the wave data used in this thesis are presented. Then, section 3 presents statistical methods used for wave characterization. The last section concludes this chapter.

2.1 Wind waves

Many different waves determine the dynamics of the ocean. Each type of ocean wave is characterized by its wave period (figure 2.1). For instance, capillary waves have a period of fewer than 0.1 seconds, infra gravity waves can have a period between 30 seconds and five minutes, and ordinary tide waves have a period between 12 and 24 hours (Ardhuin and Orfila, 2018). In this thesis, we are interested in gravity waves generated by the wind, which typically have a period between 1 and 30 seconds.

The instantaneous local wind is necessary for forming wind waves, but it is insufficient. Instead, wind duration and the distance over which the wind is blowing are essential for achieving considerable wave heights (Ardhuin and Orfila, 2018). The distance over which the wind is blowing in a constant direction is called the fetch. Waves that are generated and observed at the fetch are called wind seas, which typically have a period between 1 and 8 seconds. When wind seas leave their generation area, they form swells. Swells generally have a period between 8 and 30s and can travel over long distances. The sea surface is

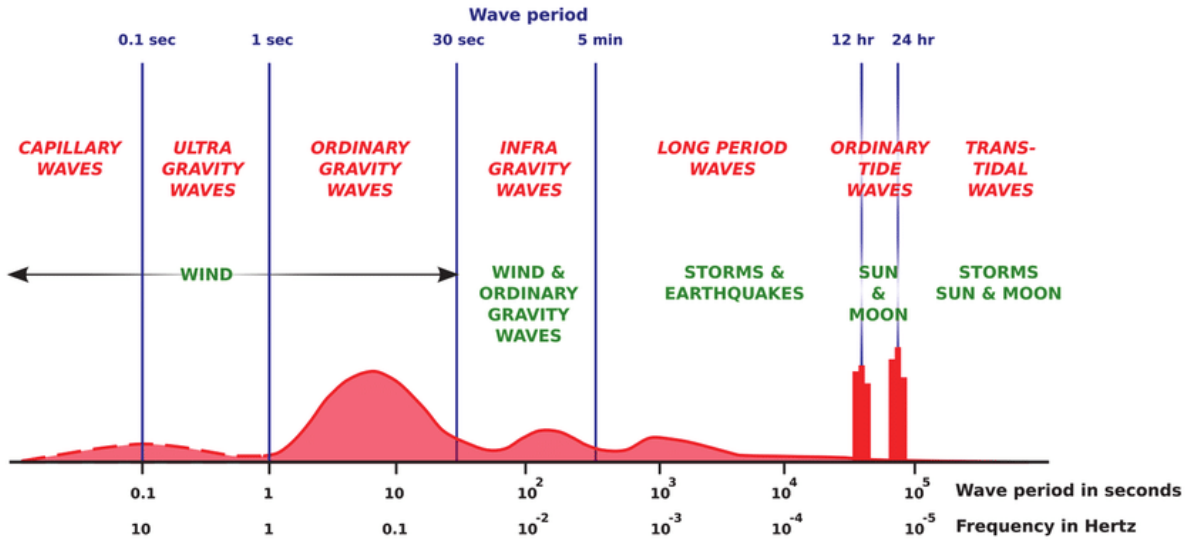

Figure 2.1 – Classification of ocean waves by wave period. Illustration source: (Ardhuin and Orfila, 2018)

characterized by a superposition of different swells generated from distant areas and wind sea generated by local wind (Ardhuin and Orfila, 2018). The statistical description of the sea surface at a given time and location is called the sea state.

Sea state characterization is needed for many applications that require extended time series with spatial resolution on the kilometer scale. In this study, we distinguish between three methods for sea state characterization: observation methods, numerical models, and statistical models.

2.2 Sea state observation

Sea state observation methods can be categorized into three categories: voluntary observing ships (VOS), *in situ* measurements, and satellite observations.

- **voluntary observing ships:**

VOS is an international program where ships are hired to record and transmit weather observations at sea. The VOS data are observations with the longest historical record, dating back to 1870, making them more useful for climate and extreme studies than any other source of wave observations (Ardhuin et al., 2019). However, VOS data are subject to biases, are unevenly distributed across the ocean, and are inhomogeneous in temporal sampling (Gulev et al., 2003).

- ***in-situ* measurements:**

in-situ measurements provide high-quality measurements of sea state parameters such as significant wave height, wave period, and wave direction using buoys and offshore platforms. However, *in-situ* measurements are mostly concentrated near coastal areas, most of which are found in Western Europe and North America.

- **Satellite observations:**

Satellite remote sensing of the ocean is another source of sea state data. From space, satellites use active sensors such as altimeters to estimate significant wave height, while other parameters such as wave period and direction can be estimated from synthetic aperture radar images (Timmermans et al., 2020). Starting in 1985 with GEOSAT (GEOdetic SATellite), the satellites provide global and quasi-continuous coverage with fine spatial resolution. However, the temporal resolution of the waves measurements at a given point is low, given that a single satellite measures waves with a time step that depends on its orbit.

2.3 Numerical models

Sea state observation techniques do not provide a complete knowledge of the sea state in space and time (Ardhuin et al., 2019). Numerical wave models are an alternative source of wave data that makes it possible to study waves with a high spatial and temporal resolution.

The ocean surface is characterized by waves of different periods and directions of propagation. This phenomenon is known as the principle of superposition. Therefore, numerical models describe the sea state using a two-dimensional spectrum. The first spectral numerical wave model was developed by the French Weather Service in 1956 (Gelci, Cazalé, and Vassal, 1957). From this point on, numerical models have undergone numerous improvements regarding mathematical models used for simulation, physical representation of the wave phenomena, validation methods, and computation and storage capacities (Laugel, 2013).

Numerical wave models are based on the energy balance equation (Thomas and Dwarakish, 2015)

$$\frac{\partial E(f,x,t,\theta)}{\partial t} = \mathrm{In} + \mathrm{Nl} + \mathrm{Dis} \tag{2.1}$$

where $E(f,x,t,\theta)$ is the two-dimensional spectrum that depends on the frequency f,

direction of propagation θ, geographic coordinates x and time t. In represents the wind energy influencing the waves, Nl accounts for non-linear wave interactions and Dis for dissipation. The first-generation wave models did not consider non-linear wave interactions and dissipation. The second-generation wave models are developed from the wind fields and account for the non-linear interactions. Finally, the third-generation wave models have improved the process of modeling the physics relevant for the characterization of the sea state in two dimensions (frequency and direction).

The sea state is generally described by synthetic statistics derived based on the moments $m_n(x,t)$ of the spectrum $E(f,x,t,\theta)$ given by

$$m_n(x,t) = \int_0^{\infty} \int_0^{2\pi} f^n E(f,x,t,\theta) df d\theta. \tag{2.2}$$

In this study, we are interested in the significant wave height (H_s) parameter defined as

$$H_s(x,t) = 4\sqrt{m_0(x,t)} \tag{2.3}$$

Numerical wave models are used for either forecasting (Remya et al., 2022), hindcasting (Boudière et al., 2013), or downscaling (Hemer, Katzfey, and Trenham, 2013). Hindcasting is the process of reconstructing past sea state conditions. Numerical downscaling methods are used to derive fine resolution long-term future projections of ocean wave parameters by numerically downscale coarse predictions from global ocean-atmospheric models to local scales.

The numerical wave data used in this study is the Homere hindcast database (Boudière et al., 2013) developed at IFREMER (French National Institute for Ocean Science). Homere is based on the third-generation wave model WAVEWATCH III on a destructured grid covering the English Channel and Bay of Biscay (figure 2.2) area from 1994 to 2021. The wind forcing considered is the CFSR (Climate Forecast System Reanalysis) wind.

2.4 Statistical models

Statistical methods can be an alternative to numerical methods, given that they are computationally inexpensive. As for numerical methods, sea state characterization statistical methods can be used for three purposes: hindcasting, forecasting, or downscaling. For hindcasting, statistical methods such as multiple linear regression (Campos et al., 2018)

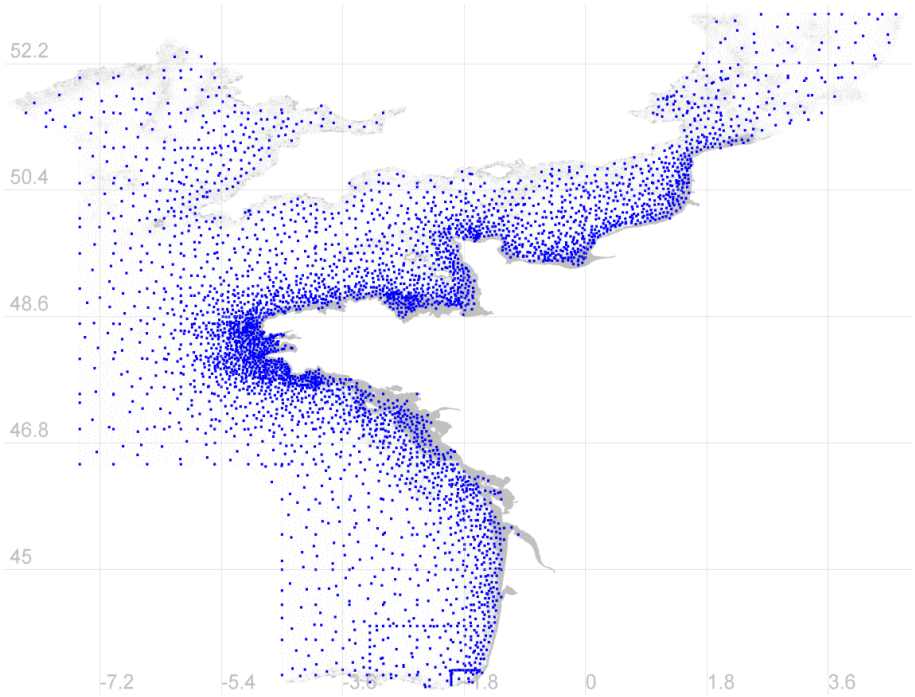

Figure 2.2 – Homere nodes in the English Channel and Bay of Biscay. Figure source: https://marc.ifremer.fr/produits/rejeu_d_etats_de_mer_homere.

are used to reconstruct past wave conditions. For forecasting, time series forecasting methods, such as autoregressive (AR) and autoregressive moving average methods (ARMA) (Ge and Kerrigan, 2016; Soares, Ferreira, and Cunha, 1996; Agrawal and Deo, 2002), are usually used to predict short-term future sea state parameters. Machine and deep learning methods, such as artificial neural networks, are also used to forecast sea state parameters (Deo and Naidu, 1998; Duan et al., 2016). On the other hand, sea state statistical downscaling methods refer to statistical methods that use large-scale variables to characterize a local-scale sea state parameter.

In chapter 2, we reviewed statistical downscaling methods used in the literature. In this chapter, we focus on SD methods used for ocean waves. SD models of ocean waves aim to statistically construct sea state parameters using large-scale conditions (predictor) to study historical and future trends of local-scale wave conditions (predictand). The wind mainly generates waves; however, sea level pressure (SLP) is usually used as a large-scale predictor in SD, given that the isobars well represent the wind direction, and wind speed is proportional to the pressure gradient (Camus et al., 2014a). For instance, Wang, Feng, and Swail (2012) used a multivariate regression model to reconstruct significant

wave height trends in the 20th century using sea level pressure as a predictor. Wang and Swail (2006) reconstructed historical and future mean seasonal H_s in the North Atlantic and North Pacific using linear regression with a redundancy analysis between SLP and H_s. The reason why most sea state SD studies (Camus et al. (2014a), Camus et al. (2014b), Cagigal et al. (2020), Laugel (2013)) use SLP instead of wind fields is that sea wind is not as well represented in GCMs as SLP. However, Wang, Swail, and Cox (2010) pointed out that "it is sufficient to use the wind-based predictor alone to represent the relationship between atmospheric conditions and H_s. However, the wind-based predictor values must be standardized to diminish the effects of both model climate and variability biases". Therefore, in this study, we focus on modeling the relationship between wind conditions and H_s.

As stated in section 2.1, the sea is characterized by a superposition of different swells generated from distant areas and wind sea generated by local wind. Therefore, to model the relationship between wind and waves, both local and global wind conditions need to be considered. Moreover, swells generated from distant areas might take several days to reach the target point. Consequently, waves at the target point depend on wind conditions at distant areas with temporal lag that can be days; therefore, preprocessing methods are needed (Camus et al. (2014a), Casas-Prat, Wang, and Sierra (2014)). Furthermore, the wind data consists of two components (zonal and meridional); therefore, statistical modeling of the relationship between wind and waves is challenging given the amount of data to be considered.

2.5 Conclusions

In this chapter, different methods for describing the sea state were discussed. First, sea state observation techniques were presented, and as Ardhuin et al. (2019) pointed out, observation techniques cannot meet engineers' requirements in the near future because they are limited in space and time. Two alternatives were then discussed, namely numerical and statistical models. Numerical models are generally more accurate and provide multivariate sea state parameters (significant wave height, period, and direction), but they are more computationally intensive than statistical models. Moreover, statistical models can well reproduce the observed (Wang, Swail, and Cox, 2010) and future (Laugel et al., 2014) wave climate.

It should be noted that the three approaches complement each other and that statisti-

cal models cannot currently replace numerical models. For example, observations cannot be used for all climate studies, but they are essential to calibrate numerical and statistical hindcasting models. Furthermore, to the best of our knowledge, statistical hindcasting/forecasting/downscaling methods cannot be used at locations in the ocean where no (observed or hindcasted) wave data are available. Therefore, numerical models are powerful tools that provide wave data with high spatial and temporal resolution from which statistical methods can learn.

In the following chapters, we focus on modeling the relationship between wind conditions over the North Atlantic and significant wave height at a location in the Bay of Biscay. The proposed methods are statistical downscaling models that link wind conditions and waves. In reality, depending on the availability of wind data, the proposed methods can also be used for hindcasting or forecasting applications; however, we have chosen to refer to the proposed methods as statistical downscaling models.

Chapter 3

Modeling the Space-Time Relation between Wind and Significant Wave Height: a Statistical Approach

Contents

Note: The results of this chapter are submitted for publication as S.Obakrim, P.Ailliot, V.Monbet, and N.Raillard, Statistical modeling of the space-time relation between wind and significant wave height[1].

1. The preprint can be found in https://doi.org/10.1002/essoar.10510147.1

3.1 Introduction

High-quality wave data is essential for many marine applications, such as designing coastal and offshore structures and planning marine operations. In the previous chapter, we discussed different methods for sea state characterization, namely, observations, numerical, and statistical methods. 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 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 data source and provide decades of records, although they are computationally expensive and sensitive to the quality of forcing fields (wind, currents, and water levels) (Roland and 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.

Various studies have compared SD and numerical models for ocean wave parameters and other climate variables. Wang, Swail, and Cox (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, and their study shows that statistical downscaling approaches can reproduce the present climatology and future projections. In addition, due to their low computational complexity, SD models allow for the estimation of uncertainties associated with the choice of general circulation models (GCMs) or climate scenarios. However, there are still some challenges in modeling the relationship between wind and sea state parameters using statistical methods, namely:

- Waves depend on both local and global wind conditions

The surface wind generates wind waves. However, it is not only the local wind that defines local waves, and wind from distant regions generates waves that may reach the target point. Therefore, SD models have to consider both wind sea and swells, which is challenging in swell-dominated areas (Hemer et al., 2012). To address this issue, we use a local and a global predictor to account for wind sea and swells, respectively (Casas-Prat, Wang, and Sierra, 2014; Camus et al., 2014a).

- Wind conditions are multicollinear and multidimensional

As discussed in the first chapter, the large-scale wind variables are multicollinear; thus, regularization methods such as ridge regression can be beneficial. Furthermore, the wind conditions are characterized by two components (zonal and meridional), which might be challenging to consider directly in a statistical model. Dimensionality reduction methods such as principal component analysis are typically used as a preprocessing step to reduce the dimension of the large-scale variables (Laugel et al., 2014; Camus et al., 2014a; Camus et al., 2014b). To address the issue of multidimensionality in this study, we introduce the wind projection, which consists of retaining only the fraction of wind blowing towards the target point. The proposed preprocessing step allows using only one variable for each grid point, reducing the dimension of the predictor by half.

- The relationship between wind and waves is not instantaneous

Wind from distant regions generates waves that may take days to reach the target point. Thus, the relationship between wind and waves is not instantaneous. Therefore, it is necessary to consider lagged wind conditions to understand the wave dynamics at a particular target location. The optimal lag at each grid point is interpreted as the travel time required for the waves to reach the target point (Camus et al., 2014a). 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 ESTELA method was used in various studies to define the temporal coverage of predictors used in SD (Camus et al. (2014a), Hegermiller et al. (2017)). The present study uses a data-driven approach to define the wave generation area. It is based on estimating waves' travel time from each source to the target point (optimal lag) using the maximum correlation between the significant wave height and wind conditions. Therefore, this method is not computationally expensive, and only wind and H_s data at the target point are needed.

This study provides a framework for the wind to waves relationship using an entirely statistical approach. Based on weather types, the statistical downscaling model links the space-time wind fields over the North-Atlantic (predictors) and the significant wave height (predictand) at a single site located in the Bay of Biscay off the French coast. The weather types are constructed using a regression-guided clustering algorithm, and then a linear regression model is fitted between the wind conditions and H_s at each weather type. 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 the one hand, is that it provides an entirely data-driven approach that estimates the travel time of waves from any source point to a target point, which is essential for the definition of predictors. On the other hand, it proposes a regression-guided clustering algorithm that accounts for both global and local climate to construct weather types.

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

3.2 Data

The atmospheric data used in this work to construct predictors is extracted from the Climate Forecast System Reanalysis (CFSR) (Saha et al., 2010a). 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 3.1).

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 3.1) at (45.2°N, 1.6°W). Waves at this point are related to both large-scale conditions in the North Atlantic (swells) and local conditions (wind seas) (Charles et al., 2012b). Swell conditions are generally dominant; however, the highest H_s are generated by strong local storms. To validate and interpret the results of the SD method, we consider the energy spectral partitioning, which identifies different wave systems. 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

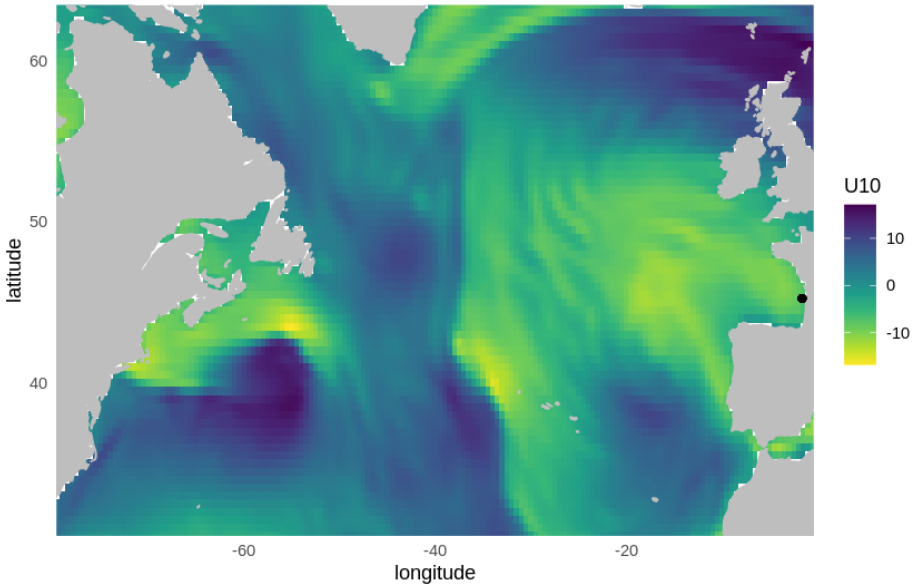

Figure 3.1 – CFSR zonal component in the considered area in 1994-01-01 00h:00. The black point represents the point of interest.

of 23 years, from 1994 to 2016.

3.3 Local predictor

Wind speed, duration, and the fetch impact the characteristics of the wind sea (Ardhuin and 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 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

$$H_s(t) = \beta_0^{(\ell)} + X^{(\ell)}(t)\beta^{(\ell)} + \epsilon^{(\ell)}(t) \tag{3.1}$$

is considered. Where $X^{(\ell)}$ is the local predictor:

$$X^{(\ell)}(t) = \{U(t), U^2(t), U^3(t), U^2(t)F(t), U(t-1), U^2(t-1), U^3(t-1), U^2(t-1)F(t-1)\} \tag{3.2}$$

$\beta_0^{(\ell)}$ and $\beta^{(\ell)}$ are model coefficients, and $\epsilon^{(\ell)}(t)$ is the model error. Model 3.1 contains polynomial terms and interactions between local variables to consider nonlinear 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:

$$r = \frac{\sum_{t=1}^{n}(\hat{H}_s(t) - \overline{\hat{H}_s})(H_s(t) - \overline{H_s})}{\sigma_{\hat{H}_s}\sigma_{H_s}} \tag{3.3}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{H}_s(t) - H_s(t))^2}{n}} \tag{3.4}$$

$$BIAS = \frac{\sum_{t=1}^{n}(\hat{H}_s(t) - H_s(t))}{n} \tag{3.5}$$

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.

Results of the local model 3.1 are shown in Figure 3.2. 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 at predicting large values of H_s, which can be explained by the fact that extremes are mainly generated by local wind.

3.4 Global predictor

In order to take swells into account, a global predictor which describes wind conditions 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

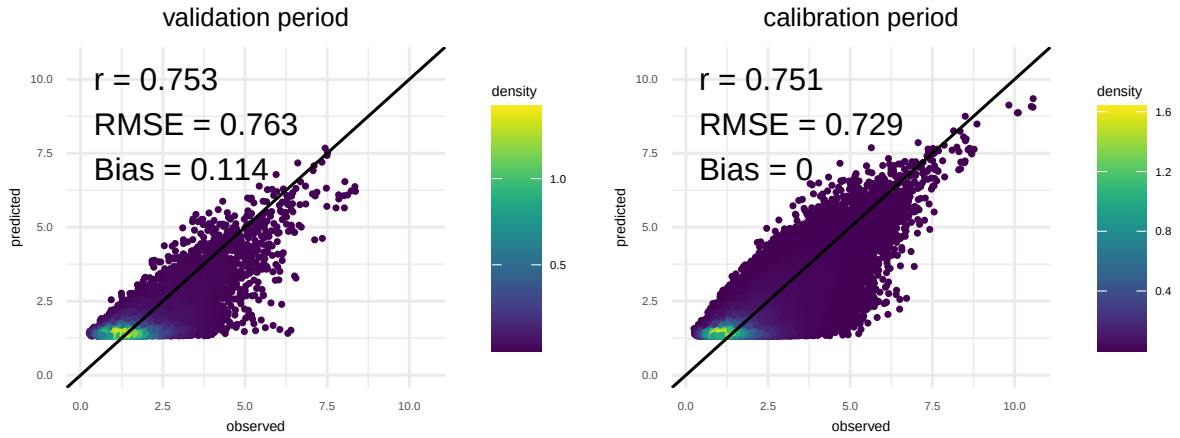

Figure 3.2 – Results of the local model 3.1 in the validation and calibration period.

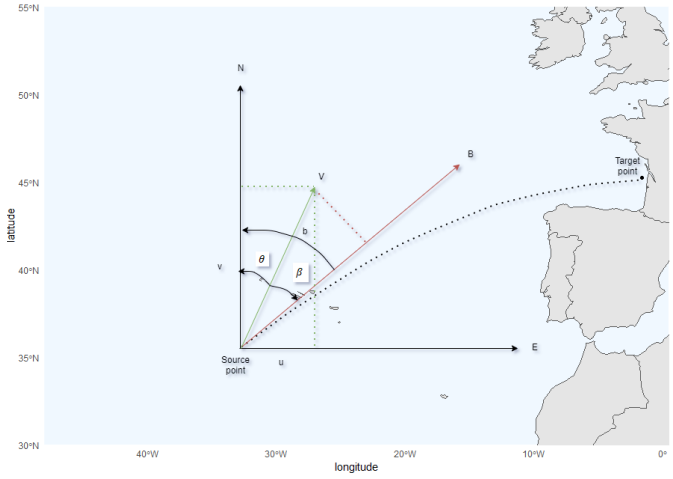

Figure 3.3 – Wind projection representation. The original wind vector V at each source point is projected into the component B defined by the bearing b of the target point from the source point in a great circle path (black dashed line). The great circle is drawn arbitrarily to explain the method and may not be the actual circle path.

to the strong correlation between wind conditions at closed locations in space and time. This section defines the global predictor related to the spatiotemporal domain of the wave generation area.

3.4.1 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.

3.4.2 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 3.3) using the equation:

$$W = U\, cos^{2s}\left(\frac{1}{2}(b-\theta)\right) \tag{3.6}$$

where W is the projected wind, U is the wind speed, s the spread parameter (Young, 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 3.4 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.

3.4.3 Temporal coverage

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

$$C_g = \frac{gT}{4\pi} \tag{3.7}$$

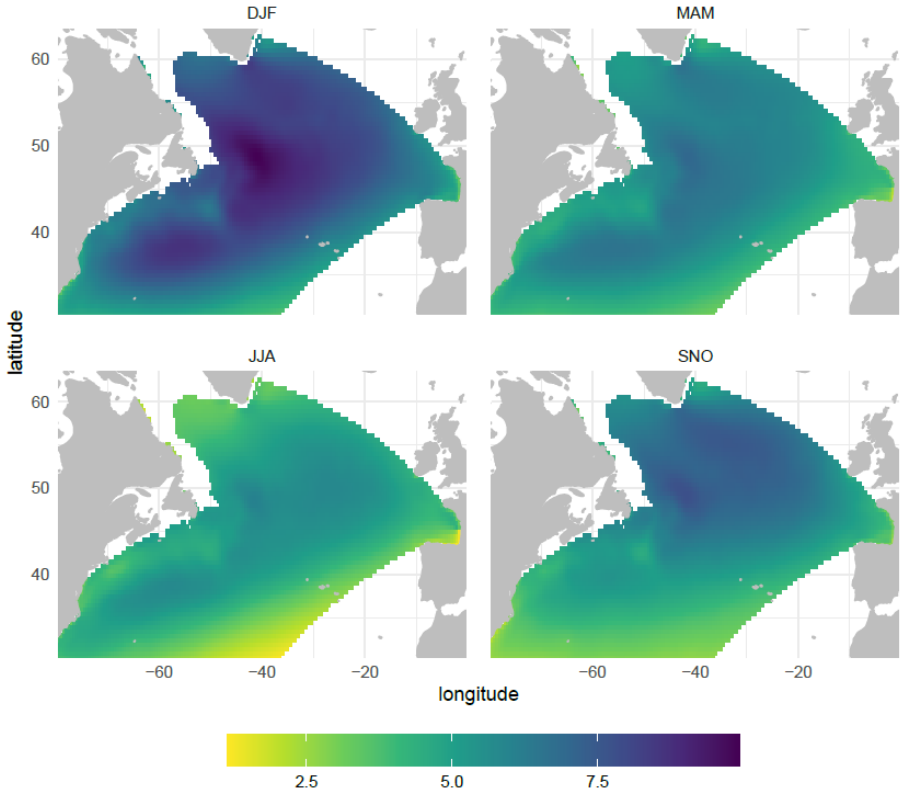

Figure 3.4 – Mean projected wind in the winter (DJF), spring(MAM), summer (JJA), and autumn (SNO).

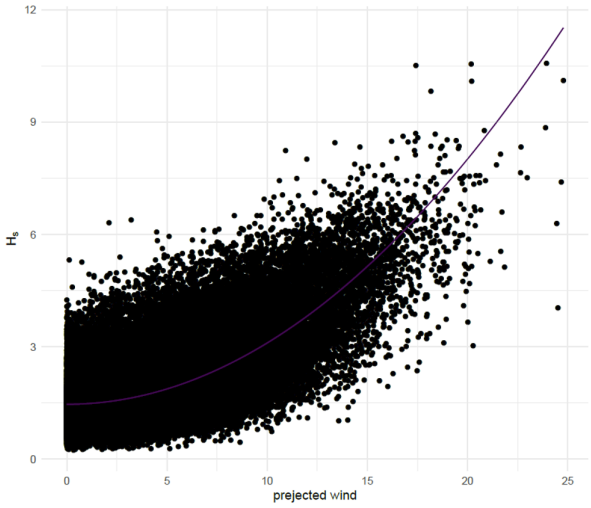

Figure 3.5 – Projected wind at point located in (45.5°N, 3.5°W) versus H_s and the estimated curve line using the model $H_s = aW^2 + b$

where g is the gravitational velocity and T is 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 and 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

$$X_j^{(g)}(t; t_j, \alpha_j) = \frac{1}{2\alpha_j+1} \sum_{i=t-t_j-\alpha_j}^{t-t_j+\alpha_j} W_j^2(i), \qquad (3.8)$$
$$t_j + \alpha_j + 1 \leq t \leq t_j - \alpha_j + n$$

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 even though 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 3.5) so that in equation (3.8) 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 be non-polynomial and computationally unfeasible due to the combinatorial explosion. Therefore,

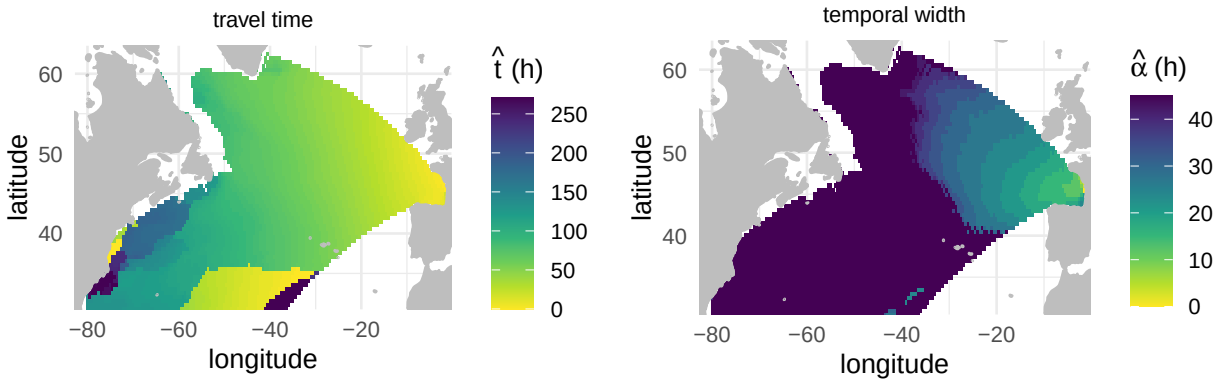

Figure 3.6 – Estimated travel time of waves and the temporal width using equation 3.9

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

$$(\hat{t}_j, \hat{\alpha}_j) = arg \max_{t_j, \alpha_j} \Big(corr(H_s, X_j^{(g)}(t_j, \alpha_j)) \Big). \tag{3.9}$$

Figure 3.6 shows the estimated travel time of waves and the temporal width. Globally, the two parameters are spatially smooth and interpretable, and 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 seven and half days) to reach the target point. These waves travel at a velocity of $8.7m/s$; thus, according to the dispersion equation (3.7), they have an average period of $11.1s$. On the 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 nine 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 five 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.

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 in these areas have negligible contributions to the H_s observed at the target location.

3.5 Wind-waves model

3.5.1 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

$$H_s(t) = X^{(\ell)}(t)\beta^{(\ell)} + X^{(g)}(t)\beta^{(g)} + \epsilon(t) \tag{3.10}$$

where $\beta^{(\ell)}$ and $\beta^{(g)}$ are local coefficients and global coefficients, respectively. Here $\beta^{(\ell)}$ is not necessarily the same as in equation (3.1). $X_t^{(\ell)}$ is the local predictor defined in equation (3.2), $X_t^{(g)}$ the global predictor defined in equation (3.8), and $\epsilon(t)$ is the model error.

3.5.2 Model fitting

Model (3.10) can be fitted using the least squares method; given by

$$(\hat{\beta}) = (X^T X)^{-1} X^T H_s \tag{3.11}$$

where $X = (X^{(\ell)}, X^{(g)})$ and $\hat{\beta} = (\hat{\beta}^{(\ell)T}, \hat{\beta}^{(g)T})^T$. The least-squares estimates in equation (3.11) 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 and Kennard, 1970) minimizes the penalized residual sum of squares

$$\arg\min_{\beta} \left\| X^{(g)}\beta^{(\ell)} + X^{(g)}\beta^{(g)} - H_s \right\|^2 + \lambda \|\beta^{(g)}\|^2 \tag{3.12}$$

where $\lambda \geq 0$ is the regularization parameter. Remark that the regularization is not applied to the parameters associated with the local predictor. The parameter λ allows to take into consideration the bias-variance trade-off.

3.5.3 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

$$X_{\beta^{(g)}}(t,j) = X_j^{(g)}(t)\hat{\beta}_j^{(g)}. \tag{3.13}$$

We expect swell systems coming from contributions from distant areas, whereas wind sea will be associated with local contributions. A natural question that arises is whether we can identify these wave systems by using $X_{\beta^{(g)}}$. 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 each class's linear regression model (3.10). Therefore, Model (3.10) now becomes

$$H_s(t) = X^{(\ell)}(t)\beta_i^{(\ell)} + X^{(g)}(t)\beta_i^{(g)} + \epsilon_i(t), \quad \forall t \in I_i \;\; i = 1, ..., K \tag{3.14}$$

where $\beta_i^{(\ell)}$ and $\beta_i^{(g)}$ 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.

3.5.4 The case of two weather types

The hyper-parameters of the model (3.14) are λ, the number of WTs K, and the K regularization parameters λ_ks associated with the different weather types (given that, at each weather type, ridge regression is fitted). Given the number of hyper-parameters, it is not computationally feasible to explore all possible combinations and optimize them simultaneously using cross-validation, as 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, and finally λ_ks are fixed for all weather types.

The most usual approach to choosing the regularization parameter λ of the ridge regression consists in performing cross-validation and taking the value of λ, which minimizes a prediction error, typically the RMSE. In the current work, we also intend to obtain a physically interpretable model in addition to forecast accuracy. Interpretability will be quantified as follows. First, the k-means clustering algorithm is used on the contributions $X_{\beta^{(g)}}$ to identify the leading two clusters. The resulting clusters are then compared with

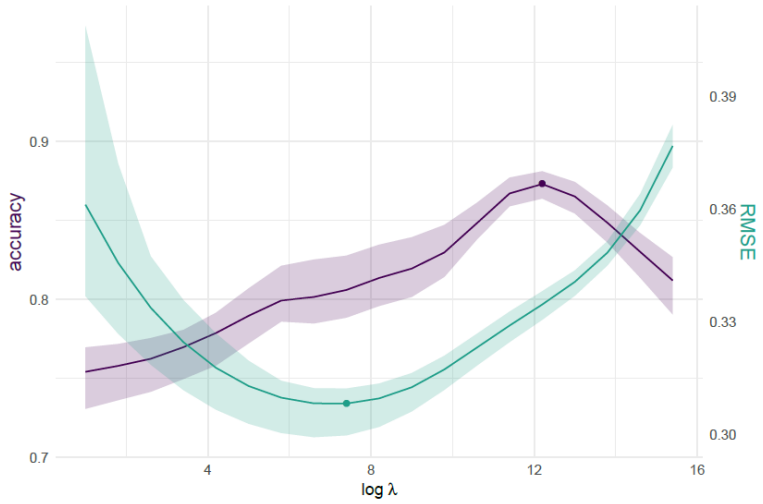

Figure 3.7 – Results of cross-validation: RMSE (green line) and classification accuracy (purple line) versus the logarithm of λ. The red and blue dots correspond to the minimum RMSE and maximum accuracy, respectively. The interval for each criterion is defined as the its minimum and maximum.

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 clusterings is measured using the classification accuracy

$$\text{accuracy} = \text{correct predictions}/\ \text{sample size} \tag{3.15}$$

Figure 3.7 shows that the value of λ that gives the optimal classification accuracy is greater than that of the optimal RMSE. Figure 3.8 shows the estimated global coefficients $\beta^{(g)}$ 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 (0.32m to 0.35m).

Figure 3.9 shows the 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 6.1). To understand the difference between the two clusters, we define the

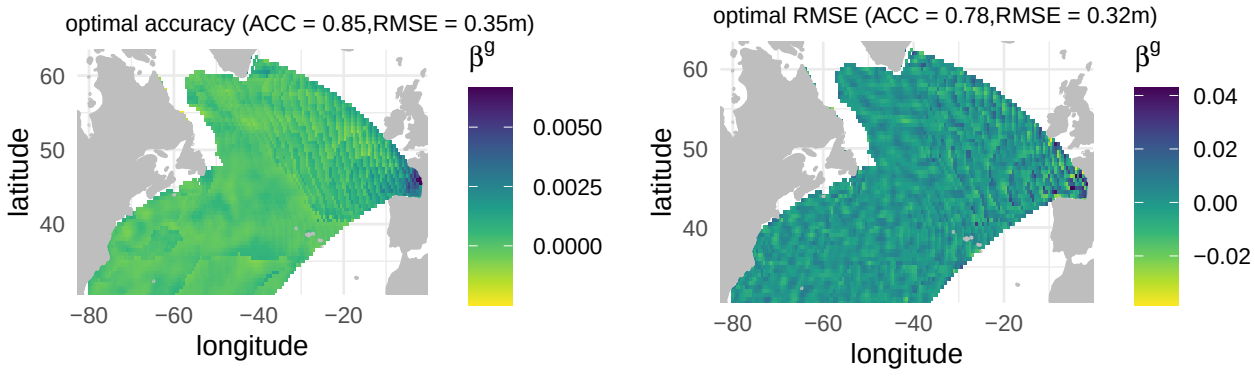

Figure 3.8 – Estimated global coefficients $\beta^{(g)}$ using ridge regression with λ that gives the maximum accuracy (left panel) and minimum RMSE (right panel).

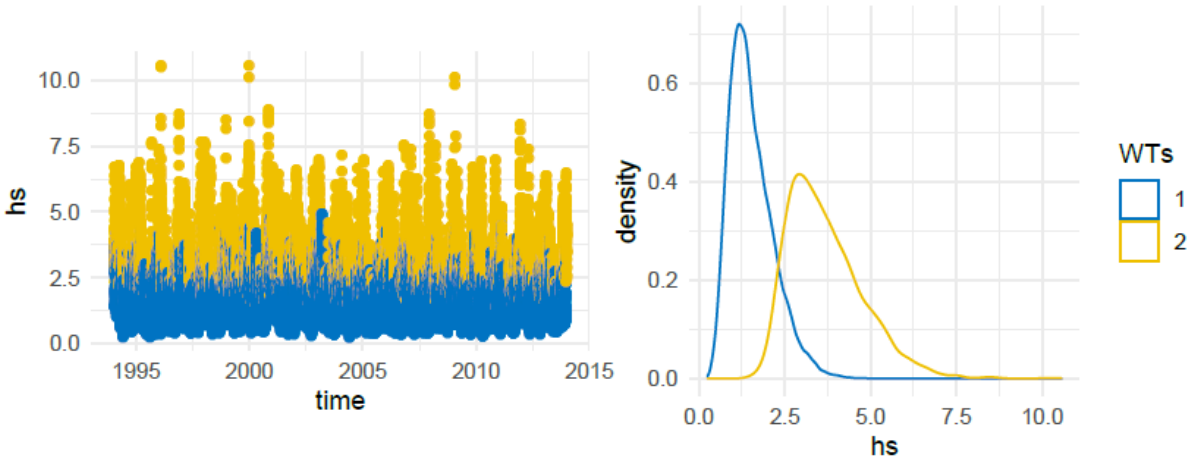

Figure 3.9 – Time series of H_s depending on the clusters (left panel) and empirical density (right panel) in the calibration period.

classes	1	2
swell	47074	6388
wind sea	974	3904

Table 3.1 – Contingency table of k-means clusters (1 and 2) and Homere sea states classes (swell and sea state) in the calibration period.

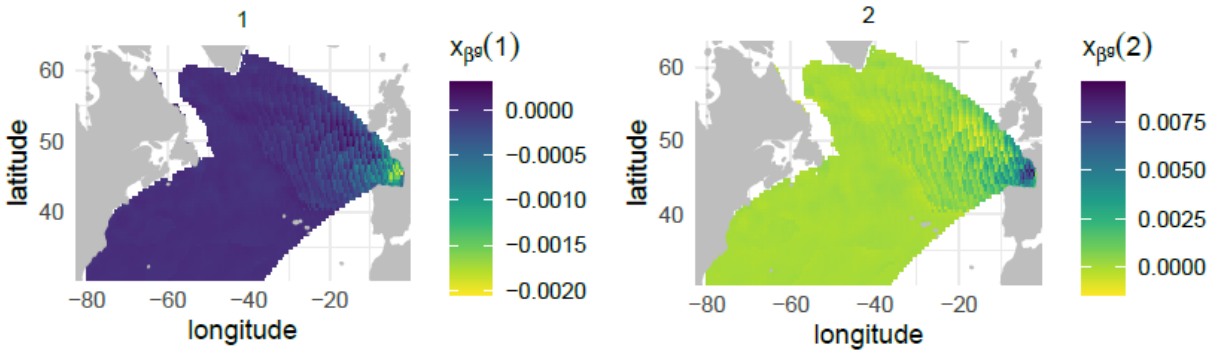

Figure 3.10 – Mean of $X_{\beta(g)}$ minus the global mean for the cluster 1 (left panel) and cluster 2 (right panel).

anomaly of $X_{\beta(g)}$ in each cluster 1 and 2 as $x_{\beta(g)}(1)$ and $x_{\beta(g)}(2)$, respectively

$$\begin{aligned} x_{\beta(g)}(1) &= \bar{X}_{\beta(g)}(1) - \bar{X}_{\beta(g)} \\ x_{\beta(g)}(2) &= \bar{X}_{\beta(g)}(2) - \bar{X}_{\beta(g)} \end{aligned} \tag{3.16}$$

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 3.10). Grid points far from the target point contribute more, which is expected when swell systems dominate. In contrast, in the second cluster, generally associated with wind sea, local wind contributes more than the global mean in H_s.

3.6 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 (3.14) is fitted using the calibration period and evaluated using the validation period. Figure 3.11 illustrates the RMSE of H_s as a function of the number of WTs. The optimal number of WTs is 5, and the RMSE decreases significantly from 1 to 5 WTs.

Figure 3.12 shows the time series of H_s and its empirical density as a function of

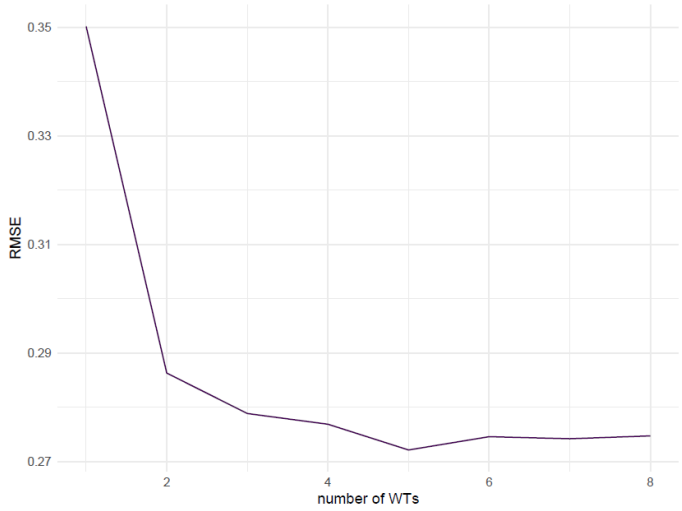

Figure 3.11 – RMSE versus the number of WTs for the validation period.

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 corresponds to extremes. In increasing order, the other clusters (2 to 4) correspond to intermediate values H_s. The bottom right panel of Figure 3.12 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 1st to the 4th and the 5th WT are equal to zero. This means that the probability of being in extreme sea states after being in the first WT is zero.

Figure 3.13 shows the mean of $X_{\beta(g)}$ at each WT where

$$x_{\beta g}(i) = \bar{X}_{\beta(g)}(i) - \bar{X}_{\beta(g)},\ i = 1, .., 5 \tag{3.17}$$

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 fourth one can be considered a composition of wind sea and swells given that local and far source 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.

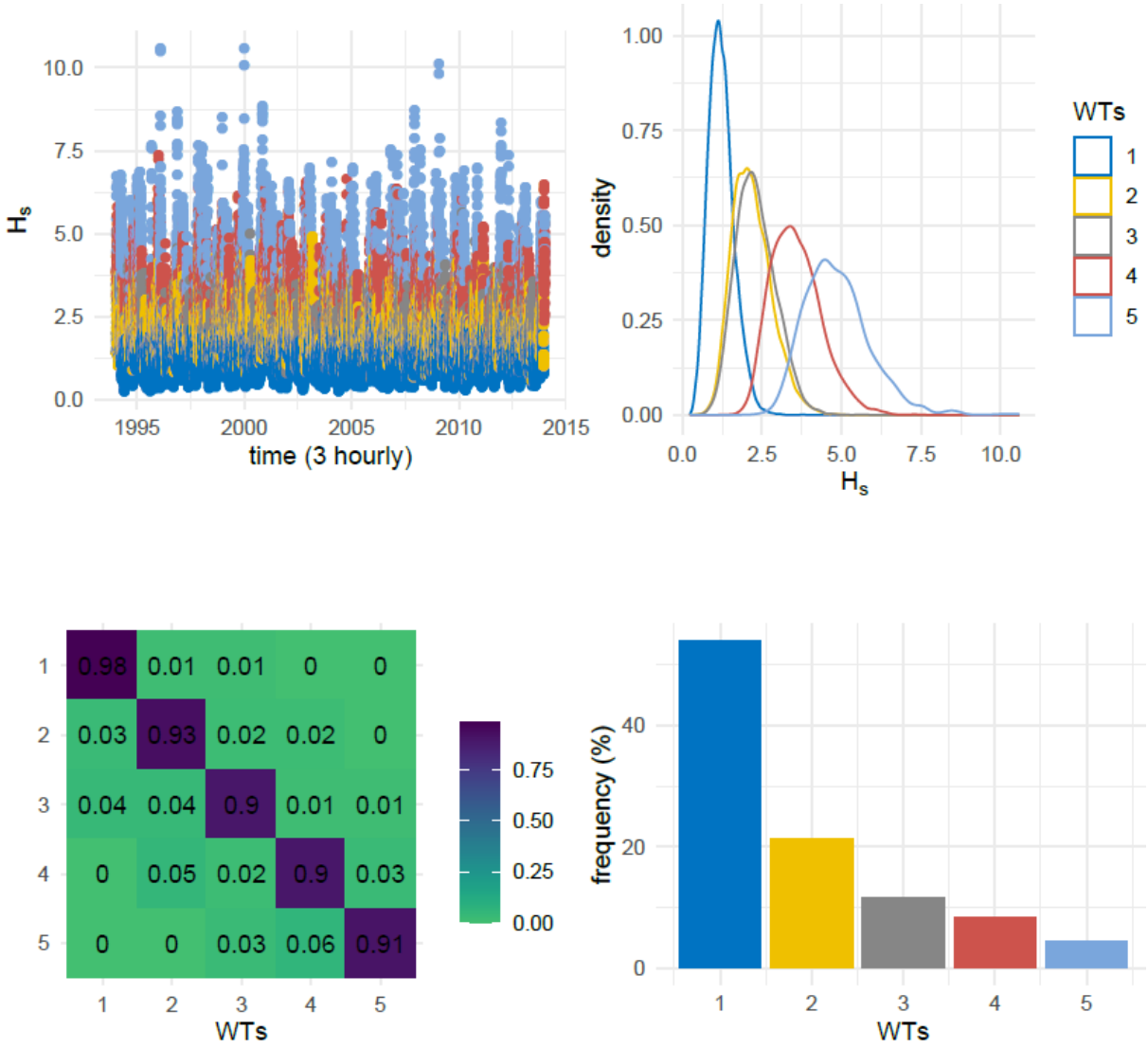

Figure 3.12 – Top left panel: time series of H_s as a function of WTs. Top right: empirical density of H_s as a function of WTs. Bottom left: transition matrix of WTs. Bottom right: Frequency of occurrence of WTs. All figures correspond to the calibration period.

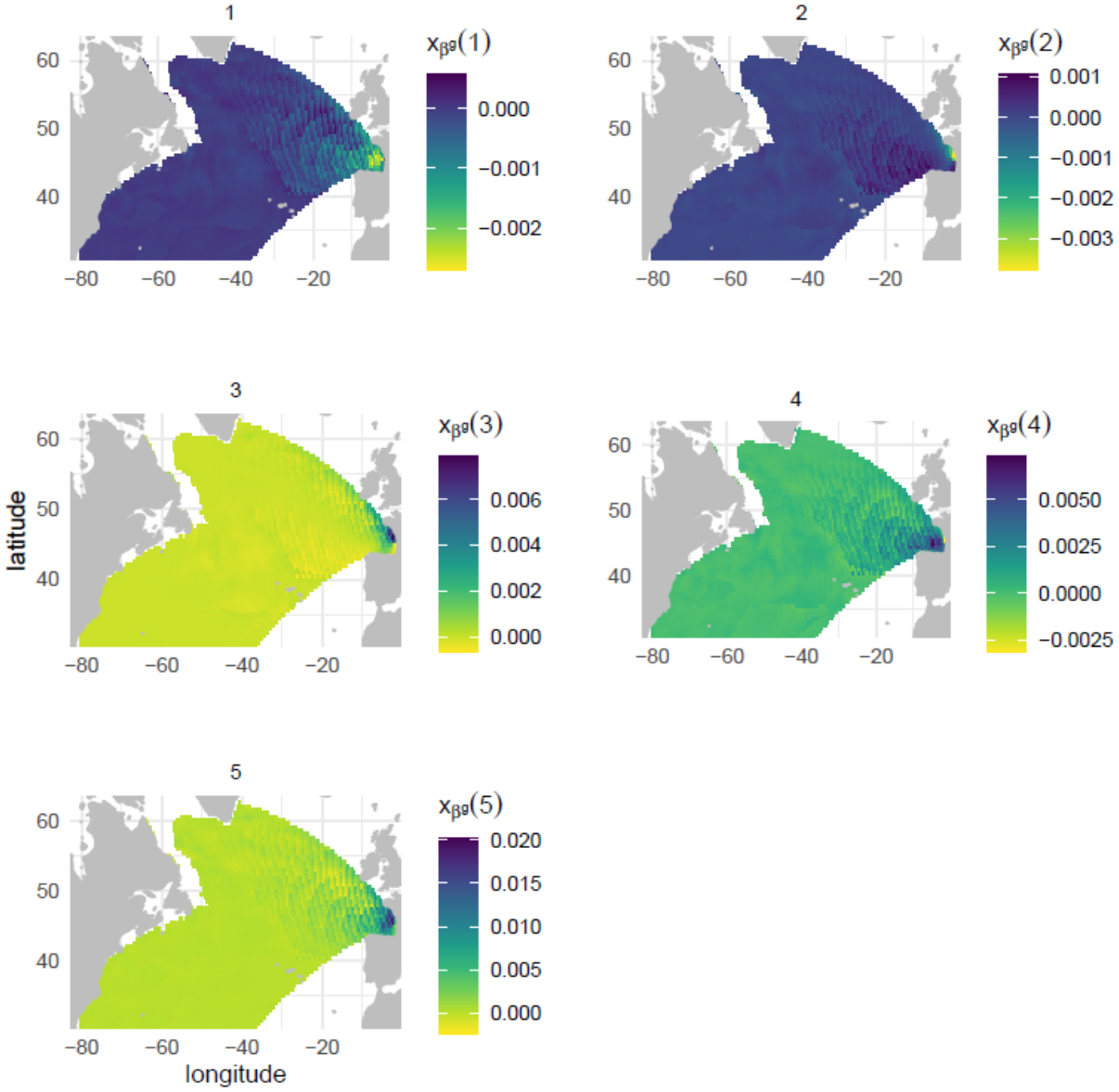

Figure 3.13 – Mean of $X_{\beta^{(g)}}$ minus the global mean for the five WTs.

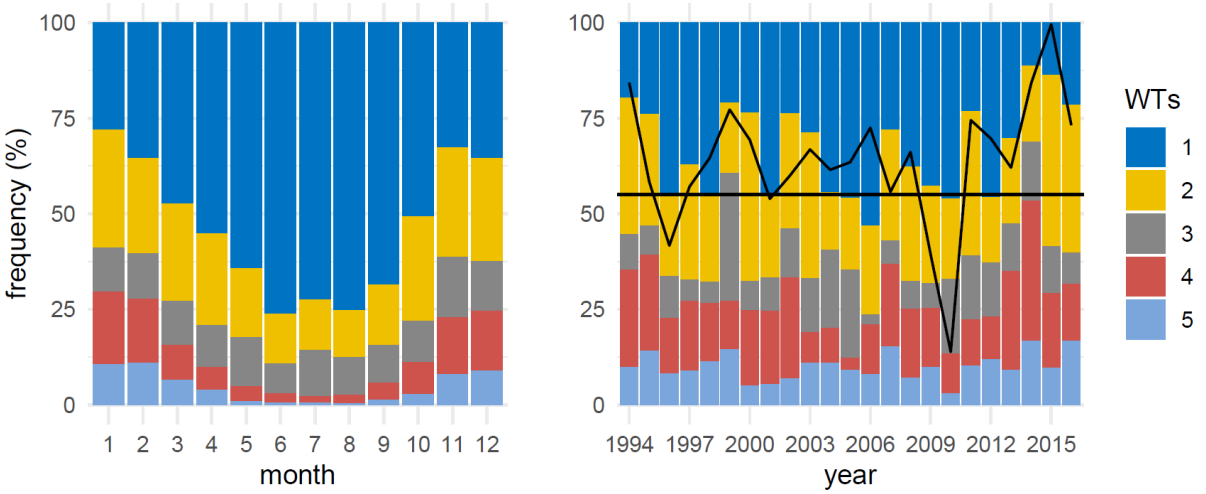

Figure 3.14 – Monthly and annual (in December-January-February) frequency occurrence of WTs in the calibration period. The continuous black line corresponds to the mean annual winter (DJF) time series of the NAO (North Atlantic Oscillation) index, and the horizontal black line indicates when NAO is less or greater than zero. When the continuous black line is below the horizontal line, the NAO is less than zero.

The monthly variability of WTs is shown in the left panel of figure 3.14. 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 long-term winter variability of frequency of occurrence of WTs is shown in the right panel of figure 3.14. The continuous black line corresponds to the mean annual winter of NAO index (Barnston and 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 to be related to the NAO index. For example, the winter of 2010 experienced fewer extreme waves, and the NAO index was less than zero. In contrast, the most extreme sea states were observed in 2014, where the NAO was greater than zero.

Figure 3.15 and 3.16 show results of model (3.14). 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 3.2, it appears that considering the global predictor is essential to explain the variability of H_s. Figure 3.17 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 consider source points that

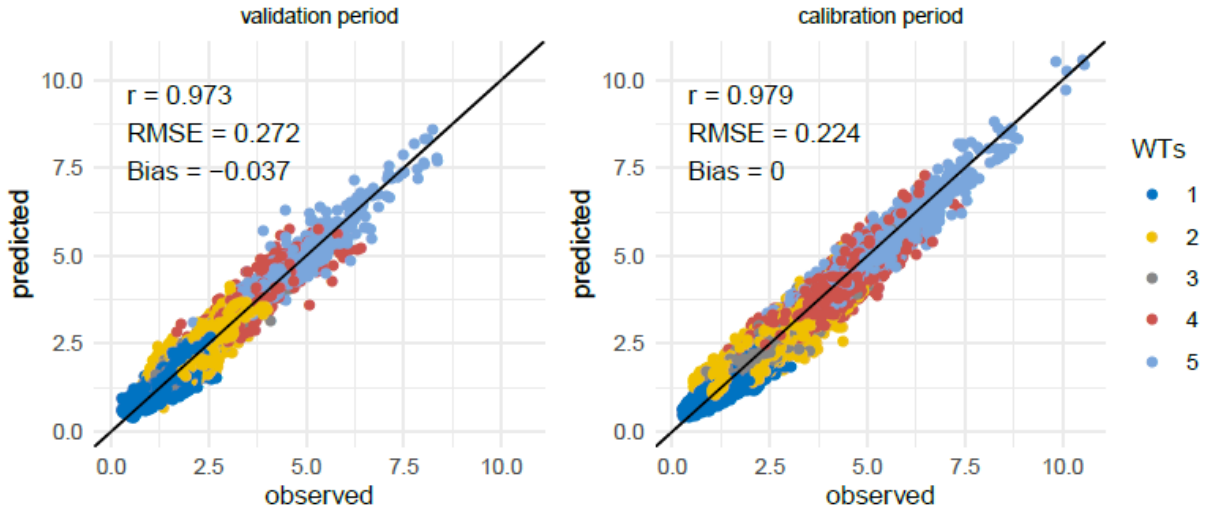

Figure 3.15 – Observed versus predicted values of H_s using the model (3.14) in the validation and calibration period.

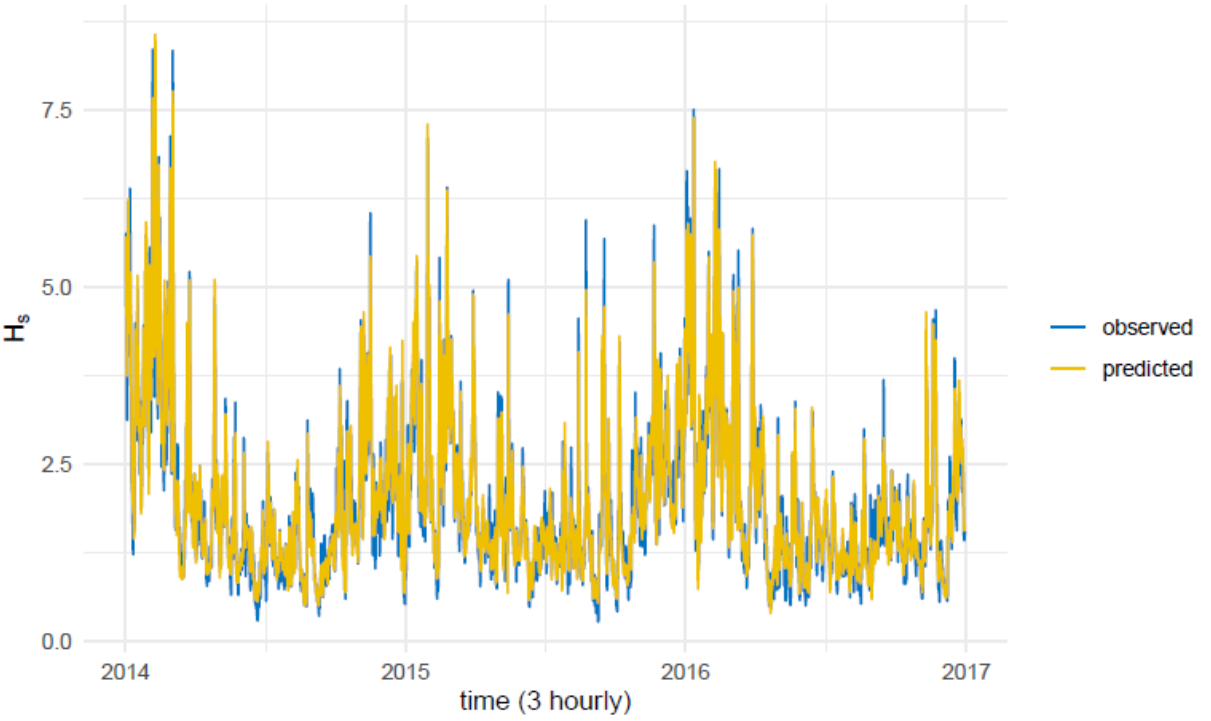

Figure 3.16 – Time series of observed and predicted values of H_s in the validation period.

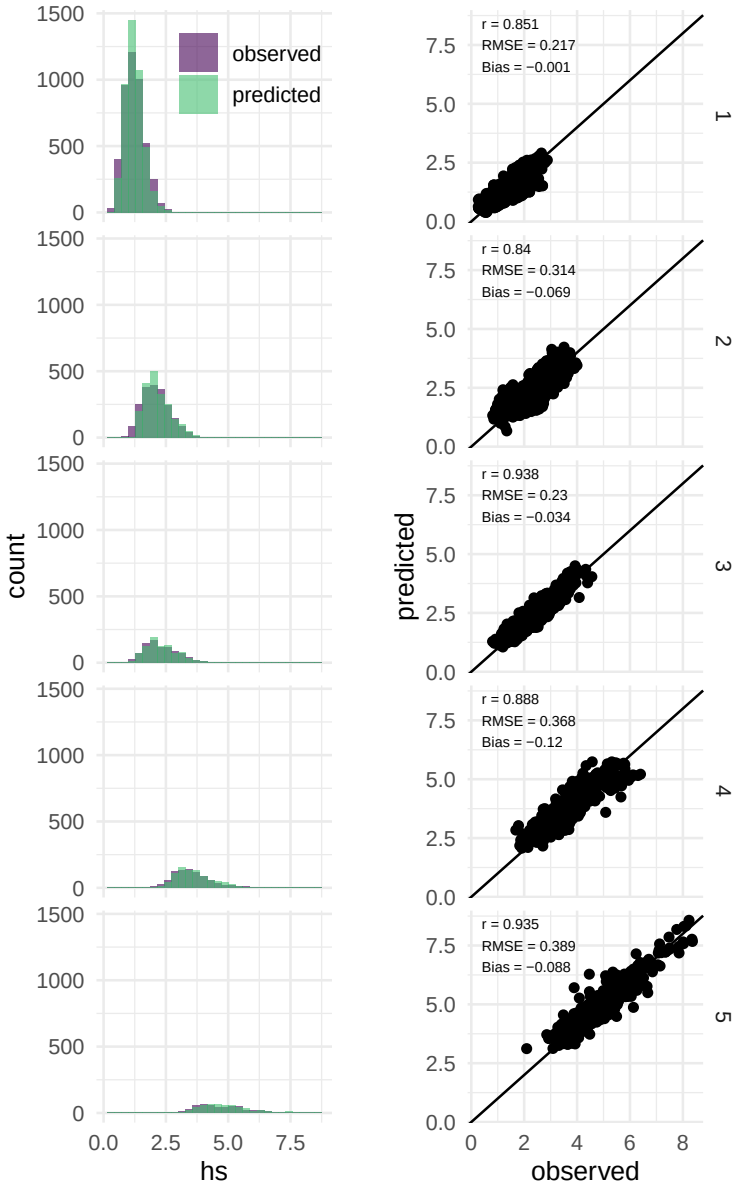

Figure 3.17 – Left panel: histogram of observed versus predicted H_s at each WT. Right panel: scatter plot of observed versus predicted H_s. Both in the validation period.

cover the swell generation, as seen in Figure 3.13. In contrast, in WT 4 and 5, the model considers mainly local source points as waves are mainly generated by local wind (Figure 3.13).

3.7 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, 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 predictors to predict H_s using a weather-types-based 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 chapter is based on observed weather types constructed using a clustering algorithm. As discussed in chapter 1, weather types can also be considered as latent variables and can be estimated using the EM (Expectation-Maximization) algorithm, where variables are evaluated based on the prediction of H_s, which can lead to optimal estimations.

Chapter 4

EM Algorithm for Generalized Ridge Regression with Spatial Covariates

Contents

Note: The results of this chapter are submitted for publication as S.Obakrim, P.Ailliot, V.Monbet, and N.Raillard, EM algorithm for generalized Ridge regression with spatial covariates[1].

4.1 Preface

In the previous chapter, we used Ridge regression to construct the link function between the North Atlantic wind conditions and H_s in the Bay of Biscay. It is clear that the

1. The preprint can be found in https://doi.org/10.48550/arXiv.2208.04754

use of Ridge regression allows us to deal with multicollinearity and improve our downscaling model's predictive ability. However, the regression model used in the previous chapter does not incorporate the fact that the covariates (wind conditions) exhibit a spatial structure. In this case, assuming that the regression coefficients also have a spatial structure is appropriate. This is usually done with the generalized Ridge (Wieringen, 2015) or the generalized LASSO (Tibshirani and Taylor, 2011), which allows incorporating any prior on the structure of the regression coefficients. However, these methods need the selection of the regularization hyper-parameters, which is usually done with cross-validation. In this chapter, we propose an Expectation-Maximization algorithm to estimate the parameters of generalized Ridge, focusing on spatial applications. The proposed method is applied to the problem of downscaling the significant wave height at the Bay of Biscay.

4.2 Abstract

The generalized Ridge penalty is a powerful tool for dealing with overfitting and for high-dimensional regressions. The generalized Ridge regression can be derived as the mean of a posterior distribution with a Normal prior and a given covariance matrix. The covariance matrix controls the structure of the coefficients, which depends on the particular application. For example, it is appropriate to assume that the coefficients have a spatial structure in spatial applications. This study proposes an expectation-maximization algorithm for estimating generalized Ridge parameters whose covariance structure depends on specific parameters. We focus on three cases: diagonal (when the covariance matrix is diagonal with constant elements), Matérn, and conditional autoregressive covariances. A simulation study is conducted to evaluate the performance of the proposed method, and then the method is applied to predict ocean wave heights using wind conditions.

4.3 Introduction

Consider an experiment where we have the data $\{y, X\}$, of n observations of a continuous variable Y and $n \times d$ matrix of covariates X. Suppose that Y is related to X via a linear model

$$Y = X\beta + \epsilon, \tag{4.1}$$

where β are model coefficients and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the model error. We suppose that the intercept is either included in β (so that the first column of X is a vector of 1) or that Y and X are centered. The least squares estimates are the best linear unbiased estimates of the parameters β. However, in the case of multicollinearity or high-dimensionality, penalized linear regression methods, like Ridge regression, are needed to control the variance. Ridge estimator of the problem (4.1) is

$$\hat{\beta}_{\lambda}^{Ridge} = \arg\min_{\beta} -\ell(\beta, \sigma^2) + \lambda \|\beta\|^2 \tag{4.2}$$

where λ is the regularization parameter and $\ell(\beta, \sigma^2)$ is the log-likelihood of the model (4.1). High values of λ permit to reduce the variance and increase the bias of the model. A good model should have a trade-off between variance and bias (Hastie et al., 2009). In order to find a trade-off between bias and variance, the hyperparameter λ needs to be selected.

Boonstra, Mukherjee, and Taylor (2015) classified methods for selecting λ into goodness-of-fit-based and likelihood-based methods. Goodness-of-fit-based methods define a goodness of fit criterion (such as the mean squared error) and minimize it in terms of λ. The most common goodness-of-fit-based method is the k-fold cross-validation which consists of partitioning observations into k groups and estimating β k times for each λ leaving out one group. For each λ, a goodness of fit score is calculated, and λ with the maximum score value is chosen. The typical choice of k is 5 and 10, while setting $k = n$ leads to leave-one-out cross-validation (LOOCV). LOOCV leads to a better estimation of λ; however, it is computationally expensive given that it requires fitting the model n times (Patil et al., 2021). Generalized cross-validation (GCV) (Golub, Heath, and Wahba, 1979) is an approximation of LOOCV that does not require fitting n models. GCV uses a weighted version of the predicted residual error sum of squares (PRESS) statistic (Allen, 1974) as a goodness of fit criterion. One of the problems with goodness-of-fit-based methods is the selection of the grid of λ, which influences the estimation.

Assuming that $Y|\beta \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, Ridge regression can be derived as the mean of a posterior distribution with the prior $\beta \sim \mathcal{N}(0_d, \sigma^2\lambda^{-1} I_d)$ (Wieringen, 2015) and as in Bayesian hierarchical linear regression, likelihood-based methods maximize the likelihood with respect to σ^2 and λ using for instance an iterative method (Lee and Nelder, 1996). Unlike goodness-of-fit-based methods, the advantage of likelihood-based approaches is, on the one hand, that they do not require grid selection for the regularization parameters.

On the other hand, likelihood-based methods can be generalized to consider any form of prior for the coefficients β. In some applications, the regression coefficients can be penalized differently, or a joint penalization of the coefficients is required. For example, in spatial statistics, where predictors have a spatial structure, it is reasonable to suppose that coefficients have a spatial structure. To do that, the generalized Ridge (Wieringen, 2015) can be used. Generalized Ridge extends the equation (4.2) by replacing the term $\lambda\|\beta\|^2$ to $\beta^T\Delta\beta$, where Δ is called the penalty matrix. In general, Δ depends on some regularization parameters (see, e.g., Goeman (2008) and Hemmerle (1975)); however, when the number of the regularization parameters is greater than 1, goodness-of-fit-based methods struggle with the problem of combinatorial explosion. Generalized Ridge in the hierarchical linear model framework, is equivalent to suppose that $\beta \sim \mathcal{N}(0_d, \Sigma_\theta)$ where Σ_θ is a covariance matrix that depends on some parameters θ. Note that Σ_θ corresponds to the inverse of the penalty matrix Δ. The classical Ridge is a special case of this model when the covariance matrix Σ_θ is diagonal, and θ is the usual regularization parameter λ.

Considering β as a hidden variable, Bishop and Nasrabadi (2006) proposed an expectation-maximization (EM) algorithm to find the maximum likelihood estimation (MLE) of parameters of a Bayesian linear regression model. The EM algorithm (Dempster, Laird, and Rubin, 1977b) is a method for estimating the parameters of a model with hidden variables. The EM algorithm alternates between two steps: the expectation and maximization steps. The E-step calculates the conditional expectation of the log-likelihood given the observations and current parameters. In the M-step, the parameters are estimated by maximizing the conditional expectation of the log-likelihood calculated in the E-step. In this study, we extend the algorithm in Bishop and Nasrabadi (2006) and propose an EM algorithm to estimate the parameters of hierarchical linear regression when $\beta \sim \mathcal{N}(0, \Sigma_\theta)$. At first, we study the case where Σ_θ is diagonal with constant elements, which corresponds to the classical Ridge in equation (4.2) and the problem studied by (Bishop and Nasrabadi, 2006). Then, we consider the case where the coefficients β have a spatial structure, especially when Σ_θ is the Matérn or the conditional autoregressive (CAR) covariance. A simulation study is done to assess the performance of the method. Then, the proposed method is applied to oceanographic data where the response variable represents a wave parameter in a location in the Bay of Biscay, and X represents wind conditions over the North Atlantic (Obakrim et al., 2022b).

This paper is organized as follows. The proposed method and its special cases are presented in Section 2. Then, a simulation study is conducted in Section 3 to assess

the performance of the proposed method. In section 4, we apply the methodology to oceanography data. Finally, this study is concluded in Section 5.

4.4 Proposed method

As stated in the introduction, Ridge regression can be viewed as a hierarchical linear model where $\beta \sim \mathcal{N}(0_d, \sigma^2\lambda^{-1}I_d)$. When there is a structure on the coefficients, it is unreasonable to consider all possible covariance functions as possible candidates for β. Therefore, we suppose that the covariance of β depends on some parameters θ, so that $\beta \sim \mathcal{N}(0_d, \Sigma_\theta)$. This motivates using the EM algorithm to find the maximum likelihood estimation of the parameters, where the model parameters are then $\Theta = (\sigma^2, \theta)$. The proposed method is described in this section, and three special cases of the covariance Σ_θ (the diagonal, Matérn, and CAR) are studied.

4.4.1 EM algorithm for generalized Ridge

Consider the linear model (4.1) and assume that β is a latent variable that follows a normal distribution. We define the regression model hierarchically as

$$\begin{aligned} \beta &\sim \mathcal{N}(0_d, \Sigma_\theta) \\ Y \mid \beta, \Theta &\sim \mathcal{N}(X\beta, \sigma^2 I_n) \end{aligned} \tag{4.3}$$

where $\Theta = (\sigma^2, \theta)$. Note that for simplicity, we assume that the mean of β is zero. The EM algorithm for the case where β has a non-zero mean will be presented in the Appendix.

Given a sample $y = (y_1, ..., y_n)$, the complete log-likelihood is expressed as

$$\begin{aligned} \ln p(y, \beta; \Theta) &= \ln p(y \mid \beta; \sigma^2) + \ln p(\beta; \theta) \\ &= -\frac{1}{2}\left(d\ln(2\pi) + \ln(|\Sigma_\theta|) + \beta^T\Sigma_\theta^{-1}\beta + n\ln(2\pi) + n\ln(\sigma^2) + \frac{1}{\sigma^2}\|y - X\beta\|^2\right) \end{aligned} \tag{4.4}$$

Maximum likelihood estimation consists of maximizing (4.4) with respect to the parameters Θ. This is usually done with the Expectation-Maximization algorithm in the latent variable context. The EM algorithm alternates between the E-step and M-step. In the E-step, the expectation $Q(\Theta|\Theta^{(t)})$ of the complete likelihood with respect to the posterior distribution of the latent variable β and the parameters $\Theta^{(t)}$ from the previous iteration

t is calculated. In the M-step, the quantity $Q(\Theta|\Theta^{(t)})$ is maximized with respect to the parameters Θ.

The E-step and M-step are defined as follows

- E-step:

$$Q(\Theta|\Theta^{(t)}) = \mathbb{E}(\ln p(y, \beta; \Theta) \mid y, \Theta^{(t)}). \tag{4.5}$$

The posterior distribution of the latent variable β is a normal distribution with mean $\mu_{\beta|y}$ and covariance matrix $\Sigma_{\beta|y}$ such that

$$\begin{cases} \Sigma_{\beta|y} = (\Sigma_\theta^{-1} + \frac{1}{\sigma^2} X^T X)^{-1} \\ \mu_{\beta|y} = (X^T X + \sigma^2 \Sigma_\theta^{-1})^{-1} X^T y. \end{cases} \tag{4.6}$$

Note that $\mu_{\beta|y}$ defined in (4.6) is a generalized Ridge estimator (see e.g. Wieringen (2015)) solution of the optimization problem

$$\mu_{\beta|y} = \arg\min_\beta \frac{\|y - X\beta\|^2}{\sigma^2} + \beta^T \Sigma_\theta^{-1} \beta \tag{4.7}$$

Therefore,

$$Q(\Theta|\Theta^{(t)}) = -\frac{1}{2}\left(\ln(|\Sigma_\theta|) + \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)})) + \ln(\sigma^2) + \frac{1}{\sigma^2}\mathbb{E}(\|y - X\beta\|^2 \mid y, \Theta^{(t)})\right) + C \tag{4.8}$$

where C is a constant and

$$\begin{cases} \mathbb{E}(\beta\beta^T|y; \Theta^{(t)}) = \Sigma_{\beta|y} + \mu_{\beta|y}\mu_{\beta|y}^T \\ \mathbb{E}(\|y - X\beta\|^2|y; \Theta^{(t)}) = \|y\|^2 - 2y^T X \mu_{\beta|y} + \mathrm{Tr}(X^T X \mathbb{E}(\beta\beta^T|y; \Theta^{(t)})) \end{cases} \tag{4.9}$$

- M-step:

The maximization step computes

$$\Theta^{(t+1)} = \arg\max_\Theta Q(\Theta|\Theta^{(t)}) \tag{4.10}$$

which leads to the following updates of the parameters σ^2 and θ

$$\begin{aligned} \sigma^{2,(t+1)} &= \frac{1}{n}(\|y\|^2 - 2y^T X \mu_{\beta|y} + \mathrm{Tr}(X^T X \mathbb{E}(\beta\beta^T|y; \Theta^{(t)}))) \\ \theta^{(t+1)} &= \arg\max_\theta \ \ln(|\Sigma_\theta^{-1}|) - \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)})) \end{aligned} \tag{4.11}$$

4.4.2 Special cases

The M-step in equation (4.11) requires the maximization of $Q(\Theta|\Theta^{(t)})$ over the parameters of the covariance Σ_θ. In this study, we will explore three cases. First, we consider the case where Σ_θ is diagonal. Then, the case where β has a spatial structure, especially when the parametric covariance is the Matérn covariance function. Finally, we consider the conditional autoregressive model (CAR).

4.4.2.1 Diagonal case

In the classical Ridge, the covariance matrix of the coefficients β is supposed to be diagonal such that

$$\Sigma_\theta = \sigma_\beta^2 \mathbf{I_d}. \tag{4.12}$$

The M-step of the covariance in (4.11) becomes

$$\sigma_\beta^{2,(t+1)} = \arg\max_{\sigma_\beta^2} \; -d\ln(\sigma_\beta^2) - \frac{1}{\sigma_\beta^2}\mathrm{Tr}(\mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)})). \tag{4.13}$$

Setting the derivatives with respect to σ_β^2 to zero, we obtain the M-step

$$\sigma_\beta^{2,(t+1)} = \frac{\mathrm{Tr}(\mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)}))}{d}. \tag{4.14}$$

Note that $\frac{1}{\sigma_\beta^2}$ corresponds to the regularization parameter λ in equation (4.1). As stated in the introduction, Ridge regression requires the selection of the regularization parameter. Therefore, the EM algorithm can be an alternative to cross-validation for estimating Ridge coefficients along with the regularization parameter. A comparison of the two methods (cross-validation and EM algorithm) is given in the Appendix.

4.4.2.2 Spatial covariance functions

In spatial statistics applications, one may assume that β has a spatial structure. One way to do that is to assume that β has a parametric covariance function. There are many choices of covariance functions that are widely used for Gaussian processes and kriging (Schulz, Speekenbrink, and Krause, 2018). In this study, we focus on the stationary Matérn

covariance, which has the form

$$K(h;\phi,\kappa) = \frac{\sigma_\beta^2}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{h}{\phi}\right)^{\kappa} K_\kappa\left(\frac{h}{\phi}\right) \tag{4.15}$$

where h is the distance between two points, Γ is the Gamma function, and K_κ is the modified Bessel function (Abramowitz, Stegun, and Romer, 1988). The Matérn function is parameterized by the variance parameter σ_β^2, the range parameter ϕ, and the smoothness parameter κ. The range parameter ϕ controls the decay rate with distance, with larger values of ϕ corresponding to more strongly correlated variables, and the smoothness parameter κ controls the mean-square differentiability of the spatial process.

The M-step of the covariance of β in (4.12) becomes

$$(\sigma_\beta^{2,(t+1)}, \theta^{(t+1)}) = \arg\max_{\sigma_\beta^2,\theta} \ \ln(|R_\theta^{-1}|) - d\ln(\sigma_\beta^2) - \frac{1}{\sigma_\beta^2}\mathrm{Tr}(R_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Phi^{(t)})) \tag{4.16}$$

where R_θ is the Matérn correlation and $\theta = (\phi, \kappa)$. Since the variance parameter is constant and following Bachoc (2013), the optimization of the variance parameter σ_β^2 can be carried out separately with the correlation parameters ϕ and κ. Therefore,

$$\begin{aligned} \sigma_\beta^{2,(t+1)} &= \frac{\mathrm{Tr}(R_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Phi^{(t)}))}{d} \\ \theta^{(t+1)} &= \arg\max_{\theta} \ \ln(|R_\theta^{-1}|) - d\ln(\mathrm{Tr}(R_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Phi^{(t)}))). \end{aligned} \tag{4.17}$$

The solution to the optimization problem in equation (4.17) cannot be done analytically; therefore, numerical optimization algorithms are used. This study uses the quasi-Newton method L-BFGS-B to optimize the parameters. Given the difficulties in estimating Matérn parameters (Kaufman and Shaby, 2013), we a priori fix the smoothness parameter as $\frac{3}{2}$, which gives the classical $\frac{3}{2}$-Matérn covariance function.

4.4.2.3 Conditional autoregressive model

The M-step in equation (4.10) requires the inversion of the covariance matrix, which can be challenging for large matrices. This problem is widely discussed in Gaussian processes literature (Ambikasaran et al., 2015; Storkey, 1999). Therefore, it can be numerically advantageous to parameterize the precision matrix (inverse of the covariance matrix) instead of the covariance matrix. This is motivated by the fact that the precision

matrix $P_\theta = \Sigma_\theta^{-1}$ can be approximated by a sparse matrix (Tajbakhsh, Aybat, and Del Castillo, 2020). In fact, the off-diagonal elements of the precision matrix correspond to the conditional covariance between two variables given the remaining variables. Therefore, conditionally independent variables have zero values in the precision matrix.

Gaussian Markov random fields (GMFs) are widely used in spatial statistics (Cressie and Wikle, 2015). GMFs models have a Markov property making them computationally and theoretically suitable (Rue, 2001). Furthermore, (Rue and Tjelmeland, 2002) demonstrated that a GMF model can approximate a Gaussian field with a Matérn correlation function and other families of correlation functions. Conditional autoregressive (CAR) models are classes of GMFs with well-defined joint Gaussian distribution (Cressie and Kapat, 2008). This subsection will study cases where the coefficients β have the CAR model property. The joint distribution of a CAR is expressed as

$$\beta \sim \mathcal{N}(0, \tau^2(I_d - \alpha H)^{-1}\Phi). \tag{4.18}$$

The distribution of β depends on unknown parameters α and τ^2, and many types of CAR models depend on the choice of the matrix H and Φ. Following (Besag, York, and Mollié, 1991), in this study, we consider the Weighted CAR (WCAR) model where

$$\Phi = \text{diag}(|N_1|^{-1}, ..., |N_d|^{-1}) \tag{4.19}$$

where $|N_i|$ is the number of neighbors of location i and $H = \left(\frac{a_{ij}}{|N_i|}\right)_{d\times d}$; $i, j = 1, ..., d$, where a_{ij} is the (i, j) element of the adjacency matrix $A = (a_{ij})_{d\times d}$, where $a_{ij} = a_{ji} = 1$ if and only if location i and j are neighbors and otherwise $a_{ij} = 0$. Putting $P_\theta = \tau^{-2}(I_d - \alpha H)\Phi^{-1}$, the second part of the M-step in the equation (4.11) becomes

$$\theta^{(t+1)} = \arg\max_\theta \ \ln(|P_\theta|) - \text{Tr}(P_\theta \mathbb{E}(\beta\beta^T \mid y, \Phi^{(t)})) \tag{4.20}$$

where $\theta = (\tau^2, \alpha)$.

As for the Matérn covariance, the solution to the optimization problem (4.20) cannot be done analytically, and the numerical optimization algorithm L-BFGS-B is used. Note that the optimization of the variance parameter τ^2 can also be carried out separately with the parameter α.

Remark that this leads to a spatial extension of the fused Ridge method proposed in

(Goeman, 2008). When $\alpha = 1$, we obtain

$$\frac{1}{\tau^2}\beta^T\Phi^{-1}(I_d - \alpha H)\beta = \frac{1}{2\tau^2}\sum_{(i,j)|a_{ij}=1}(\beta_i - \beta_j)^2. \tag{4.21}$$

This shows that any spatial coefficient variations will be penalized when solving (4.7). In this case, replacing the L2 norm with the L1 norm leads to the fused LASSO method proposed in (Tibshirani et al., 2005). However, the matrix $(I_p - \alpha H)$ is semi-positive definite when $\alpha = 1$ and thus Σ_θ is degenerate. Hereafter we impose the constraints $|\alpha| < 1$ to ensure that the precision matrix is positive definite. Another strategy would consist of adding a regular Ridge penalty (e.g., the discussion in Wieringen (2015)).

4.5 Simulation study

In this section, a simulation study is conducted to assess the performance of the proposed method for estimating model parameters for the three cases: diagonal, Matérn, and CAR.

4.5.1 Setup

This study focuses on using the proposed method for spatial applications. Therefore, we consider a 15×15 regular spatial grid in a square domain $[1, 15]^2$ where each location j has a covariate x_j. We generate $X = (x_{ij})_{n\times d}$ of n independent and identically distributed observations from a multivariate normal distribution with zero mean and a Matérn covariance with some arbitrary parameters $(\sigma_x^2, \phi_x, \kappa_x) = (6, 2, 3/2)$. Then, the coefficients β, kept the same for all observations, are simulated using either the diagonal, Matèrn, or CAR case. Finally, for a given σ^2, Y is simulated from the normal distribution according to equation (4.3).

The parameters chosen for each case are:

- Diagonal: $\sigma^2 = 36$ and $\sigma_\beta^2 = 7$
- Matérn: $\sigma^2 = 36$, $\sigma_\beta^2 = 0.1$ and $\phi = 4$
- CAR: $\sigma^2 = 36$, $\tau^2 = 1$ and $\alpha = 0.9$

The parameters are chosen so that the results of the three methods are comparable. For the CAR model, we consider four neighbors to construct the adjacency matrix, and we chose $\alpha = 0.9$ to sufficiently smooth the resulting coefficients.

The EM algorithm is initialized with an arbitrary set of parameters, and the E-step and M-step are repeated until no further improvement can be made to the likelihood value or to limit the computational cost until a maximum number of iterations is reached. The computation time for one iteration on an i5-7500 CPU and 16Go computer is 0.16, 3, and 1.8 seconds for diagonal, Matérn, and CAR, respectively.

4.5.2 Results

At first, one simulation is done for each case (diagonal, Matérn, and CAR) with $n = 800$. The parameters are estimated using the EM algorithm presented in the previous section. Figure 6.1 shows the first simulation results. Left panels correspond to the true β, and right panels correspond to the estimated β using the EM algorithm. For all the cases, the EM algorithm does well in estimating the parameters, especially the variance σ^2.

To assess the influence of the sample size on the estimations, for each case, we perform 100 independent random simulations for each sample size varying from 50 to 850. For each simulation, the EM algorithm is used to estimate the parameters. Figure 6.2 shows the normalized root mean square error $NRMSE_\beta$ and $NRMSE_y$ for the three cases where

$$\begin{aligned} NRMSE_\beta &= \frac{\sqrt{\frac{1}{d}\sum_j^d (\beta_j - \hat{\beta}_j)^2}}{\hat{\sigma_\beta}} \\ NRMSE_y &= \frac{\sqrt{\frac{1}{n'}\sum_i^{n'} (y_i - \hat{y}_i)^2}}{\hat{\sigma}_y} \end{aligned} \tag{4.22}$$

where $\hat{\beta}_j$ and $\hat{y}_i$ are the estimated β_j and y_i and $\hat{\sigma_\beta}$ and $\hat{\sigma}_y$ are the sample standard deviation of β and y, respectively. $NRMSE_y$ is calculated in a test set (which is not used in the estimation) of size $n' = \frac{n}{2}$. For the three cases, $NRMSE_\beta$ and $NRMSE_y$ decrease as the sample size increases.

To evaluate the parameter estimates, we compare the EM estimates with the maximum likelihood estimates of the parameters, hereafter referred to as MLE, knowing the true β. More precisely, the MLE estimates are defined as

$$\Theta_{\text{MLE}} = \arg\max_{\Theta} -\frac{1}{2}\left(\ln(|\Sigma_\theta|) + \beta_{\text{true}}^T \Sigma_\theta^{-1} \beta_{\text{true}} + n\ln(\sigma^2) + \frac{1}{\sigma^2}\|y - X\beta_{\text{true}}\|^2\right) + C \tag{4.23}$$

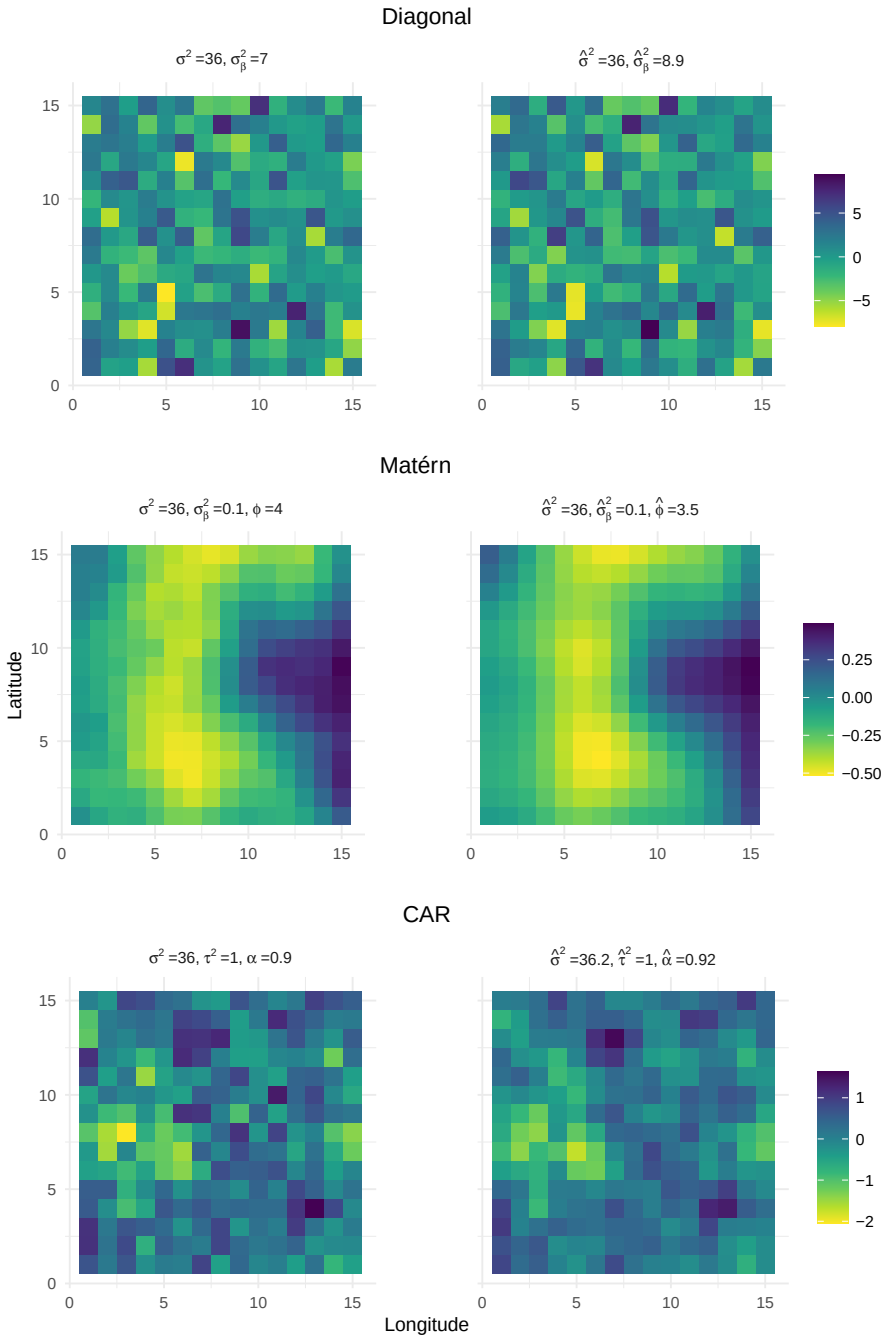

Figure 4.1 – Simulation results for the three cases (diagonal, CAR, and Matérn). The left panels correspond to the true β coefficients with the true parameters given in section 3.1, and the right panels correspond to the β estimated when the sample size $n = 800$.

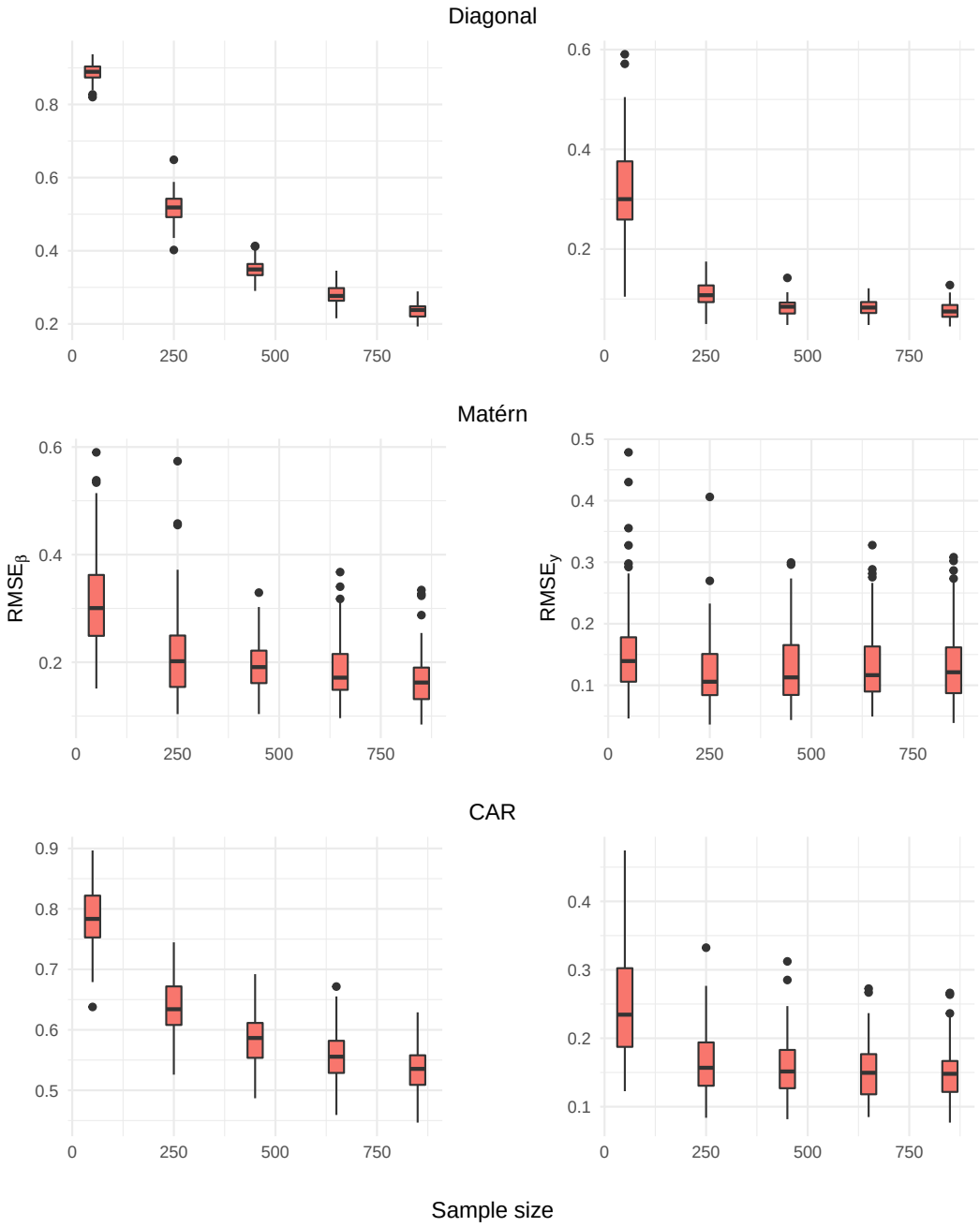

Figure 4.2 – Results of $RMSE_\beta$ (left panels) and $RMSE_y$ (right panels) for the diagonal, CAR, and Matérn case as a function of the sample size varying from 50 to 850.

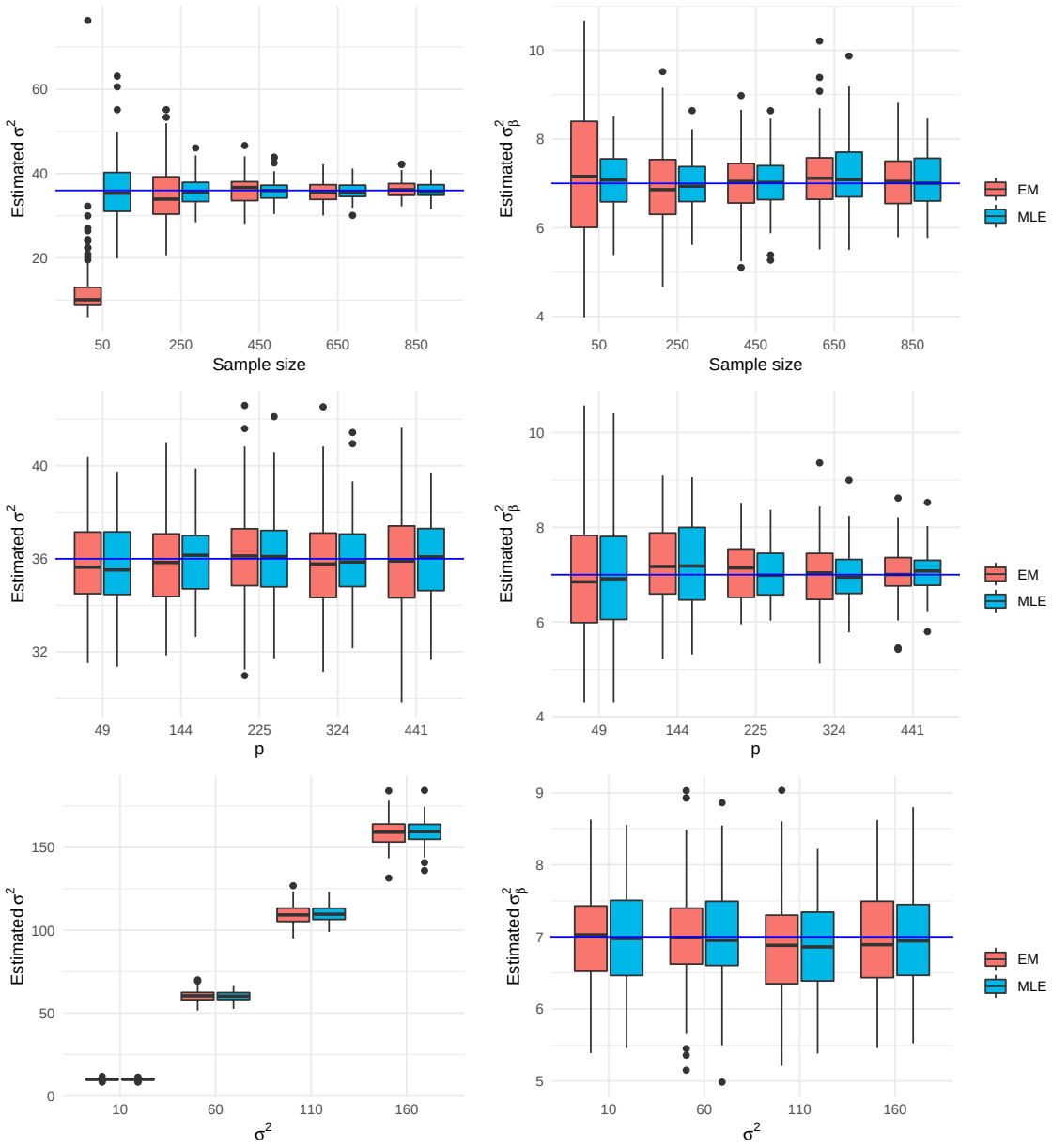

Figure 4.3 – Estimated parameters in the case where the covariance of β is diagonal as a function of the sample size, the dimension of X, d, and the variance σ^2. Red boxes correspond to EM estimates and the blue ones to MLE estimates. The blue line corresponds to the true value of the parameter σ^2 and σ^2_β, which are equal to 36 and 7, respectively.

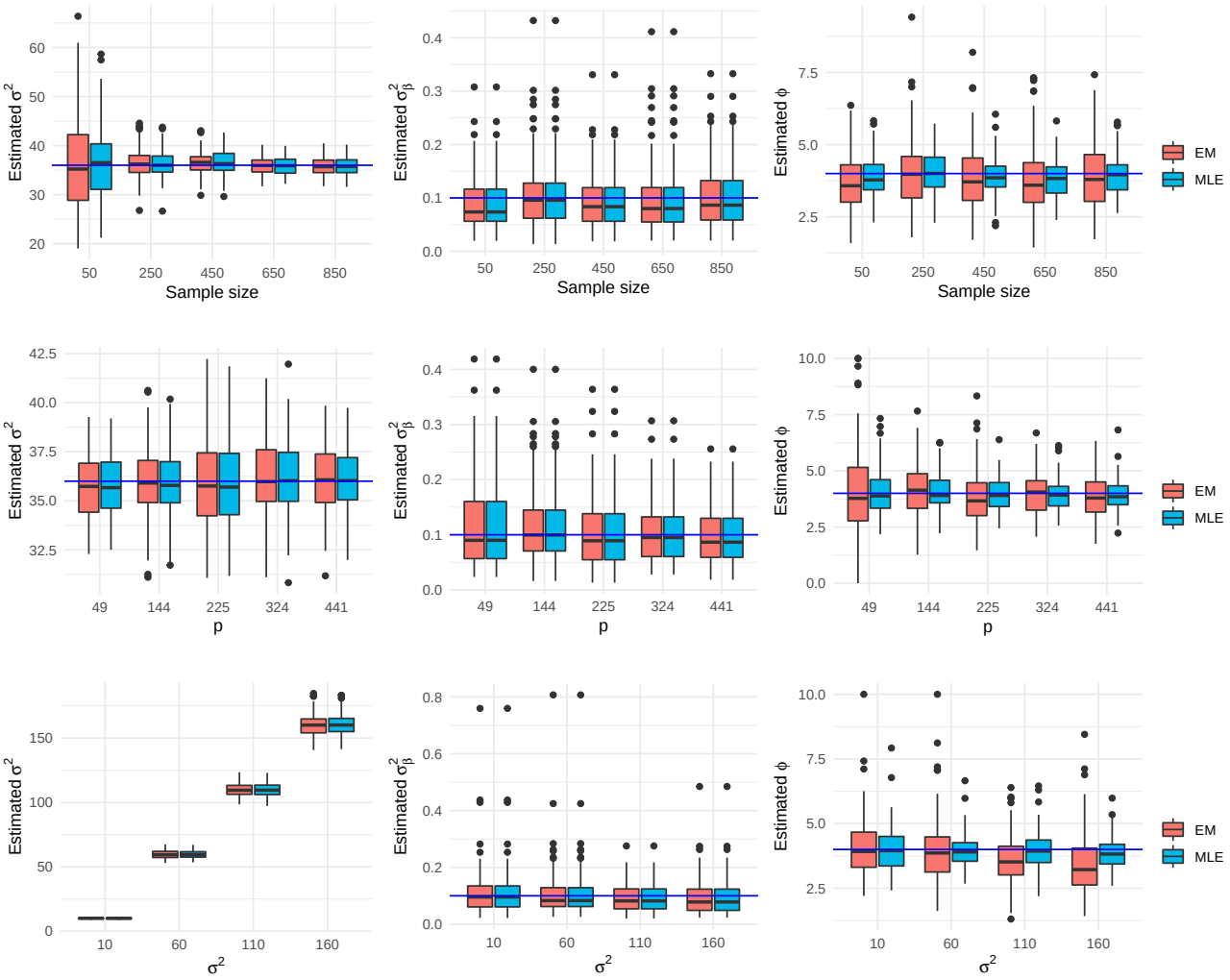

Figure 4.4 – Estimated parameters in the case where the covariance of β is the Matérn as a function of the sample size, the dimension of X, d, and the variance σ^2. Red boxes correspond to EM estimates and the blue ones to MLE estimates. The blue line corresponds to the true value of the parameter σ^2, σ^2_β, and ϕ, which are equal to 36, 0.1, and 4, respectively.

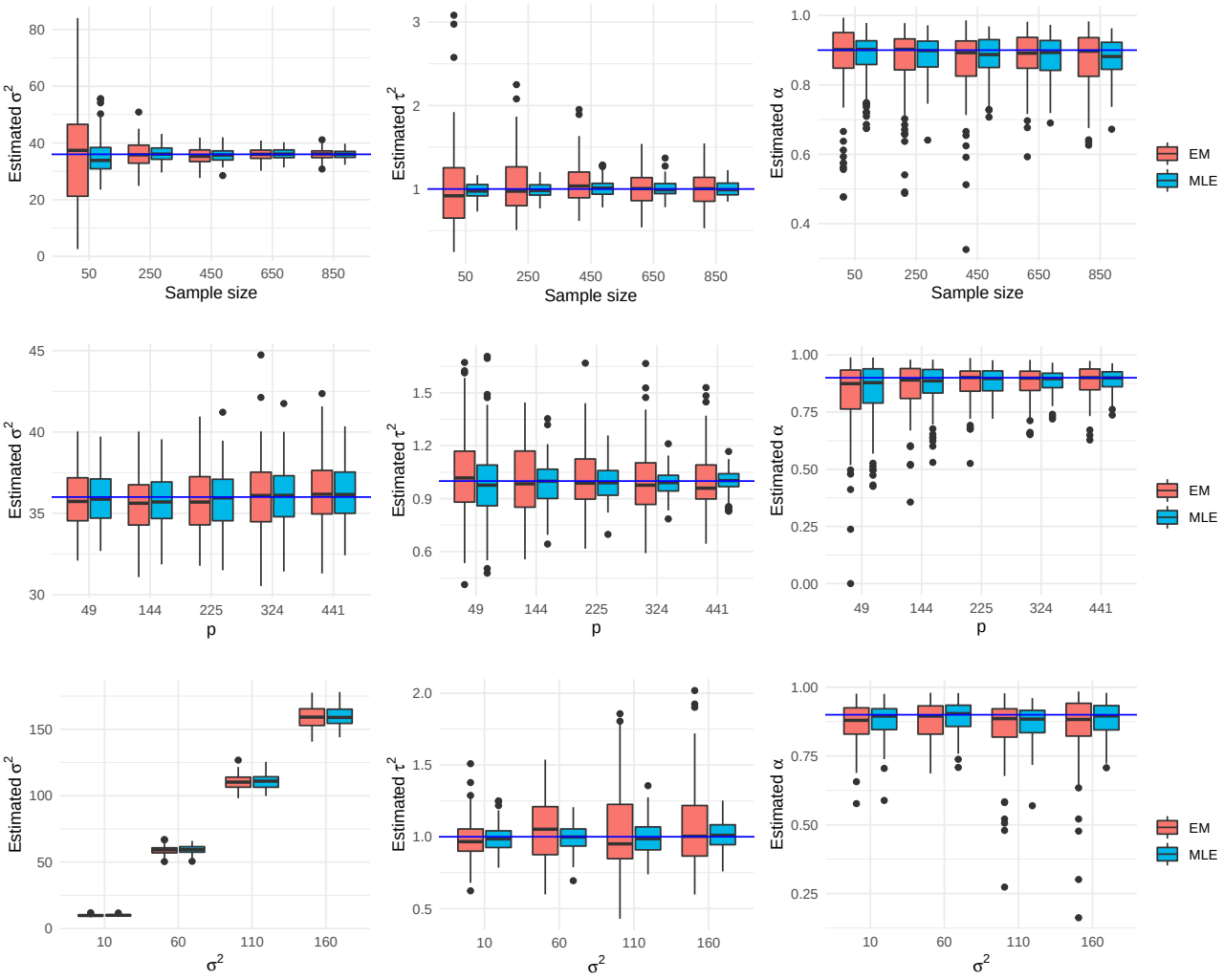

Figure 4.5 – Estimated parameters in the case where the covariance of β is the CAR as a function of the sample size, the dimension of X, d, and the variance σ^2. Red boxes correspond to EM estimates and the blue ones to MLE estimates. The blue line corresponds to the true value of the parameter σ^2, σ^2_β, and α, which are equal to 36, 1, and 0.9, respectively.

where β_{true} is the true β simulated for each case with the parameters given in section 3.1. Along with the sample size, we are also interested in how the estimates behave when varying the dimension of X, d, and the variance parameter σ^2. Note that in practice, Θ_{MLE} cannot be found directly, given that the true β is not observed (latent variable). Therefore, we expect the EM algorithm to provide less accurate estimates than MLE. However, we expect that by varying the sample size, the dimension, and the variance σ^2, the estimations asymptotically will be close to MLE estimates.

Figures 6.3, 6.4 and 6.5 show boxplots of EM (red) and MLE (blue) estimates for the diagonal, Matérn and CAR cases as a function of sample size, dimension d, and variance σ^2. For the diagonal case, the estimate of σ^2 seems to converge to the true value of the parameter (blue line) when the sample size n increases as it does in the usual linear regression model. Note that the estimate of the spatial variance σ_β^2 does not seem to converge to the true value of the parameter as the sample size increases, but when n is large enough, EM and MLE seem to provide similar results. This is not unexpected since both methods are based on a single sample of the d-dimensional field β. As expected, the dimension d also affects the estimate of the parameter σ_β^2, which converges towards the true value as d increases; however, no significant change is observed for σ^2 when d increases. The effect of the variance σ^2 on the estimation of σ_β^2 is small, and we observe that for σ^2 larger than 100, the EM and MLE tend to underestimate σ_β^2. Similar behavior can be observed for the Matérn case: the variance parameter σ^2 seems to converge towards the actual value with increasing sample size. However, there is no significant change in the other parameters (the variance σ_β and the range ϕ). The dimension d mainly influences the parameters σ_β and ϕ, which describe the spatial structure of the d-dimensional field β, and as d increases, the estimates converge to the actual values. As for the diagonal case, the EM algorithm underestimates the parameters σ_β and ϕ when the variance σ^2 increases. Finally, for the CAR case, the sample size influences the parameters σ^2 and τ^2, but only slightly the correlation parameter α, which is mainly influenced by the dimension d. The variance σ^2 has a significant influence on τ^2, but only a small one on α. To summarize:

- The sample size n mainly influences the estimation of the variance of the residuals σ^2
- The parameters which describe the spatial structure of β are mainly influenced by the dimension d
- As the variance σ^2 increases, EM underestimates the parameter σ_β^2 of the diagonal and Matérn case, and the range parameter ϕ

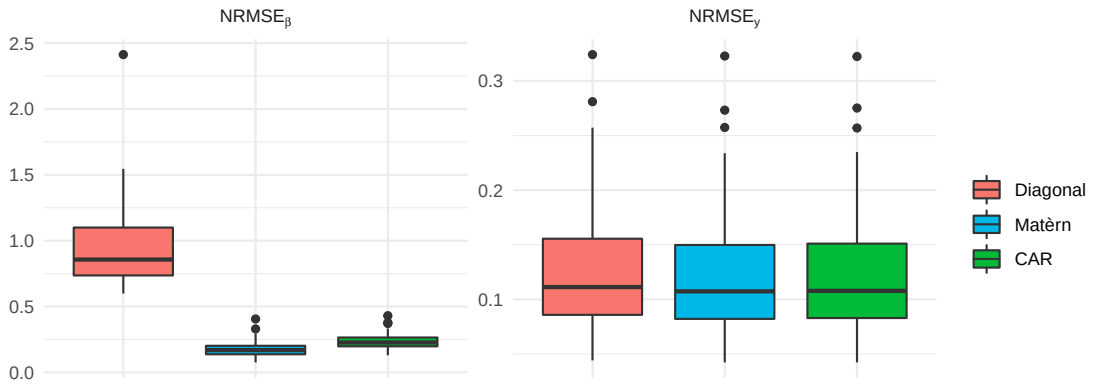

Figure 4.6 – Results of the estimations when the true beta is simulated from Matérn with the parameters $\sigma^2 = 36$, $\sigma^2_\beta = 0.1$ and $\phi = 4$ and sample size $n = 800$. The left panel correspond to $NRMSE_\beta$ and the right one for $NRMSE_y$.

- EM estimates are close to MLE estimates in most cases when the sample size and the dimension d are large enough and the variance σ^2 is small

Another interesting aspect that needs to be studied is when the coefficients β are simulated using one covariance and estimated using another covariance model. To do that, we perform 100 independent simulations of β using the Matérn covariance function, and we estimate the parameters using the three cases: diagonal, CAR, and Matérn. Figure 4.6 shows the results of $NRMSE_\beta$ and $NRMSE_y$ of the experiment. It is clear that using the Matérn covariance for the estimation gives better results in terms of $NRMSE_\beta$. Not surprisingly, the diagonal case is the worst model for estimating the coefficients. However, in terms of $NRMSE_y$, there is a small difference between the three methods.

4.6 Application

The proposed method is applied to the problem of predicting the significant wave height (H_s) at a location in the Bay of Biscay using wind conditions over the North Atlantic (figure 4.7), where the significant wave height is the average height of the highest third of the waves, a key measure of wave height that provides information about wave energy. The data used for H_s comes from the Homere hindcast database (Boudière et al., 2013), and the wind data comes from Climate Forecast System Reanalysis (CFSR)

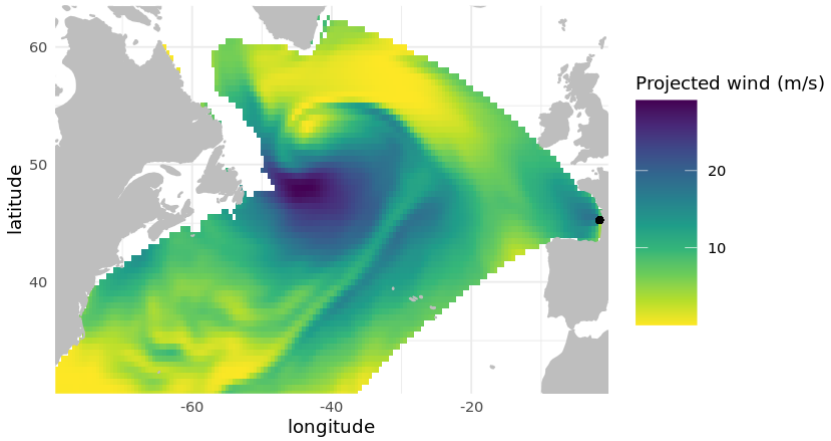

Figure 4.7 – CFSR projected wind in the North Atlantic in 1994-01-01 00h:00. The black point represents the target point.

(Saha et al., 2010b). The wind data are pre-processed before being used as a predictor (see (Obakrim et al., 2022b) for the pre-processing procedure). We consider 23 years of H_s and wind data from 1994 to 2016 with a temporal resolution of 3 hours.

The regression problem is of the form

$$H_s(t) = \sum_{j=1}^{d} X_j(t)\beta_j + \epsilon(t) \quad t = 1, ..., n \tag{4.24}$$

where $X_j(t)$ is the predictor at time t and location j defined as

$$X_j(t; t_j, \alpha_j) = \frac{1}{2\alpha_j+1} \sum_{i=t-t_j-\alpha_j}^{t-t_j+\alpha_j} W_j^2(i), \tag{4.25}$$
$$t_j + \alpha_j + 1 \leq t \leq t_j - \alpha_j + n$$

where W_j is the projected wind (figure 4.7) defined as

$$W_j = U_j \, cos\left(\frac{1}{2}(b_j - \theta_j)\right) \tag{4.26}$$

U_j is the wind speed, b_j is the great circle bearing, and θ_j is the wind direction at location j. α_j controls the length of the time window, and t_j is the mean travel time of waves which are estimated using the maximum correlation between H_s and the predictor

$$(\hat{t}_j, \hat{\alpha}_j) = arg \max_{t_j, \alpha_j} \Big(corr(H_s, X_j^g(t_j, \alpha_j)) \Big). \tag{4.27}$$

Method	r	RMSE(m)	bias(m)
Diagonal	0.941	0.414	-0.0004
Matérn	0.956	0.354	-0.04
CAR	0.957	0.352	-0.06

Table 4.1 – Quantitative comparison of the diagonal, Matérn, and CAR methods in the validation set using the correlation (r), root mean square error (RMSE), and bias.

Let $X = X_1, ..., X_d$ be the predictor which has the size 67088×5651. Since the predictor has a spatial structure. It is reasonable to assume that the coefficients β also have a spatial structure so that nearby locations have close contributions to the waves at the target point. This assumption is equivalent to suppose that $\beta \sim \mathcal{N}(0, \Sigma_\theta)$. For the covariance Σ_θ, we will consider the cases of Matérn and CAR. For comparison, we also consider the diagonal case even though it does not consider any structure between coefficients.

The model's parameters (equation 4.24) are estimated using data from 1994 to 2013, and the model is evaluated in terms of correlation, RMSE, and bias, using a validation set from 2014 to 2016. Figure 4.8 shows the results of estimating β and the covariance parameters using the EM algorithm when the covariance structure is assumed to be diagonal, Matérn and CAR. Not surprisingly, the coefficients estimated with the diagonal covariance show no physical spatial structure. Therefore, the assumption that close locations have close coefficients cannot be taken into account using the diagonal case. This motivates using the Matérn and CAR covariances. The Matérn and CAR covariances give the smoothest coefficients with a clear spatial structure. In addition, locations close to the target point have larger coefficients. Therefore, the obtained coefficients are more physically interpretable and take into account our assumption about the covariance. Note that the CAR method is less expensive numerically than the Matérn, which involves inverting the covariance matrix at each iteration of the optimization algorithm used in the M-step.

Table 6.1 shows the results of the quantitative comparison between the three methods for predicting significant wave height in the validation set using correlation (r), root mean square error (RMSE), and bias. In terms of correlation and RMSE, the diagonal method is the less accurate method. Therefore, adding the spatial structure in the covariance is advantageous in predicting the significant wave height. The CAR and Matérn methods lead to close results regarding r, RMSE, and bias.

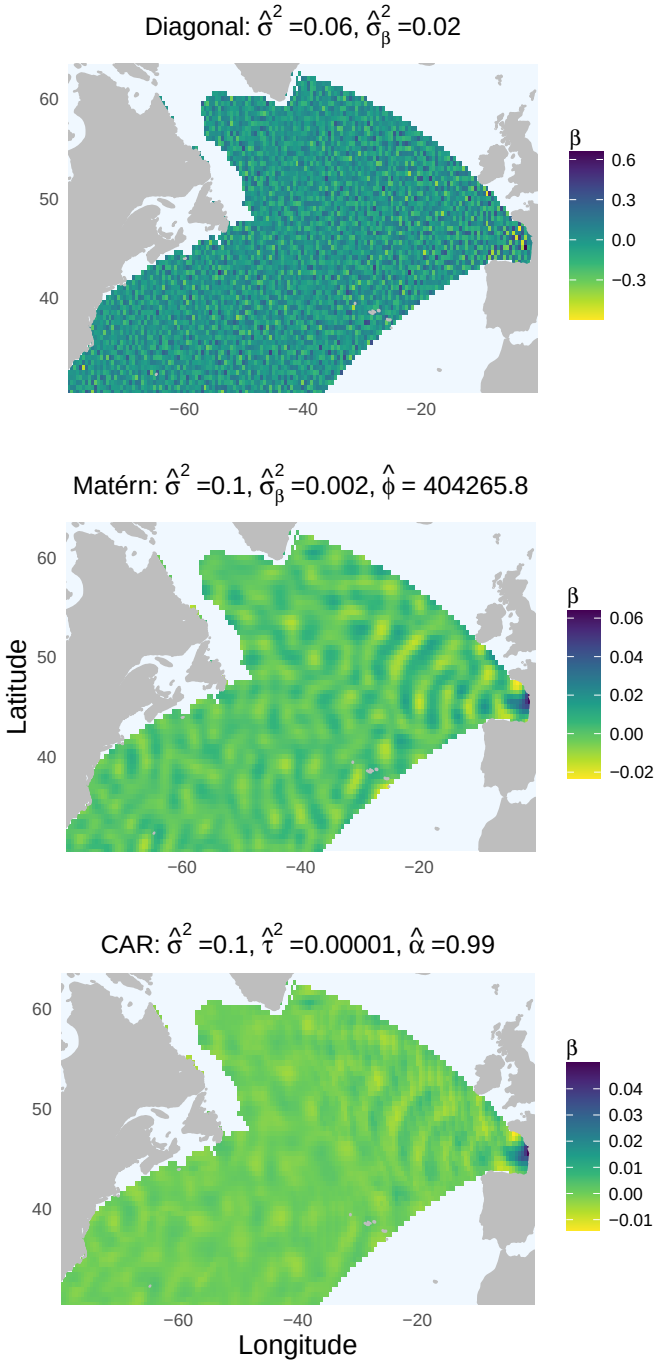

Figure 4.8 – The coefficients β estimated using the EM algorithm with diagonal, Matérn, and CAR covariance.

4.7 Summary

This study proposed an EM algorithm for estimating generalized Ridge regression with spatial covariates. We have studied three cases: the diagonal, Matérn, and the CAR case. A simulation study is carried out to evaluate the performance of the algorithms, and the EM algorithm successfully estimates the parameters in all cases. We have studied the influence of the sample size, dimension of X, and the variance σ^2 on the estimation. The sample size mainly influences the variance parameter σ^2. The range parameter of the Matérn and correlation parameter of the CAR are mainly influenced by dimension d.

The proposed method is applied to the problem of downscaling the significant wave height in the Bay of Biscay using wind conditions over the North Atlantic. The Matérn method gives smooth coefficients with a clear spatial structure; however, the CAR method slightly outperforms the Matérn method in terms of RMSE. The Matérn covariance is clearly a better choice for spatial applications. However, estimating the parameters requires the inversion of the covariance matrix at each iteration of the optimization method in the M-step, which may be a computational bottleneck in many applications. To address this issue, instead of parameterizing the covariance matrix, one can parameterize the precision matrix directly as we did with the CAR method.

.1 Comparison between cross-validation and EM

As stated in section 2, the EM algorithm can be used as an alternative for cross-validation for estimating Ridge regression. In this section, we perform a simulation study to compare the two approaches and use the same simulation procedure discussed in section 3.1. Given the same covariates X (presented in section 3.1) we perform 50 independent random samples of coefficients β using the diagonal method (with parameters $\sigma^2 = 36$ and $\sigma_\beta^2 = 7$). For each simulation, we estimate the coefficients using the EM algorithm and the cross-validation method. Figure 5.11 shows the box plot of $NRMSE_\beta$ and $NRMSE_y$. The EM algorithm outperforms cross-validation in estimating the coefficients β and predicting y.

The comparison we performed here is for the Gaussian case; therefore, it is straightforward that the EM algorithm will outperform cross-validation. To see how the two approaches behave in the non-Gaussian case, we simulate the response variable Y using

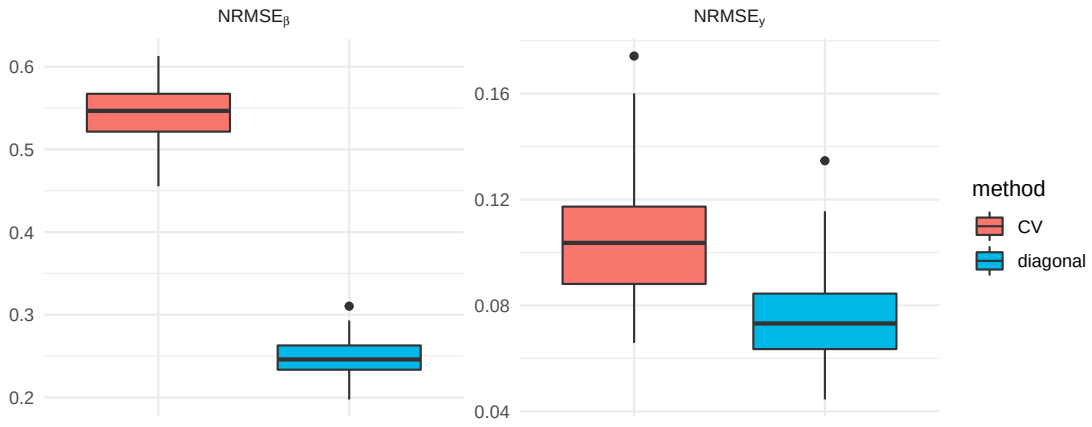

Figure 9 – Results of estimating Ridge regression with the EM algorithm and 10-fold cross-validation in the Gaussian case.

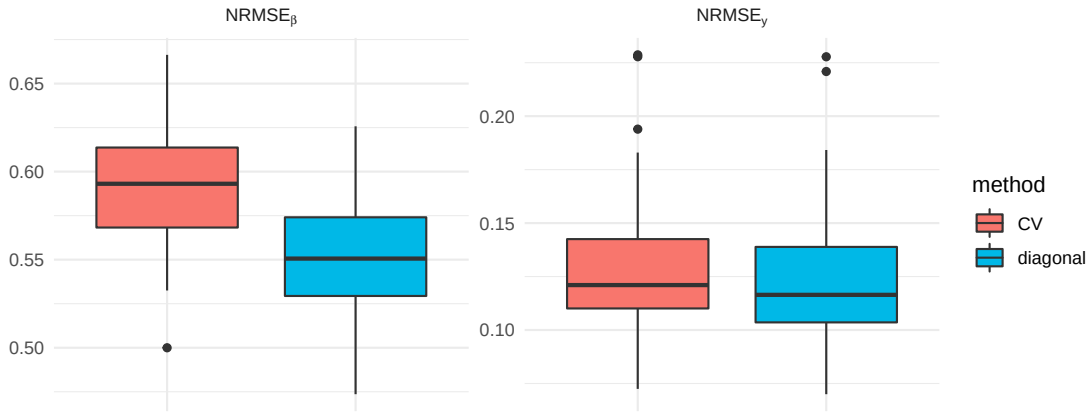

Figure 10 – Results of estimating Ridge regression with the EM algorithm and 10-fold cross-validation in the non-Gaussian case.

the model

$$Y = X\beta + \epsilon, \quad \text{where } \epsilon \sim U(2, 30) \tag{.1.1}$$

Where $U(2, 30)$ is the uniform distribution on the interval $[2, 30]$. Figure 10 shows the estimation results using the EM algorithm and cross-validation. The EM algorithm still outperforms cross-validation in both $NRMSE_\beta$ and $NRMSE_y$; however, the difference between the two methods here is small than in the Gaussian case.

.2 The case where β has a non-zero mean

In this section, we consider the case where β has a non-zero mean as defined by the hierarchically model

$$\begin{aligned} \beta &\sim \mathcal{N}(\mu_\xi, \Sigma_\theta) \\ Y \mid \beta, \Theta &\sim \mathcal{N}(X\beta, \sigma^2 I_n) \end{aligned} \tag{.2.1}$$

where $\Theta = (\sigma^2, \mu_\xi, \theta)$.

The complete log-likelihood is expressed as

$$\begin{aligned} \ln p(y, \beta; \Theta) &= \ln p(y \mid \beta; \sigma^2) + \ln p(\beta; \theta) \\ &= -\frac{1}{2}\left(\ln(|\Sigma_\theta|) + \beta^T\Sigma_\theta^{-1}\beta - 2\beta^T\Sigma_\theta^{-1}\mu_\xi + \mu_\xi^T\Sigma_\theta^{-1}\mu_\xi + n\ln(\sigma^2) + \frac{1}{\sigma^2}\|y + X\beta\|^2\right) + C \end{aligned} \tag{.2.2}$$

Where C is a constant. In the M-step, the quantity $Q(\Theta|\Theta^{(t)})$ is maximized with respect to the parameters Θ.

- E-step:

$$Q(\Theta|\Theta^{(t)}) = \mathbb{E}(\ln p(y, \beta; \Theta) \mid y, \Theta^{(t)}). \tag{.2.3}$$

The posterior distribution of the latent variable β is a normal distribution with mean $\mu_{\beta|y}$ and covariance matrix $\Sigma_{\beta|y}$ such that

$$\begin{cases} \Sigma_{\beta|y} = (\Sigma_\theta^{-1} + \frac{1}{\sigma^2}X^TX)^{-1} \\ \mu_{\beta|y} = \Sigma_{\beta|y}(\Sigma_\theta^{-1}\mu_\xi + \frac{1}{\sigma^2}X^Ty). \end{cases} \tag{.2.4}$$

Therefore,

$$\begin{aligned} Q(\Theta|\Theta^{(t)}) = &-\frac{1}{2}\Big(\ln(|\Sigma_\theta|) + \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)})) - 2\mu_{\beta|y}^T\Sigma_\theta^{-1}\mu_\xi + \mu_\xi^T\Sigma_\theta^{-1}\mu_\xi \\ &+ n\ln(\sigma^2) + \frac{1}{\sigma^2}\mathbb{E}(\|y - X\beta\|^2 \mid y, \Theta^{(t)}) + C \end{aligned} \tag{.2.5}$$

where

$$\begin{cases} \mathbb{E}(\beta\beta^T|y;\Theta^{(t)}) = \Sigma_{\beta|y} + \mu_{\beta|y}\mu_{\beta|y}^T \\ \mathbb{E}(\|y - X\beta\|^2|y;\Theta^{(t)}) = \|y\|^2 - 2y^TX\mu_{\beta|y} + \mathrm{Tr}(X^TX\mathbb{E}(\beta\beta^T|y;\Theta^{(t)})) \end{cases} \tag{.2.6}$$

- M-step:

The maximization step computes

$$\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta|\Theta^{(t)}) \tag{.2.7}$$

which leads to the following updates of the parameters

$$\begin{aligned} &\sigma^{2,(t+1)} = \frac{1}{n}(\|y\|^2 - 2y^TX\mu_{\beta|y} + \mathrm{Tr}(X^TX\mathbb{E}(\beta\beta^T|y;\Theta^{(t)}))) \\ &(\xi^{(t+1)}, \theta^{(t+1)}) = \arg\max_{\xi,\theta}\ \ln(|\Sigma_\theta^{-1}|) - \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)})) + 2\mu_{\beta|y}^T\Sigma_\theta^{-1}\mu_\xi^{(t)} - \mu_{\xi^{(t)}}^T\Sigma_\theta^{-1}\mu_\xi^{(t)} \end{aligned} \tag{.2.8}$$

4.3 Conclusions

In this chapter, we proposed an EM algorithm to estimate the parameters of the generalized Ridge regression. Using the simulation study and the application to wave height prediction, we have shown that the inclusion of spatial structure on the regression coefficients when the covariates have spatial structure improves the accuracy and interpretability of the predictive model.

Chapter 5

Maximum Likelihood Estimation of a Mixture of Generalized Ridge Regression

Contents

Note: The results of this chapter are submitted for publication as S.Obakrim, P.Ailliot, V.Monbet, and N.Raillard, Maximum likelihood estimation of a mixture of generalized ridge regression.

5.1 Preface

In Chapter 3, we proposed a model based on weather types to predict significant wave height using wind conditions. The weather types were constructed using a regression-guided clustering algorithm for accounting for large and small-scale conditions. Then, for each weather type, a Ridge regression model is fitted between H_s and the wind conditions. The model was shown to predict H_s and be interpretable effectively. In the conclusion of the chapter, we pointed out that the weather types can be considered as a hidden variable and the overall model based on the weather types can be considered as a mixture model.

Motivated by the ideas of chapter 3 and 4, the fundamental goals of this chapter are:

- Create a weather types-based model that does regression and classification at the same time
- The model must be able to make predictions of future weather types
- The model must perform a regularization to overcome the multicollinearity of the covariates
- The regularization hyperparameters must be learned by the model without using conventional hyperparameter selection methods such as cross-validation

Therefore, we propose a mixture of generalized Ridge experts for regression tasks with heterogeneous data and multicollinear covariates. The proposed method is then applied to the problem of predicting H_s using wind conditions.

5.2 Abstract

Mixture of experts are powerful tools for modeling heterogeneous data that appear in many applications such as environment, economics, and business. In this study, we focus on using mixture of experts for regression purposes and consider the case where the covariates are multicollinear. In the case of multicollinearity or high-dimensionality, penalized methods are needed, and generalized Ridge is a powerful penalization method for this purpose. In addition, the generalized Ridge allows incorporating any prior on the covariance structure of the regression coefficients, which is useful, for example, in spatial applications. The generalized Ridge may have more than one hyperparameter, and in the context of mixture modeling, estimating these hyperparameters using conventional hyperparameter selection methods, such as cross-validation, can be computationally challenging. This study proposes a variational expectation-maximization algorithm for fitting

generalized Ridge expert mixtures. A simulation study is carried out to evaluate the proposed method's performance, and the method is applied to predict ocean waves using wind conditions.

5.3 Introduction

Heterogeneous data is common in many fields such as the environment, economy, and business. Mixture models are commonly used in statistics and machine learning for regression and classification to model heterogeneous data. For example, mixtures of normal distributions (Day, 1969) are often used to find clusters in data. For regression tasks, the mixture of linear regression models, also known as finite mixture of regression, partitions the data into classes and fits a linear regression model in each group (DeSarbo and Cron, 1988). Given a set of covariates X and a response variable Y, the mixture of linear regression relates Y to X via a linear model such that

$$\mathbb{E}(Y|X, Z = k) = X\beta_k, \;\; k = 1, ..., K \tag{5.3.1}$$

where Z is a discrete hidden variable that determines the class and K is the number of classes. The mixture of regressions model in equation (5.3.1) assumes that the classes depend only on the response variable Y and that the covariates carry no information about the classes. Therefore, the mixture model cannot predict future classes from new observations of the covariates. In general, the average of the predictions on the classes with fixed weights is used as a prediction of the response variable Y (Hoshikawa, 2013).

Mixture of experts (MoE), introduced by Jacobs et al. (1991), are mixture models that assume that the class membership depends on the covariates X. Therefore, MoE models are capable of predicting the response variable as well as the class membership. MoEs have two components: several experts, that may be regressors or classifiers, and a gate network that partitions the input space into regions in which the experts are specialized. The multinomial logit model is usually used as a gating network (Geweke and Keane, 2007) and the inference of MoEs is generally done using the Expectation-Maximization algorithm (Jacobs et al., 1991), MCMC (Meeds and Osindero, 2005), or variational methods (Yuan and Neubauer, 2008). Detailed surveys on MoE models can be found in: Yuksel, Wilson, and Gader (2012); Masoudnia and Ebrahimpour (2014); Nguyen and Chamroukhi (2018).

Although MoE models are powerful tools for regression and classification, their application can be complex when the covariates are multicollinear and high dimensional. Therefore, regularized likelihood is typically used to select relevant covariates (Fraley and Raftery, 2007; Ramamurti and Ghosh, 1997); however, these methods require the selection of the regularization hyper-parameters, which may have a high computational cost. The Ridge or L^2 penalty (Hastie et al., 2009) is a powerful tool for dealing with multicollinearity and high dimensionality. Moreover, the generalized Ridge penalty allows for jointly penalizing the coefficients and permits including any prior on the structure of the coefficients (Wieringen, 2015). By assuming that the coefficients are hidden variables, Obakrim et al. (2022a) proposed an Expectation-Maximization algorithm for estimating the parameters of a linear regression model. In this study, we extend their algorithm for estimating the parameters of a mixture of generalized Ridge regression.

The remainder of this paper is structured as follows. The mixture of generalized Ridge regression model is presented in the second section. Then in the third section, the method used for model inference is detailed. In the fourth section, a simulation study is conducted to assess the performance of the proposed method, and in the fifth section, our proposed method is applied to predict wave heights in the Bay of Biscay using wind conditions over the North Atlantic. Finally, this study is concluded in the last section.

5.4 Mixture of generalized Ridge experts

Mixture of experts are a set of expert models for regression or classification and a gate that divides the heterogeneous input space into a set of homogeneous regions. The experts are individual models that are specialized in each region defined by the gate network. In this study, we consider mixture of generalized Ridge regression experts where the gating network is the classical multinomial model.

Consider an experiment where we have the data $\{y, X\}$, of n independent identically distributed (i.i.d.) observations of a continuous variable Y and $n \times d$ matrix of covariates X. Suppose that Y is related to X via a mixture of linear models, where the covariates might be multicollinear or high-dimensional. The generalized Ridge penalty (Wieringen, 2015) is a powerful tool for dealing with multicollinearity and high-dimensionality. The generalized Ridge regression can be derived as the mean of a posterior distribution with a Normal prior and a given covariance matrix Wieringen (2015). Thus, we define the mixture of

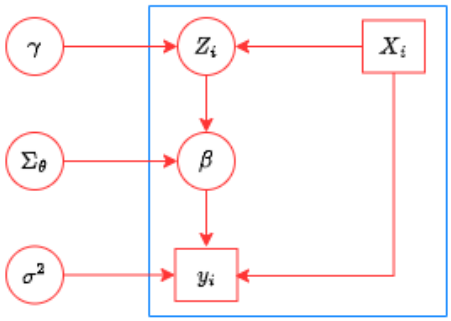

Figure 5.1 – Graphical representation of the mixture of generalized Ridge experts. X_i and y_i are observed variables, Z_i and β are hidden variables and γ, Σ_θ and σ^2 are the parameters of the model.

generalized Ridge regression experts (MoR) model hierarchically as (see figure 5.1)

$$\begin{aligned} &Z_i \sim M(1, p_i), \;\; p_i = (p_{i1}, ..., p_{iK})^T, \;\; i = 1, ..., n \\ &\beta_k \sim \mathcal{N}(0, \Sigma_{\theta k}), \;\; k = 1, .., K \\ &y_i | \beta_k, Z_i = k \sim \mathcal{N}(X_i \beta_k, {\sigma_k}^2) \\ &p_{ik} = \frac{\exp(X_i \gamma_k)}{\sum_{l=1}^{K} \exp(X_i \gamma_l)} \end{aligned} \tag{5.4.1}$$

where Z is a multinomial variable that determines the class, K is the total number of classes, $\beta = \{\beta_1, ..., \beta_K\}$, of size $d \times K$, are experts coefficients for each class, $\Sigma_\theta = \{\Sigma_{\theta_1}, ..., \Sigma_{\theta_K}\}$, of size $d \times d \times K$, are the covariance matrices of the regression coefficients at each class k, $\sigma^2 = (\sigma_1^2, ..., \sigma_K^2)$ are the variances of the residuals at each class k, and $\gamma = \{\gamma_1, ..., \gamma_K\}$, of size $d \times K$, are the coefficients of the gating network. Figure 5.1 shows a graphical representation of the MoR model. We consider Z_i and β as hidden variables and Σ_θ, γ, and σ^2 as the parameters of the model. Note that for a matter of simplicity, we suppose that all the β_k's have a zero mean.

5.5 Model inference

Many methods have been proposed in the literature for inferring mixture of expert models (see Yuksel, Wilson, and Gader (2012) for a review of inference methods). In this study, we propose a variational Expectation-Maximization (EM) algorithm for estimating the parameters of the MoR model.

5.5.1 EM algorithm

The complete likelihood of the model (5.4.1) is expressed as

$$\begin{aligned}\mathcal{L}(Y, Z, \beta; \Theta) &= p(Y|Z, \beta; \Theta)p(\beta|Z)p(Z) \\ &= p(\beta)\prod_{i=1}^{n} p(y_i|z_i, \beta)p(z_i) \\ &= p(\beta)\prod_{i=1}^{n}\prod_{k=1}^{K} p(y_i|z_i, \beta)^{\omega_{ik}} p_{ik}^{\omega_{ik}}\end{aligned} \tag{5.5.1}$$

where $\Theta = \{\gamma, \Sigma_\theta, \sigma^2\}$ are the model parameters and $\omega_{ik} = 1$ if $Z_i = k$ and $\omega_{ik} = 0$ otherwise. Therefore, the complete log-likelihood is expressed as

$$\ln \mathcal{L}(Y, Z, \beta; \Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \omega_{ik} \ln \Phi(y_i; X_i\beta_k, \sigma_k) + \sum_{k=1}^{K} \ln \Phi(\beta_k; 0, \Sigma_{\theta k}) + \sum_{i=1}^{n}\sum_{k=1}^{K} \omega_{ik} \ln p_{ik} \tag{5.5.2}$$

where Φ is the multivariate normal distribution density function. Starting with an initial guess of the parameters $\Theta^{(0)}$, the EM algorithm alternates between the E-step and M-step until convergence. In the E-step, the expectation $Q(\Theta|\Theta^{(t)})$ of the complete likelihood with respect to the posterior distribution of the latent variables Z and β and the parameters $\Theta^{(t)}$ from the previous iteration t, is calculated. In the M-step, the quantity $Q(\Theta|\Theta^{(t)})$ is maximized with respect to the parameters Θ.

The E-step requires the computation of the expectation of the complete log-likelihood with respect to the posterior distribution $p(Z, \beta|Y; \Theta)$ of the latent variables Z and β. $p(Z, \beta|Y; \Theta)$ has the form

$$p(Z, \beta|Y; \Theta) = \frac{p(Z; \Theta)p(Y|Z, \beta; \Theta)p(\beta|Z; \Theta)}{\sum_Z p(Z; \Theta)\int_\beta p(Y|Z, \beta; \Theta)p(\beta|Z; \Theta)}. \tag{5.5.3}$$

The calculation of this distribution requires integrating over all possible values of Z and β, which is intractable. To address this issue, we propose a variational EM algorithm.

5.5.2 Variational EM

In this subsection, we first recall the idea and motivation behind the variational EM algorithm; then, we present the variational approximation we propose to address the

problem of intractability in the E-step. Theoretical results on variational inference in general can be found in Blei, Kucukelbir, and McAuliffe (2017) and on variational EM in particular in Beal (2003). The variational EM algorithm has also proven itself in practice (see, e.g., El Assaad et al. (2016); Bernardo et al. (2003); Kounades-Bastian et al. (2016)).

The log-likelihood function of the model (5.4.1) is expressed as

$$\ln p(Y;\Theta) = \ln \sum_{Z} \int_{\beta} p(Y, Z, \beta; \Theta) \tag{5.5.4}$$

which is intractable. An alternative view of the EM algorithm (Bishop and Nasrabadi, 2006) is motivated by the fact that

$$\ln p(Y;\Theta) = \mathcal{L}(q,\Theta) + KL(q||p) \tag{5.5.5}$$

where

$$\begin{aligned} \mathcal{L}(q,\Theta) &= \sum_{Z} \int_{\beta} q(Z,\beta) \ln \frac{p(Y,Z,\beta;\theta)}{q(Z,\beta)} \\ KL(q||p) &= -\sum_{Z} \int_{\beta} q(Z,\beta) \ln \frac{p(Z,\beta|Y;\theta)}{q(Z,\beta)} \end{aligned} \tag{5.5.6}$$

where $q(Z,\beta)$ is a distribution over the latent variables Z and β and $KL(q||p)$ is the Kullback-Leibler (KL) divergence between $q(Z,\beta)$ and the posterior $p(Z,\beta|Y;\Theta)$. Given that $KL(q||p) \geq 0$, $\mathcal{L}(q,\Theta)$ is a lower bound of $\ln p(Y;\Theta)$. $\mathcal{L}(q,\Theta)$ is also called the variational free energy, which can be expressed as follows (Neal and Hinton, 1998)

$$\mathcal{L}(q,\Theta) = \mathbb{E}_q(\ln \mathcal{L}(Y,Z,\beta;\Theta)) + H(q) \tag{5.5.7}$$

where q is the distribution over the latent variables, $\mathbb{E}_q(.)$ denotes the expectation with respect to the distribution q and $H(q)$ is the entropy of q. The E-step maximises the lower bound $\mathcal{L}(q,\Theta)$ with respect to the distribution $q(Z,\beta)$. The lower bound is maximized when the KL divergence is equal to zero, which corresponds to the case where $q(Z,\beta)$ is equal to the posterior distribution $p(Z,\beta|Y;\Theta)$. The EM algorithm starts with an initial guess of the parameters, $\Theta^{(0)}$, and repeatedly applies the following two steps until convergence:

- E-step: $q^{(t)} = \arg\max_q \mathcal{L}(q,\Theta^{(t)})$
- M-step: $\Theta^{(t+1)} = \arg\max_\Theta \mathcal{L}(q,\Theta)$

Given that the posterior distribution $p(Z, \beta|Y; \Theta)$ is intractable, variational approximation can be used.

In this study, we propose a variational approximation that simplifies the E-step by imposing some constraints on the distribution q. We assume that the distribution q can be factorized into

$$q(Z, \beta) = \prod_{i,l} q_Z(Z_{il}) \prod_{l} q_\beta(\beta_l) \tag{5.5.8}$$

where q_Z and q_β are the distribution over the latent variables Z_{il} and β, respectively. Along with the factorization assumption in equation (5.5.8), we also assume that β_k's are independent. Therefore, the distribution q is characterized by the mean μ_{q_k}, the covariance Σ_{q_k} and τ_{ik} where $\tau_{ik} = p_Z(Z_i = k)$ and $q_{\beta_k} \sim \mathcal{N}(\mu_{q_k}, \Sigma_{q_k})$.

The lower bound in (5.5.7) becomes

$$\mathcal{L}(q, \Theta) = \mathbb{E}_{q_Z, q_\beta}(\ln \mathcal{L}(Y, Z, \beta; \Theta)) + H(q) \tag{5.5.9}$$

where

$$\begin{aligned}\mathbb{E}_{q_Z, q_\beta}(\ln \mathcal{L}(Y, Z, \beta; \Theta)) = &\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \mathbb{E}_{q_\beta}(\ln \Phi(y_i; X_i\beta_k, \sigma_k)) + \sum_{k=1}^{K} \mathbb{E}_{q_\beta}(\ln \Phi(\beta_k; 0, \Sigma_{\theta_k})) \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \ln p_{ik}\end{aligned} \tag{5.5.10}$$

and

$$H(q) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \ln \tau_{ik} + \frac{1}{2}\sum_{k=1}^{K}(p(1 + \ln 2\pi) + \ln |\Sigma_{q_k}|). \tag{5.5.11}$$

We have

$$\begin{aligned}\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \mathbb{E}_{q_\beta}(\ln \Phi(y_i; X_i\beta_k, \sigma_k)) &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \ln {\sigma_k}^2 - \frac{1}{2}\sum_{k=1}^{K} \frac{1}{{\sigma_k}^2} \mathbb{E}_{q_\beta}((y - X\beta_k)^T \tau_k (y - X\beta_k)) + C \\ &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \ln {\sigma_k}^2 - \frac{1}{2}\sum_{k=1}^{K} \frac{1}{{\sigma_k}^2} (y^T \tau_k y - 2y^T \tau_k X \mu_{q_k} \\ &\quad + \mathrm{Tr}(X^T \tau_k X \Sigma_{qk})) + C\end{aligned} \tag{5.5.12}$$

where $\tau_k = \text{diag}(\tau_{ik})$ and C is a constant. And

$$\sum_{k=1}^{K} \mathbb{E}_{q_\beta}(\ln \Phi(\beta_k; 0, \Sigma_{\theta_k})) = -\frac{1}{2}\sum_{k=1}^{K}(\ln|\Sigma_{\theta k}| + \text{Tr}(\Sigma_{\theta_k}^{-1}\Sigma_{q_k}) + \mu_{q_k}\Sigma_{\theta k}\mu_{q_k}^T) + C. \quad (5.5.13)$$

Therefore

$$\begin{aligned}\mathcal{L}(q,\Theta) = &-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}(\ln {\sigma_k}^2 + \ln\tau_{ik} - 2\ln p_{ik}) \\ &-\frac{1}{2}\sum_{k=1}^{K}(\frac{n}{{\sigma_k}^2}(y^T\tau_k y - 2y^T\tau_k X\mu_{q_k}) + \text{Tr}((\frac{1}{{\sigma_k}^2}X^T\tau_k X + \Sigma_{\theta_k}^{-1})\Sigma_{q_k}) - \ln|\Sigma_{q_k}|) + C\end{aligned} \quad (5.5.14)$$

Given some initial values $\Theta^{(0)}$, ${\mu_{qk}}^{(0)}$ and ${\Sigma_{qk}}^{(0)}$, the variational E-step can be performed using

$$\begin{aligned}&\arg\max_{\tau}\mathcal{L}(q,\Theta^{(t)}) \\ &\text{subject to}\sum_{k=1}^{K}\tau_{ki} = 1,\ i = 1,...,n\end{aligned} \quad (5.5.15)$$

$$\arg\max_{\mu_q}\mathcal{L}(q,\Theta^{(t)}) \quad (5.5.16)$$

$$\arg\max_{\Sigma_q}\mathcal{L}(q,\Theta^{(t)}). \quad (5.5.17)$$

After calculating the derivatives and equating to zero, the updates for the parameters in the variational E-step are

$${\tau_{ik}}^{(t)} = \frac{{V_{ik}}^{(t-1)}}{\sum_{l=1}^{K}{V_{il}}^{(t-1)}} \quad (5.5.18)$$

$${\Sigma_{qk}}^{(t)} = (\frac{1}{{\sigma_k}^{2(t-1)}}X^T\tau_k^{(t)}X + {\Sigma_{\theta_k}^{(t-1)}}^{-1})^{-1} \quad (5.5.19)$$

$${\mu_{q_k}}^{(t)} = \frac{1}{{\sigma_k}^{2(t-1)}}{\Sigma_{qk}}^{(t)}X\tau_k^{(t)}y \quad (5.5.20)$$

where

$${V_{ik}}^{(t-1)} = p_{ik}^{(t-1)}e^{-\frac{1}{2}(\ln{\sigma_k}^{2(t-1)} + \frac{1}{{\sigma_k}^{2(t-1)}}(y_i^2 - 2y_iX_i{\mu_{q_k}}^{(t-1)} + X_i{\Sigma_{q_k}}^{(t-1)}X_i^T + (X_i{\mu_{q_k}}^{(t-1)})^2))} \quad (5.5.21)$$

After the variational E-step, in the M-step, we maximize $\mathcal{L}(q,\Theta)$ with respect to the

parameters Θ. The updates of the parameters are

$$\sigma_k^{2,(t)} = \frac{1}{\mathrm{Tr}(\tau_k{}^{(t)})}(y^T \tau_k{}^{(t)} y - 2y^T \tau_k{}^{(t)} X \mu_{q_k}{}^{(t)} + \mathrm{Tr}(X^T \tau_k{}^{(t)} X \Sigma_{q_k}{}^{(t)})) \tag{5.5.22}$$

$$\Sigma_{\theta_k}{}^{(t)} = \arg\max_{\Sigma_{\theta_k}} \mathcal{L}(q^{(t)}, \Theta) \tag{5.5.23}$$

$$\gamma^{(t)} = \gamma^{(t-1)} - \left[\frac{\partial^2 \mathcal{L}(q^{(t)}, \Theta)}{\partial \gamma \partial \gamma^T}\right]^{-1} \frac{\partial \mathcal{L}(q^{(t)}, \Theta)}{\partial \gamma} \tag{5.5.24}$$

where $\gamma = (\gamma_1^T, ..., \gamma_K^T)^T$. Equation (5.5.24) corresponds to a single Newton-Raphson update of γ of the multinomial logit model. The expression of the gradient and Hessian are well known (see, for example, (Chamroukhi, 2010)) and have the form

$$\begin{aligned} \frac{\partial \mathcal{L}(q^{(t)}, \Theta)}{\partial \gamma} &= \tilde{X}^T(\tilde{\tau} - \tilde{P}) \\ \frac{\partial^2 \mathcal{L}(q^{(t)}, \Theta)}{\partial \gamma \partial \gamma^T} &= -\tilde{X}^T \tilde{\mathrm{T}} \tilde{X}^T \end{aligned} \tag{5.5.25}$$

where $\tilde{X}$ is a matrix of order $Kn \times Kp$ defined as

$$\tilde{X} = \begin{bmatrix} X & 0 & 0 & \dots & 0 \\ 0 & X & 0 & \dots & 0 \\ \vdots & 0 & & \ddots & \vdots \\ 0 & \dots & & 0 & X \end{bmatrix} \tag{5.5.26}$$

and $\tilde{\tau}$ and $\tilde{P}$ are vectors of length $n \times K$, formed by concatenating τ_k and p_k ($p_k = (p_{1k}, ..., p_{nk})$), respectively. Finally, $\tilde{\mathrm{T}}$ is a matrix of order $nK \times nK$ which is defined as

$$\tilde{\mathrm{T}} = \begin{bmatrix} \mathrm{T}_{11} & \mathrm{T}_{12} & \dots & \mathrm{T}_{1K} \\ \mathrm{T}_{21} & \mathrm{T}_{22} & \dots & \mathrm{T}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{T}_{K1} & \mathrm{T}_{K2} & \dots & \mathrm{T}_{KK} \end{bmatrix} \tag{5.5.27}$$

where the matrices T_{kh} are diagonal matrices of order $n \times n$ such that $diag(\mathrm{T}_{kh})_i = p_{ik}(\delta_{kh} - p_{ih})$, where δ_{kh} is equal to 1 if $k = h$ and 0 otherwise.

5.5.3 Note on the covariances $\Sigma_{\theta k}$

The covariances $\Sigma_{\theta k}$'s are the main parameters that control the structure of the regression coefficients β_k at each class k. Their choice depends, therefore, on the application at hand. For instance, when we believe that there is no structure between the regression coefficients, one can choose $\Sigma_{\theta k} = \sigma^2_{\theta_k} I_d$, which corresponds to fitting the classical Ridge regression in each class k. However, the coefficients may exhibit a structure in many applications, as in spatial applications. Obakrim et al. (2022a) proposed an EM algorithm for estimating generalized Ridge for spatial application. The study considered three cases: when the covariance is diagonal constant, Matérn, and conditional autoregressive (CAR). In this study, we focus on using our proposed method for spatial applications, and we consider the CAR covariance as it is less computationally expensive than the Matérn. The CAR permits to avoid the covariance matrix's inversion during the maximization step in equation (5.5.20) by directly parameterizing the precision matrix. The precision matrix of the CAR at each class k, which depends on the parameters ${\tau_k}^2$ and α_k, is defined as follows

$$P_{\theta_k} = {\tau_k}^{-2}(I_d - \alpha_k H)\Phi^{-1} \tag{5.5.28}$$

where

$$\Phi = \text{diag}(|N_1|^{-1}, ..., |N_d|^{-1}) \tag{5.5.29}$$

where $|N_i|$ is the number of neighbors of location i and $H = \left(\frac{a_{ij}}{|N_i|}\right)_{d\times d}$; $i, j = 1, ..., d$, where a_{ij} is the (i, j) element of the adjacency matrix $A = (a_{ij})_{d\times d}$, where $a_{ij} = a_{ji} = 1$ if and only if location i and j are neighbors and otherwise $a_{ij} = 0$.

5.5.4 Variational EM algorithm training details

Mixture models suffer from locally optimal solutions; therefore, the solution depends on the initial values of the EM (or variational EM) algorithm (Shireman, Steinley, and Brusco, 2017). Our proposed algorithm is initialized as follows: at first, we use the K-means algorithm on the covariates X to find K clusters in the data, then the parameters γ_k k=1,..., K, are initialized as follows:

$$\gamma_k = (X^T X)^{-1} X^T \hat{z}_k, \;\; k = 1, ..., K \tag{5.5.30}$$

where $\hat{z_{ik}} = 1$ if the observation i is in the K-means cluster k and $\hat{z_{ik}} = 0$ otherwise. In the case when the matrix $(X^T X)$ is ill-conditioned, one may add a constant of regu-

larization. After initializing γ_k's, the probabilities p_{ik} are calculated. We have chosen to initialize the parameters γ_k k=1,..., K, in this way instead of directly fitting a multinomial model, because the multinomial model is more computationally expensive and numerical experiments show that our initialization method leads to results close to an initialization with a multinomial model. The other parameters (${\sigma_k}^2$ and covariance parameters θ_k for $k = 1, .., K$) can be initialized randomly or fixed arbitrary.

After initialization, the variational E-step and M-step are repeated until until the stopping criterion is met or until a maximum number of iterations is reached. The stopping criterion chosen in this study is the root mean square error (RMSE) between observed and predicted response variable Y. A summary of the variational EM algorithm used in this study is presented in Algorithm 2.

Algorithm 2: Variational EM algorithm

Input: Observed response variable y, a matrix of covariates X (of size n and $n \times d$, respectively) , and a number of classes K

Initialization: Initialize the parameters $\Theta^{(0)} = (\gamma^{(0)}, \Sigma_\theta^{(0)}, \sigma^{2,(0)})$ using the procedure described in subsection 5.5.4

repeat

  VE-step

  for $k = 1$ *to* K **do**

    for $i = 1$ *to* n **do**

      Compute the probability τ_{ik} using the equation (5.5.18)

    Compute the covariance Σ_{qk} and the mean μ_{qk} using the equations (5.5.19) and (5.5.20), respectively

  VM-step

  for $k = 1$ *to* K **do**

    Compute the variance σ_k^2 using the equation (5.5.22)

    Compute the covariance $\Sigma_{\theta k}$ using the equation (5.5.23)

    Update the parameters γ_k using the equation (5.5.24)

until *the stopping criterion is met or a maximum number of iterations is reached*;

5.6 Simulation study

In this section, we perform a simulation study to assess the performance of the proposed method. We will focus on using the method for spatial applications; therefore, we consider the CAR covariance structure for the regression coefficients presented in the previous section.

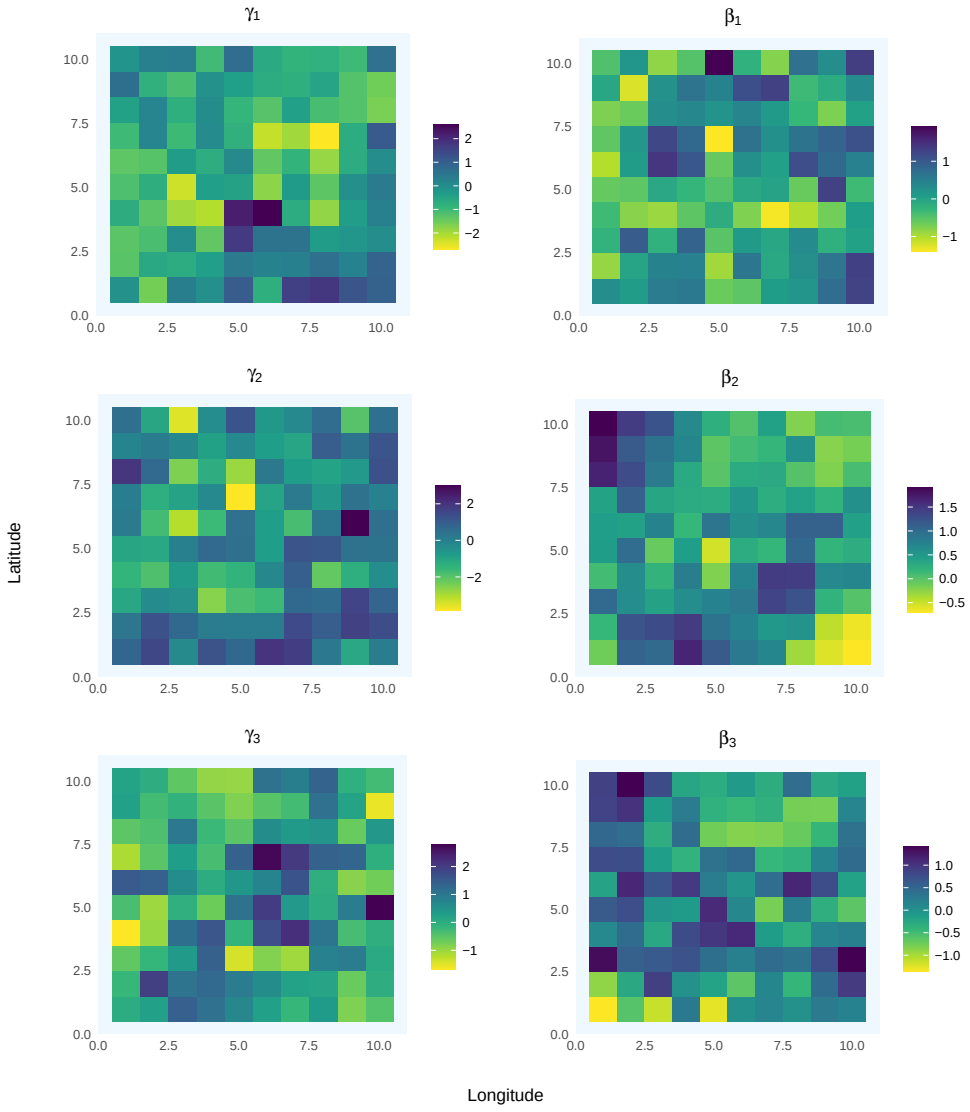

Figure 5.2 – An example of the simulated gate coefficients γ (left panels) and experts coefficients β (right panels).

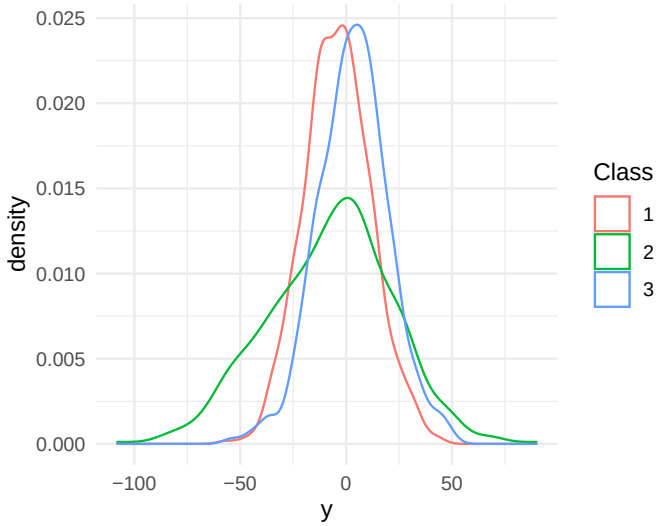

Figure 5.3 – Empirical density of a simulated response variable Y as a function of the classes 1, 2, and 3.

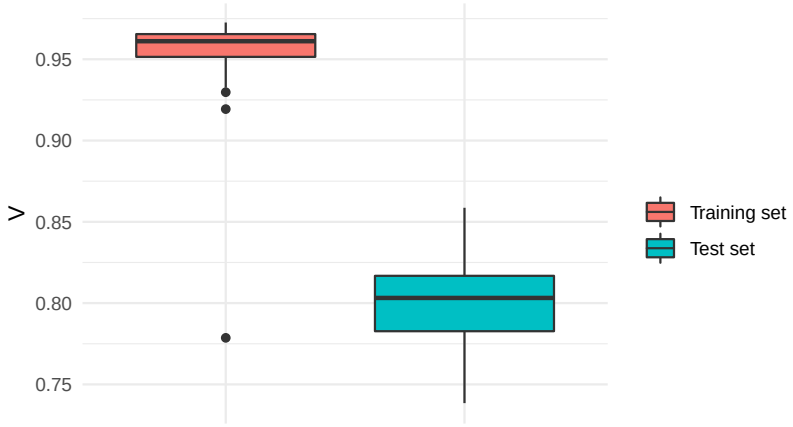

Figure 5.4 – Results of Cramér's V, from 100 simulations, in the training (red box) and test (blue box) sets, resulting from the comparison of the true and estimated classes using the mixture of generalized Ridge experts.

5.6.1 Setup

For all the simulations, we consider three classes ($K = 3$), and we proceed as follows:

- We consider a 10×10 regular spatial grid in a square domain $[1, 10]^2$. Then we generate $X = (x_{ij})_{n \times d}$ of n independent and identically distributed observations from a multivariate normal distribution with zero mean and CAR covariance with the parameters $(\tau_X^2, \alpha_X) = (8, 0.94)$. Hence, each location j has a covariate x_j.
- The γ_k are simulated in each class using a multivariate normal distribution of zero mean and CAR covariance with parameters: $(\tau_{\gamma_1}^2, \alpha_{\gamma_1}) = (2.0.99)$, $(\tau_{\gamma_2}^2, \alpha_{\gamma_2}) = (5, 0.8)$, and $(\tau_{\gamma_3}^2, \alpha_{\gamma_3}) = (3, 0.9)$ for the first, second, and third class, respectively. Then, we calculate the probabilities $p_{ik}, i = 1, ..., n,\ k = 1, 2, 3$.
- The β_k coefficients are simulated at each class from a multivariate normal distribution of zero mean and CAR covariance with parameters: $(\tau_1^2, \alpha_1) = (2.0.8)$, $(\tau_2^2, \alpha_2) = (0.5, 0.99)$, and $(\tau_3^2, \alpha_3) = (1, 0.9)$ for the first, second, and third class, respectively.
- Finally, Y is simulated, conditionally on β_k's, Z, and X, from a normal distribution of variances $\sigma_1^2 = 3$, $\sigma_2^2 = 7$, and $\sigma_3^2 = 2$ at the first, second, and third class, respectively.

The parameters are chosen so that they are different at each class and the variance of the predictions $X\beta_k$ is sufficiently large that the $sigma^2{}_k$. Figure 5.2 shows an example of simulated coefficients γ and β and figure 5.3 shows the empirical density of a simulated response variable Y as a function of the classes.

5.6.2 Parameter estimation

To evaluate the parameters estimation, we consider three validation measures: the normalized root mean square error (RMSE) of the response y ($NRMSE_y$), normalized RMSE of the coefficients β_k in class k ($NRMSE_{\beta_k}$), and Cramér's V (V), defined respectively as

$$\begin{aligned} NRMSE_y &= \frac{\sqrt{\frac{1}{n}\sum_i^n (y_i - \hat{y}_i)^2}}{\hat{\sigma}_y} \\ NRMSE_{\beta_k} &= \frac{\sqrt{\frac{1}{d}\sum_j^d (\beta_{kj} - \hat{\beta_{kj}})^2}}{\hat{\sigma}_{\beta_k}} \\ V &= \sqrt{\frac{\chi^2/n}{K-1}} \end{aligned} \tag{5.6.1}$$

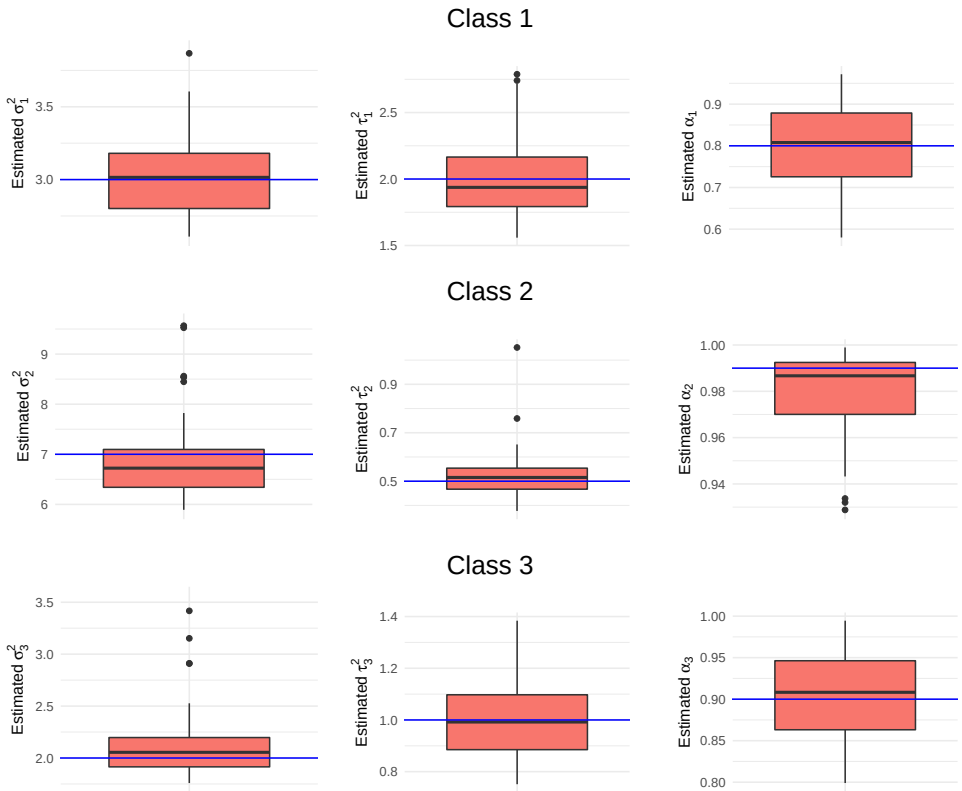

Figure 5.5 – Results of estimating the parameters ${\sigma_k}^2$, ${\tau_k}^2$, and α_k^2 for each class k=1,2,3 using MoR with CAR covariance on 100 simulations. The blue line corresponds to the true value of the parameters.

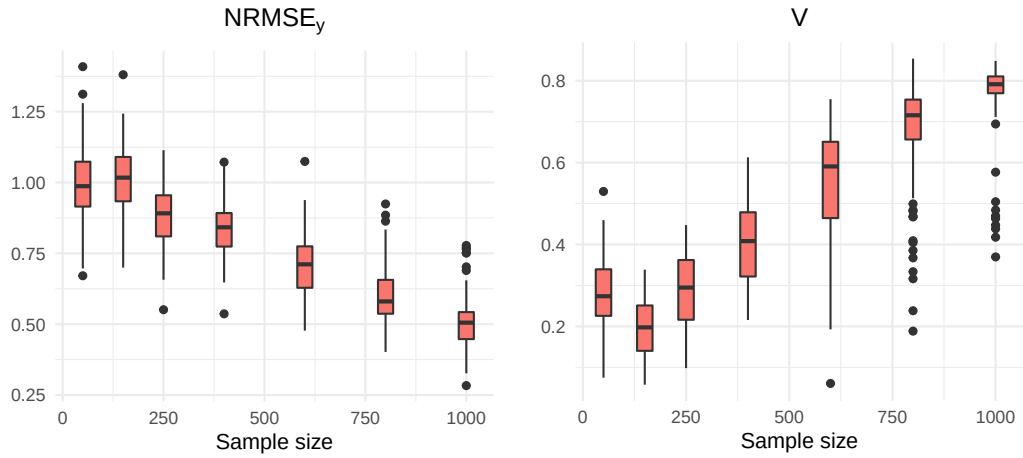

Figure 5.6 – Results of $NRMSE_y$ (left panel) and the Cramér's V (right panel) as a function of the sample size varying from 50 to 1000.

where $\hat{y}_i$ is the predicted y_i, $\hat{\sigma}_y$ the sample standard deviation of y, $\hat{\beta_{kj}}$ is the estimated β_k, $\hat{\sigma}_{\beta_k}$ the sample standard deviation of β_k, and χ^2 is the Pearson's chi-squared statistic between the observed and predicted classes. Note that the Cramér's V is an association measure for categorical variables, with a value between 0 and 1 (0 corresponds to no association and 1 to complete association between variables).

We perform 100 independent simulations using the methodology presented in the subsection 5.6.1, with the sample size $n = 1000$, and for each simulation, we estimate the parameters using our proposed method. Figure 5.4 shows the Cramér's V in the training set (red box) and a test set (blue box) of size 1000. Where the training set is the data used for estimating the parameters which has $n = 1000$ observations, and the test set generated independently which is not used o estimate the parameters. The Cramér's V has values around 0.96 and 0.8 in the training and test set, respectively. There appears to be some overfitting in the class estimation, which may be due to the fact that the gate network coefficients (γ) are not penalized.

Figure 5.5 shows the results of estimating σ_k^2 and the CAR parameters τ_k^2 and α_k for the classes $k = 1, 2, 3$, where the blue line corresponds to the true value of the parameters. generally, the parameters are well estimated in all classes.

To evaluate the sample size's impact on the parameters' estimation, we perform 100 independent simulations for each sample size, n, varying from 50 to 1000. Figure 5.6 shows the $NRMSE_y$ and the Cramér's V, both in the test set, as a function of the sample size.

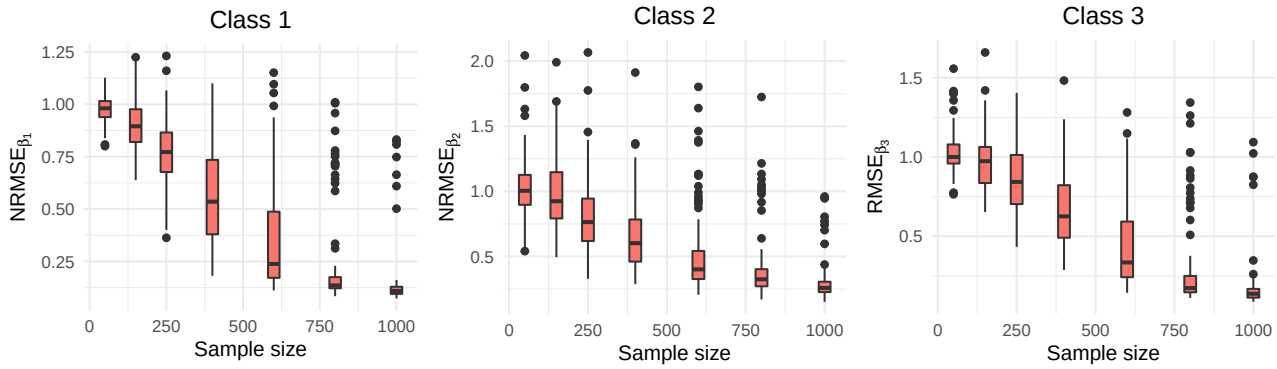

Figure 5.7 – Results of $NRMSE_{\beta_1}$, $NRMSE_{\beta_2}$, and $NRMSE_{\beta_3}$ as a function of the sample size varying from 50 to 1000.

As expected, $NRMSE_y$ decreases, and V increases as the sample size increases. Figure 5.7 shows the results of $RNMSE_{\beta_k}$ for each class $k = 1, 2, 3$. For the three classes, $RNMSE_{\beta_k}$ decreases as the sample size increases. Figure 5.8 shows the influence of the sample size on the estimation of the parameters σ_k^2 and the CAR covariances τ_k^2 and α_k for the classes $k = 1, 2, 3$, where the blue line corresponds to the true value of the parameters. All the parameters in the three classes converge to the true value as the sample size increases.

5.6.3 Selection of the number of classes

One important hyper-parameter that needs to be selected in mixture models is the number of classes K. Different methods have been used in the literature for this purpose, ranging from information-based to cross-validation methods (see McLachlan and Rathnayake (2014) for a review). In this study, we use 10-fold cross-validation to select the number of classes. We partition the data into ten groups, and we estimate the parameters using the MoR (CAR) model for each number of classes K, ranging from 1 to 6, leaving out one group. For each K, we calculate the mean $NRMSE_y$, and the number of classes with the minimum $NRMSE_y$ is chosen. Figure 5.9 shows the cross-validation results, where the black line is the mean $NRMSE_y$ of the ten validation groups, and the interval corresponds to the mean $NRMSE_y$ plus and minus its standard deviation. The optimal number of classes that gives the minimum $NRMSE_y$ is $K = 3$, corresponding to the actual number of classes.

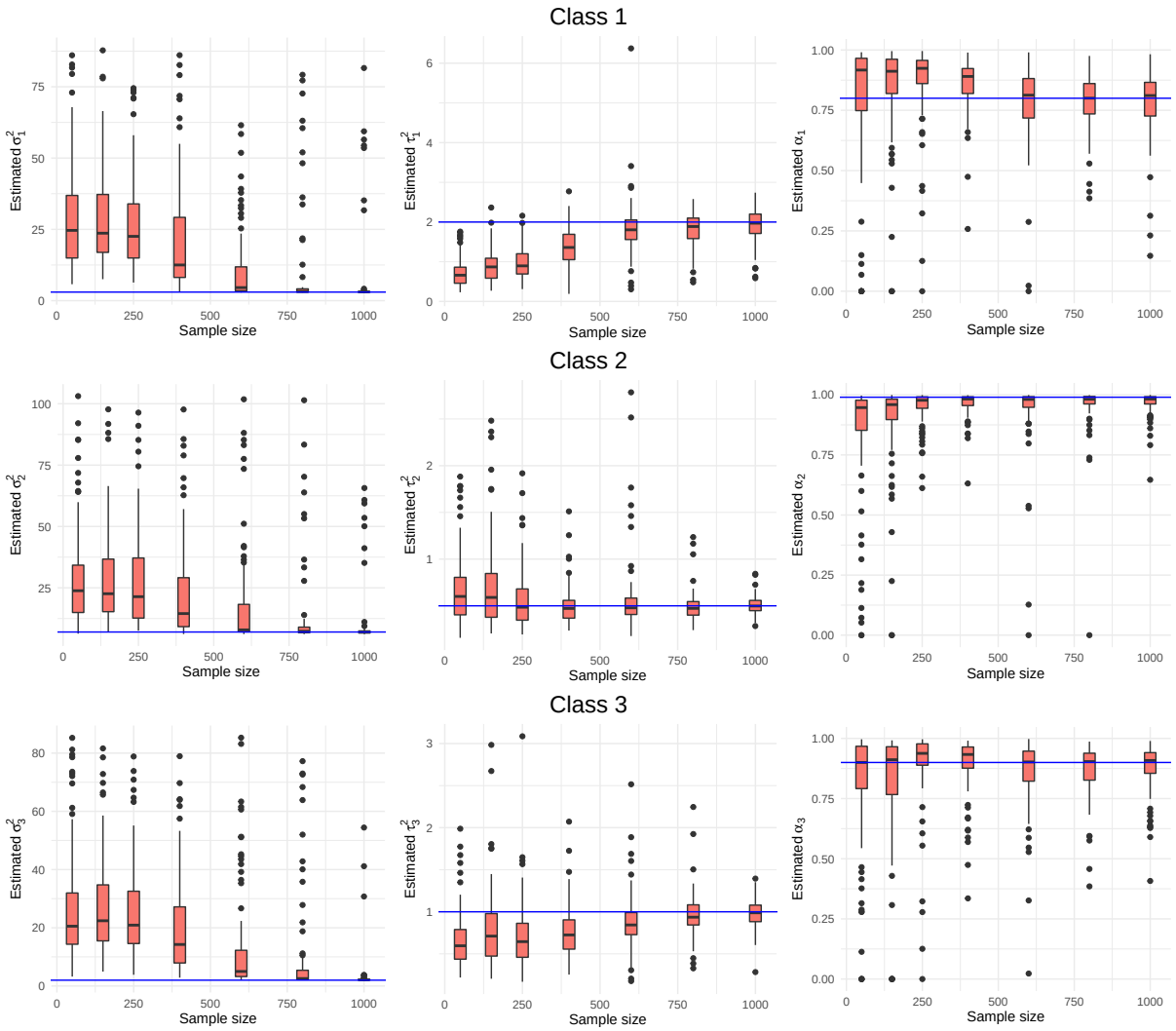

Figure 5.8 – Estimated parameters ${\sigma_k}^2$, ${\tau_k}^2$, and α_k^2 for each class k=1,2,3 as a function of the sample size varying from 50 to 1000. The blue line corresponds to the true value of the parameter.

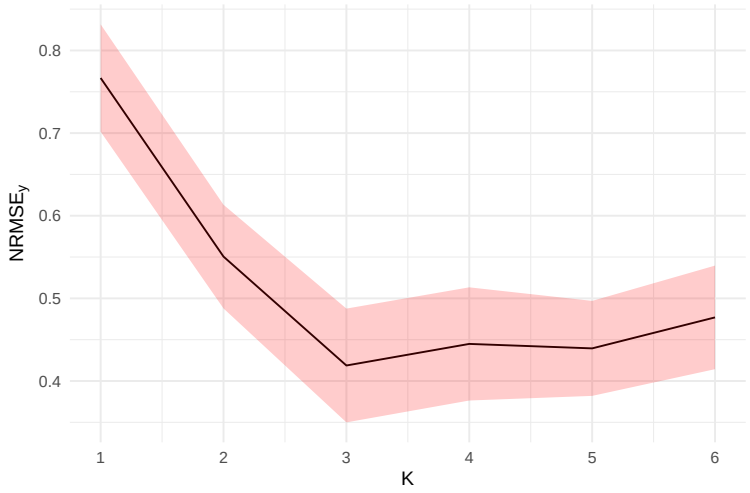

Figure 5.9 – 10-fold cross-validation results for selecting the optimal number of classes K.The black line corresponds to the mean of $NRMSE_y$ in the ten folds, and the red interval corresponds to the mean of $NRMSE_y$ minus and plus the standard deviation of $NRMSE_y$ in the folds.

5.6.4 Comparison to other methods

In this subsection, we compare our proposed method with two other approaches: the finite mixture regression (FMR) and the mixture of linear regression experts (MoE). Unlike our proposed model, in the finite mixture model, defined in equation (5.3.1), the covariates carry no information about the class membership. In this study, we use the Flexmix package in R (Leisch, 2004) to fit the finite mixture regression model. On the other hand, the mixture of linear regression experts corresponds to fitting a linear regression without regularization on each class, and the class membership is determined using the multinomial logit model. The difference between FMR and MoE is that the class membership in MoE depends on the covariates X, but in FMR it does not. The main difference between these two methods and MoR is that we use regularization of the regression coefficients in MoR, but not in MoE nor FMR. For MoR model, we use two covariance methods: the diagonal and CAR noted MoR (diagonal) and MoR (CAR), respectively.

To compare the methods, we perform 100 independent simulations using the same methodology presented in the subsection 5.6.1, and we estimate the parameters using FMR, MoE, MoR (diagonal), and MoR (CAR). Figure 5.10 shows the results of the comparison. The left panel corresponds to $NRMSE_y$ in a test set of size 1000, and the right panel corresponds to the Cramer's V in the training set of size 1000. Note that we

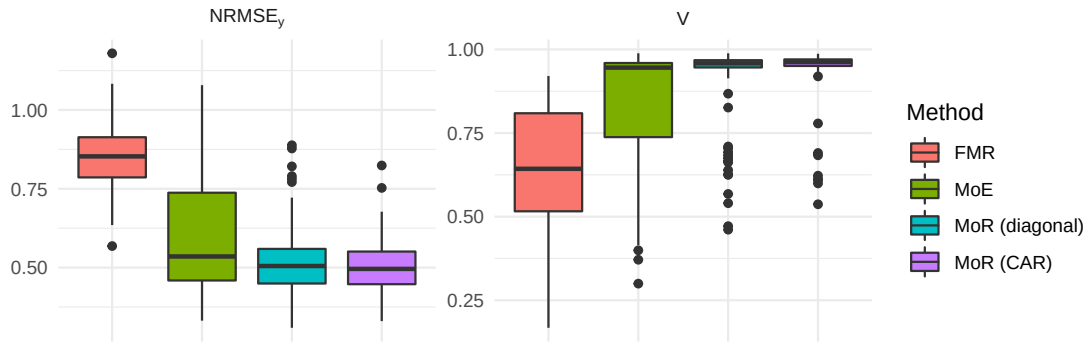

Figure 5.10 – Results of the validation measures ($NRMSE_y$ and Cramér's V) of 100 simulations for finite mixture regression (FMR), mixture of regression experts without regularization (MoE), mixture of generalized Ridge experts with diagonal constant covariance matrix (MoR (diagonal)), and mixture of generalized Ridge experts with CAR covariance (MoR (CAR)).

used the Cramer's V in the training set because the finite mixture model predicts only constant weights of classes. For the two validation measures, the FMR is the worst model. Then comes the MoE model, which has the highest uncertainty in predicting the response variable. The MoR with diagonal and CAR covariance are the best models in predicting Y and finding the classes whit the CAR model is slightly better than the diagonal one.

5.7 Application

In this study, we use the proposed method to predict the significant wave height (H_s) at a location in the Bay of Biscay using wind conditions over the North Atlantic. The significant wave height is the average height of the highest third of the waves, which provides essential information about wave energy. (Obakrim et al., 2022b) used a weather-types regression-based approach for the same problem. Weather typing involves finding the leading atmospheric circulation patterns influencing waves at the target location. After constructing the weather types, using a regression-guided clustering algorithm, Ridge regression is fitted at each class between the response variable H_s and wind conditions. Thus, their model is equivalent to the equation (5.3.1), but the classes were formed a priori using a clustering algorithm. In this study, we construct the weather types (classes) in a statistically optimal way using our proposed method.

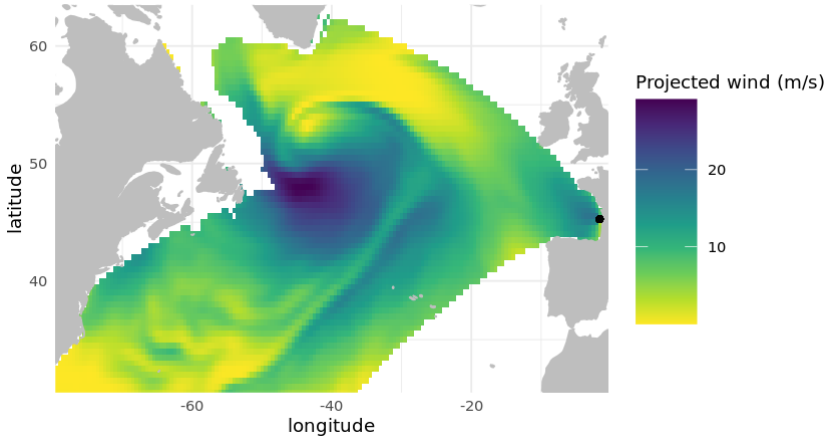

Figure 5.11 – CFSR projected wind in the North Atlantic in 1994-01-01 00h:00. The black point represents the target point.

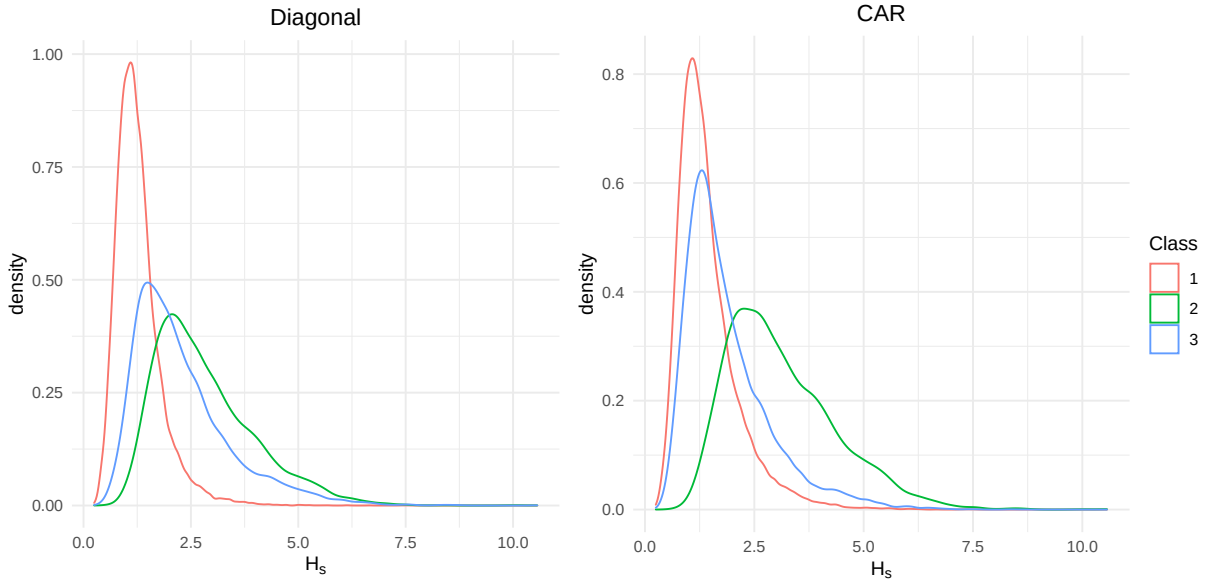

Figure 5.12 – The empirical density of H_s as a function of the estimated classes (1,2, and 3) using the mixture of generalized Ridge with the diagonal (left panel) and CAR (right panel) covariances.

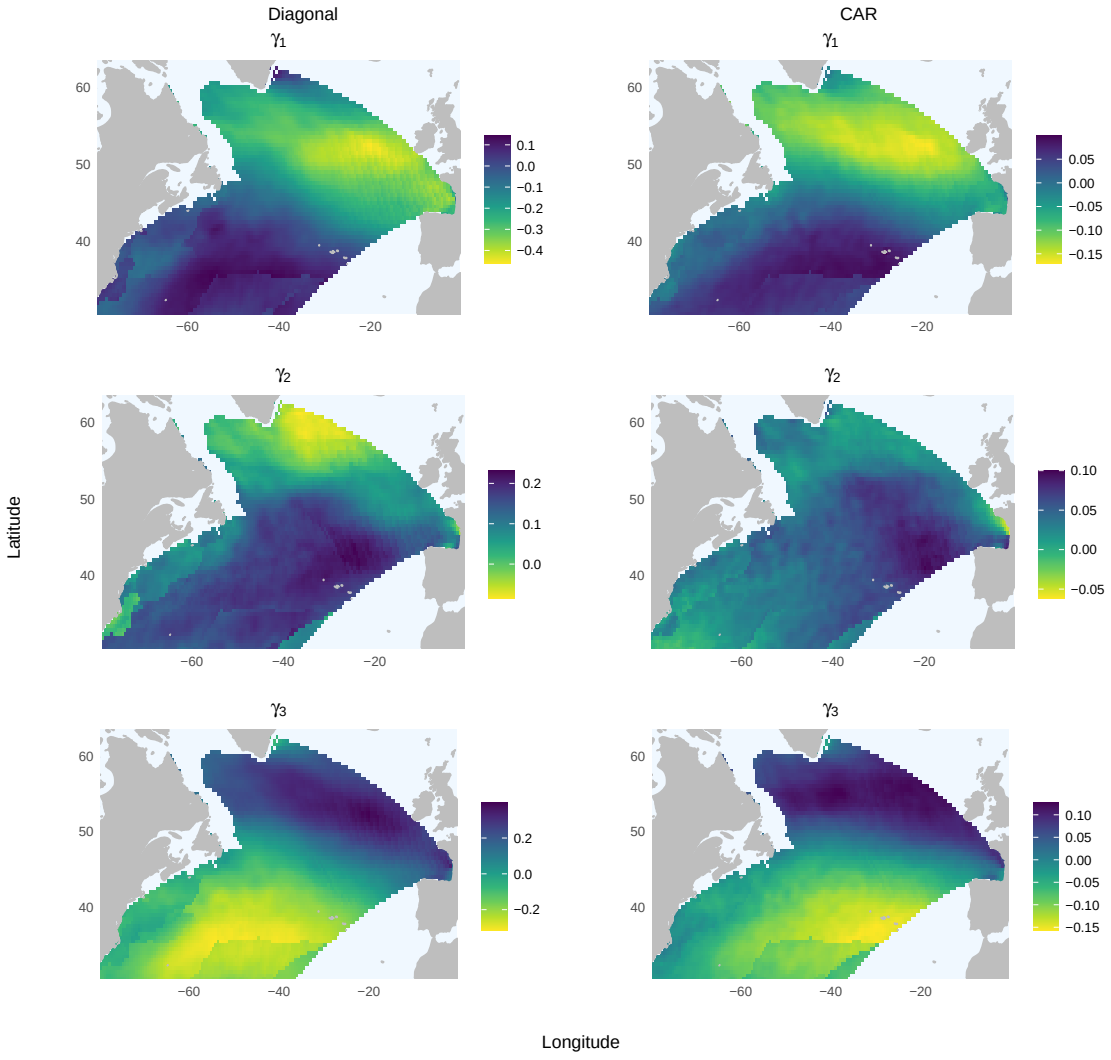

Figure 5.13 – Estimated parameters γ_1, γ_2, and γ_3 for the class 1, 2, and 3, respectively for the diagonal (left panel) and CAR (right panel) cases.

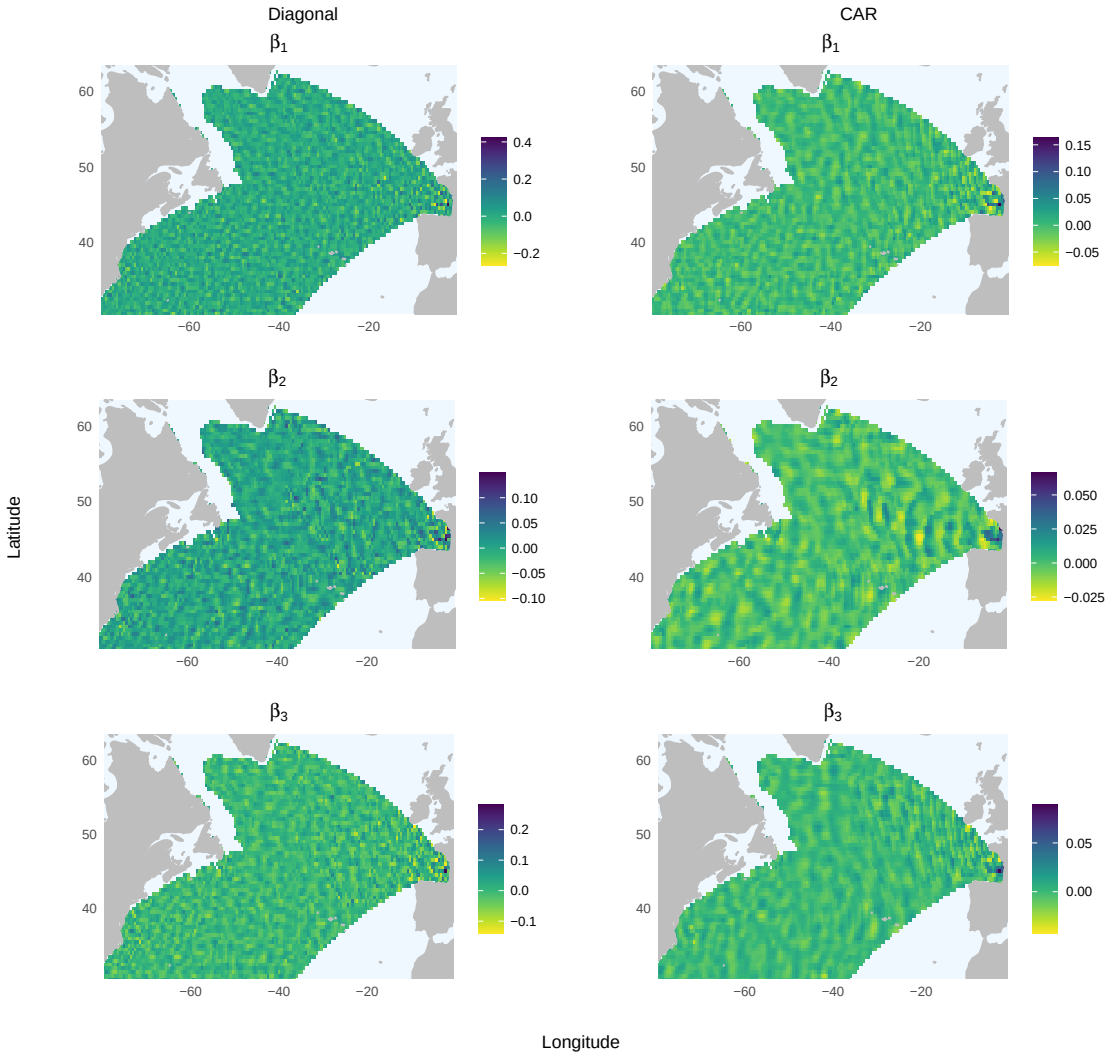

Figure 5.14 – Estimated regression coefficients β_1, β_2, and β_3 for the class 1, 2, and 3, respectively for the diagonal (left panel) and CAR (right panel) cases.

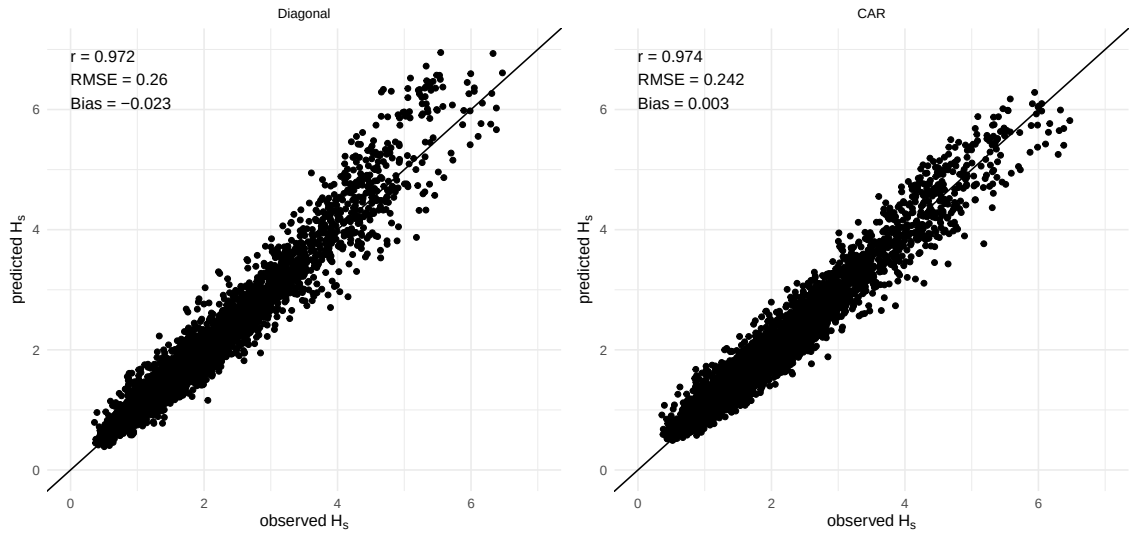

Figure 5.15 – Observed versus predicted H_s in the test set for the diagonal (left panel) and CAR (right panel) covariance cases.

The data used for H_s comes from the Homere hindcast database (Boudière et al., 2013), and the wind data comes from Climate Forecast System Reanalysis (CFSR) (Saha et al., 2010b). The wind data are pre-processed before being used as a predictor (see (Obakrim et al., 2022b) for the pre-processing procedure). We consider 23 years of H_s and wind data from 1994 to 2016 with a temporal resolution of 3 hours. Since the data here are time series, we uses the terms $H_s(t)$, $X(t)$, and $Z(t)$ to refer to H_s, the covariates X, and the class membership variable Z at time t, respectively.

We consider the hierarchical mixture of Ridge experts as defined in equation (5.4.1). The response variable H_s is of size $n = 67088$ and the matrix of covariates $X = X_1, ..., X_d$ is of size 67088×5651. At a given time t, the covariates are defined as $X(t) = X_1(t), ..., X_d(t)$, where $X_j(t)$ is the covariate at time t and location j defined as

$$X_j(t; t_j, \alpha_j) = \frac{1}{2\alpha_j+1} \sum_{i=t-t_j-\alpha_j}^{t-t_j+\alpha_j} W_j^2(i), \qquad t_j + \alpha_j + 1 \leq t \leq t_j - \alpha_j + n \tag{5.7.1}$$

where W_j is the projected wind (figure 5.11) defined as

$$W_j = U_j \, cos\left(\frac{1}{2}(b_j - \theta_j)\right) \tag{5.7.2}$$

Method	r	RMSE(m)	bias(m)
Obakrim et al. (2022b)	0.973	0.272	-0.03
MoR (Diagonal)	0.972	0.26	-0.02
MoR (CAR)	0.974	0.242	-0.003

Table 5.1 – Quantitative comparison of the method developed in Obakrim et al. (2022b), MoR with diagonal and CAR covariances, in the test set using the correlation (r), root mean square error (RMSE), and bias.

U_j is the wind speed, b_j is the great circle bearing, and θ_j is the wind direction at location j. α_j controls the length of the time window, and t_j is the mean travel time of waves which are estimated using the maximum correlation between H_s and the predictor

$$(\hat{t}_j, \hat{\alpha}_j) = arg \max_{t_j, \alpha_j} \Big(corr(H_s, X_j^g(t_j, \alpha_j)) \Big). \tag{5.7.3}$$

The hierarchical mixture of generalized Ridge experts, defined in equation (5.4.1), is fitted to the data, and the parameters are estimated using the proposed variational EM algorithm. For the covariances $\Sigma_{\theta k}$, $k = 1, .., K$, we consider the diagonal and CAR cases. The model's parameters are estimated using data from 1994 to 2013, and the model is evaluated using a test set of the data from 2014 to 2016. Cross-validation results show that $K = 3$ is the optimal number of classes. Figure 5.12 shows the empirical density of H_s as a function of the classes obtained by fitting the mixture of generalized Ridge experts with the diagonal and CAR covariances in the left and right panel, respectively. The classes obtained are physically interpretable and depend on the severity of the sea state. The first class corresponds to waves with low height, the second class to moderate wave heights, and the third class to high wave heights. Figure 5.13 shows the parameters γ_1, γ_2, and γ_3 for the diagonal (left panel) and CAR (right panel). The results of figure 5.13 can be interpreted as follows: for both the diagonal and CAR cases, the first class corresponds to waves coming from the southwest of the target point, and the second class corresponds to waves generated by the wind in the middle of the North Atlantic, and the third class corresponds to waves generated in the north. Figure 5.14 shows the estimated regression coefficients β_1, β_2, and β_3 for the diagonal and CAR cases. As expected, the CAR coefficients have a more smooth spatial structure than the diagonal.

Figure 5.15 shows the scatter plot of the observed versus predicted H_s in the test set using the MoR model with the diagonal (left panel) and CAR covariance (right panel). In terms of correlation (r), root mean square error (RMSE), and bias, the MoR model

with CAR covariance is better than that with diagonal covariance. A comparison between our proposed methods (MoR (Diagonal) and MoR (CAR)) with the method developed in Obakrim et al. (2022b) for the same site is shown in table 6.1. Therefore, the MoR with CAR covariance is the best model for predicting the significant wave height.

5.8 Summary

In this study, we proposed an algorithm for estimating the parameters of a mixture of generalized Ridge regression. We showed that using the EM algorithm is problematic given that the posterior distribution in the E-step is intractable; therefore, we proposed a variational approximation of the E-step. The simulation study shows that the variational EM algorithm can estimate the model's parameters.

The proposed method is applied to predicting the significant wave height at a location in the Bay of Biscay, using wind conditions over the North Atlantic. The resulting classes are physically interpretable and correspond to different wave systems. The proposed method does well in predicting H_s, and the comparative study shows that our method performs better than the method proposed in Obakrim et al. (2022b).

In this work, we have shown that the proposed mixture of generalized Ridge experts can solve multicollinearity and incorporate any covariance structure of the regression coefficients without estimating the regularization hyperparameters using conventional hyperparameter selection methods. However, in our model, the gate network parameters (γ_k, k=1,..., K) are not penalized, which can lead the gate network to overfit (as can be seen in Figure 5.4). Thus, future research could investigate the possibility of including a regularization for the gate network in the same manner as for the regression coefficients. (β_k, k=1,..., K).

5.9 Conclusions

In this chapter, we proposed a mixture of generalized Ridge experts model fitted using a variational EM algorithm. The model can perform regression and classification and predict future class membership. In addition, the model performs regularization and can incorporate any covariance structure of the regression coefficients without the need to estimate the regularization hyperparameters using cross-validation.

The method is applied to create a weather-types-based regression model for predicting the significant wave height from wind conditions. The resulting weather types are physically interpretable and correspond to different wave systems. Furthermore, the model predicts H_s well and is better than the model developed in chapter 3. Note that in chapter 3 model, we used two predictors: the local predictor and the global predictor. However, in this chapter, we have only considered the global predictor because if it is combined with the local predictor, it will be challenging to find a suitable covariance structure for the regression coefficients.

Chapter 6

Modeling the Space-Time Relation between Wind and Significant Wave Height: a Deep Learning Approach

Contents

Note: Part of the results of this chapter are published as M. Michel, S.Obakrim, N.Raillard, P.Ailliot, and V.Monbet, Deep learning for statistical downscaling of sea states[1]*. The other part of the results are presented in the Climate Informatics international conference*[2] *and accepted for publication as S.Obakrim, V.Monbet, N.Raillard, and P.Ailliot, Learning the spatiotemporal relationship between wind and significant wave height using deep learning*[3]

6.1 Preface

In the previous chapters, we developed several statistical methods for modeling the relationship between the significant wave height and wind conditions. As discussed in

1. The article can be found in https://doi.org/10.5194/ascmo-8-83-2022
2. https://ncics.org/news/events/ci2022/
3. The preprint can be found in https://doi.org/10.48550/arXiv.2205.13325

Chapter1, deep learning methods are gaining attention in the climate community and were used in many studies for downscaling climate variables such as precipitation and temperature. To the best of our knowledge, deep learning methods have not been yet used for sea state downscaling; therefore, in Michel et al. (2022), we developed a downscaling model for sea state parameters using a convolutional neural network model. The model is based on the predictors (local and global) defined in Chapter3, and the model does well in predicting H_s.

As for now, the methods developed in this thesis are based on the predictors defined in Chapter 3, where a preprocessing step is used to define the temporal structure of the global predictor. The preprocessing step is based on estimating the optimal lagged wind conditions (interpreted as the travel time of waves) using the maximum correlation between H_s and wind conditions. The objective of this chapter is to construct the link function between H_s and wind conditions without this preprocessing step, using deep learning.

6.2 Abstract

Ocean wave climate significantly impacts near-shore and off-shore human activities, and its characterization can help design ocean structures such as wave energy converters and sea dikes. Therefore, engineers need long time series of ocean wave parameters. Numerical models are a valuable source of ocean wave data; however, they are computationally expensive. Consequently, statistical and data-driven approaches have gained increasing interest in recent decades. Using a two-stage deep learning model, this work investigates the spatiotemporal relationship between North Atlantic wind and significant wave height (H_s) at an off-shore location in the Bay of Biscay. The first step uses convolutional neural networks (CNNs) to extract the spatial features that contribute to H_s. Then, long short-term memory (LSTM) is used to learn the long-term temporal dependencies between wind and waves.

6.3 Introduction

Characterization of wave climate is required for many marine applications, such as designing coastal and off-shore structures and planning ship operations. Wind waves are generated by the surface wind, with the local wind creating the wind sea and wind from

distant areas creating waves that propagate and form swells (Young (1999)). Waves in the Bay of Biscay depend on local and large-scale wind conditions in the North Atlantic (Charles et al. (2012a)); however, swells generally dominate the sea state. Swells travel large distances and take up to five days to cross the Atlantic from Cape Hatteras to the Bay of Biscay (Ardhuin and Orfila (2018)). Consequently, waves observed at a given location depend on wind conditions over the North Atlantic in a time window of several days, and it is challenging to reproduce this complex spatiotemporal relationship using machine learning. This work aims to propose a deep learning approach that learns this relationship.

The advantage of deep learning methods (Goodfellow, Bengio, and Courville (2016)) lies in their ability to build hierarchical representations of predictors. In particular, in the case of spatial data, convolutional neural networks (CNNs) allow for learning complex spatial features from the data (Gu et al. (2018)). Moreover, long short memory (LSTMs) (Hochreiter and Schmidhuber (1997)) have proven to be very successful in predicting time series and sequence data. In this work, we propose a non-expensive data-driven approach that learns the underlying spatiotemporal structure of the relationship between wind and waves using a two-stage model based on CNNs and LSTM.

This paper is organized as follows. Section 2 presents the problem of downscaling ocean waves and related works. Section 3 describes the data used in this work. Section 4 presents the proposed two-stage model, the architecture, and the training process. Section 5 discusses the results of this work. Finally, Section 6 presents the conclusions and future work directions.

6.4 Problem statement and related work

The problem of improving the spatial resolution of climate variables is known under the name of downscaling (Maraun et al. (2010)). Downscaling approaches attempt to construct a numerical or statistical link between large-scale and local-scale variables. The advantage of statistical downscaling (SD) over numerical models is primarily in terms of computational efficiency. A rigorous comparison of the two approaches can be found in (Wang, Swail, and Cox (2010); Laugel et al. (2014)).

In the case of ocean waves, wind (Obakrim et al. (2022b)) or sea level pressure (SLP) (Camus et al. (2014a)) are commonly used to downscale ocean wave parameters. However, to establish a link function between the wind (or SLP) and the local ocean wave param-

eters, it is necessary to consider a large spatial and temporal coverage and, consequently, many potential explanatory variables that are highly correlated. Some methods determine the wave generation area for any ocean location worldwide. For example, ESTELA (Pérez et al. (2014)) is a numerical model that uses spectral information to select the fraction of energy that travels to the target point from selected source points. The ESTELA method can be used to design statistical downscaling methods. For instance, Camus et al. (2014a), and Hegermiller et al. (2017) used the ESTELA method to define the predictors used in their SD model.

Obakrim et al. (2022b) proposed a data-driven approach that determines the wave generation area by estimating the travel time of waves generated in each considered source point that reaches the target point. Then, the predictors were defined based on the wave generation area, and finally, a SD model based on weather types was built.

As far as we know, the existing methods for SD of ocean wave parameters define a priori the spatiotemporal structure of the predictors, and then the SD model is built using these predictors. This study proposes a deep learning approach that automatically learns the spatiotemporal relationship between wind and waves.

6.5 Data preparation

The Climate Forecast System Reanalysis (CFSR) (Saha et al. (2010a)) hourly wind data is considered in this study as a predictor. CFSR is a global reanalysis developed by the National Centers for Environmental Prediction (NCEP) that covers the period from 1979 to the present with an hourly time step and a spatial resolution of 0.5°by 0.5°. The historical H_s data is extracted from the hindcast database HOMERE (Boudière et al. (2013)) at the target location with spatial coordinates (45.2°N, 1.6°W) located in the Bay of Biscay. The temporal resolution of both wind and H_s data is up-scaled to 3-hourly data. The period from 1994 to 2016 is considered in this study, leading to a dataset with $n = 67208$ observations.

Instead of using both zonal and meridional components as a predictor, we use the projected wind (Obakrim et al. (2022b)) defined at each location j and time t, as

$$W_j(t) = U_j(t)\, cos^2\left(\frac{1}{2}(b_j - \theta_j(t))\right) \tag{6.5.1}$$

where $W_j(t)$ is the projected wind, $U_j(t)$ is the wind speed, $\theta_j(t)$ is the wind direction,

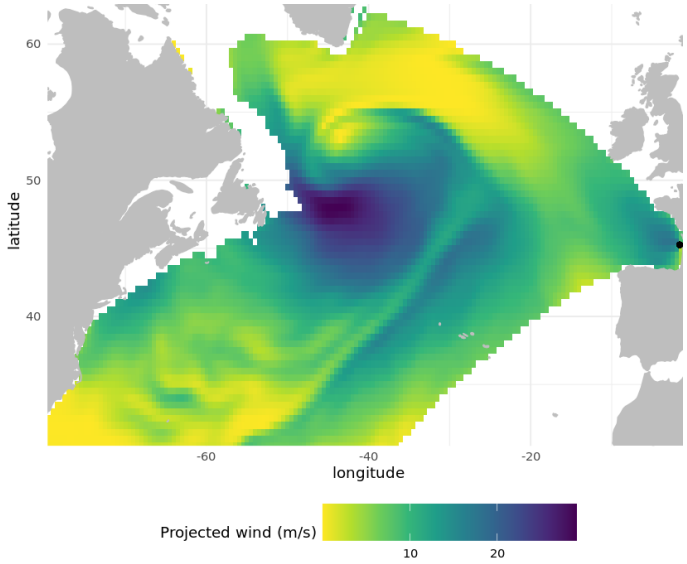

Figure 6.1 – The projected wind defined in (6.5.2) in 1994-01-01 00h:00. The black point represents the target point

and b_j is the great circle bearing from the source point j to the target point. Under the assumption that waves travel in great circle paths, grid points whose paths are blocked by land are neglected (Figure 6.1). Therefore, we define the global predictor at time t as

$$X^{(g)}(t) = (W_1^2(t), ..., W_p^2(t)) \tag{6.5.2}$$

where $p = 5651$ is total number of grid points.

Following (Obakrim et al. (2022b)), in order to capture the wind sea, we also define the local predictor as

$$X^{(\ell)}(t) = \{U(t), U^2(t), U^3(t), U^2(t)F(t), U(t-1), U^2(t-1), U^3(t-1), U^2(t-1)F(t-1)\} \tag{6.5.3}$$

where $U(t)$ is the wind speed at the target point and $F(t)$ is the fetch length at time t, calculated as the minimum of the distance from the target point to shore in the direction from which the wind is blowing and $500km$. The fetch has an important effect on wind sea characteristics (Ardhuin and Orfila (2018)); therefore, it is commonly used to construct empirical wind wave models.

6.6 Proposed methodology

As mentioned in the last section, state-of-the-art statistical methods for downscaling wave parameters usually use a preprocessing step to create features that consider the wave generation area. This study proposes a deep learning approach that automatically extracts these features. Since waves may take several days to reach the target point, the history and current wind can be used to predict H_s. An example of this type of model could have the following form

$$H_s(t) = f(X^{(g)}(t - t_{max}), ..., X^{(g)}(t)) \tag{6.6.1}$$

where t_{max} can be interpreted as the maximum travel time of the waves and will be referred to as such in the following. However, this approach can be computationally challenging given the dimension of the predictor (5651 in our case). Instead, in this study, we propose to use current wind conditions to estimate current and future H_s.

In order to describe the complex spatiotemporal relationship between wind and H_s, we propose the following two-stage model

$$\begin{aligned} &\textbf{1}^{\textbf{st}}\ \textbf{stage:}\ [H_s(t|X^{(g)}(t)), ..., H_s(t + t_{max}|X^{(g)}(t))] = f(X^{(g)}(t)) + \epsilon(t), \quad f : \mathbb{R}^p \to \mathbb{R}^{t_{max}} \\ &\textbf{2}^{\textbf{nd}}\ \textbf{stage:}\ H_s(t) = g(X^{(g)}(t), f(X^{(g)}(t - t_{max})), ..., f(X^{(g)}(t))) + \epsilon\prime(t), \quad g : \mathbb{R}^{t_{max} * t_{max} + 8} \to \mathbb{R} \end{aligned} \tag{6.6.2}$$

where the notation $H_s(t_1|X^{(g)}(t_2))$ represents the contribution of wind conditions at time t_2 in H_s at time t_1. ϵ and $\epsilon\prime$ are the errors of the 1st stage and 2nd stage, respectively. The 1st stage estimates the current and future H_s using current wind conditions. The 2nd stage estimates H_s using the past predictions obtained from the 1st stage. Along with the local predictor $X^{(g)}$, the input for the 2nd stage is a $t_{max} * t_{max}$ matrix of the form

$$\begin{pmatrix} \hat{H}_s(t - t_{max}|X^{(g)}(t - t_{max})) & \dots & \hat{H}_s(t|X^{(g)}(t - t_{max})) \\ \vdots & \ddots & \vdots \\ \hat{H}_s(t|X^{(g)}(t)) & \dots & \hat{H}_s(t + t_{max}|X^{(g)}(t)) \end{pmatrix} \tag{6.6.3}$$

where $\hat{H}_s(t_1|X^{(g)}(t_2))$ represents the prediction, obtained from the 1st stage, of the contribution of wind conditions at time t_2 in the H_s at time t_1. When $t_1 = t_2$, this prediction represents the wind sea (first column of the matrix in equation (6.6.3)); for $t_1 > t_2$, on

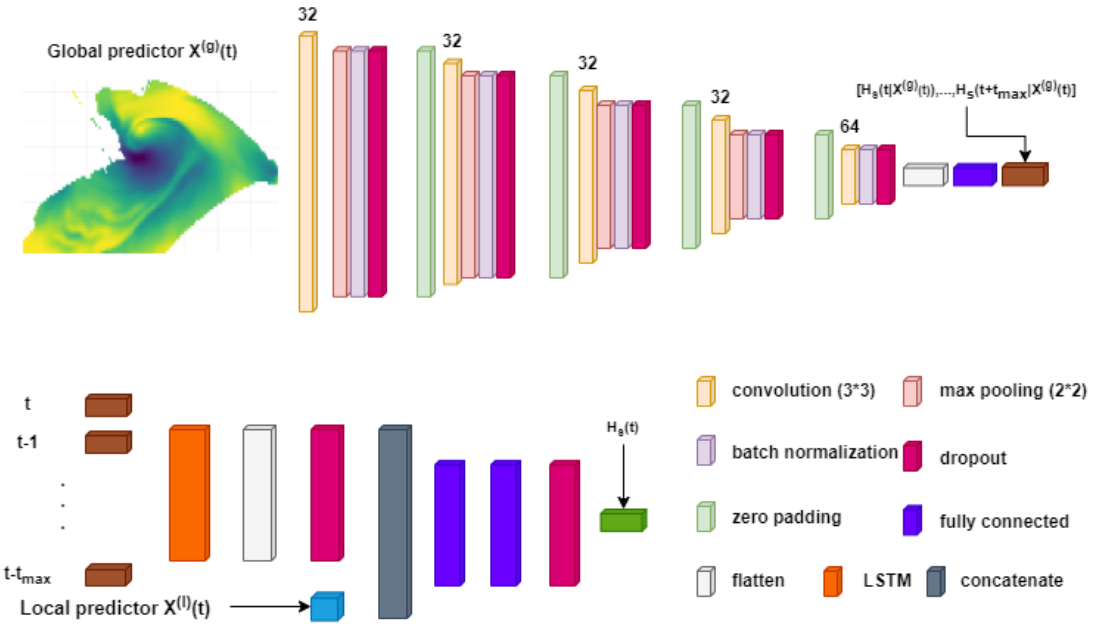

Figure 6.2 – Architecture of the two-stage model in equation (6.6.2)

the other hand, the prediction represents the H_s caused by swells.

The general structure of the model is shown in Figure 6.2. The 1st stage consists of a series of 3*3 convolutions followed by the ReLU activation function, 2*2 max pooling layer, Batch Normalisation, then a flatten followed by a dense layer. The 2nd stage starts with an LSTM layer that learns the long-term dependencies of the $(t - t_{max}, ..., t)$ outputs of the 1st stage. The output of the LSTM layer is then concatenated with the local predictor X^l and fed into two fully connected layers. The dropout layer is used in both stages to prevent the network from overfitting. The loss function chosen in this study is the mean squared error (MSE) which is expressed as

$$\begin{aligned} MSE(1^{\text{st}}\text{ stage}) &= \frac{1}{t_{max}} \sum_{i=0}^{t_{max}} \frac{1}{n - t_{max} - 1} \sum_{t=1}^{n-t_{max}} (H_s(t+i) - \hat{H}_s(t+i|X^{(g)}(t)))^2 \\ MSE(2^{\text{nd}}\text{ stage}) &= \frac{1}{n} \sum_{t=1}^{n} (H_s(t) - \hat{H}_s(t))^2 \end{aligned} \tag{6.6.4}$$

Where n is the total number of observations and $\hat{H}_s$ is the prediction of H_s. The Keras framework with Tensorflow backend (Chollet et al. (2015)) is used in this work to train the model on an Nvidia K80s GPU using the Adam optimizer (Kingma and Ba (2014)) and mini-batches of 64.

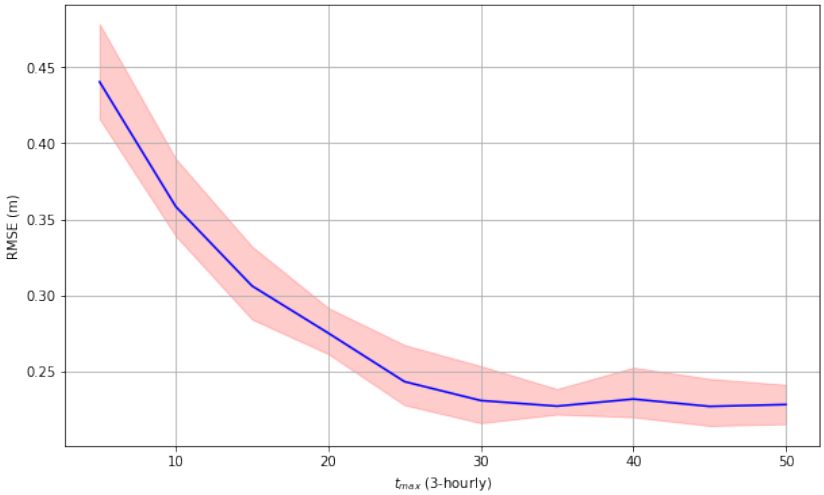

Figure 6.3 – Results of cross-validation using different values of t_{max}. The blue line represents the mean of RMSE, and the red interval represents the minimum and maximum RMSE

6.7 Results

The period from 1994 to 2011 is used to train the two-stage model and the period from 2012 to 2014 serves as the validation period. The measures chosen in this paper to validate the analysis are the correlation coefficient (r), the root mean square error (RMSE), and the bias. Different values for the maximum travel time of waves t_{max} are tested, and the results of k-fold cross-validation (with $k = 5$) are shown in Figure 6.3. The RMSE stabilises approximately at $t_{max} = 30 \times 3h$, which corresponds to about 3.3 days, and the gain is substantial compared to using $t_{max} = 5$. This means that wind conditions over a time window of at least 3.3 days must be considered to characterize the wave climate at the target location. In the following, the value of t_{max} is chosen equal to 30.

Figure 6.4 shows the scatter plot of observed versus predicted values of H_s using the two-stage model (6.6.2). The RMSE in the validation period equals $0.21m$ for an H_s of mean $1.9m$ and standard deviation $1.1m$. The model performs well in predicting H_s and accounts for both wind and swell. The validation measures in the calibration and validation periods are almost the same. This means that the model does not overfit the training data and generalizes well the relationship between wind and waves. Furthermore, the seasonality of H_s is well captured by the two-stage model, as shown in Figure 6.5.

A comparison of the two-stage model with two other statistical approaches is made in Table 6.1. The first approach, described in (Obakrim et al. (2022b)), is based on weather

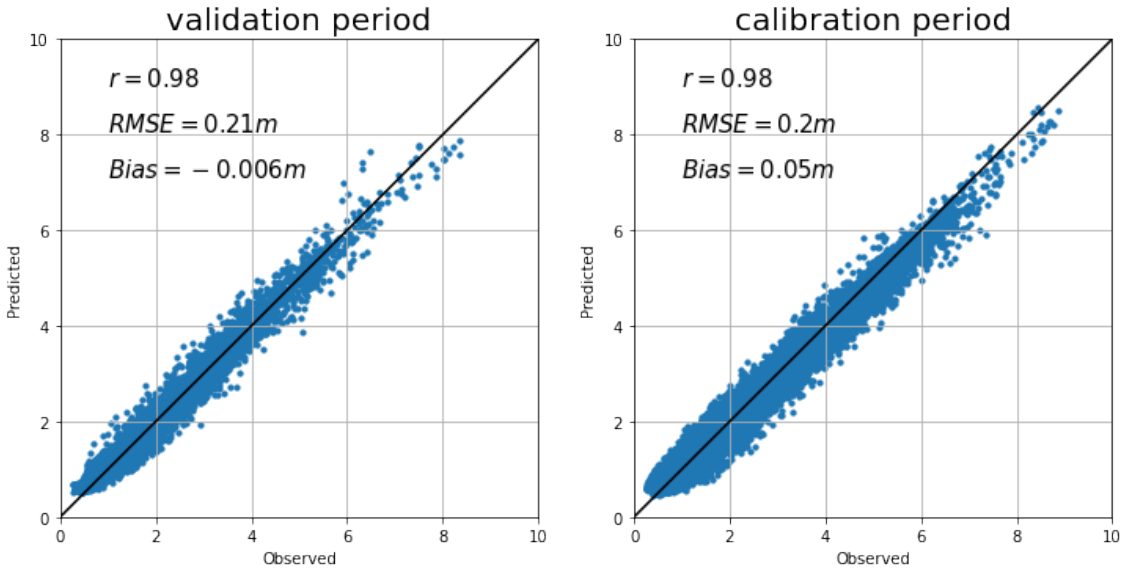

Figure 6.4 – Observed versus predicted H_s in the validation period (left panel) and calibration period (right panel)

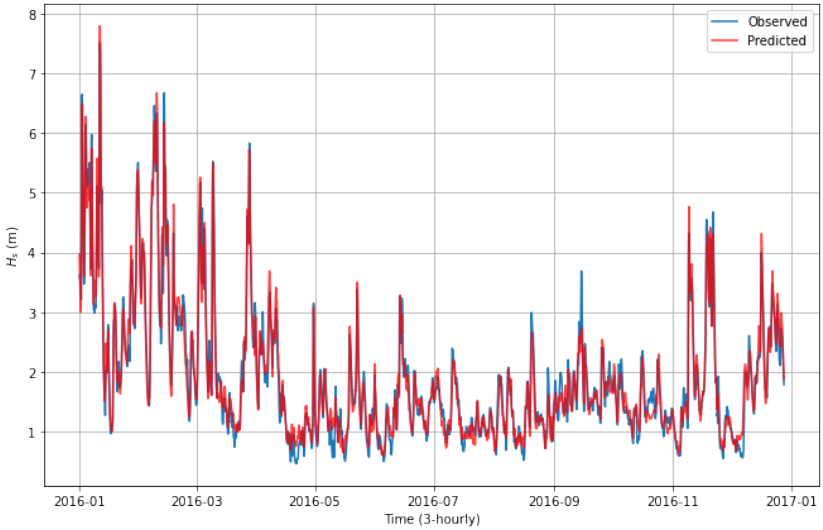

Figure 6.5 – Time series of observed (blue line) and predicted (red line) H_s in 2016

Method	r	RMSE(m)	bias(m)
two-stage model	0.98	0.21	-0.006
weather types	0.97	0.27	-0.03
H-CNN	0.97	0.27	-0.04

Table 6.1 – Comparison of the two-stage model, weather types, and H-CNN methods.

types (Maraun et al. (2010)). As for the present work, the local and global predictors were considered. However, to reduce the dimension of the predictor, a single predictor is extracted at each spatial location j to predict H_s at time t. It is defined a priori as

$$X_j^{(g)}(t; t_j, \alpha_j) = \frac{1}{2\alpha_j+1} \sum_{i=t-t_j-\alpha_j}^{t-t_j+\alpha_j} W_j^2(i), \quad t_j + \alpha_j + 1 \leq t \leq t_j - \alpha_j + n \tag{6.7.1}$$

where t_j is the travel time of waves, α_j controls the length of the time window, and W_j is the projected wind at location j. The parameters t_j and α_j were estimated using the maximum correlation between h_s and the global predictor. The second method (Michel et al. (2022)) uses CNNs to predict H_s using the same predictors as in (Obakrim et al. (2022b)). Thus, the main difference with the approach proposed in this work is that the temporal dimension of the global predictor is reduced a priori using the preprocessing step based on the maximum correlation described above. The numerical results in Table 6.1 indicate that the two-stage model significantly outperforms the other two methods in terms of the validation measures.

6.8 Summary

This study proposes a two-stage model based on deep learning to predict H_s using wind conditions. The model can automatically learn the underlying spatiotemporal structure of the relationship between wind and waves. The model does well in predicting H_s and is computationally inexpensive (about 5min using a computer of 30GB RAM, two cores CPU, and a 16GB GPU). The proposed methodology is based on two stages which are trained separately. A natural question for future work is whether we can estimate the parameters jointly using back-propagation and eventually speed up the training process and improve the results. Future work also includes using the method to predict other sea state parameters, such as wave direction and period.

The proposed method can be used for climate and weather studies at any ocean location worldwide. For nearby locations, one can train only the 2^{nd} stage at each location, using the weights of one location as initialization for the others and leaving the 1^{st} stage the same. The model can also learn from buoy data instead of hindcast data and eventually fill in the gaps and complete historical data.

6.9 Conclusions

In this chapter, we studied the potential of using deep learning models for downscaling the significant wave height using wind conditions. We developed a two-stage model based on convolutional neural networks and long short term memory deep learning models capable of predicting H_s without a preprocessing step that defines the temporal structure of the predictors as in Chapter 3. Furthermore, the proposed model predicts well H_s and outperforms the statistical methods developed in the previous chapters.

Chapter 7

Conclusions

Contents

7.1 Summary

In this thesis, we investigated the use of statistical and deep learning methods for modeling the relationship between wind conditions and the significant wave height. At first, we developed a weather-types-based regression method that predicts H_s using wind conditions. The weather types model predicts well H_s; however, the individual regression models for each weather type do not consider that the covariates (wind conditions) have a spatial structure. Therefore, in Chapter 4, we developed a new method for estimating the parameters of generalized Ridge regression that can incorporate any covariance structure of the regression coefficients, and we focused on the use of spatial covariances such as Matérn and conditional autoregressive (CAR). Then, in Chapter 5, we combined the ideas of Chapters 3 and 4 and proposed a mixture of generalized Ridge experts, which is estimated using a variational EM algorithm. The mixture of generalized Ridge experts is used as a weather-types-regression-based model for downscaling H_s, and the model outperforms the model proposed in Chapter 3. Finally, in Chapter 6, we investigated the use of deep learning for downscaling sea state parameters and proved the potential of these methods in modeling the spatiotemporal relationship between wind and waves.

The main findings of this work can be summarized as follows:

- Taking into account lagged wind conditions is important in order to statistically model the relationship between wind and waves
- Using a local and a global predictor is beneficial in constructing the link function between wind and waves (Chapter 3 and 5)
- Lagged wind conditions can be considered either by using a preprocessing step as in Chapter 3 or can be learned automatically using deep learning (Chapter 6)

Method	local predictor	r	RMSE(m)	bias(m)
Weather types (Chapter 3)	yes	0.973	0.272	-0.03
GR-Diagonal (Chapter 4)	no	0.941	0.414	-0.0004
GR-Matérn (Chapter 4)	no	0.956	0.354	-0.04
GR-CAR (Chapter 4)	no	0.957	0.352	-0.06
MoR-Diagonal (Chapter 5)	no	0.972	0.26	-0.02
MoR-CAR (Chapter 5)	no	0.974	0.242	-0.003
H-CNN (Michel et al. (2022))	yes	0.972	0.271	-0.04
Two-stage model (Chapter 6)	yes	0.98	0.21	-0.006

Table 7.1 – Quantitative comparison of the method developed in this thesis. The first method is the weather-types-based model developed in Chapter 3, GR-diagonal, GR-Matérn, and GR-CAR are the methods developed in Chapter 4 (GR for generalized Ridge), MoR-Diagonal and MoR-CAR are the methods proposed in Chapter 5, H-CNN is the deep learning method developed in Michel et al. (2022). Finally, the two-stage model is the method proposed in 6. The local predictor column indicates whether the model takes into account the local predictor.

- It is beneficial, both in terms of prediction accuracy and interpretability, to consider that the regression coefficients have a spatial structure when the covariates have a spatial structure
- A weather-types-based-regression model can be constructed using mixture of regression models, especially mixture of experts, given that they allow for future predictions of weather types
- Constructing weather types using a mixture model allows the weather types to be evaluated based on the prediction of H_s, which leads to optimal estimations (Chapter 5)

A comparison between all the methods developed in this thesis is shown in table 7.1. Regarding the validation measures (correlation r, RMSE, and bias), the two-stage model outperforms the other methods; However, in terms of interpretability, it is the least interpretable given the complex model architecture. On the other hand, the mixture of generalized Ridge experts provides good results and physically interpretable weather types, yet in terms of computational complexity, it is the most expensive model.

7.2 Perspectives

The objective of this thesis was to study the use of statistical and machine learning methods for sea state characterization. We believe that our work opens new research

avenues listed in the following non-exhaustive list:

- The methods proposed in this study were only used for significant wave height downscaling; however, it would be interesting to verify if they also perform well for other sea state parameters, such as wave period and direction. Furthermore, The downscaling methods can be extended to nearby locations by adjusting only the local predictor and keeping the same global predictor for nearby locations. It is also possible to apply our proposed methods to any ocean location worldwide; however, the quality of the model will depend on the quality of the available wind and wave data and the tidal conditions at the target point.
- Investigate the use of the proposed methods for operational applications such as long-term sea state monitoring, short-term forecasting or hindcasting.
- The EM algorithm proposed for generalized Ridge regression is limited to linear regression with Gaussian errors. A natural question that arises is whether the algorithm can be extended to be used for generalized linear models.
- The proposed mixture of generalized Ridge experts has been shown to work well in both simulations and applications. However, the convergence of the variational EM algorithm used to estimate the parameters is not guaranteed, and it would be interesting to investigate the theoretical convergence of the algorithm.
- While this thesis work is oriented to wave parameters prediction, some of our proposed methods, such as the EM algorithm for generalized Ridge and mixture of generalized Ridge experts are domain-free. They can be applied to any domain, and it could be interesting to investigate their performance compared to other methods.

Bibliography

Abramowitz, Milton, Irene A Stegun, and Robert H Romer (1988), *Handbook of mathematical functions with formulas, graphs, and mathematical tables*.

Agrawal, JD and MC Deo (2002), « On-line wave prediction », *in*: *Marine structures* 15(*1*), pp. 57–74.

Ailliot, Pierre and Valérie Monbet (2012), « Markov-switching autoregressive models for wind time series », *in*: *Environmental Modelling & Software* 30, pp. 92–101.

Ailliot, Pierre et al. (2015), « Stochastic weather generators: an overview of weather type models », *in*: *Journal de la Société Française de Statistique* 156(*1*), pp. 101–113.

Allen, David M (1974), « The relationship between variable selection and data agumentation and a method for prediction », *in*: *technometrics* 16(*1*), pp. 125–127.

Ambikasaran, Sivaram et al. (2015), « Fast direct methods for Gaussian processes », *in*: *IEEE transactions on pattern analysis and machine intelligence* 38(*2*), pp. 252–265.

Ardhuin, Fabrice and Alejandro Orfila (2018), « Wind waves », *in*: *New Frontiers in Operational Oceanography*, pp. 393–422.

Ardhuin, Fabrice et al. (2019), « Observing sea states », *in*: *Frontiers in Marine Science*, p. 124.

Bachoc, François (2013), « Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer experiments », PhD thesis, Université Paris-Diderot-Paris VII.

Baño-Medina, Jorge, Rodrigo Manzanas, and José Manuel Gutiérrez (2020), « Configuration and intercomparison of deep learning neural models for statistical downscaling », *in*: *Geoscientific Model Development* 13(*4*), pp. 2109–2124.

Barnston, Anthony G and Robert E Livezey (1987), « Classification, seasonality and persistence of low-frequency atmospheric circulation patterns », *in*: *Monthly weather review* 115(*6*), pp. 1083–1126.

Beal, Matthew James (2003), *Variational algorithms for approximate Bayesian inference*, University of London, University College London (United Kingdom).

Bellone, Enrica, James P Hughes, and Peter Guttorp (2000), « A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts », *in*: *Climate research* 15(*1*), pp. 1–12.

Benestad, Rasmus E, Deliang Chen, and Inger Hanssen-Bauer (2008), *Empirical-statistical downscaling*, World Scientific Publishing Company.

Bernardo, JM et al. (2003), « The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures », *in*: *Bayesian statistics* 7(*453-464*), p. 210.

Besag, Julian, Jeremy York, and Annie Mollié (1991), « Bayesian image restoration, with two applications in spatial statistics », *in*: *Annals of the institute of statistical mathematics* 43(*1*), pp. 1–20.

Bishop, Christopher M and Nasser M Nasrabadi (2006), *Pattern recognition and machine learning*, vol. 4, 4, Springer.

Bitner-Gregersen, Elzbieta M et al. (2016), « Sea state conditions for marine structures' analysis and model tests », *in*: *Ocean Engineering* 119, pp. 309–322.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017), « Variational inference: A review for statisticians », *in*: *Journal of the American statistical Association* 112(*518*), pp. 859–877.

Boé, J et al. (2006), « A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling », *in*: *Journal of Geophysical Research: Atmospheres* 111(*D23*).

Boé, J et al. (2007), « Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies », *in*: *International Journal of Climatology: A Journal of the Royal Meteorological Society* 27(*12*), pp. 1643–1655.

Boonstra, Philip S, Bhramar Mukherjee, and Jeremy MG Taylor (2015), « A small-sample choice of the tuning parameter in ridge regression », *in*: *Statistica Sinica* 25(*3*), p. 1185.

Boudière, Edwige et al. (2013), « A suitable metocean hindcast database for the design of Marine energy converters », *in*: *International Journal of Marine Energy* 3, e40–e52.

Cagigal, Laura et al. (2020), « A multivariate, stochastic, climate-based wave emulator for shoreline change modelling », *in*: *Ocean Modelling* 154, p. 101695.

Campos, RM et al. (2018), « Extreme wind-wave modeling and analysis in the south Atlantic ocean », *in*: *Ocean Modelling* 124, pp. 75–93.

Camus, Paula et al. (2014a), « A method for finding the optimal predictor indices for local wave climate conditions », *in*: *Ocean Dynamics* 64(*7*), pp. 1025–1038.

Camus, Paula et al. (2014b), « A weather-type statistical downscaling framework for ocean wave climate », *in*: *Journal of Geophysical Research: Oceans* 119(*11*), pp. 7389–7405.

Cannon, Alex J and Paul H Whitfield (2002), « Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models », *in*: *Journal of Hydrology* 259(*1-4*), pp. 136–151.

Casas-Prat, Mercè, Xiaolan L Wang, and Joan P Sierra (2014), « A physical-based statistical method for modeling ocean wave heights », *in*: *Ocean Modelling* 73, pp. 59–75.

Cawley, Gavin C et al. (2003), « Statistical downscaling with artificial neural networks. », *in*: *ESANN*, pp. 167–172.

Chamroukhi, F. (2010), « Hidden process regression for curve modeling, classification and tracking », Ph.D. Thesis, Université de Technologie de Compiègne, URL: https://chamroukhi.com/FChamroukhi-PhD.pdf.

Charles, Elodie et al. (2012a), « Climate change impact on waves in the Bay of Biscay, France », *in*: *Ocean Dynamics* 62(*6*), pp. 831–848.

Charles, Elodie et al. (2012b), « Present wave climate in the Bay of Biscay: spatiotemporal variability and trends from 1958 to 2001 », *in*: *Journal of Climate* 25(*6*), pp. 2020–2039.

Chollet, François et al. (2015), *keras*.

Cressie, Noel and Prasenjit Kapat (2008), « Some diagnostics for Markov random fields », *in*: *Journal of computational and graphical statistics* 17(*3*), pp. 726–749.

Cressie, Noel and Christopher K Wikle (2015), *Statistics for spatio-temporal data*, John Wiley & Sons.

Day, Neil E (1969), « Estimating the components of a mixture of normal distributions », *in*: *Biometrika* 56(*3*), pp. 463–474.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977a), « Maximum likelihood from incomplete data via the EM algorithm », *in*: *Journal of the Royal Statistical Society: Series B (Methodological)* 39(*1*), pp. 1–22.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977b), « Maximum likelihood from incomplete data via the EM algorithm », *in*: *Journal of the Royal Statistical Society: Series B (Methodological)* 39(*1*), pp. 1–22.

Deo, MC and C Sridhar Naidu (1998), « Real time wave forecasting using neural networks », *in*: *Ocean engineering* 26(*3*), pp. 191–203.

DeSarbo, Wayne S and William L Cron (1988), « A maximum likelihood methodology for clusterwise linear regression », *in*: *Journal of classification* 5(*2*), pp. 249–282.

Duan, WY et al. (2016), « A hybrid EMD-SVR model for the short-term prediction of significant wave height », *in*: *Ocean Engineering* 124, pp. 54–73.

El Assaad, Hani et al. (2016), « A variational Expectation–Maximization algorithm for temporal data clustering », *in*: *Computational Statistics & Data Analysis* 103, pp. 206–228.

Fernández-Montes, Sonia et al. (2013), « Spring and summer extreme temperatures in Iberia during last century in relation to circulation types », *in*: *Atmospheric Research* 127, pp. 154–177.

Flecher, Cedric et al. (2010), « A stochastic daily weather generator for skewed data », *in*: *Water Resources Research* 46(*7*).

Fraley, Chris and Adrian E Raftery (2007), « Bayesian regularization for normal mixture estimation and model-based clustering », *in*: *Journal of classification* 24(*2*), pp. 155–181.

Ge, Ming and Eric C Kerrigan (2016), « Short-term ocean wave forecasting using an autoregressive moving average model », *in*: *2016 UKACC 11th International Conference on Control (CONTROL)*, IEEE, pp. 1–6.

Gelci, R, H Cazalé, and J Vassal (1957), « Sea state forecasting. The spectral method (In French) », *in*: *Bulletin d'information du Comité d'Océanographie et d'Etude des Côtes* 9, pp. 416–435.

Geweke, John and Michael Keane (2007), « Smoothly mixing regressions », *in*: *Journal of Econometrics* 138(*1*), pp. 252–290.

Goeman, Jelle J (2008), « Autocorrelated logistic ridge regression for prediction based on proteomics spectra », *in*: *Statistical Applications in Genetics and Molecular Biology* 7(*2*).

Golub, Gene H, Michael Heath, and Grace Wahba (1979), « Generalized cross-validation as a method for choosing a good ridge parameter », *in*: *Technometrics* 21(*2*), pp. 215–223.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016), *Deep learning*, MIT press.

Gu, Hao et al. (2019), « Blind channel identification aided generalized automatic modulation recognition based on deep learning », *in*: *IEEE Access* 7, pp. 110722–110729.

Gu, Jiuxiang et al. (2018), « Recent advances in convolutional neural networks », *in*: *Pattern recognition* 77, pp. 354–377.

Gulev, Sergey K et al. (2003), « Assessment of the reliability of wave observations from voluntary observing ships: Insights from the validation of a global wind wave climatology based on voluntary observing ship data », *in*: *Journal of Geophysical Research: Oceans* 108(*C7*).

Hasselmann, Klaus F et al. (1973), « Measurements of wind-wave growth and swell decay during the Joint North Sea Wave Project (JONSWAP). », *in*: *Ergaenzungsheft zur Deutschen Hydrographischen Zeitschrift, Reihe A*.

Hastie, Trevor et al. (2009), *The elements of statistical learning: data mining, inference, and prediction*, vol. 2, Springer.

Hegermiller, CA et al. (2017), « A multimodal wave spectrum–based approach for statistical downscaling of local wave climate », *in*: *Journal of Physical Oceanography* 47(*2*), pp. 375–386.

Hemer, Mark A, Jack Katzfey, and Claire E Trenham (2013), « Global dynamical projections of surface ocean wave climate for a future high greenhouse gas emission scenario », *in*: *Ocean Modelling* 70, pp. 221–245.

Hemer, Mark A et al. (2012), « Advancing wind-waves climate science: The COWCLIP project », *in*: *Bulletin of the American Meteorological Society* 93(*6*), pp. 791–796.

Hemmerle, William J (1975), « An explicit solution for generalized ridge regression », *in*: *Technometrics* 17(*3*), pp. 309–314.

Hessami, Masoud et al. (2008), « Automated regression-based statistical downscaling tool », *in*: *Environmental modelling & software* 23(*6*), pp. 813–834.

Hochreiter, Sepp and Jürgen Schmidhuber (1997), « Long short-term memory », *in*: *Neural computation* 9(*8*), pp. 1735–1780.

Hoerl, Arthur E and Robert W Kennard (1970), « Ridge regression: Biased estimation for nonorthogonal problems », *in*: *Technometrics* 12(*1*), pp. 55–67.

Hong, Song-You and Masao Kanamitsu (2014), « Dynamical downscaling: Fundamental issues from an NWP point of view and recommendations », *in*: *Asia-Pacific Journal of Atmospheric Sciences* 50(*1*), pp. 83–104.

Hoshikawa, Toshiya (2013), « Mixture regression for observational data, with application to functional regression models », *in*: *arXiv preprint arXiv:1307.0170*.

Huang, Jian et al. (2018), « A constructive approach to l0 penalized regression », *in*: *The Journal of Machine Learning Research* 19(*1*), pp. 403–439.

Hurrell, James W et al. (2003), « An overview of the North Atlantic oscillation », *in*: *Geophysical Monograph-American Geophysical Union* 134, pp. 1–36.

Jacobs, Robert A et al. (1991), « Adaptive mixtures of local experts », *in*: *Neural computation* 3(*1*), pp. 79–87.

Kaufman, CG and Benjamin Adam Shaby (2013), « The role of the range parameter for estimation and prediction in geostatistics », *in*: *Biometrika* 100(*2*), pp. 473–484.

Keller, Denise E et al. (2017), « Testing a weather generator for downscaling climate change projections over Switzerland », *in*: *International Journal of Climatology* 37(*2*), pp. 928–942.

Kerbiriou, Marie-Aurélie et al. (2007), « Influence of sea-states description on wave energy production assessment », *in*: *Proceedings of the 7th European Wave and Tidal Energy Conference, Porto, Portugal, Sept*, pp. 11–13.

Kingma, Diederik P and Jimmy Ba (2014), « Adam: A method for stochastic optimization », *in*: *arXiv preprint arXiv:1412.6980*.

Kounades-Bastian, Dionyssos et al. (2016), « A variational EM algorithm for the separation of time-varying convolutive audio mixtures », *in*: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(*8*), pp. 1408–1423.

Kullback, Solomon (1997), *Information theory and statistics*, Courier Corporation.

Laugel, Amélie (2013), « Climatologie des états de mer en Atlantique nord-est: analyse du climat actuelet des évolutions futures sous scénarios de changement climatique par descente d'échelle dynamique et statistique », PhD thesis, Paris Est.

Laugel, Amélie et al. (2014), « Wave climate projections along the French coastline: dynamical versus statistical downscaling methods », *in*: *Ocean Modelling* 84, pp. 35–50.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015), « Deep learning », *in*: *nature* 521(*7553*), pp. 436–444.

Lee, Youngjo and John A Nelder (1996), « Hierarchical generalized linear models », *in*: *Journal of the Royal Statistical Society: Series B (Methodological)* 58(*4*), pp. 619–656.

Leisch, Friedrich (2004), « Flexmix: A general framework for finite mixture models and latent glass regression in R », *in*.

Maraun, Douglas et al. (2010), « Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user », *in*: *Reviews of geophysics* 48(*3*).

Masoudnia, Saeed and Reza Ebrahimpour (2014), « Mixture of experts: a literature survey », *in*: *Artificial Intelligence Review* 42(*2*), pp. 275–293.

McLachlan, Geoffrey J and Suren Rathnayake (2014), « On the number of components in a Gaussian mixture model », *in*: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4(*5*), pp. 341–355.

Meeds, Edward and Simon Osindero (2005), « An alternative infinite mixture of Gaussian process experts », *in*: *Advances in neural information processing systems* 18.

Michel, Marceau et al. (2022), « Deep learning for statistical downscaling of sea states », *in*: *Advances in Statistical Climatology, Meteorology and Oceanography* 8(*1*), pp. 83–95.

Neal, Radford M and Geoffrey E Hinton (1998), « A view of the EM algorithm that justifies incremental, sparse, and other variants », *in*: *Learning in graphical models*, Springer, pp. 355–368.

Nguyen, Hien D and Faicel Chamroukhi (2018), « Practical and theoretical aspects of mixture-of-experts modeling: An overview », *in*: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8(*4*), e1246.

Obakrim, Said et al. (2022a), *EM algorithm for generalized Ridge regression with spatial covariates*, DOI: 10.48550/ARXIV.2208.04754, URL: https://arxiv.org/abs/2208.04754.

Obakrim, Said et al. (2022b), « Statistical modeling of the space-time relation between wind and significant wave height », *in*: *Earth and Space Science Open Archive*, p. 20, DOI: 10.1002/essoar.10510147.2, URL: https://doi.org/10.1002/essoar.10510147.2.

Patil, Pratik et al. (2021), « Uniform consistency of cross-validation estimators for high-dimensional ridge regression », *in*: *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 3178–3186.

Peña-Angulo, D et al. (2016), « The influence of weather types on the monthly average maximum and minimum temperatures in the Iberian Peninsula », *in*: *Atmospheric Research* 178, pp. 217–230.

Pérez, Jorge et al. (2014), « ESTELA: a method for evaluating the source and travel time of the wave energy reaching a local area », *in*: *Ocean Dynamics* 64(*8*), pp. 1181–1191.

Permatasari, Sri Maulidia, Anik Djuraidah, and Agus M Soleh (2017), « Statistical Downscaling with Gamma Distribution and Elastic Net Regularization: Case Study: Monthly Rainfall 1981-2013 at Indramayu », *in*: *The 2nd International Conference On Applied Statistics (ICAS 2016)*, pp. 121–129.

Ramamurti, Viswanath and Joydeep Ghosh (1997), « Regularization and error bars for the mixture of experts network », *in*: *Proceedings of International Conference on Neural Networks (ICNN'97)*, vol. 1, IEEE, pp. 221–225.

Rasp, Stephan, Michael S Pritchard, and Pierre Gentine (2018), « Deep learning to represent subgrid processes in climate models », *in*: *Proceedings of the National Academy of Sciences* 115(*39*), pp. 9684–9689.

Remya, PG et al. (2022), « Indian Ocean wave forecasting system for wind waves: development and its validation », *in*: *Journal of Operational Oceanography* 15(*1*), pp. 1–16.

Roland, Aron and Fabrice Ardhuin (2014), « On the developments of spectral wave models: numerics and parameterizations for the coastal ocean », *in*: *Ocean Dynamics* 64(*6*), pp. 833–846.

Rue, Hååvard and Hååkon Tjelmeland (2002), « Fitting Gaussian Markov random fields to Gaussian fields », *in*: *Scandinavian journal of Statistics* 29(*1*), pp. 31–49.

Rue, Håvard (2001), « Fast sampling of Gaussian Markov random fields », *in*: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(*2*), pp. 325–338.

Saha, Suranjana et al. (2010a), « The NCEP climate forecast system reanalysis », *in*: *Bulletin of the American Meteorological Society* 91(*8*), pp. 1015–1058.

Saha, Suranjana et al. (2010b), « The NCEP climate forecast system reanalysis », *in*: *Bulletin of the American Meteorological Society* 91(*8*), pp. 1015–1058.

Sailor, DJ et al. (2000), « A neural network approach to local downscaling of GCM output for assessing wind power implications of climate change », *in*: *Renewable energy* 19(*3*), pp. 359–378.

Scher, Sebastian (2018), « Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning », *in*: *Geophysical Research Letters* 45(*22*), pp. 12–616.

Schulz, Eric, Maarten Speekenbrink, and Andreas Krause (2018), « A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions », *in*: *Journal of Mathematical Psychology* 85, pp. 1–16.

Sha, Yingkai et al. (2020), « Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: Daily maximum and minimum 2-m temperature », *in*: *Journal of Applied Meteorology and Climatology* 59(*12*), pp. 2057–2073.

Shireman, Emilie, Douglas Steinley, and Michael J Brusco (2017), « Examining the effect of initialization strategies on the performance of Gaussian mixture modeling », *in*: *Behavior research methods* 49(*1*), pp. 282–293.

Soares, C Guedes, AM Ferreira, and C Cunha (1996), « Linear models of the time series of significant wave height on the Southwest Coast of Portugal », *in*: *Coastal Engineering* 29(*1-2*), pp. 149–167.

Storkey, Amos J (1999), « Truncated covariance matrices and Toeplitz methods in Gaussian processes », *in*: *1999 Ninth International Conference on Artificial Neural Networks ICANN 99.(Conf. Publ. No. 470)*, vol. 1, IET, pp. 55–60.

Sungkawa, Iwa, Anita Rahayu, et al. (2019), « Extreme rainfall prediction using bayesian quantile regression in statistical downscaling modeling », *in*: *Procedia Computer Science* 157, pp. 406–413.

Tajbakhsh, Sam Davanloo, Necdet Serhat Aybat, and Enrique Del Castillo (2020), « On the Theoretical Guarantees for Parameter Estimation of Gaussian Random Field Models: A Sparse Precision Matrix Approach », *in*: *Journal of Machine Learning Research* 21(*217*), pp. 1–41.

Teutschbein, Claudia and Jan Seibert (2012), « Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods », *in*: *Journal of hydrology* 456, pp. 12–29.

Thomas, T Justin and GS Dwarakish (2015), « Numerical wave modelling–A review », *in*: *Aquatic procedia* 4, pp. 443–448.

Tibshirani, Robert et al. (2005), « Sparsity and smoothness via the fused lasso », *in*: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(*1*), pp. 91–108.

Tibshirani, Ryan J and Jonathan Taylor (2011), « The solution path of the generalized lasso », *in*: *The annals of statistics* 39(*3*), pp. 1335–1371.

Timmermans, BW et al. (2020), « Global wave height trends and variability from new multimission satellite altimeter products, reanalyses, and wave buoys », *in*: *Geophysical Research Letters* 47(*9*), e2019GL086880.

Tolman, Hendrik L et al. (2009), « User manual and system documentation of WAVEWATCH III TM version 3.14 », *in*: *Technical note, MMAB Contribution* 276, p. 220.

Tracy, Barbara et al. (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.

Vidaurre, Diego, Concha Bielza, and Pedro Larranaga (2013), « A survey of L1 regression », *in*: *International Statistical Review* 81(*3*), pp. 361–387.

Vrac, Mathieu, Katharine Hayhoe, and Michael Stein (2007), « Identification and intermodel comparison of seasonal circulation patterns over North America », *in*: *International Journal of Climatology: A Journal of the Royal Meteorological Society* 27(*5*), pp. 603–620.

Vrac, Mathieu, Michael Stein, and Katharine Hayhoe (2007), « Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing », *in*: *Climate Research* 34(*3*), pp. 169–184.

Wang, Xiaolan L, Yang Feng, and VR Swail (2012), « North Atlantic wave height trends as reconstructed from the 20th century reanalysis », *in*: *Geophysical Research Letters* 39(*18*).

Wang, Xiaolan L, Val R Swail, and Andrew Cox (2010), « Dynamical versus statistical downscaling methods for ocean wave heights », *in*: *International Journal of Climatology: A Journal of the Royal Meteorological Society* 30(*3*), pp. 317–332.

Wang, XL and VR Swail (2006), « Historical and possible future changes of wave heights in northern hemisphere oceans », *in*: *Atmosphere-ocean interactions* 2(*2*), p. 240.

Ward Jr, Joe H (1963), « Hierarchical grouping to optimize an objective function », *in*: *Journal of the American statistical association* 58(*301*), pp. 236–244.

Wieringen, Wessel N van (2015), « Lecture notes on ridge regression », *in*: *arXiv preprint arXiv:1509.09169*.

Wilby, Robert L et al. (1998), « Statistical downscaling of general circulation model output: A comparison of methods », *in*: *Water resources research* 34(*11*), pp. 2995–3008.

Xue, Yongkang et al. (2014), « A review on regional dynamical downscaling in intraseasonal to seasonal simulation/prediction and major factors that affect downscaling ability », *in*: *Atmospheric research* 147, pp. 68–85.

Yamashita, Rikiya et al. (2018), « Convolutional neural networks: an overview and application in radiology », *in*: *Insights into imaging* 9(*4*), pp. 611–629.

Yarnal, Brent and Brent Frakes (1997), « Using synoptic climatology to define representative discharge events », *in*: *International Journal Of Climatology: A Journal Of The Royal Meteorological Society* 17(*3*), pp. 323–341.

Young, Ian R (1999), *Wind generated ocean waves*, Elsevier.

Yuan, Chao and Claus Neubauer (2008), « Variational mixture of Gaussian process experts », *in*: *Advances in neural information processing systems* 21.

Yuksel, Seniha Esen, Joseph N Wilson, and Paul D Gader (2012), « Twenty years of mixture of experts », *in*: *IEEE transactions on neural networks and learning systems* 23(*8*), pp. 1177–1193.

Zorita, Eduardo and Hans Von Storch (1999), « The analog method as a simple statistical downscaling technique: Comparison with more complicated methods », *in*: *Journal of climate* 12(*8*), pp. 2474–2489.

Zucchini, Walter and Peter Guttorp (1991), « A hidden Markov model for space-time precipitation », *in*: *Water Resources Research* 27(*8*), pp. 1917–1923.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : Downscaling statistique et changement climatique en zone côtière

Mot clés : Descente d'échelle, État de mer, Ridge généralisée, Mélange d'experts, Algorithme EM, Apprentissage profond

Résumé : Le climat des vagues océaniques a un impact significatif sur les activités humaines, et sa compréhension est importante sur le plan socio-économique et environnemental. Dans cette thèse, nous nous intéressons à la caractérisation des paramètres d'état de mer tels que la hauteur significative des vagues (H_s) en utilisant des méthodes statistiques et d'apprentissage profond. En particulier, nous nous intéressons à la modélisation de la relation entre les conditions de vent de l'Atlantique Nord et les paramètres d'état de la mer à un endroit situé dans le Golfe de Gascogne. Étant donné la multidimensionnalité des données de vent et la relation décalée en temps entre les conditions de vent et les vagues, nous proposons d'abord un cadre général pour sélectionner les covariables pertinentes qui influencent la hauteur significative des vagues. Après l'étape de prétraitement, un modèle de régression basé sur les types de temps est proposé pour modéliser la relation entre le vent et les vagues. Les types de temps sont construits à l'aide d'un algorithme de classification puis, pour chaque type de temps, une régression de Ridge est ajustée entre les conditions de vent et la hauteur significative des vagues. Le modèle prédit bien H_s, mais il présente certaines limites, à savoir : (i) la régression de Ridge ne tient pas compte du fait que les covariables ont une structure spatiale ; et (ii) les types de temps sont construits a priori à l'aide d'un algorithme de classification et ils ne sont pas évalués en fonction de la prédiction de H_s. Par conséquent, nous proposons un algorithme d'espérance-maximisation (EM) pour estimer les paramètres de la régression de Ridge généralisée avec des covariables spatiales, puis, pour tenir compte les points (i) et (ii), nous proposons un mélange d'experts de Ridge généralisés estimés à l'aide d'un algorithme EM variationnel. Ce modèle est utilisé comme modèle de régression basé sur les types de temps et ses performances sont supérieures à celles du modèle original. Finalement, la dernière partie de cette thèse est consacrée au développement de méthodes d'apprentissage profond pour la prédiction des paramètres de l'état de la mer.

Title: Statistical downscaling and climate change in the coastal zone

Keywords: Downscaling, Sea state, Generalized Ridge, Mixture of experts, EM algorithm, Deep learning

Abstract: Ocean wave climate has a significant impact on human activities, and its understanding is socioeconomically and environmentally important. In this thesis, we are interested in characterizing sea state parameters such as significant wave height (H_s) using statistical and deep learning methods. In particular, we are interested in modeling the relationship between North Atlantic wind conditions and sea state parameters at a location in the Bay of Biscay. Given the multidimensionality of the wind data and the time-lagged relationship between wind conditions and waves, we first propose a general framework to select the relevant covariates that influence the significant wave height. After the preprocessing step, a regression model based on weather types is proposed to model the relationship between wind and waves. The weather types are constructed using a clustering algorithm, and then, for each weather type, a Ridge regression is fitted between the wind conditions and the significant wave height. The model predicts H_s well; however, it has some limitations, namely: (i) Ridge regression does not take into account that the covariates have a spatial structure; and (ii) the weather types are constructed a priori using a clustering algorithm, and they are not evaluated based on the prediction of H_s. Therefore, we propose an expectation-maximization (EM) algorithm to estimate the parameters of the generalized Ridge regression with spatial covariates. Then, to account for (i) and (ii), we propose a mixture of generalized Ridge experts estimated using a variational EM algorithm. This model is used as a weather-types-based regression model, and its performance is better than that of the original model. Finally, the last part of this thesis is devoted to developing deep learning methods for sea state parameters prediction.