Wind-Wave Probabilistic Forecasting based on Ensemble Predictions

Maxime FORTIN

Kongens Lyngby 2012
IMM-MSc-2012-86

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk IMM-PhD-2012-86

Summary

Wind and wave forecasts are of a crucial importance for a number of decision-making problems. Nowadays, considering all potential uncertainty sources in weather prediction, ensemble forecasting provides the most complete information about future weather conditions. However, ensemble forecasts tend to be biased and underdispersive, and are therefore uncalibrated. Calibration methods were developed to solve this issue. So far, these methods are usually applied on univariate weather forecasts and do not take any possible correlation into account. Since wind and wave forecasts have to be jointly taken into account in some decision-making problems, e.g. offshore wind farm maintenance, we propose in this thesis a bivariate approach, generalizing existing univariate calibration methods to jointly calibrated ensemble forecasts. A other method using the EPS-prescribed correlation in order to recover the dependence lost during the marginal calibration is also proposed. Even if the univariate performance of the marginal calibration is preserved, results confirm the need for bivariate approaches. Contrary to the univariate approach, the bivariate calibration method generates correlated bivariate forecasts, though it appears to be too sensitive to outliers when estimating necessary model parameters. Jointly calibrated distributions are too wide and therefore overdispersive. The different calibration methods are tested on ECMWF ensemble predictions over the offshore platform FINO$_1$ located in the North Sea close to the German shore.

Preface

This thesis was prepared at the department of Informatics and Mathematical Modelling at the Technical University of Denmark in fulfilment of the requirements for acquiring an M.Sc. in Informatics. My supervisors are Pierre Pinson and Henrik Madsen from the IMM department at DTU.

The thesis deals with wind and wave ensemble forecast calibration. A univariate calibration method is employed and the use of a bivariate approach is investigated. The different methods are tested on the ECMWF ensemble forecasts over the offshore measurement platforms $FINO_1$ located in the North Sea.

Lyngby, 03-August-2012

Maxime FORTIN

Acknowledgements

This project wouldn't have been possible without my supervisor Pierre Pinson that I sincerely would first like to thank for his ongoing support and the time he took to help me when needed. Working with him has been a great pleasure and of an enormous interest.

I would also like to thank Henrik Madsen for having permitted me to work among his group at DTU IMM.

Finally, I am grateful to everyone of the department of statistics of DTU Informatics for the ideas they suggested me and their warm welcome.

Contents

Chapter 1

Introduction

Wind and wave forecasts are employed in various areas where they are of substantial economic value. Several decision-making problems e.g. offshore wind farm maintenance planning, require joint forecasts of wind AND waves as input, since these variables jointly impact the potential cost of inappropriate decisions. It is expected that forecasts have to be of the highest possible quality in order to maximise their usefulness when making decisions.

During the past few decades, weather forecasts have evolved from classical deterministic weather maps and their time evolution, to advanced probabilistic approaches informing about potential likely paths for the development of the weather. Indeed, considering all potential uncertainty sources in weather prediction, probabilistic forecasts comprise today the forecast product that provides the most complete information about future weather. However, probabilistic forecasts as direct output from Numerical Weather Prediction (NWP) models, i.e ensemble forecasts, tend to be subject to biases and dispersion errors: they are uncalibrated.

This project aims at investigating statistical post-processing methods for ensemble predictions of wind and waves, in order to maximise their quality and potential usefulness as input to decision-making in a stochastic optimization environment. The purpose of such statistical approaches is to probabilistically calibrate the forecasts for these variables, by reducing the predicted errors of the ensemble mean and by providing reliable information about the future evolution of wind speed and wave height. Since these two variables are linked and that their joint forecast is of particular interest to some forecast users, the use of univariate but also bivariate calibration methods is envisaged.

1.1 What's the point in forecasting?

Nowadays, weather forecasts are used in numerous sectors of activity. Especially, wind and wave forecasting is of particular interest for a number of decision-making problems. Of course, wind and wave forecasts can be used on an everyday basis in order to choose the most appropriate coat for the following day or to know about the best day of the week for kite-surfing. More importantly, wind and wave forecasts allow some companies to make substantial savings and governments to save human lives.

Figure 1.1: Illustration of sectors using wind and wave forecasts

The most known use of wind and wave forecasts is for public safety. In the state of Florida, for instance, hurricane-related forecasts are of crucial importance for the safety of the inhabitants. Thanks to accurate forecasts, appropriate preventive measures can be taken to guarantee people safety and avoid casualties.

Wind and wave forecasts are used in the energy sector, especially now that onshore and offshore wind energy has taken a leading role in the development of renewable energy solutions (Pinson et al., 2007, 2012). Energy production monitoring or management requires very precise information about the environment around an energy production site of interest, in near-real time and for the coming minutes to days. For example, the last-minute cancellation of an offshore wind farm maintenance operation because of rough weather can be highly costly. Weather forecasts are similarly used for energy trading, where traders need visibility into weather changes that will impact demand and prices sufficiently in advance. In the actual context of world energy policy changing and led by renewable energy sources, the importance of weather forecasts is growing fast. They will be a necessity for some countries like Denmark that expects to produce 100% of its energy from renewable resources in 2050 (Mathiesen et al., 2009).

Wind and wave forecasts are also used for sailing or maritime transport applications e.g. ship routing (Hinnenthal, 2008), where they generally support finding the “best route” for ships. For most transits this will mean the minimum transit time while avoiding significant risk for the ship. The goal is not to avoid all adverse weather conditions but instead to find the best balance between minimizing transit time, fuel consumption and not placing the vessel at risk with weather damage and crew injury.

1.2 Ensemble forecasting

Short-term weather forecasts for lead times of a few hours are usually issued based on purely statistical methods e.g. from time series analysis. From 6 hours to days ahead, the most accurate type of weather forecasts are the Numerical Weather Predictions (NWP). Unlike statistical methods, they are flow dependent and therefore much more efficient for appraising medium-range weather evolutions. Employing a NWP model entails relying on computer resources to solve fluid dynamics and thermodynamics equations applied to the Atmosphere, in order to predict atmospheric conditions hours to days ahead. An estimated initial state of the Atmosphere, built by collecting observational data all over the world, is necessary as a starting point for NWP models. However, the Atmosphere can never be completely and perfectly observed due to measurement accuracy and limited observational coverage. Thus the initial state of NWP models will always be slightly different from the true initial state of the Atmosphere. In the early 60's, Edward Lorenz showed that the dynamics of the Atmosphere were highly sensitive to initial conditions, that is, two slightly different atmospheric states could finally result, in the future, in two very different sets of weather conditions. This is also known as the "butterfly effect", naively saying that the motion of the flies of a butterfly could lead to a storm on other side of the world. Such considerations on the sensitivity of initial conditions in numerical weather prediction was the origin for the subsequent development of ensemble forecasting. Ensemble forecasting is based upon the idea that forecast error characteristics, resulting from a combination of initial conditions errors and model imperfections, can be estimated. It is assumed that initial conditions uncertainties can be modelled by perturbing the initial state and that model deficiencies can be represented by a stochastic parametrisation of NWP models (see Chapter 2.2.1).

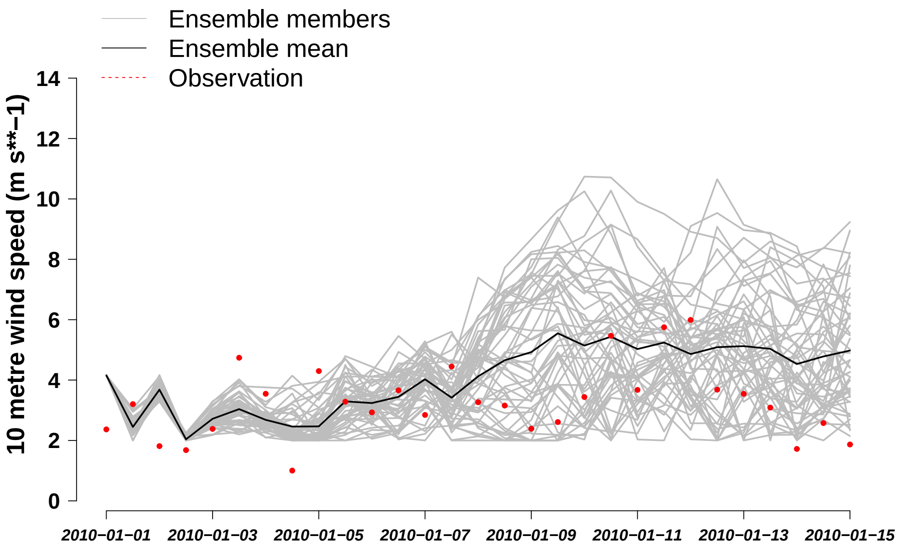


Figure 1.2: Example of ensemble forecast trajectories of surface wind speed issued at time t for lead times between $+06h$ ahead and $+168h$ ahead. The predicted ensemble mean is represented by the black line, the ensemble members by the grey lines and the observations in red points.

In practice, an ensemble forecast is composed of N different forecasts, referred to as ensemble members, issued at the same time t for the same future time $t + k$ (hence with k denoting the lead time). Ensemble members follow their own trajectories. Each of them is relevant in the sense that it obeys to the same physics equations and the same parametrization. Figure 1.2 shows an example with 51 ensemble members of surface wind speed from the European Centre for Medium-range Weather Forecasts (ECMWF). Ensemble forecasts from ECMWF will be further introduced in Chapter 2.2.1. It can be seen from that figure that slightly different initial states and model parametrization can lead to completely different scenarios. For instance for lead time $k = +168$ h, the different members predict wind speed between 2 m.s^{-1} and approximately 9 m.s^{-1}.

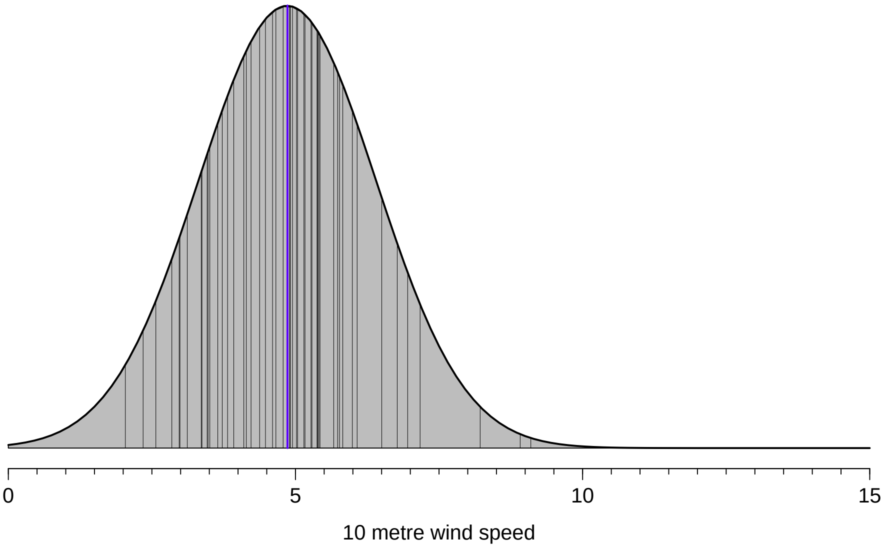


Figure 1.3: Example of ensemble forecast issued at a time t for a certain time $t+k$. The vertical black lines represent the ensemble members and the blue one represents the ensemble mean, the curve represents the probabilistic density function computed from the 51 forecasts

Ensemble forecasts for a given meteorological variable, location and lead time, can be seen as a sample of a distribution. Then, ensemble forecasts can be translated into a predictive density estimated from the N different forecasts of the ensemble prediction system. However, only dealing with predictive densities for each meteorological variable, location and lead time, implies that the trajectory structure of ensemble members is lost. In more general terms, it is their spatio-temporal and multivariate dependencies that are lost.

1.3 Ensemble Forecast Verification and Calibration

Considering all potential sources of uncertainty in the forecasting process, it appears relevant to present forecasts as probabilistic (predictive densities or ensemble predictions) instead of single-valued. However, the information provided by ensemble and probabilistic forecasts has to be reliable. Let us take the example of the forecast "there is a probability 0.95 that the event A occurs". This probability suggests that the event A will most likely occur, but a question remains : "Should we trust this forecast?". To be reliable, over all times this specific forecast is issued, the corresponding event A should occur approximately 95% of the times, so as to verify that the probabilities communicated are verified in practice. If the event A occurs only 50% of the times when a probability of 0.95 is predicted, then this probabilistic information may be useless and the ensemble forecast system can not be trusted.

Unlike for deterministic forecasts, ensemble forecast verification is not straightforward. Specific scores and diagnostic tools have to be used to assess their quality. Unfortunately, ensemble forecasts tend to be biased and underdispersive, that is, the ensemble mean shows systematic errors while the ensemble spread is generally lower than it should be, hence leading to observations often falling outside of the ensemble range. Figure 1.2 is a good example of this issue: for the short to early-medium range (+06 h to +48 h), the ensemble range is too small and observations often fall out of this range. Ensemble forecast are in that case referred to as *underdispersive*. Finding and correcting the origins of these anomalies directly into the numerical model is not an easy task, since error sources might include model resolution, physics parametrizations and also the ensemble initial states perturbation method (see Chapter 2.2.1). Solving those problems is a really difficult work considering the huge number of grid points in a numerical weather prediction model. For example, reducing a negative bias over France could lead to the creation of a positive bias over the United States of America and vice versa. Atmosphere is a really complex (nonlinear) and chaotic system. Simplifications have to be made so weather can be predicted. This is the reason why statistical post-processing methods are used to calibrate the forecasts, that is to solve the bias and under-dispersion problems. Those forecast post-processing methods are based upon the idea that analysing the past errors of a model can help to reduce the future errors of the same model. The main advantages of such methods is that they are easy to implement and location-specific.

Gneiting (Gneiting et al., 2007) illustrated the problem of ensemble forecast verification in a comprehensive way. At a time $t+k$, in view of all uncertainty sources, Nature chooses a distribution G_{t+k}, and picks one random number from that distribution to obtain the observation x_{t+k}. Of course, G_{t+k} is not observed in practise because only one outcome realizes at time $t+k$. Then at a subsequent time $t+k+\Delta t$, the distribution $G_{t+k+\Delta t}$ chosen by Nature is different from G_{t+k}, and this whatever Δt. An ensemble forecast tries to estimate G_{t+k} by predicting ensemble members $(y^{(1)}_{t+k|t}, \dots, y^{(M)}_{t+k|t})$ assumed to be a sample of a distribution $F_{t+k|t}$. However, $F_{t+k|t}$ and G_{t+k} are never identical, they differ in terms of both mean and variance. This motivates the necessity to recalibrate ensemble forecasts.

All these problems can de addressed by ensemble calibration methods. There exists a variety of calibration methods based on this same idea of correcting ensemble forecasts based on recent past errors. Some of the most employed techniques include Ensemble dressing (Bröcker and Smith, 2008), Bayesian Model Averaging (Raftery et al., 2005; Sloughter et al., 2010), Ensemble model Output Statistics (Gneiting et al., 2005; Thorarinsdottir and Gneiting, 2008), Logistic regression (Wilks and Hamill, 2007), etc.

The method of ensemble dressing is the simplest ensemble calibration method. It is based upon the idea that a kernel should be assigned to every debiased ensemble members. The overall predicted probability density function is the normalized sum of all the individual member kernels. Ensemble dressing is then a kernel density smoothing approach. Most usually chosen kernels are of the Gaussian type, though others could also be employed.

The Bayesian Moving Averaging (BMA) was introduced by Raftery in 2005 (Raftery et al., 2005). It is a more sophisticated version of the ensemble dressing technique, it also assigns a kernel to every ensemble members but with different weights, the weight depending on the skill of the ensemble member during the training period. The predicted probability density function is then the weighted sum of all the member kernels.

The Ensemble Model Output Statistics (EMOS) technique was introduced by Gneiting in 2005 (Gneiting et al., 2005) as an extension to conventional linear regression usually applied for deterministic forecasts. The approach is to construct linear models with the ensemble mean and the ensemble spread as predictors, the parameters correcting those two first moments are estimated through a training period.

Those calibration methods have been widely used for univariate forecasts like 2 m temperature, precipitation, wind speed and wind direction, etc... However, a few studies have used those methods to calibrate bivariate forecasts such as u and v components of wind (Pinson, 2012; Schuhen et al., 2012), or other type of variable like temperature and precipitation (Möller et al., 2012). Those kinds of methods allow joint calibration of two variables, so their relationship can be taken into account.

This thesis aims at calibrating wind and wave ensemble forecasts. Since the two variables are strongly correlated and jointly used by some users, we propose an univariate but also a bivariate calibration method using the EMOS approach, so the correlation of the two variables can be respected.

1.4 Structure

The report is organised in 5 chapters. In Chapter 2, we present the type of data used for the study of wind and wave forecast at the $FINO_1$ measurement site. We explain the relationship that links waves to wind and briefly analyse the different patterns of those two variables on the site of interest. We also describe the computation of ensemble forecasts at ECMWF.

In Chapter 3, several methods to assess probabilistic forecasts are presented. We firstly list the different scores and tools used to verify univariate forecasts, and then we continue with the multivariate assessment methods.

In Chapter 4, we describe the calibration methods employed to correct the ensemble mean and the ensemble spread, going from univariate to bivariate approaches.

In Chapter 5, the results are discussed. We first expose the univariate calibration and show the importance of multivariate calibration because of the existing correlation of wind and waves.

And finally, in chapter 6, we summarise the results and improvements of the method and discuss the perspectives.

Chapter 2

Data

This master thesis deals with point ensemble forecasting calibration. Thus, every kind of data (observations, analysis and forecasts) is specific of one location only. The location of interest is the $FINO_1$ offshore measurement site located in the German North Sea ($54°01'N$, $06°35'E$) close to the offshore wind farms Borkum Riffgrund and Borkum West. This measurement site is part of a research project of 3 offshore measurement platforms in the North sea and the Baltic sea (see figure 2.1).

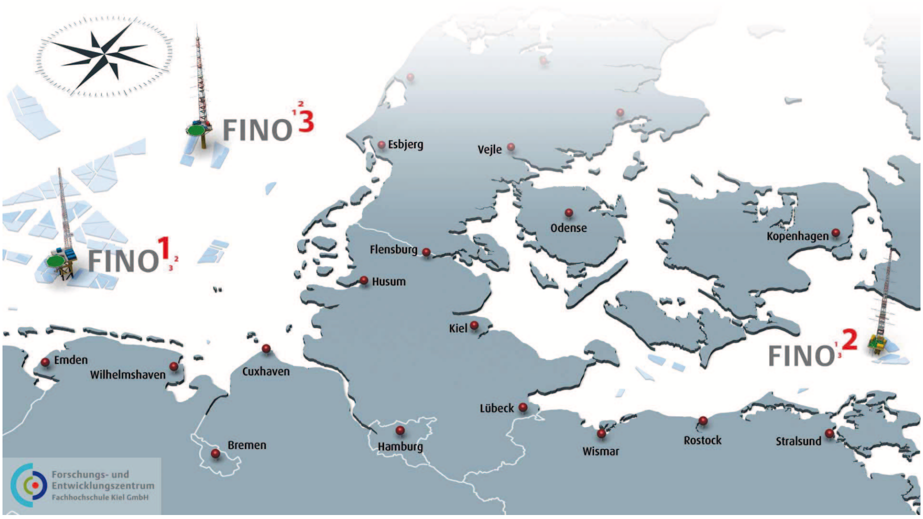


Figure 2.1: FINO project : 3 measurement sites on the North sea and the Baltic sea

2.1 Observations

$FINO_1$ collects meteorological, oceanographic and biological data. Among all these different types of data, a buoy on the $FINO_1$ site provides wave observations (direction, height, period) with a time resolution of 30 minutes, and a measurement mast provides wind speed and direction data at eight different height levels (from 33 m to 100 m).

On the north-west side of the mast, classic wind vanes are installed at 33, 50, 70 and 90 m height and high-resolution ultrasonic anemometers (USA) are installed at the intermediate levels (40, 60 and 80 m) to determine the wind direction and speed with a time resolution of approximately 10 minutes. The $FINO_1$ research platform has been providing the highest continuous wind measurement in the offshore area world-wide since September 2003.

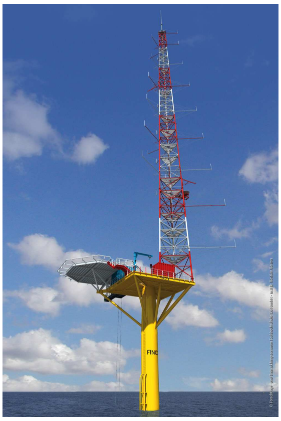
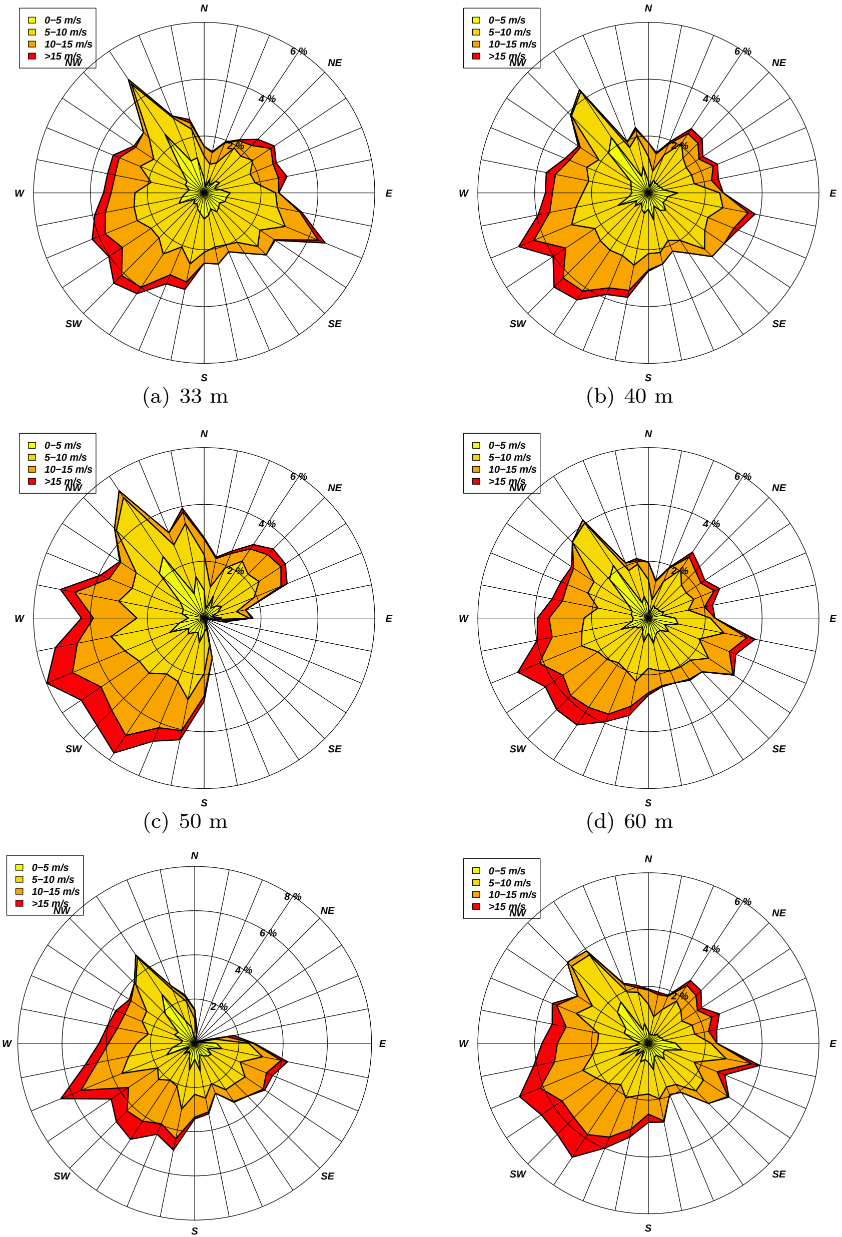


Figure 2.2: $FINO_1$ mast with wind sensors from 33 m to 100 m

2.1.1 10 m Wind Speed

The $FINO_1$ mast measures wind speed and direction over a 100 m high column. For this study, wind speed observations were subject to the quality control procedure proposed by Baars (Baars, 2005).



(a) 33 m

(b) 40 m

(c) 50 m

(d) 60 m

(e) 70 m

(f) 80 m

Figure 2.3: Wind Roses at different heights of the $FINO_1$ mast (frequency depending on direction and speed) over the period January 2010 - December 2011. Wind speed below to 5 m.s^{-1} respresented in yellow, from 5 to 10 m.s^{-1} in gold, from 10 to 15 m.s^{-1} in orange and over 15 m.s^{-1} in red. Frequency of occurence of the radial axis are 2,4 and 6%.

The wind rose is the method of graphically presenting the wind conditions, direction and speed, over a period of time at a specific location. Over the analysis period, observed wind directions are sorted into 32 different bins (every 11.75°) and observed wind speeds are sorted into 4 bins (0-5,5-10,10-15,+15 m.s^{-1}). The corresponding frequency of occurrence of each bin is then represented on a circular axis, the resulting figure is called a rose. The wind-roses represented on figure 2.3 show that, at $FINO_1$ winds mainly blow, like for the rest of Western Europe, from the south-west. Most of the low pressure systems, driven by strong fluxes, come from the Atlantic ocean, and this is the reason why we can notice on figure 2.3 that strong winds (>15 m.s^{-1}) come mainly from this direction. North-westerly winds are also quite frequent in $FINO_1$. These winds are soft and never reach 15 m.s^{-1}. We can notice that for heights of 50 and 70 m, $FINO_1$ observations are incorrect, a mask effect most likely caused by the measurement mast can be identified on the corresponding wind roses, hiding the sensor from south-easterly winds at 50 m and from north-easterly winds at 70 m.

Our study deals with surface wind forecasts, that is the wind speed observed 10 m above the ground, but unfortunately the $FINO_1$ mast does not measure wind speed at lower height than 33 m. Forecasts verification against $FINO_1$ observations are of a crucial importance. Indeed, observations are the closest sources of data from reality. Plus it has been proved that disparities exist if performing forecast verification against analysis or against observations (Pinson and Hagedorn, 2012). Indeed forecast quality verified against observations tend to be higher than if verifying against analysis. So we want to compare 10 m wind speed forecast with $FINO_1$ mast observations. We decide to extrapolate observed wind speed from the lowest measurement levels to the 10 m height using the logarithmic wind speed profile generally used in the boundary layer. The mean wind speed is assumed to increase as a logarithmic function with the height and to be null at the ground. Thus we use the following equation:

$$U(z) = u_* ln(\frac{z}{z_0}) \tag{2.1}$$

with z_0 the roughness length depending on the nature of the terrain (over ocean $z_0 \sim 10^{-2}$ m), and u_* the friction (or shear) velocity (ms^{-1}). It exists more complex and more realistic versions of this assumption taking into account the atmospheric thermal stability (Tambke et al., 2004) as the Mounin Obukhov theory:

$$U(z) = u_* ln(\frac{z}{z_0}) + \psi(z, z_0, L) \quad \text{with} \quad L = \frac{u_*^3}{\kappa \frac{g}{T} \overline{w'T'}} \tag{2.2}$$

with ψ the stability term, L the Mounin-Obukhov length depending on the stability, g the acceleration due to gravity, T the temperature, $\overline{w'T'}$ the heat fluxes, and κ a constant. Unfortunately, no sensor at $FINO_1$ that could inform us about the temperature profile and heat fluxes are available. This is the reason why we use equation (2.1) which does not take into account atmospheric stability, so the atmosphere is assumed neutral.

In order to find the optimal level combination to extrapolate the wind speed at the 10 m height, we first estimate the error of the extrapolation of wind speed at 33 m height using the higher measurements. After comparison to observed 33 m wind speed data, it appears that the optimal level combination using equation (2.1), with a mean absolute error of 0.18 m.s^{-1} (corresponding to a relative error of 4%) was the use of the 40 and 60 m wind speed data. This choice is also consistent with the figure 2.3 where it has previously been noticed that wind speed at 50 m and 70 m are affected by a mask effect significantly decreasing the reliability of the measurements at those heights. Measurements data above 70 m have not been tested for the extrapolation to guarantee a certain degree of relevance of the logarithmic law for wind speed profiles. Thus, this approximation seems relevant enough for our study and is applied for the 10 m wind speed extrapolation every time the 33 m, 40 m and 60 m wind speeds are all available.

2.1.2 Significant Wave Height

Waves represents the vertical motion of the sea surface resulting of the surface wind stress action. Indeed when the wind blows on the water surface, some energy is transferred to the ocean and converted into potential energy : waves. There exists two type of waves : the wind-wave and the swell. Wind-waves are waves directly created by local winds. They appear to be chaotic and turbulent with a small wavelength. Contrary to wind-waves, swell is a wave that has been created by winds earlier and far away from the site of interest. After creation, waves travel through ocean and become less chaotic, less turbulent, this type of wave is called swell.

Significant wave height, also called $H_{\frac{1}{3}}$ is a variable statistically computed from wave height in order to characterise the sea state. It represents the mean wave height (trough to crest) of the highest third of the waves. It is widely used in oceanography. Contrary to pure wave height, the hourly average of $H_{\frac{1}{3}}$ is not equal to zero and $H_{\frac{1}{3}}$ is non negative.

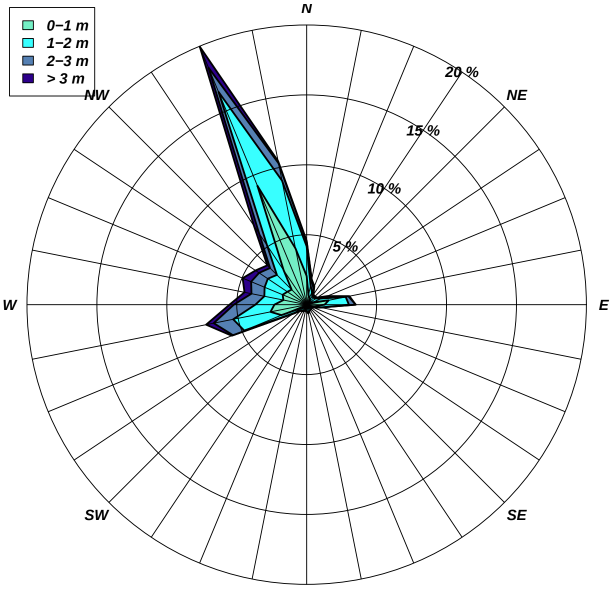


Figure 2.4: Wave rose showing the frequency of wave direction as a function of wave height

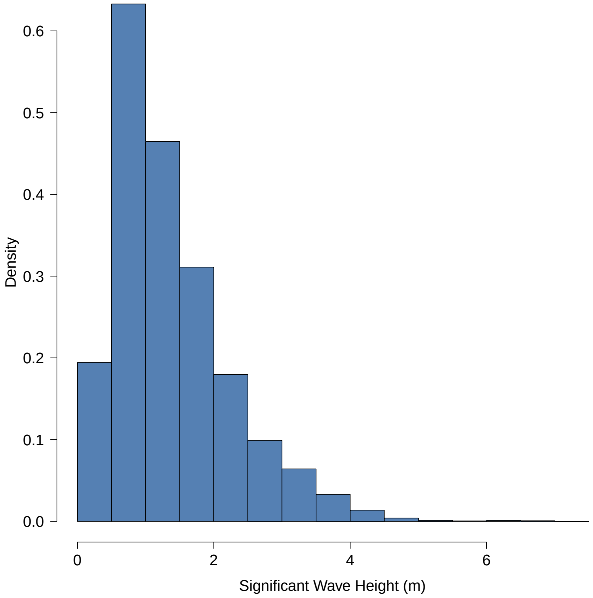


Figure 2.5: Histogram of Significant wave height

Figure 2.4 shows that, at $FINO_1$, waves mostly come from North-north-west, they are essentially swell coming from the Atlantic ocean. Some waves come also from west or east, they are mainly wind-waves. As indicated by the figure 2.5, most of the wave heights do not exceed 4 m.

2.1.3 Wind and Wave Interaction

Winds and waves are linked to each other. They are the results of a strong interaction and are therefore correlated.

Figure 2.6 shows the scatterplot of observed 10 m wind speed and significant wave height at the $FINO_1$ station. It summarises the complex relationship that exists between surface wind speed and significant wave height. It can be noticed that for strong winds ($\|\vec{u}\|$ >6 m.s^{-1}), the correlation of the two meteorological variables is high, it confirms the theory introduced previously saying that waves are created by winds and especially that wind-waves are directly influenced by the local wind when it is strong. The stronger winds, the higher the waves. For soft winds ($\|\vec{u}\|$ <6 m.s^{-1}), the correlation is close to zero. Indeed, when

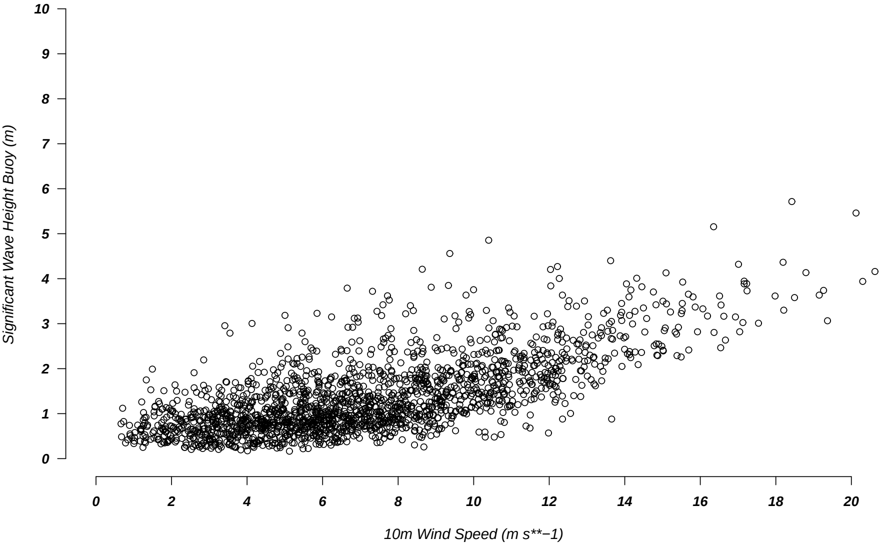


Figure 2.6: Scatterplot of observed 10 m wind speed and significant wave height at $FINO_1$ station over the entire period 2010-2011.

the wind does not blow strongly, wind-waves are not present, only swell, not influenced by local winds, can be observed.

2.1.4 Availability

The main issue with the use of observations is the inconstant availability. Contrary to analysis data from NWP models, it is subject to measurement errors and maintenance periods.

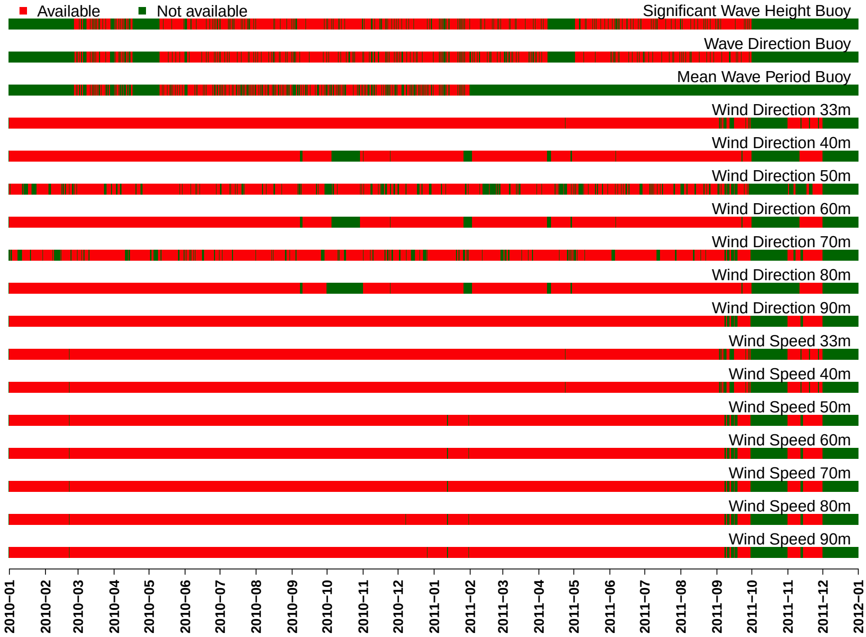


Figure 2.7: $FINO_1$ mast and buoy measurements availability over 2010-2011 (red symbolizes period when data is available, green symbolizes period when data is missing)

Figure 2.7 summarises the availability of the different meteorological and oceanographic measured variables at $FINO_1$ from January 2010 to December 2011. Wind observations are available for almost 80% of the 2 years data. However, for 50 m and 70 m measurement, the availability is more fluctuating than for other heights. Wave observations cover a smaller period, measurements start in March 2010 and stop in September 2011. Plus, during the measurement period, the availability varies strongly from one month to the other. We can notice that, in October 2011, no measurements are available, this problem could have been caused by maintenance operations.

2.2 Forecasts

2.2.1 Generalities about wind and wave forecasting

Forecasts are provided by the Ensemble Prediction System (EPS) from the European Center for Medium-Range Weather Forecasts (ECMWF). The prediction system belongs to the family of the numerical weather prediction models (NWP). Basically, a NWP represents space (atmosphere and land surface) in a 3D grid, and then from an initial state determined by weather conditions, predicts the future weather at every points of the grid by applying equations of fluid dynamics and thermodynamics of the atmosphere and some parametrizations. Wind at the 10 m height is the standard level for SYNOP observations (surface synoptic observations) and is then important to forecast. Wind is not directly predicted at this height because NWP model vertical levels are pressure levels.

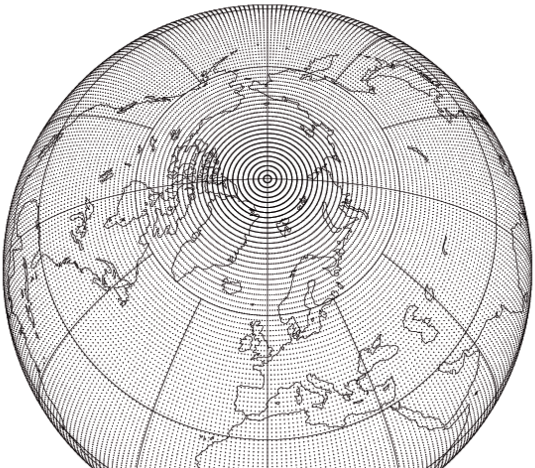


Figure 2.8: Example of a grid of a numerical weather prediction model (source `http://rda.ucar.edu`)

It is obtained by vertical interpolation between the lowest pressure level of the NWP and the surface, using Monin-Obukhov similarity theory (see equation 2.2). This procedure is appropriate over the ocean or in areas where the surface is smooth and homogeneous and therefore does not significantly influence wind speed.

Ocean waves are modelled by the wave model (WAM). This model solves the complete action density equation, including non-linear wave-wave interactions. The model has an averaged spatial resolution of 25km.

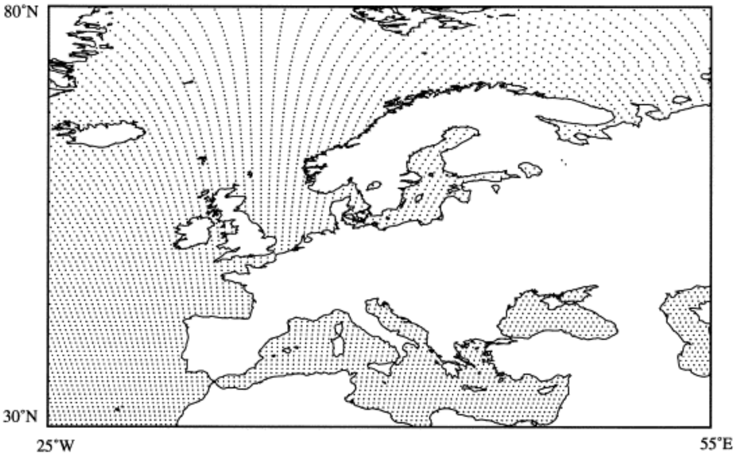


Figure 2.9: Grid used by the ECMWF operational global wave model (source (Bidlot and Holt, 1999))

The interaction of wind and waves is modelled by coupling the ECMWF atmospheric model with the wave model WAM in a two-way interaction mode. At every time step, surface winds are provided as input to the wave model, while the Charnock parameter, characterising the roughness length (Charnock, 1995), as determined by the sea state, is given to the atmospheric model and used to estimate the slowing down of the surface winds during the next coupling time step.

2.2.2 Ensemble forecasting

This thesis deals with ensemble forecasts. This type of forecasts differs from the well known deterministic forecasts that provide an unique value for a particular time at a particular location. The underlying idea of ensemble forecasting is that the initial state that initiates a NWP model is never perfectly defined because of measurement uncertainty due to sensors quality and spacial and temporal resolution, or because of the sparse observation sources all around the globe. Plus NWP models are subject to parametrisations and simplifications of dynamic and thermodynamic equations. Thus, it is obvious that a unique forecast is not a sufficient information considering all these sources of errors. This is the issue that ensemble forecasting tries to solve by simulating uncertainty of the different error sources and thus providing a probabilistic forecast instead of a deterministic.

The EPS from ECMWF is composed of 51 members: 50 "perturbed" forecasts and one control (“unperturbed“) forecast. The word “perturbed“ denotes small perturbations that are added to the control analysis (the supposed best initial state) to create different virtual initial states. These “perturbed“ members result from a complex algorithm that aims at taking into account not only initial conditions uncertainties but also uncertainties introduced by dynamics and physics representation in numerical models. This is a 3 step algorithm:

1. A singular vector technique searches for perturbations on wind, temperature or pressure that will have the maximum impact (differences with the control forecast) after 48 hours of forecast.
2. Perturbations are modified by an ensemble of data assimilations (EDA): a set of 6-hour forecasts starting from 10 different analyses differing by small perturbations on observations, temperature and stochastic physics.
3. Model uncertainty is modelled by stochastic perturbation techniques. One modifies the physical parametrisation schemes and the other modifies the vorticity tendencies modelling the kinetic energy of the unresolved scales (scales smaller than the grid model resolution).

Perturbations are extracted from these different methods, linearly combined into 25 global perturbations. Then their signs are reversed to create the 25 other perturbations (“mirror“ perturbations). These 50 perturbed analyses are used to initiate the 50 perturbed forecasts. The EPS model has a horizontal resolution of approximately 50 km with 62 vertical levels (pressure levels) between the surface and the 5 hPa level (≈35 km). The integration time step is 1800 s. This resolution is much lower than for the deterministic model (≈10 km horizontal resolution) because of computation cost. Forecasts are generated twice a day (00_{UTC} and 12_{UTC}) and have a temporal resolution of 6 hours.

Wind speed and significant wave height are not instantaneous forecasts but hourly averaged variables. That is, a forecast issued for k hours ahead of surface wind speed represents predicted wind speed average on the previous hour of interest (between $k-1$ and k hours ahead). In order to guarantee consistency with the forecast definition and resolution of the different variables, wind speed and wave height observations are also averaged on the previous hour of every forecast hours (00_{UTC}, 06_{UTC}, 12_{UTC} and 18_{UTC}). This thesis deals with point probabilistic forecasts on the $FINO_1$ offshore measurement site (Germany, North Sea Position $54°01'N$, $06°35'E$) for lead times from 6 h to 168 h ahead. Since the FINO$_1$ location is not precisely on a grid point of the numerical model, forecasts are the result of a spatial interpolation of the predicted values on the closest model grid points.

Chapter 3

Ensemble Forecast verification

In 1993, Murphy put a special emphasis on a very important question: “What is a good forecast?“ (Murphy, 1993). He distinguished three types of goodness for a forecast system that he identified as consistency, quality and value. These types of goodness are connected to each other, however every one of them points out a certain aspect of forecasts.

1. **Consistency** corresponds to the difference between the forecaster’s judgement and the forecast. It is a subjective notion that can not be assessed quantitatively.

2. **Value** corresponds to the economic benefits, or the savings, realized by the use of forecasts in a decision-making problem.

3. **Quality** denotes the correspondence between the prediction and the observation.

Quality is the measure of goodness this thesis mainly deals with. It consists of comparing predicted values with observations. The more representative of the observations the predicted values are, the better the quality. For ensemble forecasting, it is important to distinguish measures-oriented from distribution-oriented approaches. Indeed, the quality of an ensemble forecast not only consists in the correspondence between observation and one forecast value, but also between observation and the distribution provided by the ensemble forecast. This is what distinguish ensemble forecast verification from deterministic forecast verification. In order to assess forecasts quality, it exists different verification scores and graphic tools which goes from quantitative products, like bias or mean absolute error, to qualitative like PIT diagram and Rank histograms.

This chapter explains how to assess forecast quality while listing and detailing the univariate and multivariate scores/tools used in this study.

In this report, x_{t+k} denotes the observation at time $t+k$ and $y^{(j)}_{t+k|t}$ the j^{th} ensemble forecast member issued at time t for time $t+k$ (hence k denoting the lead time).

3.1 Univariate Forecasts Verification

An ensemble forecast for a given meteorological variable, location and lead time consists in a set of predicted values. This set might comprise forecasts from several NWP models or from the same models but with different initial conditions and parametrisations. From this ensemble of forecasts, different criteria can be computed (mean, median, quantiles value...). Certain of those criteria can be preferred by some users because of their sensitivity to forecast errors. This sensitivity can be represented by a loss function (or cost function) whose goal is to inform about how prediction errors impact on a score. For example the MAE and RMSE scores, which are two well-know scores do not have the same loss function. In the case of RMSE the loss function is a quadratic function whereas for MAE it is a linear functions. RMSE is therefore much more sensitive to large errors. It has been proved in the literature (Gneiting, 2011) that the mean and the median value of an ensemble forecast are specific point forecasts that respectively minimise quadratic and linear loss function.

Bias The bias is the average of errors, it indicates systematic errors.

$$Bias(k) = \frac{1}{n}\sum_{t=1}^{n} \bar{y}_{t+k|t} - x_{t+k} \tag{3.1}$$

with n the number of forecasts over the verification period, $\bar{y}_{t+k|t}$ the ensemble mean. The bias of an ensemble forecast is minimised by the ensemble mean. It is a negatively oriented score, that is the closer to zero, the better.

Mean Absolute Error (MAE) The Mean Absolute Error is the average of absolute errors,

$$MAE(k) = \frac{1}{n}\sum_{t=1}^{n} |\tilde{y}_{t+k|t} - x_{t+k}| \tag{3.2}$$

with n the number of forecasts over the verification period, and $\tilde{y}_{t+k|t}$ the ensemble median. MAE's loss function is a linear function, and so the ensemble median minimises the MAE. The MAE is a negatively oriented score with zero being the minimum value.

Root Mean Square Error (RMSE) The Root Mean Square Error is the average of the squared errors, compared to the MAE it is much more sensitive to large errors.

$$RMSE(k) = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \left(\bar{y}_{t+k|t} - x_{t+k}\right)^2} \tag{3.3}$$

with n the number of forecasts over the verification period, and $\bar{y}_{t+k|t}$ the ensemble mean. RMSE's loss function is a quadratic function, and so the ensemble mean minimises the RMSE. The RMSE is also a negatively oriented score, with zero being the minimum value. Like the Bias, the RMSE only assess the ensemble mean quality and is independent of the ensemble spread.

Continuous Rank Probabilistic Score (CRPS) The Continuous Rank Probabilistic Score is a specific score for probabilistic forecast, it assesses the quality of the entire predicted probability density function.

$$CRPS(f, x_{t+k}) = \int_x (F(x) - \mathbb{I}\{x \geqslant x_{t+k}\})^2 dx \tag{3.4}$$

Where $\mathbb{I}\{x \geqslant x_{t+k}\}$ is the heaviside step function, taking the value 1 for $x \geqslant x_{t+k}$ and 0 otherwise, f is the predictive probability density function and F the corresponding cumulative density function (cdf).

The CRPS estimates the area between the predicted cumulative density function and the cdf of the observation (heaviside). Gneiting and Raftery (Gneiting and Raftery, 2007) showed that the CRPS can be written as follows:

$$CRPS(f, x) = \mathbb{E}_f|X - x| - \frac{1}{2}\mathbb{E}_f|X - X'| \tag{3.5}$$

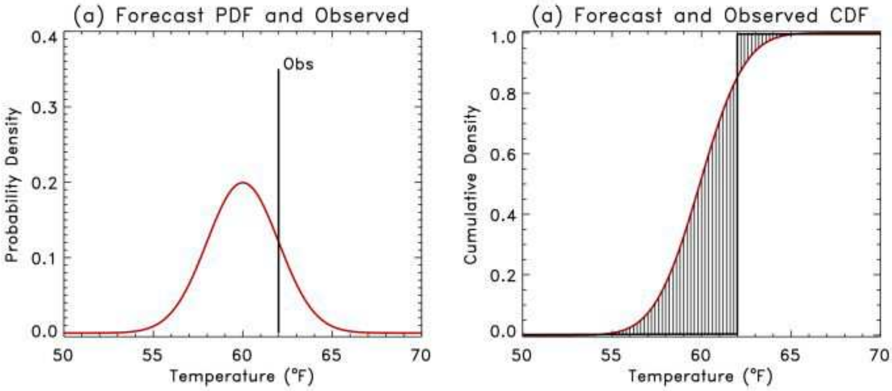


Figure 3.1: Illustration of the CRPS for one probabilistic forecast (a) pdf and observation value, (b) corresponding cdfs (source `http://www.eumetcal.org`)

Where X and X' are independent random variables with distribution f, and x_{t+k} is the observation. This score permits a direct comparison of deterministic and probabilistic forecasts considering that the cdf of a deterministic forecast would also be an heaviside function. For a ensemble forecast of M members $(y^{(}_{t+k|t}1), \ldots, y^{(}_{t+k|t}M))$ sampling a predictive distribution denoted by $\widehat{f}_{t+k|t}$, the CRPS can be computed as follows:

$$CRPS(\widehat{f}_{t+k|t}, x_{t+k}) = \frac{1}{M}\sum_{j=1}^{M}|y^{(j)}_{t+k|t} - x_{t+k}| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|y^{(j)}_{t+k|t} - y^{(i)}_{t+k|t}| \quad (3.6)$$

The CRPS is a negatively oriented score, with zero being the minimum value.

Sharpness Sharpness is a property of the forecast only, it does not depends on the observation. It characterises the ability of the forecast to deviate from the climatological probabilities. It is important for the ensemble spread not to be wider than the climatological spread and not too sharp if it leads to a loss of reliability. For an equal level of reliability, the sharper the better. Here, a way to assess sharpness is to determine the width of two equidistant quantiles from the median.

Rank Histograms The Rank Histogram (also known as Talagrand Diagram) is not a score but a tool employed to qualitatively assess ensemble spread con-

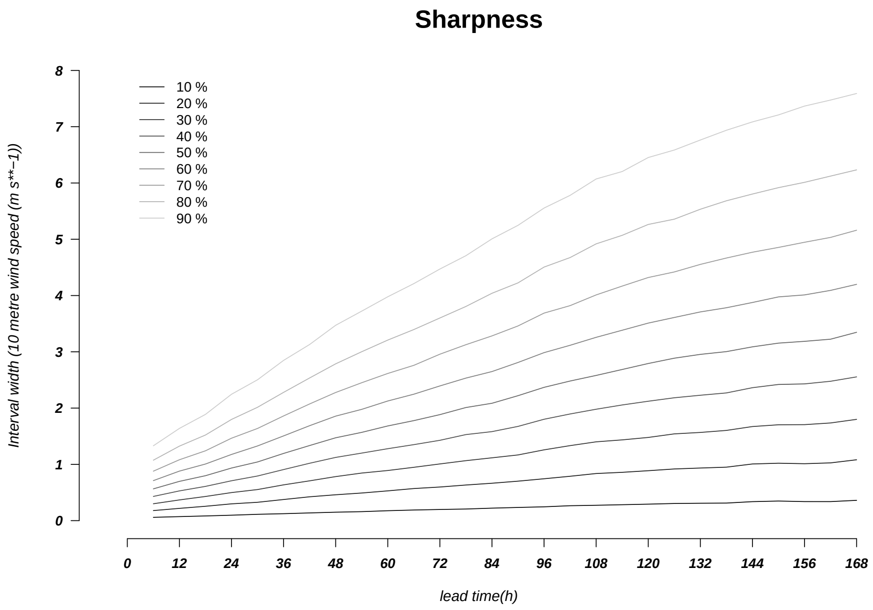


Figure 3.2: Example of sharpness assessment of the 10 m wind speed forecast from +06 to +168 h ahead. The width of the different probability intervals around the median forecast (from 10% to 90%) are drawn

sistency, and so ensemble forecast reliability. It is based on the idea that, for a perfect ensemble forecast, the observation is statistically just another members of the predicted sample, that is the probability of occurrence of observations within each delimited ranges (or bins) of the predicted variable should be equal. Rank histograms are computed in 2 steps:

1. The rank is computed looking at eventual equality between ensemble members and observation
$$s^{<} = \sum_{i=1}^{M} \mathbb{I}\{y_{t+k|t}^{(i)} < x_{t+k}\} \qquad s^{=} = \sum_{i=1}^{M} \mathbb{I}\{y_{t+k|t}^{(i)} = x_{t+k}\} \tag{3.7}$$
the rank r is an random integer picked $\{s^{<}+1, ..., s^{<}+s^{=}\}$.

2. All the ranks from the tested period are then aggregated and there respective frequency are plotted to obtain the rank histogram.

For a perfect ensemble forecast, every ensemble member is equally probable, thus every rank is equally populated, leading to a uniform histogram.

Figure 3.3 shows an example of a rank histogram. The horizontal axis represents the sorted bins of the M-members ensemble forecast system, and the vertical axis represents the probability of occurrence of observation into each bin. The 95% consistency bars have been added to the figure. Consistency bars give the potential range of empirical proportions that could be observed even if dealing with perfectly reliable probabilistic forecasts. These intervals depend on the length of the period tested and are estimated as follows,

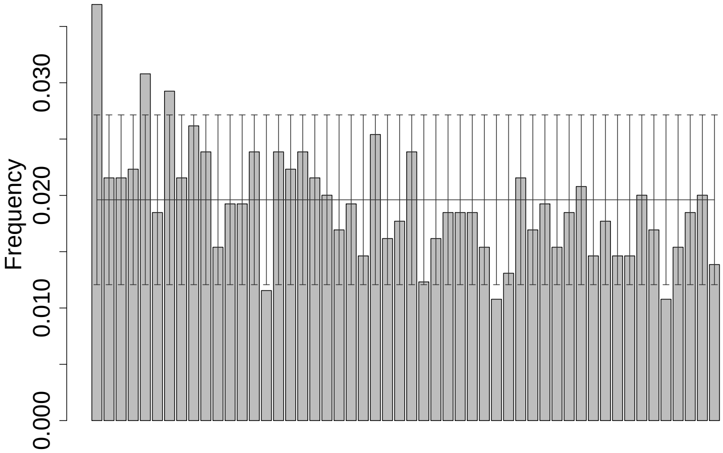


Figure 3.3: Example of a rank histogram with 95% consistency bars

$$
\begin{aligned}
I_c &= \frac{1}{M} + 1.96\frac{s}{\sqrt{n}} \qquad (3.8)\\
&= \frac{1}{M} + 1.96\sqrt{\frac{p(1-p)}{n}} \qquad (3.9)\\
&= \frac{1}{M} + 1.96\sqrt{\frac{\frac{1}{M}(1-\frac{1}{M})}{n}}
\end{aligned}
$$

with p the perfect probability of occurrence, M the number of ensemble members and n the number of valid observations. A rank histogram is considered statistically uniform if the probability of occurrence of each bins lies into the consistency bars. It exists particular shapes of rank histograms :

- If the too extreme bins are overpopulated (U shape), then the forecasts are underdispersive because most of the observations fall outside of the ensemble range.

- If the middle bins are overpopulated (bell shape), then the forecasts are overdispersive, observations do not enough fall into the extreme bins because the predicted distribution is too wide.
- If the lower bins are overpopulated, then the forecasts have a positive bias.
- If the higher bins are overpopulated, then the forecasts have a negative bias.
The figure 3.4 illustrates the previous list.

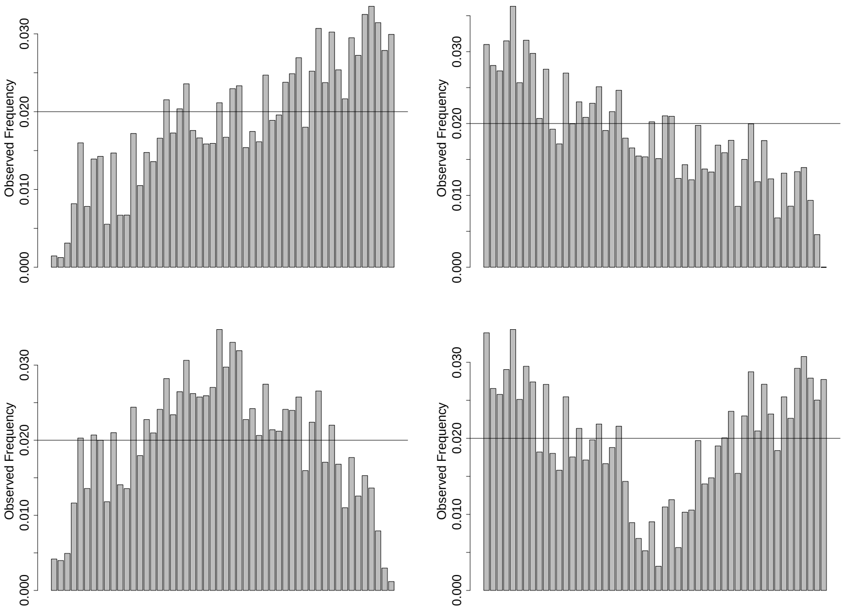


Figure 3.4: Usual kinds of rank histogram : negatively biased (top left), positively biased (top right),underdispersive (bottom left), overdispersive (bottom right)

Reliability Index It exists a way to quantitatively assess reliablity. The reliability index Δ, introduced by Delle Monache in 2006 (Delle Monache et al., 2006) quantifies the deviation of the rank histogram from uniformity.

$$\Delta_k = \sum_{j=1}^{M} \left| \xi_{k_j} - \frac{1}{M+1} \right| \tag{3.10}$$

where ξ_{k_j} is the observed relative frequency of the rank j for lead time k and M the number of ensemble members. The Reliability index is a negatively oriented score, with zero being the minimum value.

PIT Diagram The PIT diagram is equivalent to the rank histogram. It is the most transparent way to illustrate the performance and characteristics of a probabilistic forecast system. It represents the observed frequency of occurrence conditional on predicted probabilities. The x-axis represents the predicted probability and the y-axis the observed frequency. For instance, a point on the PIT diagram with coordinates $(x = 0.9, y = 0.6)$ can be interpreted as the event A predicted with a probability of 0.9 is actually observed only 6 times out of 10 over the tested period. In that case, the forecast is not reliable. To be reliable, the event A predicted with probability 0.9 should be approximately observed 9 times out of 10. For a perfect ensemble forecast system, predicted probability and observed probability should be identical, the PIT diagram should be represented by the 45° straight line. As with the rank histogram, it exists several type of PIT diagrams: if the slope is to low, then the forecasts are underdispersive and if the slope is to high they are overdispersive.

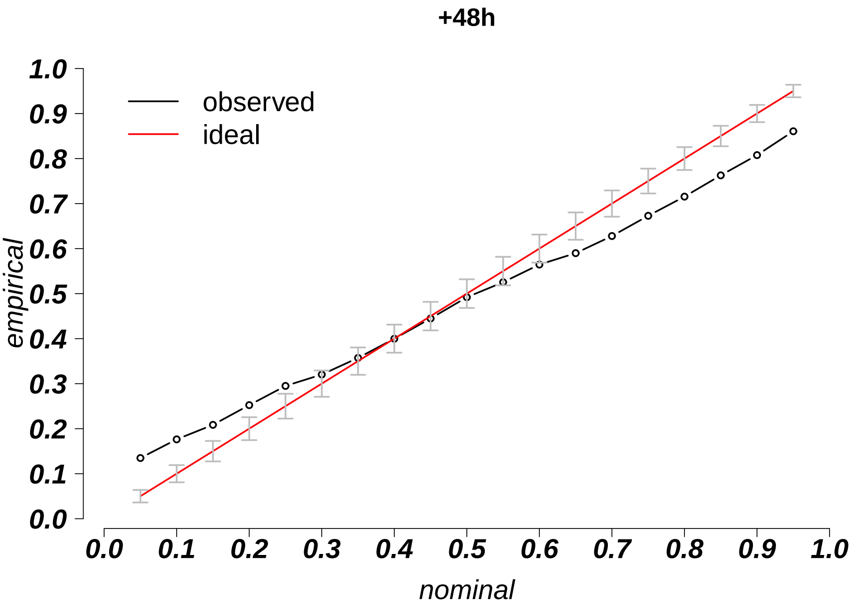


Figure 3.5: Example of a PIT diagram with the 95% consistency bars, the horizontal axis represents the predicted probability the vertical axis represents the observed probability

As well as for the rank histogram, consistency bars can be obtained thanks to the equation (3.10). However, contrary to the rank histogram all the consistency bars do not have the same width. Indeed, the perfect probability p is not constant (from 0 to 1) and the product $p(1-p)$ evolves as a quadratic function with a maximum at $p = 0.5$. Consistency bars are therefore larger for the predicted probability $p = 0.5$.

3.2 Multivariate Forecasts Verification

Multivariate forecasting needs specific multivariate verification. It exists several generalisations of univariate scores/tools permitting to assess the quality of multivariate forecasts.

Bivariate Root Mean Square Error The bivariate Root Mean Square Error is the multivariate generalisation of the RMSE and is computed as follows:

$$bRMSE(k) = \sqrt{\frac{1}{N}\sum_{t=1}^{N} \|\bar{\boldsymbol{y}}_{t+k|t} - \boldsymbol{x}_{t+k}\|^2} \tag{3.11}$$

with $\bar{\boldsymbol{y}}_{t|t-k}$ and $\boldsymbol{x}_t$ denoting the multivariate ensemble forecast mean vector and the multivariate observation vector.

Bivariate Mean Absolute Error The bivariate Mean Absolute Error is the multivariate generalisation of the MAE and is computed as follows :

$$bMAE(k) = \frac{1}{N}\sum_{t=1}^{N} \|\tilde{\boldsymbol{y}}_{t+k|t} - \boldsymbol{x}_{t+k}\| \tag{3.12}$$

with $\tilde{\boldsymbol{y}}_{t+k|t}$ and $\boldsymbol{x}_{t+k}$ the multivariate ensemble forecast median vector and the multivariate observation vector.

Energy Score (es) The Energy Score is the multivariate generalisation of the Continuous Rank Probabilistic Score. It was introduced by Geniting and Raftery in 2007 (Gneiting and Raftery, 2007) and can be computed as follows :

$$es(f, \boldsymbol{x}) = \mathbb{E}_f\|\boldsymbol{X} - \boldsymbol{x}\|^\beta - \frac{1}{2}\mathbb{E}_f\|\boldsymbol{X} - \boldsymbol{X}'\|^\beta \tag{3.13}$$

Where X and X' are independent random vectors coming from the distribution f, $\boldsymbol{x}$ is the observation vector and β represents the dimension of the problem (3.13 reduces to the CRPS formulation when $\beta = 1$). The Energy score is a negatively oriented score, with zero being the minimum value. For a given time and a give lead time, the energy score of a bivariate ensemble forecast, with $(\boldsymbol{y}^(1)_{t+k|t}, \boldsymbol{y}^(2)_{t+k|t}, \ldots, \boldsymbol{y}^(M)_{t+k|t})$ denoting the ensemble members, can be written

$$es(\widehat{f}_{t+k|t}, \boldsymbol{x}_{t+k}) = \frac{1}{M}\sum_{j=1}^{M} \|\boldsymbol{y}^{(j)}_{t+k|t} - \boldsymbol{x}_{t+k}\| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} \|\boldsymbol{y}^{(j)}_{t+k|t} - \boldsymbol{y}^{(i)}_{t+k|t}\| \quad (3.14)$$

However, while sampling a large number of outcomes from a predicted calibrated distribution the computation of this score can be costly. In order to reduce the computation effort, a Monte Carlo approximation can be used. Then the energy score can take the following form :

$$es(\widehat{f}_{t+k|t}, \boldsymbol{x}_{t+k}) = \frac{1}{K}\sum_{j=1}^{K} \|\boldsymbol{y}^{(j)}_{t+k|t} - \boldsymbol{x}_{t+k}\| - \frac{1}{2(K-1)}\sum_{i=1}^{K-1} \|\boldsymbol{y}^{(i)}_{t+k|t} - \boldsymbol{y}^{(i+1)}_{t+k|t}\| \quad (3.15)$$

where $y^{(1)}_{t+k|t}, \ldots, y^{(K)}_{t+k|t}$ is a random sample of size K=10000 picked out from the predicted probability density function $\widehat{f}_{t+k|t}$.

Multivariate Rank Histogram A multivariate rank histogram (Gneiting et al., 2008) is the multivariate generalization of rank histogram seen above. It shares the same ideas : multivariate rank histogram of a calibrated forecast should be uniform, multivariate rank histogram of underdispersive multivariate forecasts have an U shape and so on. Lets consider a multivariate ensemble forecast with the ensemble members $\boldsymbol{y}^i$ and the observations $\boldsymbol{x}$ defined by vectors that take values in $\mathbb{R}^d$: $\boldsymbol{y}^i = (y_1^{(i)}, y_2^{(i)}, \ldots, y_n^{(i)})$ and $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.

$$\boldsymbol{x} \preceq \boldsymbol{y}^{(i)} \qquad \text{if and only if} \qquad x_j \preceq y_j^{(i)} \quad \text{for} \quad j = 1, 2, \ldots, d.$$

Let suppose a bivariate ensemble forecasts as illustrated in figure 3.6, then $\boldsymbol{x} \preceq \boldsymbol{y}^{(i)}$ if and only if x belongs to the square to the left and below $\boldsymbol{y}^{(i)}$. The computation of the multivariate rank is a two step algorithm:

1. Assign pre-rank :
   We determine pre-rank ρ_j for each ensemble member:

   $$\rho_j = \sum_{k=0}^{m} \mathbb{I}\{x_k \leq x_j\} \quad (3.16)$$

   where $\mathbb{I}\{x_k < x_j\}$ denotes the indicator function $\mathbb{I}$ with the condition $x_k < x_j$. It is equal to 1 if the condition is true, 0 otherwise. Each pre-rank is an integer between 1 and $m + 1$.

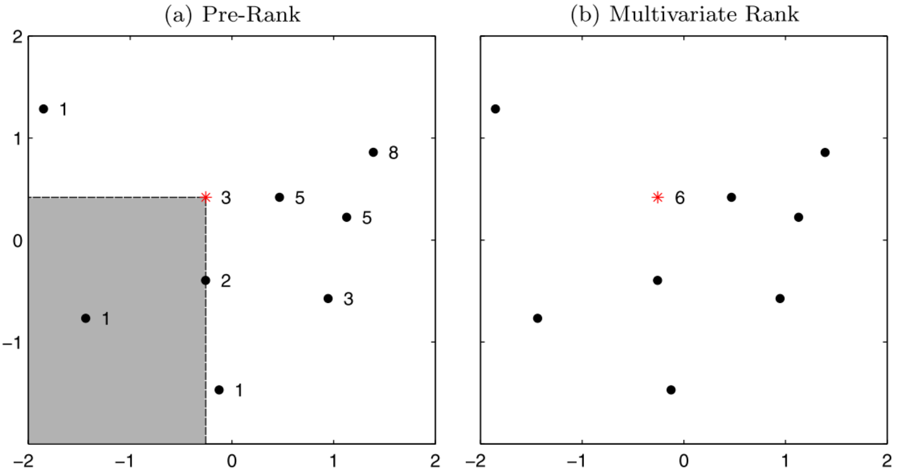


Figure 3.6: Illustration of the computation of a bivariate rank histogram for a particular ensemble forecast. (a) Ensemble forecast members and observations with associated pre-ranks. The observations pre-rank is 3 because 3 of the 9 points belong to its lower left (observations included). (b) From (a), four point have pre-rank ≤ 2, and two points have pre-rank 3, that is $s^{<} = 4$ and $s^{=} = 2$. Hence, the multivariate rank is a random outcome of the set $\{5, 6\}$ (source (Gneiting et al., 2008))

2. Find the multivariate rank :
   For the multivariate rank r, we note the rank of the observations, while possible ties are resolved at random:

$$s^{<} = \sum_{j=0}^{m} \mathbb{I}\{\rho_j < \rho_0\} \qquad \textit{and} \qquad s^{=} = \sum_{j=0}^{m} \mathbb{I}\{\rho_j = \rho_0\} \tag{3.17}$$

   The multivariate rank r is chosen from a discrete uniform distribution on the set $\{s^{<} + 1, \ldots, s^{<} + s^{=}\}$ and is an integer between 1 and $m + 1$.

3. Aggregate rank and plot multivariate rank histogram :
   We finally aggregate all multivariate ranks to plot the multivariate rank histogram and add the 95% consistency bars.

The figure 3.7 show an positively biased multivariate ensemble forecasting system of wind speed and significant wave height for 48 h ahead at $FINO_1$. Therefore, such a multivariate rank histogram suggests that the grey square at the bottom-left of figure 3.6 is rarely populated by ensemble vectors.

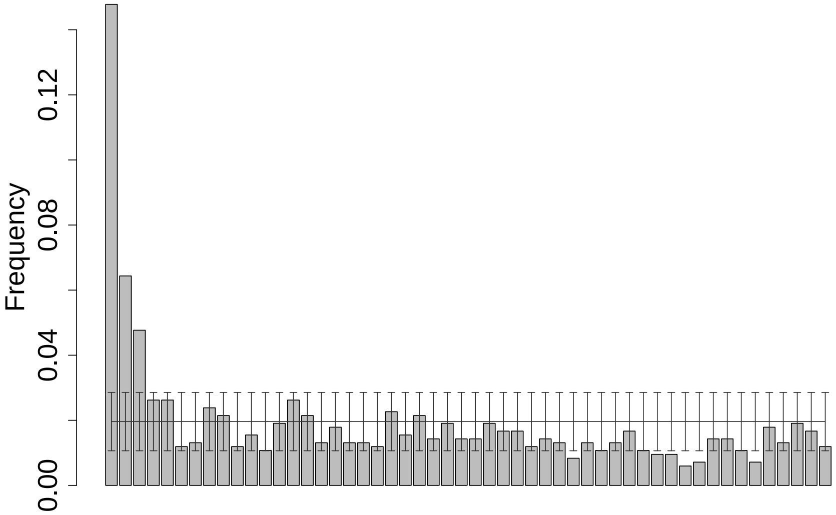


Figure 3.7: Example of Multivariate Rank Histogram for Wind speed +48 h forecast over $FINO_1$

Multivariate Reliability Index As it is the case for univariate rank histogram, a reliability index can be computed from equation (3.10) for any multivariate rank histogram in order to quantitatively assess forecasts calibration. We call this index the multivariate reliability index.

Determinant Sharpness The determinant sharpness is the multivariate generalization of the sharpness described previously. We use the same definition as employed in (Möller et al., 2012), that is,

$$DS = (det\Sigma)^{1/(2d)} \tag{3.18}$$

where Σ is the empirical covariance matrix of a multivariate ensemble forecast for a d-dimensional quantity.

3.3 Skill Score

The skill score is a way to directly compare one type of score for two different methods. For a given lead time k and a given score, the skill score is defined as:

$$SkillScore(k) = 1 - \frac{Score(k)}{Score_0(k)} \tag{3.19}$$

with $Score(k)$ the score of the tested method and $Score_0(k)$ is the score of the benchmark method. Contrary to the other score presented before, the skill score is always positively oriented, the higher the skill score, the better the tested method is compared to the benchmark. The maximum value of a skill score is 1, it would indicate a perfect score for the tested method. A skill score of 0 would indicate that the tested method and the benchmark have the same score, and a negative skill score would indicate that the tested method is worse than the benchmark.

Chapter 4

Ensemble Forecast Calibration Method

In this thesis, the main assumption is that Nature is not deterministic but probabilistic. We assume that at a time t+k, Nature chooses a distribution G_{t+k}, representing the uncertainty of the atmospheric conditions, and picks out one random number from that distribution to obtain the observation x_{t+k}. We also assume that the ensemble forecast members $y^{(j)}_{t+k|t}$ are outcomes from a predicted distribution $F_{t+k|t}$ that tries to estimate G_{t+k}. The goal of ensemble forecasting is to predict the closest distribution $F_{t+k|t}$ to the one chosen by the Nature G_{t+k}. However, ensemble forecasts tend to be uncalibrated (i.e. biased and underdispersive), which means that the mean and the spread of the predicted probability distribution show systematic errors. Therefore, we propose some calibration methods that aim at reducing those errors.

The first assumption of our model is that given a variable (wind speed or wave height), the distributions $\{G_{t_1}, G_{t_2}, \ldots, G_{t_n}\}$ chosen by the Nature to obtain the observation for the times $\{t_1, t_2, \ldots, t_n\}$ are of the same type and only differ from their parameters. The goal of our calibration method is to fit a distribution conditional on the ensemble mean and the ensemble variance. Therefore, we do a first approximation by fitting a distribution to sets of observed wind speed and significant wave height, conditional on the predicted ensemble mean being within some bins. In order to assess the representativeness of the chosen distribution for the different variables, we use quantile-quantile plots.

Quantile-Quantile plots (Q-Q plot) (Wilks, 2006) compare quantiles (inverse function of the cumulative density function) of an empirical data with quantiles of a distribution function with parameters being representative of the data. It is an indirect way of comparing two density functions. The horizontal axis represents the quantiles (dimension values) of the mathematical distribution function, and the vertical axis represents the quantiles estimated from the empirical data. A Q-Q plot of two samples coming from the same distribution would show points around the diagonal confirming that the quantiles of the two samples are similar.

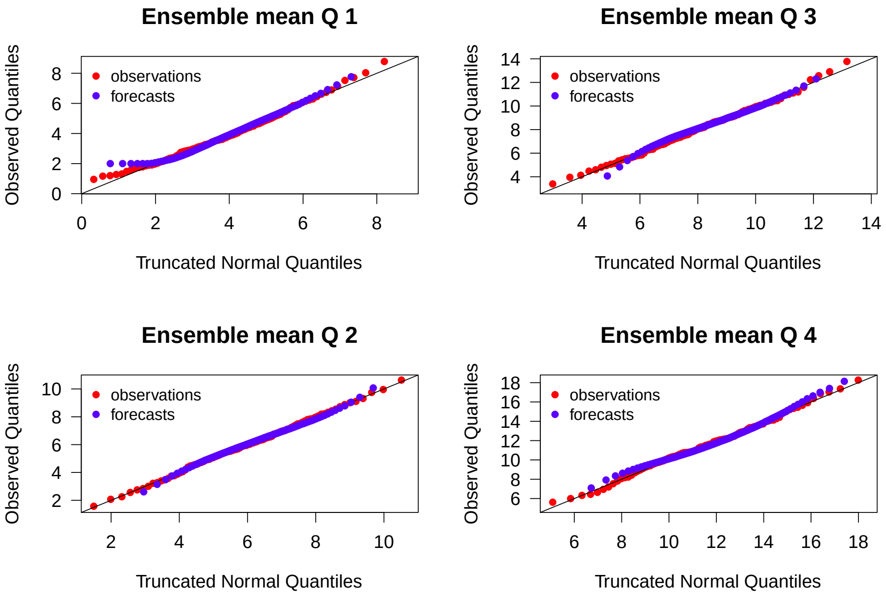


Figure 4.1: Q-Q plots for observed and predicted 10 m wind speeds conditional on the 10 m wind speed ensemble mean being in the four quartiles denoted Q1,Q2,Q3 and Q4 compared with a truncated normal distribution. Data used covers the period from January 2010 to December 2010.

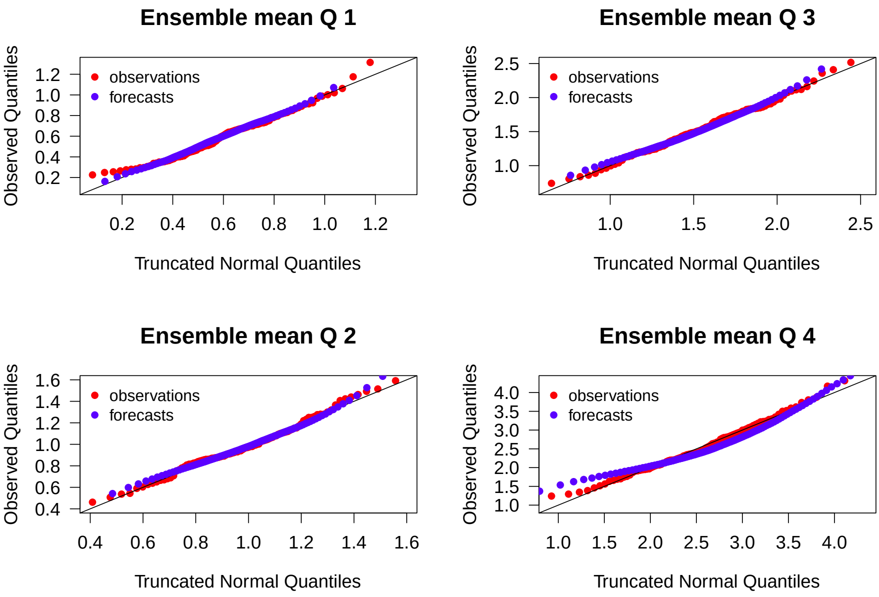


Figure 4.2: Q-Q plots for observed and predicted significant wave height conditional on the significant wave height ensemble mean being in the four quartiles denoted Q1,Q2,Q3 and Q4 compared with a truncated normal distribution. Data used covers the period from January 2010 to December 2010.

Figures 4.1 and 4.2 show the Q-Q plots comparing 10 m wind speed and significant wave height 48 hours ahead forecasts and observations of the year 2010 with the a truncated normal distribution with a lower bound at zero, fitted by the maximum likelihood technique (see Chapter 4.1). The data has been split into 4 bins corresponding to the four quartiles (25%,50%,75% and 100%) of the ensemble forecast mean. In the figures quantiles lies around the diagonal confirming that both variables sample the distributions tested. Thus, we choose to model significant wave height and 10 m wind speed as sampling truncated normal distributions with a cut off at zero for both variables. It can noticed on figure 4.1 that the 10 m wind speed ensemble members does not predict speeds below $2m.s^{-1}$, which is certainly due to the ECMWF parametrisation of this variable. However, there is no valuable reason for the wind not to blow below that speed at that height, plus our wind speed extrapolation might provide wind speeds between 0 and 2 $m.s^{-1}$. The truncation for the 10 m wind speed is therefore chosen at 0 $m.s^{-1}$ and not at 2 $m.s^{-1}$.

4.1 Univariate Calibration CAL[1]

After analysis of the QQ-plots for every lead times, we assume then that each observations x_{t+k}, as well as each ensemble member $y^{(j)}_{t+k|t}$, sample a truncated normal distribution with a cut-off at zero as employed in (Thorarinsdottir and Gneiting, 2008). Thus, the non-negativity of wind speed and significant wave height is respected.

$$x_{t+k} \sim \mathcal{N}_0(\mu_{t+k}, \sigma^2_{t+k}) \qquad y_{t+k|t} \sim \mathcal{N}_0(\hat{\mu}_{t+k|t}, \hat{\sigma}^2_{t+k|t}) \tag{4.1}$$

The truncated normal distribution is the probability distribution of a normally distributed variable whose value is either bounded below or above (or both). Here, only a lower boundary at zero is used. The distribution is defined by two parameters. The parameters μ and σ^2 are respectively the location parameter and the scale parameter. They represent the mean and the variance of the underlying normal distribution (non truncated) and are slightly different from the truncated mean and the truncated variance. The probability density function of the truncated normal distribution is given by

$$f^0(y) = \frac{1}{\sigma}\frac{\varphi\left(\frac{y-\mu}{\sigma}\right)}{\Phi\left(\frac{\mu}{\sigma}\right)} \qquad for \quad y > 0 \tag{4.2}$$

$f^0(y) = 0$ otherwise. Here, φ and Φ respectively denote the standard probability and cumulative density functions of the normal distribution.

The first two moments of the truncated normal distribution with a cut-off at zero can be computed as follows,

$$\mu_0 = \mu + \frac{\varphi(\frac{\mu}{\sigma})}{\Phi(\frac{\mu}{\sigma})}\sigma \tag{4.3}$$

$$\sigma_0^2 = \sigma^2 - \sigma^2\frac{\varphi(\frac{\mu}{\sigma})}{\Phi(\frac{\mu}{\sigma})}[1 + \frac{\mu}{\sigma}] \tag{4.4}$$

The closer the sample of the distribution to zero, the more the two first moments of the truncated normal distribution μ_0 and σ_0^2 differ from the moments of the underlying normal distribution μ and σ^2.

Given a lead time k, we assume that the location parameter of the observation μ_{t+k} is a linear function of the predicted mean $\bar{y}_{t+k|t}$. We choose not to follow exactly the idea of Thorarinsdottir (Thorarinsdottir and Gneiting, 2008) who builds a multiple linear model taking every ensemble members as predictors for the location parameter estimation. Considering the large number of ensemble members of the EPS (51 members) compared to the ensemble forecasting system used by Thorarinsdottir (8 members of the UWME), we choose to restrict

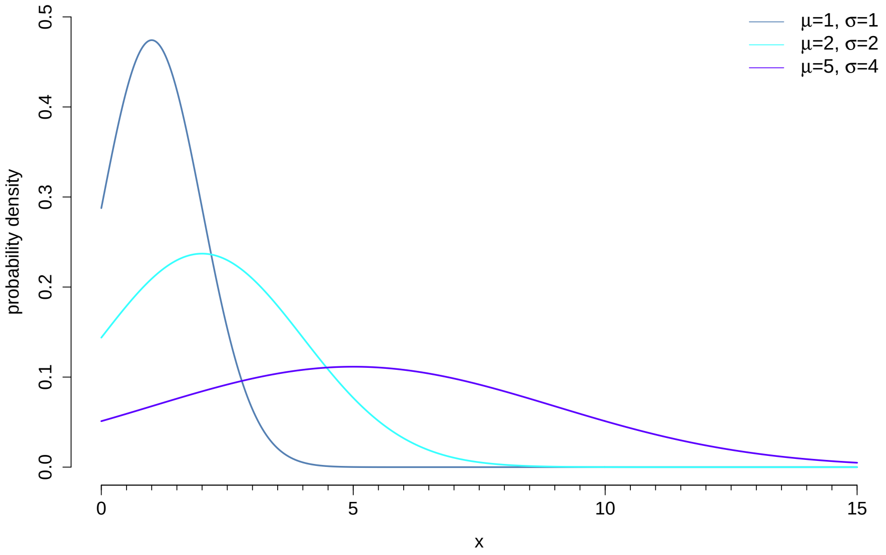


Figure 4.3: Examples of univariate truncated normal distributions with different location and spread parameters.

the predictors only to the ensemble mean. Therefore the model for the mean correction can be written as follows :

$$\mu_{t+k} = a + b\bar{y}_{t+k|t} \tag{4.5}$$

The mean correcting parameters a and b reflect the relationship between the ensemble mean and the location parameter during the training period. It has to be noted that even if wind speed and significant wave height are non negative variables, the location parameter of the truncated normal distribution does not have to be positive. Indeed, $\mu \in \mathbb{R}$ and the parameters a and b are then unconstrained.

Furthermore, the spread parameter of the observation σ^2_{t+k} is also assumed to be a linear function of the predicted variance $s^2_{t+k|t}$.

$$\sigma^2_{t+k} = c + ds^2_{t+k|t} \tag{4.6}$$

The parameters c and d reflect the relationship between the ensemble spread and the forecast error during the training period. When the correlation is important between those two, d is high and c is small, when the ensemble variance

information is not useful, d tends to be smaller and c to be higher. c and d are constrained to be non negative because of the positive nature of the second order moment. The non-negativity of the parameters c and d is guaranteed by writing $c = \gamma^2$ and $d = \delta^2$.
Thus the calibrated forecast $y^*_{t+k|t}$ should take the form of a truncated normal distribution with the corrected parameters,

$$y^*_{t+k|t} \sim \mathcal{N}_0(a + b\bar{y}_{t+k|t}, c + ds^2_{t+k|t}) \tag{4.7}$$

This model is not rigorously correct in the sense that the linear models should involve the true truncated mean and the true truncated variance instead of the location and spread parameters. However, The wind speed and the significant wave height rarely reach the lower boundary (zero), and so the location parameter μ and the truncated mean μ_0 do not significantly differs. It is also true for the spread parameter σ^2 and the truncated variance σ_0^2.

The estimation of those four correcting parameters can be done in different ways. (Gneiting et al., 2005; Thorarinsdottir and Gneiting, 2008) employed a minimum CRPS estimation, that is the parameters are estimated in such a way that they minimise the CRPS score over the estimation period. However, such a technique can not be employed for bivariate distributions. So, out of concern of consistency between the univariate and bivariate calibration methods, we choose to use maximum likelihood estimation. The maximum likelihood estimation is a popular method used to estimate the parameters of a mathematical function (often a distribution function). The parameters issued from the maximum likelihood estimation are supposed to be the most probable parameters given the observed data.
Let suppose that $(x_1, ..., x_n)$ $x_i \in \mathbb{R}^d$ is an independent and identically distributed sample coming from a statistical distribution with a probability density function $\{p_\theta : \theta \in \Theta\}$ in $\mathbb{R}^d$. θ is the parameter (scalar or vector) to be estimated. For a single observation x_i, the likelihood function is identical to the probability density function. The only difference is that the pdf is a function of the data (parameters being fixed) whereas the likelihood function is a function of the unknown parameters (data being fixed). The likelihood function of a distribution given a sample of n independent values is the product of the n individual likelihood functions :

$$\begin{aligned} \mathcal{L}(\theta|x_1, ..., x_n) &= p(x_1|\theta) \times p(x_2|\theta) \times ... \times p(x_n|\theta) \\ &= \prod_{i=1}^{n} p(x_i|\theta) \end{aligned} \tag{4.8}$$

We use a slightly different method which is a combination of maximum likelihood estimation techniques (Pawitan, 2001). Instead of having a sample of n observations from one and only statistical distribution with a probability density function p_θ, we have n independent observations $(x_1, ..., x_n)$ sampling n different statistical distributions which share the same parameter θ with the respective probability density functions $(p_{i,\theta}, ..., p_{n,\theta})$. Then the likelihood function becomes :

$$\mathcal{L}(\theta|x_1, ..., x_n) = \prod_{i=1}^{n} p_i(x_i|\theta) \tag{4.9}$$

where $p_i(x_i|\theta)$ represents the individual likelihood function of the pair (x_i,θ), equal to the probability density function of the distribution that x_i is sampling with the parameter θ.

Generally, we work with the log-likelihood function which is the sum of the logarithm of the individual likelihood functions :

$$\begin{aligned} \ln(\mathcal{L}(\theta|x_1, ..., x_n)) &= \ln(\prod_{i=1}^{n} p_i(x_i|\theta)) \\ &= \sum_{i=1}^{n} \ln(p_i(x_i|\theta)) \end{aligned} \tag{4.10}$$

In these functions (likelihood and log-likelihood), the observed values $x_1, ..., x_n$ are the fixed parameters and θ is actually the variable. The maximum likelihood estimation finds the best estimator of θ by maximising the likelihood (or log-likelihood) function:

$$\theta_{mle} = \underset{\theta\in\Theta}{\text{argmax}} \quad \mathcal{L}(\theta|x_1, ..., x_n) \tag{4.11}$$

θ_{mle} is the statistically most probable estimator of θ.

In our case, that is the correction of the ensemble mean and variance, $p_i(x_i|\theta)$ is equal to the probability density function of a truncated normal distribution $\mathcal{N}^0(a + b\bar{y}_{t+k|t}, c + ds_{t+k|t})$ seen in equation (4.2). Then, the log-likelihood function that has to be maximised can be written as follows:

$$\begin{aligned} \mathcal{L}(a, b, \gamma, \delta) &= \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma}\frac{\varphi(\frac{y-\mu}{\sigma})}{\Phi(\frac{\mu}{\sigma})}\right) \\ &= \sum_{i=1}^{n} \left(- \ln\sigma + \ln\varphi(\frac{y-\mu}{\sigma}) - \ln\Phi(\frac{\mu}{\sigma})\right) \\ &= \sum_{i=1}^{n} \left(- \frac{1}{2}\ln(\gamma^2 + \delta^2 s_i^2) \right. \\ &\qquad \left. + \ln\varphi(\frac{y - a + b\bar{y}_i}{\sqrt{\gamma^2 + \delta^2 s_i^2}}) - \ln\Phi(\frac{a + b\bar{y}_i}{\sqrt{\gamma^2 + \delta^2 s_i^2}})\right) \end{aligned} \tag{4.12}$$

$$c = \gamma^2 \quad and \quad d = \delta^2$$

with n being the length of the training period, a, b, γ and δ some unconstrained parameters. Of course, this method can only be employed if we assume that forecasts errors are independent in time. However, as explained in (Raftery et al., 2005), estimates are unlikely to be sensitive to this assumption because calibration is done for one particular time only and not several simultaneously. The optimization algorithm used to estimate the four parameters for every forecast needs starting values and the solution might be sensitive to those initial values.Therefore, we use the parameters from the previous estimation day as starting values. This allows a faster convergence of the algorithm and a better consistency in the evolution of the parameters from one day to the other. In this report, this univariate calibration method is called CAL[1].

4.2 Bivariate Calibration

We have seen in chapter 2 that surface wind speed and significant wave height are, in a certain way, correlated. Wind-waves are created by the local winds and are then strongly correlated to surface wind speed, but swell was created in another location and a few hours or days before by different winds and thus might be uncorrelated to local wind speed. Since it calibrates one variable at a time, the univariate calibration method does not take into account the correlation and therefore predicts uncorrelated bivariate forecasts. The idea of a bivariate calibration is to jointly calibrate the variables so their existing correlation is not violated.

Bivariate calibration of weather forecasts is a new field of researches that only a few scientists have investigated to this day. Pinson (Pinson, 2012) used a bivariate EMOS to calibrate the u and v components of wind. Assuming that the joint distribution of the two components of the wind is a bivariate normal distribution, he did not modify the predicted correlation but corrected the mean of each component through a linear function with the two predicted means as predictor. Möller (Möller et al., 2012) used Gaussian copula to introduce a bivariate calibration after having marginally calibrated different variables like temperature, wind speed and precipitation. She showed that the dependence lost during marginal calibration of any weather variable can be recovered using gaussian copula. The correlation of the copula is estimated from the training period. Schuhen (Schuhen et al., 2012) also jointly calibrated the u and v component. Like Pinson, she assumed that wind vectors sample a bivariate normal distribution. However, instead of using a predicted correlation she showed that correlation could be estimated through a sinusoidal function of the ensemble mean wind direction.

In this thesis, we create a bivariate calibration method that aims at calibrating 10 m wind speed and significant wave height while taking into account their predicted correlation. We assume that vectors defined by 10 m wind speed and significant wave height sample a truncated bivariate normal distribution with a lower bound at zero for both components (Horrace, 2005; Wilhelm and Manjunath, 2010). We denote Y the two components forecast vector sampling a truncated bivariate normal distribution.

$$Y \sim \mathcal{N}_2^0(\hat{\boldsymbol{\mu}}_{t+k|t}, \hat{\boldsymbol{\Sigma}}_{t+k|t}) \tag{4.13}$$

with $\hat{\boldsymbol{\mu}}_{t+k|t}$ the two components predicted mean vector with the first component for the wind ($\hat{\mu}_u$) and the second for the wave height ($\hat{\mu}_h$). $\hat{\boldsymbol{\Sigma}}_{t+k|t}$ is the predicted variance-covariance matrix with the variances on the diagonal and the covariances outside of the diagonal.

$$\hat{\boldsymbol{\mu}}_{t+k|t} = \begin{pmatrix} \hat{\mu}_{u,t+k|t} \\ \hat{\mu}_{h,t+k|t} \end{pmatrix} \tag{4.14}$$

$$\hat{\boldsymbol{\Sigma}}_{t+k|t} = \begin{pmatrix} \hat{\sigma}^2_{u,t+k|t} & \hat{\rho}_{t+k|t}\hat{\sigma}_{u,t+k|t}\hat{\sigma}_{h,t+k|t} \\ \hat{\rho}_{t+k|t}\hat{\sigma}_{u,t+k|t}\hat{\sigma}_{h,t+k|t} & \hat{\sigma}^2_{h,t+k|t} \end{pmatrix} \tag{4.15}$$

Indices $_u$ and $_h$ respectively denote the 10 m wind speed and the significant wave height component.
Given boundaries, the truncated bivariate normal distribution is then fully described by 5 parameters:

1. the mean value for the two components μ_u and μ_h
2. the variance for two components σ_u^2 and σ_h^2
3. the correlation between the two components ρ

The probability density function of the truncated bivariate normal distribution can be written as follows:

$$f^0(\boldsymbol{y}) \quad = \quad \frac{\exp\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\}}{\iint_0^\infty \exp\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\}d\boldsymbol{x}} \tag{4.16}$$

Two examples of truncated bivariate normal distribution with a cut-off at zero for both dimensions are illustrated in figure 4.4.

The models for the correction of the predicted mean and the predicted variance remain identical to the univariate calibration:

$$\mu_{u,t+k} = a_u + b_u\bar{y}_{u,t+k|t} \tag{4.17}$$

$$\mu_{h,t+k} = a_h + b_h\bar{y}_{h,t+k|t} \tag{4.18}$$

$$\sigma^2_{u,t+k} = c_u + d_u s^2_{u,t+k|t} \tag{4.19}$$

$$\sigma^2_{h,t+k} = c_h + d_h s^2_{h,t+k|t} \tag{4.20}$$

The parameters c_u,d_u,c_h and d_h have to be non-negative in order to ensure that the mean and the variance are positive. Thus we write $c_u = \gamma_u^2$,$d_u = \delta_u^2$,$c_h = \gamma_h^2$

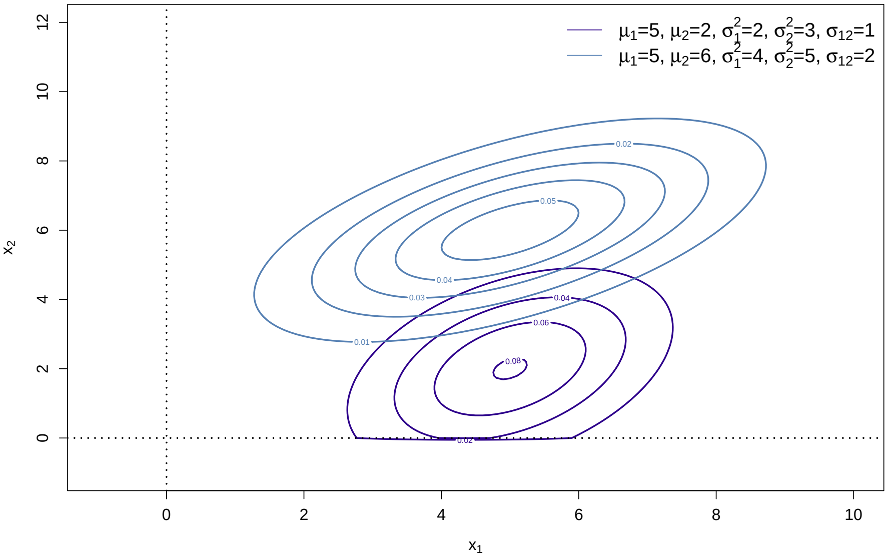


Figure 4.4: Example of a truncated bivariate normal distribution with a cut-off at zero for both dimensions.

and $d_h = \delta_h^2$ which leads to

$$\begin{aligned}
\mu_{u,t+k} &= a_u + b_u \bar{y}_{u,t+k|t} && (4.21)\\
\mu_{h,t+k} &= a_h + b_h \bar{y}_{h,t+k|t} && (4.22)\\
\sigma^2_{u,t+k} &= \gamma_u^2 + \delta_u^2 s^2_{u,t+k|t} && (4.23)\\
\sigma^2_{h,t+k} &= \gamma_h^2 + \delta_h^2 s^2_{h,t+k|t} && (4.24)
\end{aligned}$$

The last parameter to deal with is the correlation ρ. We propose here two different ways of taking the correlation into account. Firstly, we propose to estimate the correcting parameters while forcing the correlation to be equal to the predicted one, maximum likelihood technique is employed using the truncated bivariate distribution. A second approach is envisaged, we propose to marginally estimate the correcting parameter as for the univariate calibration method CAL^1 and then recover the lost dependence using the bivariate truncated distribution with the raw predicted correlation.

4.2.1 Joint Calibration method CAL^2

We choose to build a bivariate calibration method that take into account the predicted correlation without correcting it. The true correlation is assumed to be identical to the predicted one. So we write :

$$\rho_{t+k} = \hat{\rho}_{t+k|t} \tag{4.25}$$

Thus the calibration method does not violate the existing relationship between wind speed and wave height. Of course, as it has been discussed for the univariate calibration method, it is not strictly correct since the correlation of the underlying normal distribution and the correlation of the truncated normal distribution are not exactly identical. Indeed, the closer to the boundaries the ensemble members are, the more different the two correlation are. However, this assumption is relevant in the sense that ensemble members of wind speed and significant wave height are rarely close to the lower boundary (zero). The correlation being hold fixed, the calibration method has to estimated 8 parameters (4 for the mean vector correction and 4 for the variance-covariance matrix). As it has been done for the univariate calibration method, parameters a_u,b_u,c_u,d_u,a_h,b_h,c_h,d_h are estimated through maximum likelihood technique with the correlation as a fixed parameter. The function to maximise can be written as follows,

$$\mathcal{L}(\boldsymbol{\eta}) = \sum_{i=1}^{n} \left[\ln \Big(f(\boldsymbol{\eta}|\boldsymbol{x}_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \Big) \right]$$

where f denotes the density of the truncated bivariate normal predicted distribution, $\eta = (a_u, b_u, \gamma_u, \delta_u, a_h, b_h, \gamma_h, \delta_h)^T$ the parameter vector, $\boldsymbol{x}$ the observation vector, $\boldsymbol{\mu}$ the predicted mean vector, $\boldsymbol{\Sigma}$ the predicted variance-covariance matrix for all forecasts contained in the training period of n events.

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\eta}) &= \sum_{i=1}^{n} \left[\ln \Big(\frac{\exp\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_i)\}}{\iint_0^{\infty} \exp\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\} d\boldsymbol{x}} \Big) \right] \qquad (4.26) \\
&= \sum_{i=1}^{n} \left[-\ln \Big(\iint_0^{\infty} \exp\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_i)\} d\boldsymbol{x} \Big) \right. \\
&\quad \left. -\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_i) \right]
\end{aligned}$$

with

$$\begin{aligned}
-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_i) &= \begin{pmatrix} x_{u,i}-\mu_{u,i} \\ x_{h,i}-\mu_{h,i} \end{pmatrix}^T \begin{pmatrix} \sigma_{u,i}^2 & \rho_i\sigma_{u,i}\sigma_{h,i} \\ \rho_i\sigma_{u,i}\sigma_{h,i} & \sigma_{h,i}^2 \end{pmatrix}^{-1} \begin{pmatrix} x_{u,i}-\mu_{u,i} \\ x_{h,i}-\mu_{h,i} \end{pmatrix} \\
&= -\frac{(x_{u,i}-a_u+b_u\bar{y}_{u,i})^2}{2(1-\rho_i^2)(\gamma_u^2+\delta_u^2 s_{u,i}^2)} \\
&\quad -\frac{(x_{h,i}-a_h+b_h\bar{y}_{h,i})^2}{2(1-\rho_i^2)(\gamma_h^2+\delta_h^2 s_{h,i}^2)} \\
&\quad +\frac{\rho_i(x_{u,i}-a_u+b_u\bar{y}_{u,i})(x_{h,i}-a_h+b_h\bar{y}_{h,i})}{(1-\rho_i^2)\sqrt{(\gamma_u^2+\delta_u^2 s_{u,i}^2)(\gamma_h^2+\delta_h^2 s_{h,i}^2)}}
\end{aligned}$$

The correlation parameter ρ of the truncated bivariate normal distribution is fixed. The maximum likelihood estimation finds the values for the parameters under which the probability that the training data would have been observed is maximal. Since the number of parameter is higher, the estimation is much more sensitive and the computation time more costly than for the univariate method, it is even more important to find appropriate initial values to start the optimization. Therefore, it is even more important that parameters from the previous estimation are used as input to initiate the algorithm. In this report, this bivariate calibration method is called CAL^2.

4.2.2 EPS-prescribed correlation approach CAL^{1+}

In a second approach, we choose to combine marginally calibrated forecasts and use the predicted correlation to recover the dependence during the univariate calibration method. The method is similar to the one proposed by Möller (Möller et al., 2012). In order to obtain a joint predictive distribution of 10 m wind speed and significant wave height, each variable is individually calibrated as described in Chapter 4.1 without considering nor the other variable neither the correlation existing between the two of them. Then in a second step, the dependence lost during the estimation of the correcting parameters is recovered by sampling a truncated bivariate normal distribution (equation(4.16)) with the marginally estimated parameters and the correlation prescribed by the raw forecasts. Therefore, this calibration method will share the same univariate scores than the univariate calibration method but intend to improve the bivariate pattern of the forecasts. Contrary to Möller who used a correlation estimated from the training period, we use the correlation predicted by the EPS for the valid date which allow a more dynamic correction. In this report, this method is called CAL^{1+}.

Chapter 5

Results

In this chapter, the results of the different calibration methods applied on the EPS forecasts the ECMWF for the specific $FINO_1$ location are exposed. Forecasts have been tested over the 1855 forecasts events of the period 2010-2011 when both 10 m wind speed and significant wave height were both available. Calibration methods are systematically compared with the raw ensemble forecasts and the climatology benchmark. The verification tools/scores described in Chapter 3 are used to assess the different forecast types. We first assess the univariate calibration method for lead times from 6 to 168 hours ahead with univariate scores/tools and then expose the results of the bivariate calibration methods for the 48 hours ahead prediction horizon.

5.1 Benchmarking

A benchmark is defined by a very simple model that can used in forecasting. The climatology forecast is a type of probabilistic forecasts that is often used as a benchmark because of it simplicity. It consists in forecasting, the pdf of the climatology, that is a non parametric density function estimated from the past observations. The climatology forecast is by definition reliable, but not sharp. It is not flow dependent and therefore the predicted pdf is always identical.

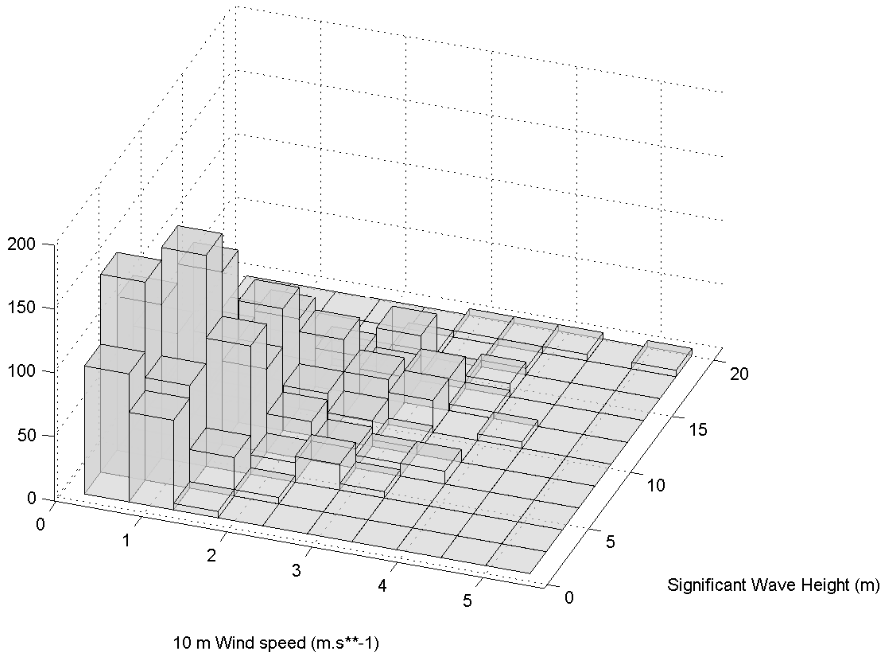


Figure 5.1: Three dimensional representation of the climatology benchmark forecast. The vertical axis represents the number of occurence per bin. This histogram has been computed with the 1855 pairs of wind speed - wave height observations available over the period 2010-2011.

The Climatology forecasts is defined by the same parameters shown in table 5.1 and the same distribution seen in figure 5.1 for every day and every prediction horizon.

μ_u	μ_h	σ_u^2	σ_h^2	ρ
7.33	1.39	12.73	0.70	0.72

Table 5.1: Parameters of the climatology benchmark computed over the entire period 2010-2011

5.2 Univariate Calibration Method

10 m wind speed and significant wave height ensemble forecasts are marginally calibrated with the calibration method CAL[1] detailed in Chapter 3.1, that is calibrated forecasts are obtained individually for every variable by fitting a truncated univariate distribution with corrected parameters. Correcting parameters are estimated from the training period and applied on every forecasts during the years 2010 and 2011.

The small amount of data suggests using a sliding window technique to estimate the correcting parameters in which the training data consists in the recent past. The optimal length of the training period depends on the variability of the atmospheric conditions on the site of interest. A short training period would suggest a fast adaptivity to seasonal variation whereas a long training period would reduce the variability of the parameter estimation. There does not exist an automatic way to find the optimal length for the training period. Thus, in order to estimate the most appropriate length, different lengths of training period from 20 days to 60 days have been tested, as proposed in (Schuhen et al., 2012) and the mean energy score (average on every lead times) has been computed for every simulation.

As suggested by the figure 5.2, the optimal length of the training period minimising the energy score of the marginally calibrated forecasts is approximatively 42 days. Thus the correcting parameters for day d are estimated on the period between day $d-1$ and $d-43$. However, the length of the training period being always equal to 42 days, if the data on day $d-10$, for instance, is not available, then the training data set covers the period from day $d-1$ to day $d-44$. Considering the fact that two ensemble forecasts are issued per days at ECMWF (one at 00_{UTC} and the other at 12_{UTC}), a 42 days training period consists in 84 forecast events.

Figure 5.3 illustrate the marginal calibration of an ensemble forecast of surface wind speed issued on the 14th of July 2010 at 00_{UTC} for the 16th of July 2010 at 00_{UTC}(48 hours ahead). The two forecasts are assumed to sample truncated normal distributions. The coefficients a_u, b_u, c_u, and d_u for the corresponding day can be seen in table 5.2. They have been estimated from the training period including 84 forecasts from the 2nd of June and the 14th of July 2010 and have been applied to the raw forecast (blue) in order to obtain the calibrated one (red). The ensemble mean has been translated from 8.2 m.s^{-1} to 7.5 m.s^{-1} whereas the corresponding wind speed observation is 7.2 m.s^{-1}. As it is indicated by the highest density of the calibrated distribution, the ensemble spread has been dilated.

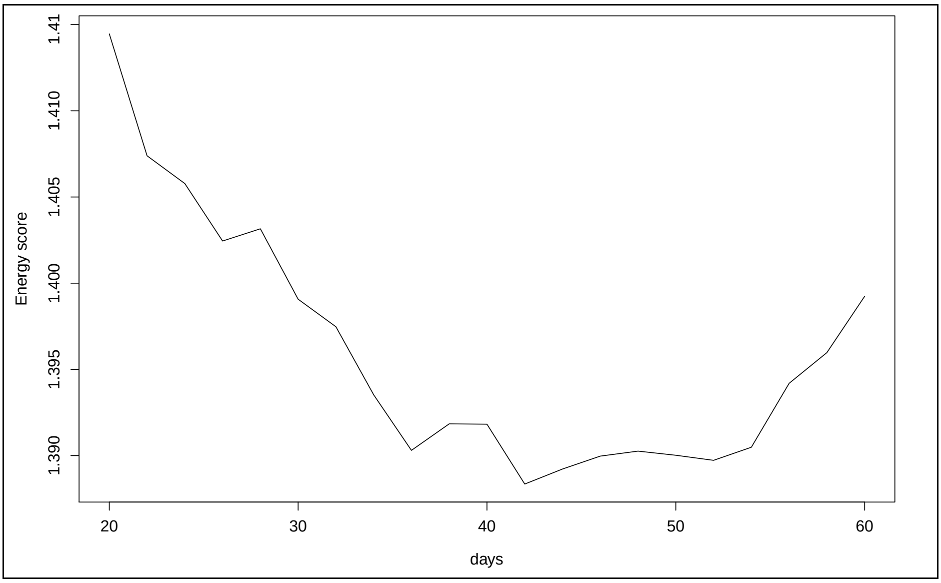


Figure 5.2: Mean Energy score (average on every lead times) of the marginally calibrated 10 m wind speed and significant wave height forecasts as a function of the length of the training period.

a_u	b_u	c_u	d_u
0.24	0.87	0.76	0.84

Table 5.2: Parameters estimates of the univariate calibration CAL^1 of a 48 hours ahead wind speed EPS forecast valid on the 16th of July 2010 at 00_{UTC}.

The first goal of any calibration method is to obtain reliable ensemble forecasts, i.e equally likely to occur ensemble members. As explained in Chapter 3, reliability can be qualitatively assessed by rank histograms and PIT diagrams. Figures 5.4 and 5.5 show a comparison of the univariate rank histograms and PIT diagrams of the different 48 hours ahead 10 m wind speed and significant wave height forecasts aggregated over the year 2010 and 2011.

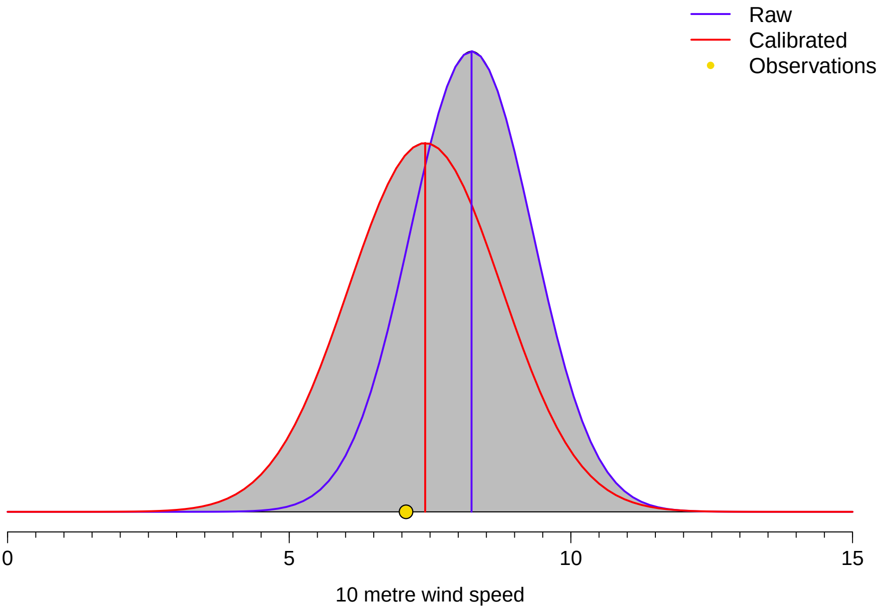
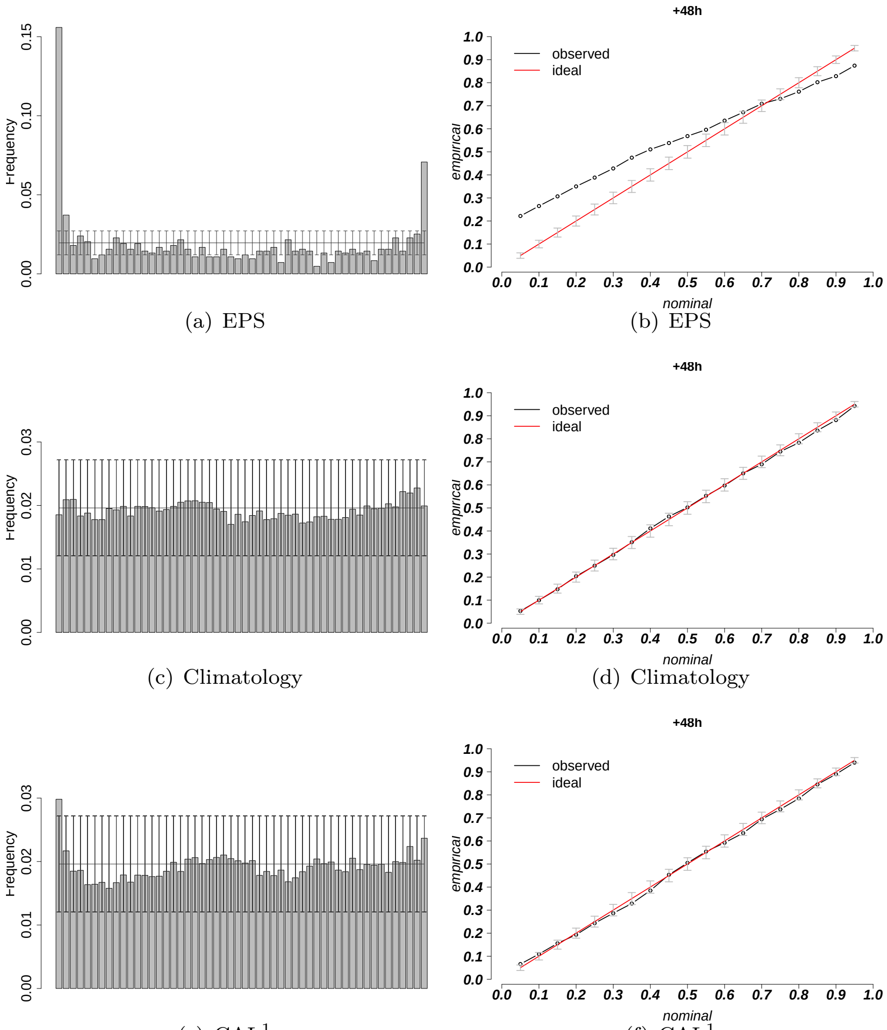


Figure 5.3: Example of univariate calibration of a 48 hours ahead ensemble forecast of 10m wind speed valid on the 16th of July 2010 at 00_{UTC}. The vertical bars represent the predicted means, the curves represent the forecast pdfs, the blue elements correspond to the EPS and the red ones to the CAL[1] forecast. The corresponding observation is symbolised by the yellow point on the x-axis.

While wind speed and wave height raw forecasts are both very underdipsersive, as indicated by the U shape of the respective rank histograms and the small slope of the corresponding PIT diagrams, the other types of forecasts seem calibrated. Indeed other rank histograms are relatively close to uniformity and PIT diagrams close to the diagonal. After calibration, ensemble forecasts provide reliable univariate probabilistic information about the future weather conditions, which means that every ensemble member is equally likely to occur and the predicted quantiles are statistically representative of the uncertainty. Apart from the first bin which is slightly overpopulated, i.e over the top of the consistency interval, every other rank lies into it. The climatology benchmark is perfectly calibrated with every bin being into the consistency interval and a PIT diagram almost indistinguishable from the diagonal. In these figures, we can not clearly compare CAL^1 reliability with the climatology reliability. However, reliability improvements can be summarised and quantitatively assessed for every lead time thanks to the reliability index.

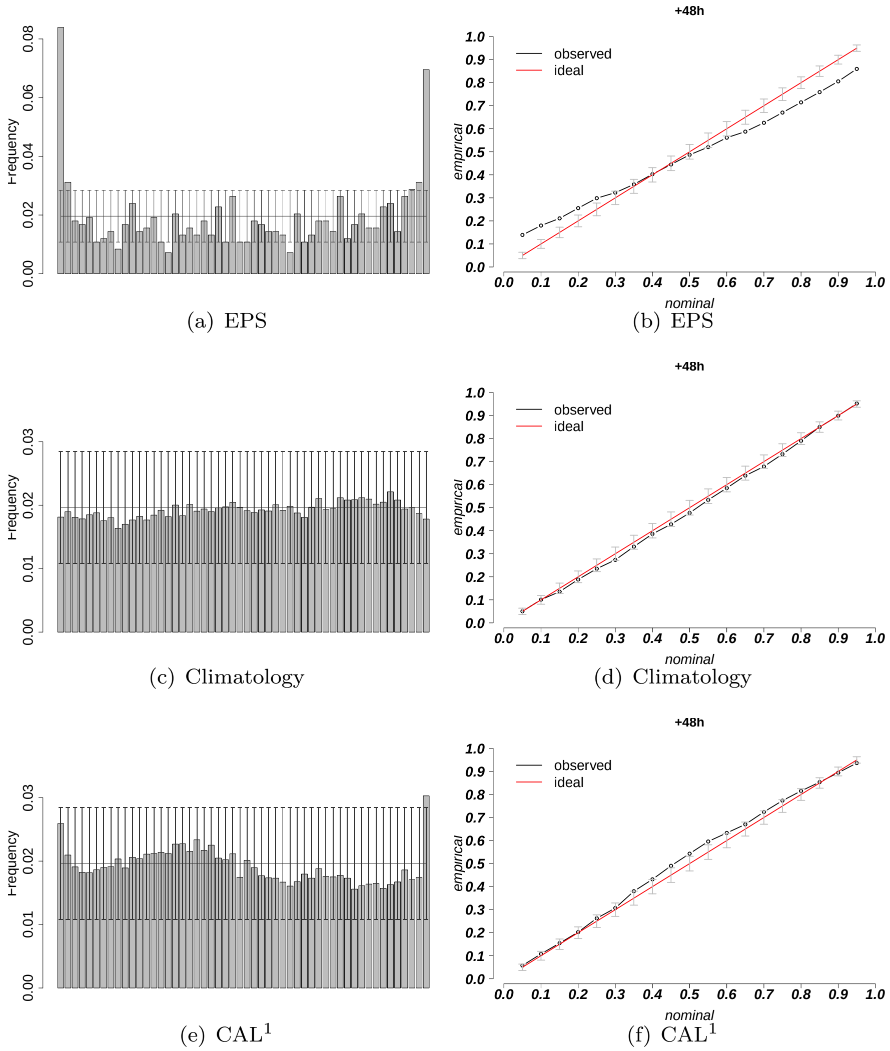


(a) EPS (b) EPS

(c) Climatology (d) Climatology

(e) CAL^1 (f) CAL^1

Figure 5.4: Rank Histograms and PIT diagrams of the 48 hours ahead 10 m wind speed (a) raw forecasts, (b) climatology, and (c) marginally calibrated forecasts



(a) EPS

(b) EPS

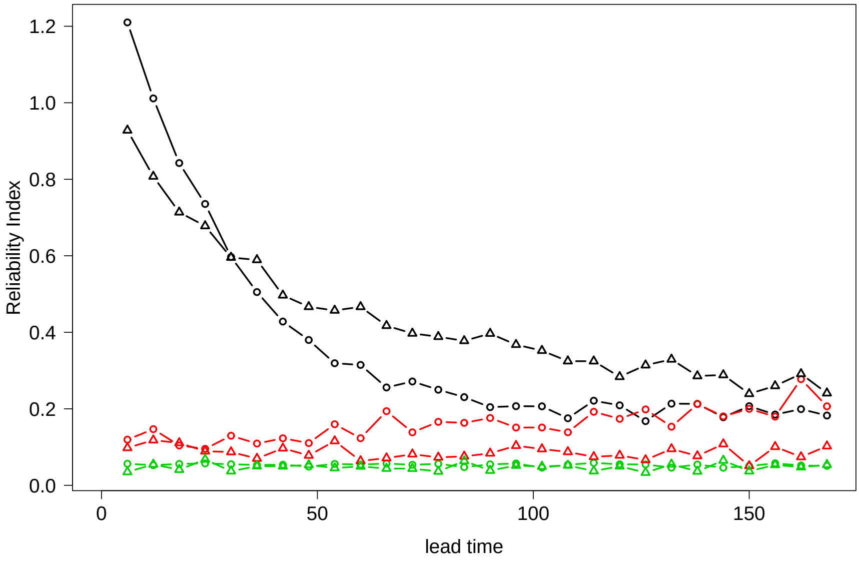
(c) Climatology

(d) Climatology

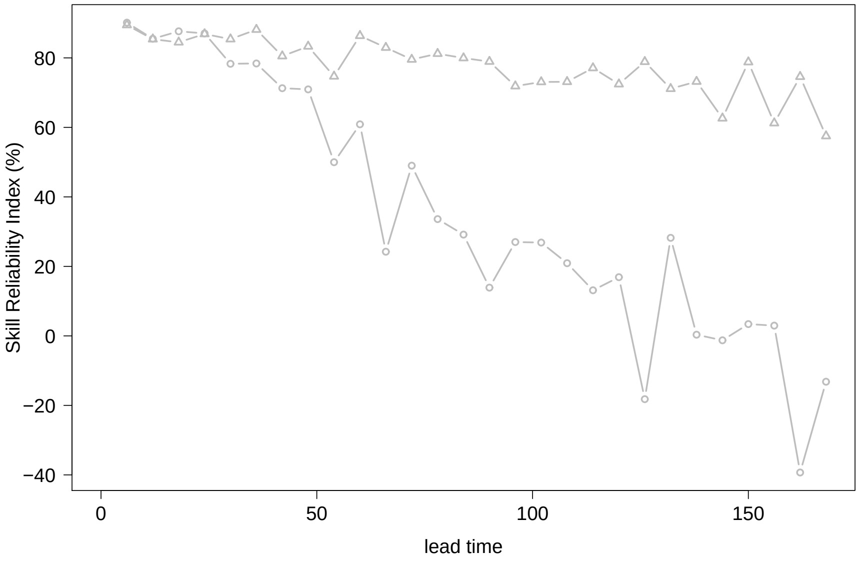
(e) CAL^1

(f) CAL^1

Figure 5.5: Rank Histograms and PIT diagrams of 48 hours ahead significant wave height (a) raw forecasts, (b) climatology, and (b) marginally calibrated forecasts



(a) Reliability Index



(b) Skill Reliability Index against raw forecasts

Figure 5.6: Quantitative reliability assessment of the different univariate forecasts types. (a) Evolution of the reliability index with the predictive horizon of the EPS (black), climatology benchmark (green) and CAL^1 forecast (red) for 10 m wind speed (triangle) and significant wave height (circle). (b) Skill reliability index of the marginally calibrated forecasts CAL^1 against the raw forecasts for 10 m wind speed (triangle) and significant wave height (circle).

Figure 5.6(b) shows the evolution of the reliability index improvements with the prediction horizon for the same types of forecasts than in the previous figures. The climatology benchmark has the best reliability index (< 0.05 for every lead time) of the three forecasts types. Forecast reliability is considerably improved by the univariate calibration method CAL^1 for both variables. Indeed, significant improvements of more than 80% can be noticed for the firsts lead times of both 10 m wind speed and significant wave height forecasts. These improvements remains more or less constant for medium range for 10 m wind speed whereas they decrease much faster for the significant wave height. In can be noticed that raw ensembles forecasts tend toward the climatology for long prediction horizon and are therefore already more or less calibrated.
After calibration, the forecasts are unbiased.

Table 5.3 shows the bias score of the different forecast types for the 48 hours ahead horizon. The bias is similar for every lead times and is therefore only exposed for one lead time (48 hours horizon). The marginally calibrated forecasts CAL^1 performs all types of forecasts. Even if significant wave height raw forecasts are much less biased (0.025 m) than the 10 m wind speed forecasts (0.355 m.s^{-1}), calibration allows bias decrease for both variables. CAL^1 even performs the climatology forecast. Indeed, the climatology benchmark has been computed over less than two years of data and is therefore less accurate.

	Significant Wave Height
EPS	0.025
Climatology	-0.023
CAL^1	0.019

	10 m Wind Speed
Raw forecast	0.355
Climatology	-0.045
CAL^1	-0.018

Table 5.3: Bias of the different forecast types for the 48 hours ahead horizon

While bias only informs about the systematic errors, RMSE assesses errors amplitude giving more weight to large errors. Improvements of the RMSE as a function of the prediction horizon are depicted in the figure 5.7 for the 10 m wind speed and the significant wave height. The skill RMSE of the marginally calibrated forecasts CAL^1 against raw forecasts has been computed over the entire period 2010-2011. Calibrated forecasts predict better ensemble means resulting resulting in improvements of the RMSE. The most significant improvements can be seen for the first lead times where calibration allows a decrease of the RMSE of approximately 5% for both variables. These improvements are more significant for the wave forecasts. It can also be noticed an oscillation in the skill RMSE for both variables. Indeed, since the EPS is issued twice a day and

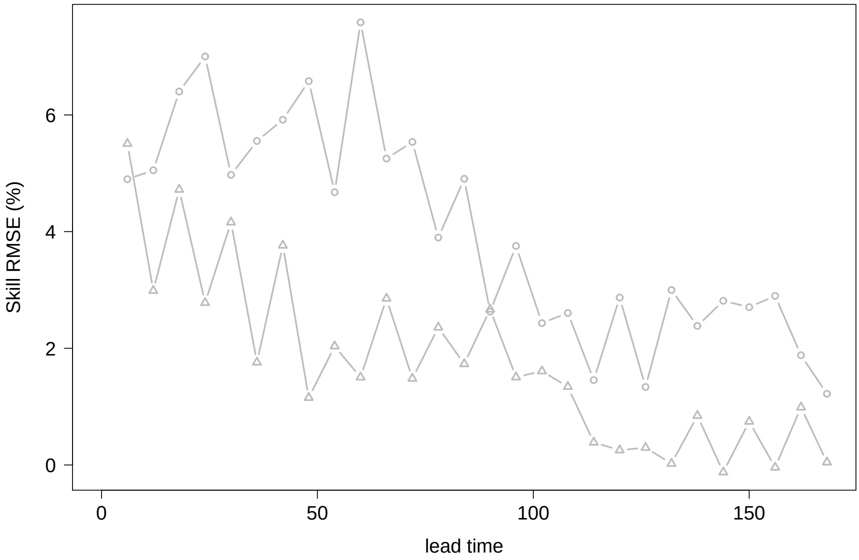


Figure 5.7: Improvement of the RMSE as a function of the lead time computed over the entire period 2010-2011. Comparison is made between marginally calibrated forecasts raw forecasts for 10 m wind speed (triangle) and significant wave height (circle).

has a time resolution of six hours, the 6, 18, 30, 42,..., 162 hours ahead forecasts are only valid at 06 and 18_{UTC} whereas the 12, 24, 36, 48,..., 168 hours ahead forecasts are only valid at 00 and 12_{UTC}. Thus, EPS forecasts might be less accurate for some specific hours of the day which could impact the improvements of calibration method.

A detailed study of the temporal variation of the RMSE is of a great interest for the understanding of the calibration method behaviour for the mean correction.

Figure 5.8 shows the evolution of the ensemble mean error of the EPS and the CAL^1 10 m wind speed forecasts over 2010-2011. The CAL^1 forecasts present a constant accuracy all over the entire period. However, the EPS bias being completely different from 2010 and 2011, the improvements induced by the univariate calibration method are much more significant during the first year than the second. The calibration method is obviously more efficient when the raw forecasts bias is stationary.

Figure 5.9 shows the evolution of the ensemble mean errors of EPS and CAL^1

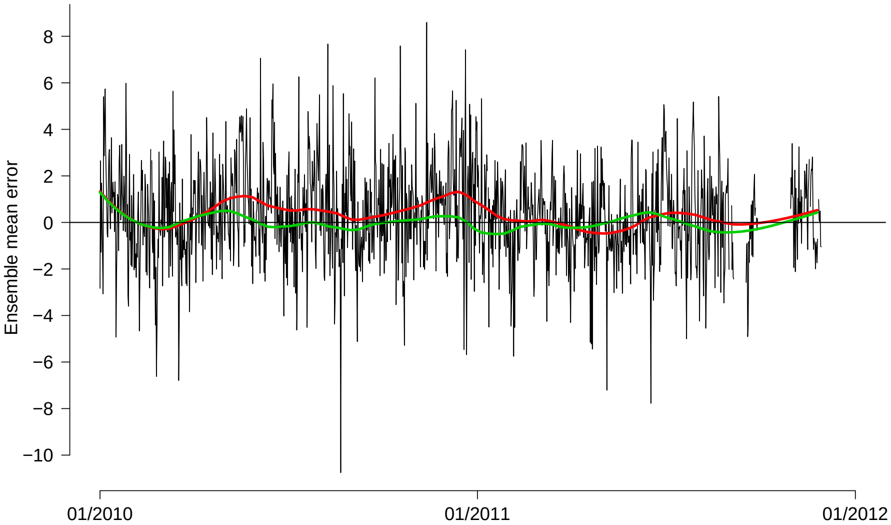


Figure 5.8: Comparison of 10 m wind speed ensemble mean errors over 2010-2011. The black line is the ensemble mean errors of the raw forecasts, the red line is the smoothed ensemble mean errors of the same forecast and the green line is the smoothed ensemble mean errors after marginal calibration.

significant wave height forecasts. It can be seen that the calibration is more efficient than for the wind speed 5.8. Indeed, the EPS bias from one month to the other is much less fluctuating than for the wind speed and is therefore easier the correct.

The CRPS is a score that is affected by both mean and variance correction. In other words, the CRPS is impacted by the entire predicted distribution and therefore allows an overall univariate assessment of the marginal calibration method CAL[1]. Figure 5.10 presents the improvements of the CRPS obtained through marginal calibration for both 10 m wind speed forecasts and significant wave height. The calibration efficiency is much more significant for the CRPS than for other scores. Indeed, after marginal calibration, the CRPS is improved of 16% to almost 20% for the shortest horizon. Even for the 60 hours ahead prediction horizon, improvements of 5% can still be seen for both variables. Calibration method improves the entire distribution prediction, that is calibrated probabilistic forecasts are more representative of the uncertainty (variance cor-

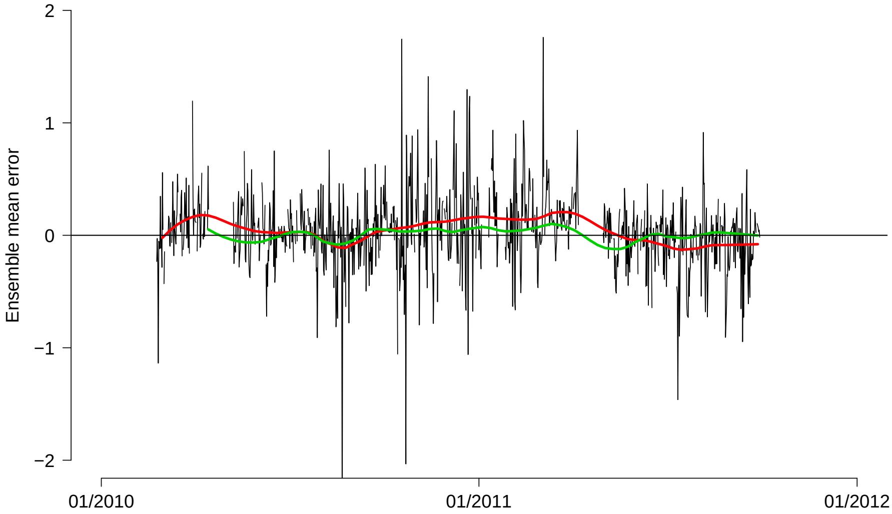
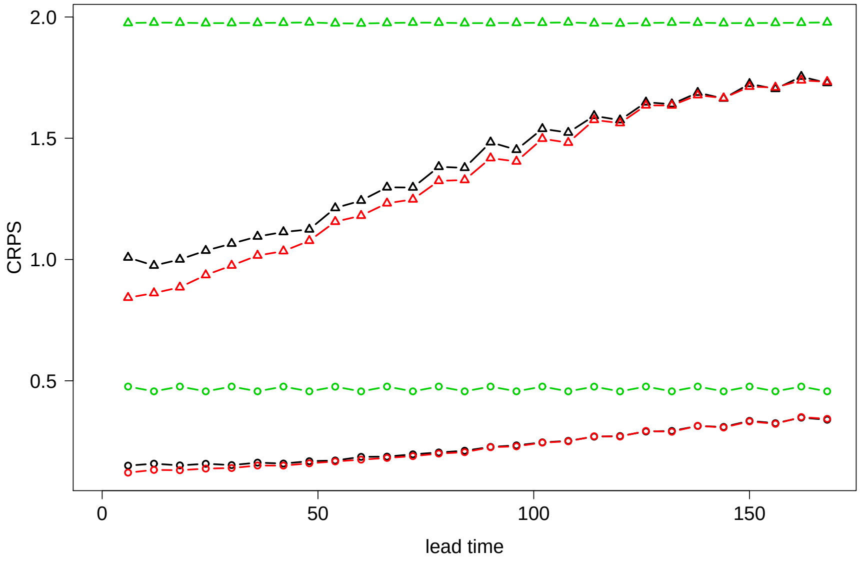


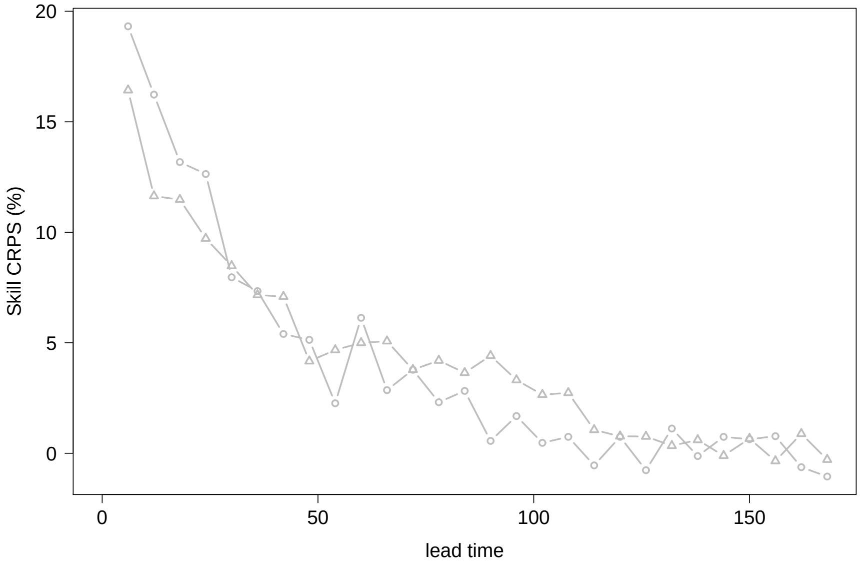
Figure 5.9: Comparison of significant wave height ensemble mean errors over 2010-2011. The black line is the ensemble mean errors of the raw forecasts, the red line is the smoothed ensemble mean errors of the same forecast and the green line is the smoothed ensemble mean errors after marginal calibration.

rection) and closer to the observation (mean correction).

All those results prove that the univariate calibration method CAL^1 is efficient. It provides reliable univariate forecasts while increasing the prediction accuracy of the ensemble mean and the ensemble variance.



(a) CRPS



(b) Skill CRPS against EPS

Figure 5.10: CRPS assessment as a function of the prediction horizon for the 10 m wind speed (triangle) and the significant wave height (circle). (a) CRPS of the EPS (black), Climatology (green) and CAL^1 (red). (b) Skill CRPS against the EPS.

5.3 Bivariate calibration method

Wind speed and wave height are interconnected. The relationship that exists between these two variables is complex, wind impacts wave which impacts back wind speed. An univariate calibration does not that into account that interaction, and marginally correcting those two variables implies the lost of their correlation. This is the reason why it is important to jointly calibrate the 10 m wind speed and the significant wave height forecasts, so the correlation is not lost. The methods described in Chapter 4.2 have been applied on the 48 hours ahead ensemble forecasts of the ECMWF from January 2010 to December 2011. The results are exposed in this section assessing the calibrated forecasts thanks to scores/tools described in Chapter 3.2.

5.3.1 Proof of concept

Firstly, we assess calibrated forecasts CAL^{1+}. Since correcting parameters are identical with the univariate calibration method CAL^{1}, these forecast types share the same univariate properties. However, the predicted dependence of the 10 m wind speed and significant wave height lost during the univariate calibration have been recovered through equation (4.16). Therefore calibrated forecasts CAL^{1+} are supposed to predict more realistic bivariate distributions and therefore to show improvements. This type of forecast calibration is similar to the one proposed by Moller (Möller et al., 2012). However, unlike Möller who used a correlation estimated from the training period, we use the correlation predicted by the EPS for the valid date. Doing so allows more dynamic bivariate distribution patterns able to completely change from one day to an other.

A first example of calibrated forecasts in a bivariate point a view is shown in figure 5.11 valid on the 19th of September 2011 at 12_{UTC}. In this example, the mean correction is not efficient and the spread correction are not significant. However, contrary to CAL^{1}, the CAL^{1+} forecast has allowed the correlation from the raw ensemble to be carried over to the distribution and is therefore more realistic considering the EPS forecasts and the corresponding observation. Indeed, recovering the correlation induces a preferential direction for the bivariate distribution. It not only leads to a sharper distribution but it correctly adjust the distribution when the observations is located in the preferential direction.

Figure 5.12 shows in three dimensions the different calibrated distributions valid

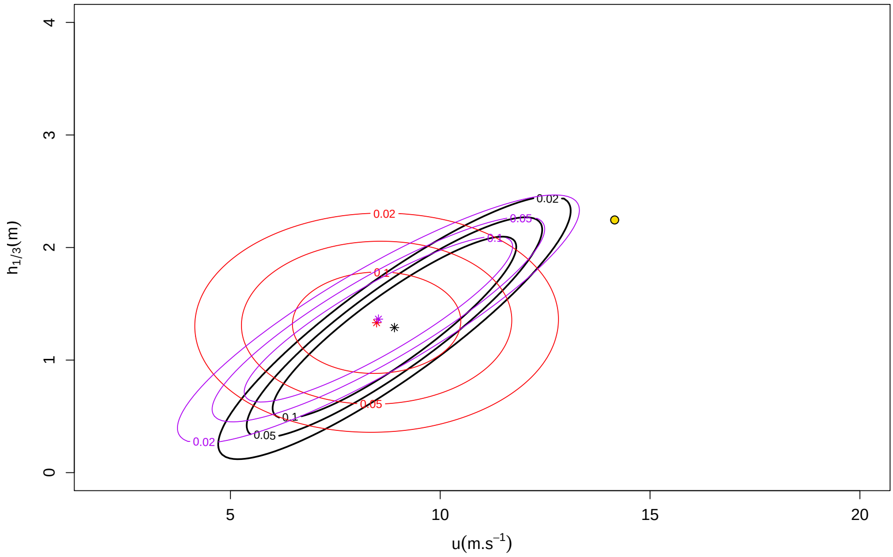
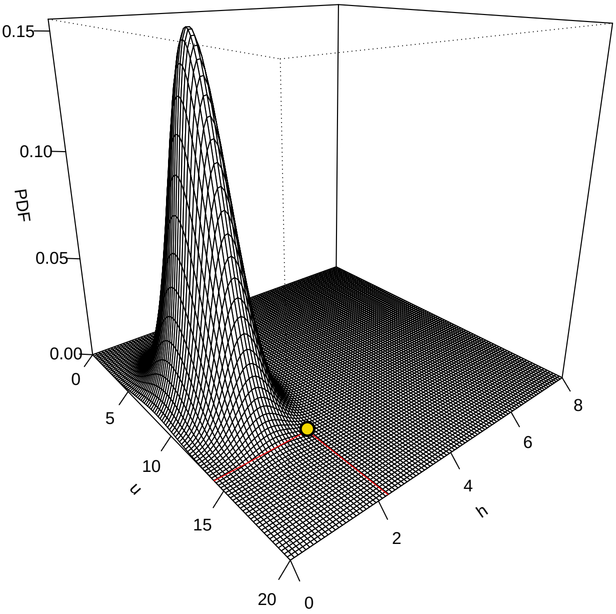
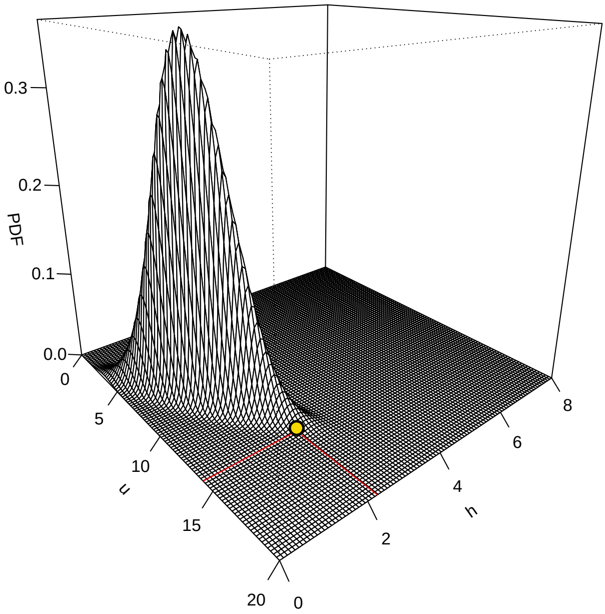


Figure 5.11: Example of 48 hours ahead raw and calibrated forecasts valid on the 19th of September 2011 at 12_{UTC}. The black, red and blue ellipses represent the EPS (black), CAL^1 (red) and CAL^{1+} (purple) distribution contours 0.1, 0.05 and 0.02 if existing. The stars represents the respective predicted mean and the yellow point symbolises the corresponding observation

on the 19th of September 2011 at 12_{UTC}. After marginal calibration the observation is still out of the margins of the distribution. The correlation of the forecasts CAL^{1+} orients the predicted distribution in a preferential direction and better covers the observation, the corresponding density values is therefore not null anymore.

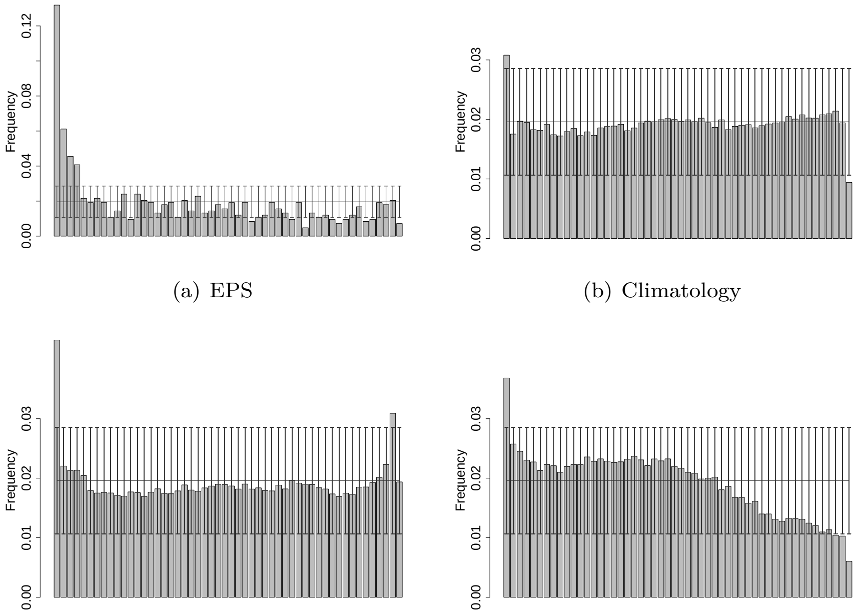


(a) CAl^1



(b) CAL^{1+}

Figure 5.12: Three dimensional example of 48 hours ahead of the (a) CAL^1 and (b) CAL^{1+} predicted distributions valid on the 19th of September 2011 at 12_{UTC}. The yellow point the corresponding observation



(a) EPS

(b) Climatology

(c) CAL^1

(d) CAL^{1+}

Figure 5.13: Multivariate rank histograms of the 48 hours ahead (a) EPS, (b) climatology, (c) CAL^1 and (d) CAL^{1+} forecasts of 10m wind speed and significant wave height.

Figure 5.13 presents the multivariate rank histograms of the EPS, the climatology, CAL^1 and CAL^{1+} forecasts computed over the entire period. Without any surprises, the raw forecast is as underdispersive in a bivariate than in an univariate point of view. The Climatology benchmark is the most reliable forecast, as indicated by a multivariate rank histogram very close the uniformity. The figure shows that the CAL^1 method improves calibration considerably but still presents an underdispersive characteristic. Indeed, the extreme bins of the corresponding multivariate rank histogram are overpopulated. This is a proof that, even with accurate predicted means and variances, an uncorrelated forecast does not cover well enough possible observations pairs. The CAL^{1+} forecast enlarge the tail of the bivariate CAL^1 distribution in a direction depending on the predicted correlation of the EPS. This action results in a less underdispersive multivariate rank histogram than the CAL^1 one. The extreme bins overpopulation are reduced. However the corresponding histogram is still not close to uniformity. Indeed, overpopulation reduction is more important on the last bins than on the first which creates a positive bias tendency as indicated by the negative slope of the multivariate rank histogram with the final bins less filled than

Calibration Method	bRMSE	bMAE	es	Δ	DS
EPS	1.998	1.752	1.179	0.481	0.355
Climatology	3.724	3.594	2.078	0.064	1.468
CAL^1	1.971	1.729	1.110	0.094	0.699
CAL^{1+}	1.973	1.733	1.114	0.229	0.51

Table 5.4: Comparison of bivariate scores of the different calibration methods for the +48h forecasts over the entire period

the others.

Table 5.4 presents the bivariate scores of the different 48 hours ahead forecasts types. The bivariate RMSE and MAE, the Energy score, the multivariate reliability index and the determinant sharpness are exposed for the raw forecasts, the climatology, the marginally calibrated forecasts CAL^1 and finally the EPS-prescribed correlation approach CAL^{1+}. The raw forecasts performs the climatology for all scores expect the reliability index. The marginally calibrated forecasts CAL^1 performs all type of forecasts. This type of forecasts is almost as reliable as the climatology, and has the best bRMSE, bMAE and es. The CAL^{1+} has similar bRMSE and bMAE with the CAL^1. Indeed, these two types of forecasts predict the same two first moments and only differ from the correlation. Thus scores like bRMSE and bMAE that does not take into account the spatial dependence of the bivariate forecasts are similar, slightly differing due to the sampling process. It can be seen that, as suggested by the multivariate rank histogram in figure 5.13, the dependence recovering process implies a decrease of reliability. The energy score is also better for the CAL^1 forecast than for the CAL^{1+}. However, the determinant sharpness of the CAL^{1+} forecasts is lower than for the CAL^1. Even if the CAL^{1+} method seems to provide more realistic and sharper bivariate distributions by conserving the raw predicted spatial pattern, it does not perform the marginal calibration method CAL^1.

We have here the proof that a bivariate calibration approach is needed to jointly calibrate forecasts while conserving the dependence of the two variable.

5.3.2 Joint Calibration

it has been proved that a bivariate calibration approach taking into account the predicted correlation while estimating the correcting parameter is needed. The results of the bivariate approach CAL^2 are exposed.

As it has been done for the univariate calibration method, the first step is to determine the optimal length of the training period. Indeed, since the likelihood technique is not applied on the same distribution than for the univariate calibration method and that the number of parameters to even has been doubled, the appropriate length of the training period for the bivariate approach could be different.

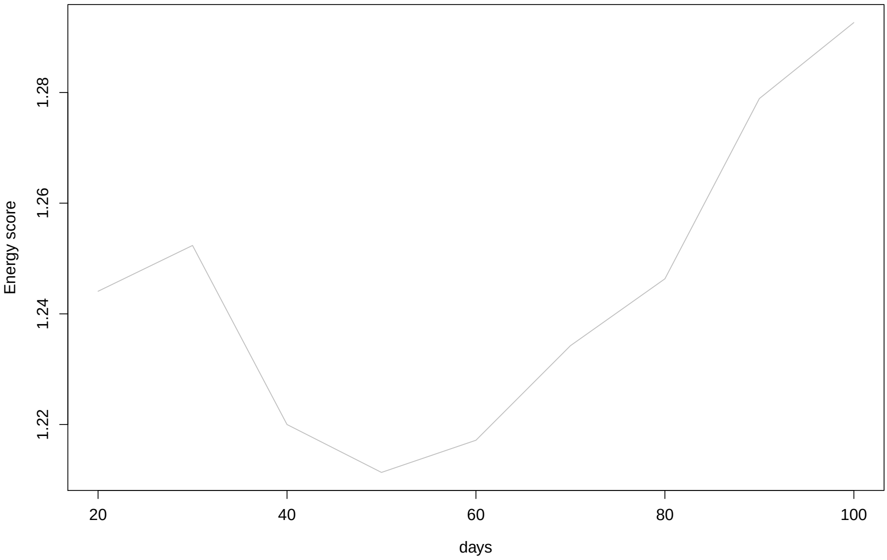


Figure 5.14: Energy score of the 48 hours ahead jointly calibrated forecasts in 2010 as a function of the length of the sliding training period.

Considering the figure 5.14, lengths of the training period for the bivariate approach between 40 and 60 days work well. Out of concern for consistency with the univariate method we choose a 42 days training period. However, both 10 m wind speed and significant wave height observation have to be available at the same time to be introduced in the training period, therefore the first and the last day of the training data could be separated by more than 100 days.

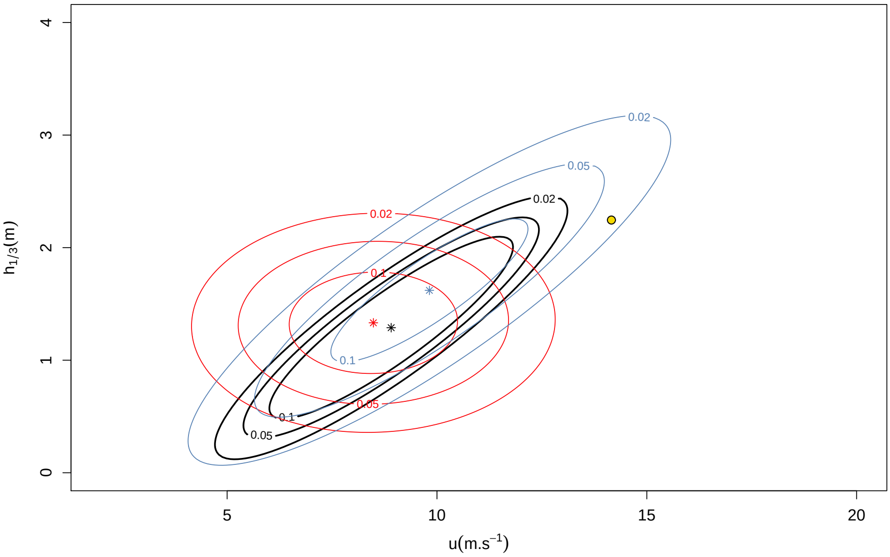


Figure 5.15: Example of 48 hours ahead raw and calibrated forecasts valid on the 19th of September 2011 at 12_{UTC}. The black, red and blue ellipses represent the EPS (black), CAL^1 (red) and CAL^2 (blue) distribution contours 0.1, 0.05 and 0.02 if existing. Stars symbol represent the respective predicted mean and the yellow point symbolises the corresponding observation

Figure 5.15 shows an example of 48 hours ahead marginally and jointly calibrated ensemble forecast valid on the 19th of September 2011 at 12_{UTC} (same date than in the figure 5.11). The univariate calibration does not take the correlation into account and therefore predicts an uncorrelated ensemble forecast for the valid date. The bivariate approach not only provides more realistic distribution while retaining the positive correlation of the raw ensemble, but also impacts the correcting parameter estimation which permits to cover the observation.

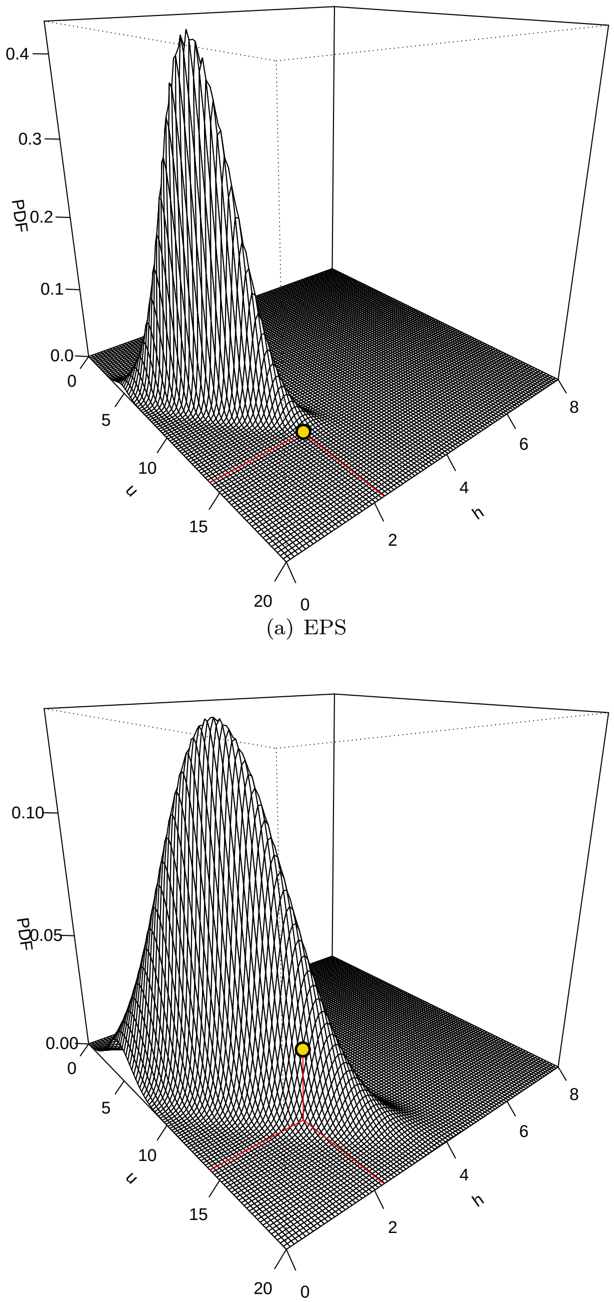
Table 5.5 presents the different parameter estimates on the the 19th of September 2011 at 12_{UTC} for both univariate and bivariate calibration methods. Differences are more significant on the spread correcting parameters than on the others. The 10 m wind speed parameter estimates differ much more from the univariate to the bivariate method than the significant wave heigh ones. The

Method	a_u	b_u	c_u	d_u	a_h	b_h	c_h	d_h
CAL^1	0.28	0.92	3.24	0.35	0.09	0.98	0.04	0.79
CAL^2	-0.49	1.12	8.09	0.71	-0.03	1.12	0.01	3.62

Table 5.5: Example of parameters estimates of the bivariate and univariate calibration method for 48 hours ahead the wind speed (a_u,b_u,c_u,d_u) and the wave height forecasts (a_h,b_h,c_h,d_h) valid on the 19th of September 2011 at 12_{UTC} (issued the 17th of September 2011 at 12_{UTC}).

spread correcting parameters are strongly amplified for both variables in the bivariate approach resulting in the wider distribution seen in figure 5.15.

The underdispersive characteristic of the EPS and the spread correction is much more obvious by visualizing the three-dimensional raw and jointly calibrated predicted distributions as it is presented in the figure 5.16. Indeed, the observation on the valid date, symbolised by the yellow point, falls outside of the margins of the raw predicted distribution with a corresponding density that can be considered as null. This forecast is considered as bad either because of an ensemble mean too far from the observation or because of a too small predicted spread. After the bivariate calibration the predictive distribution is wider and the predicted mean is closer to the observation. Thus, the jointly calibrated distribution better covers the observation, the corresponding density is not null anymore. Plus, the correlation is retained after calibration thanks to the bivariate approach.



(a) EPS

(b) CAL^2

Figure 5.16: Example of Three dimensional 48 hours ahead of (a) raw and (b) jointly calibrated forecast distributions valid on the 19th of September 2011 at 12_{UTC}. The yellow point the corresponding observation

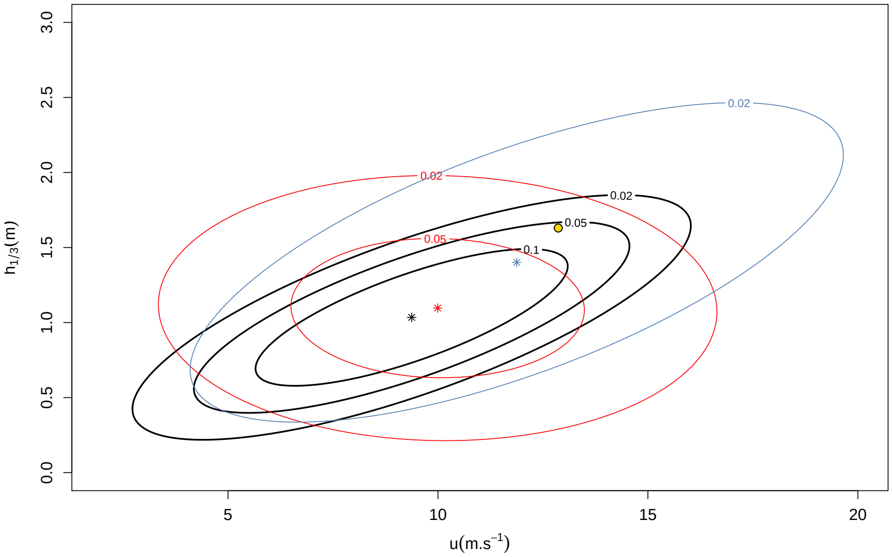


Figure 5.17: Example of 48 hours ahead raw and calibrated forecasts valid on the 18th of June 2011 at 12_{UTC}. The black, red and blue ellipses represent the EPS (black), CAL^1 (red) and CAL^2 (blue) distribution contours 0.1, 0.05 and 0.02 if existing. Stars symbols represent the respective predicted mean and the yellow point symbolises the corresponding observation

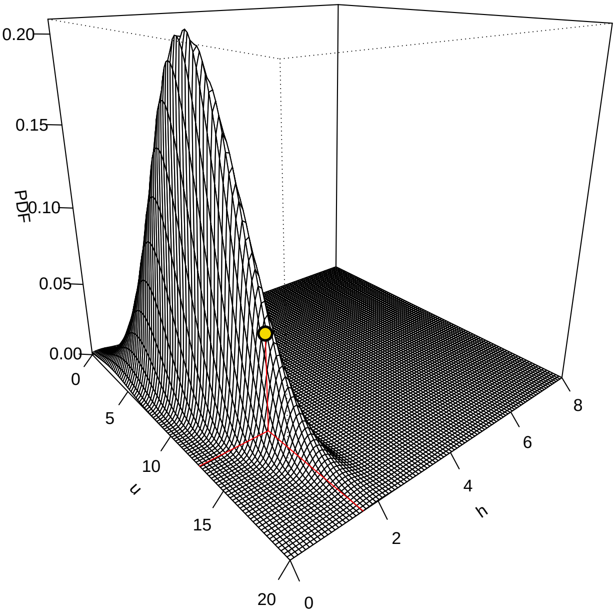
An other example of calibration valid on the 18th of June 2011 at 12_{UTC} is shown in the figure 5.17. For this example, the jointly calibrated distribution is much wider than the marginally calibrated distribution. Indeed, even though the jointly calibrated distribution predicts a better mean vector than the marginally calibrated one, the spread is extremely amplified. Being dilated, the maximum density value of the CAL^2 distribution is approximately 0.02 whereas it exceed 0.05 for the CAL^1 distribution. Table 5.6 shows the parameters estimates valid on the 18th of June 2011 at 12_{UTC} corresponding to the calibrated distribution presented in the figure 5.17. It confirms the fact that the spread correcting parameters are extremely overestimated with parameters c_u and d_u issued from the bivariate calibration method approximately five times greater than the one from the univariate calibration method. The parameter d_h from CAL^2 is also greater than d_h from CAL^1. Whereas the variance of the EPS 10 m wind speed components is 9.6 m.s^{-1}, after marginal calibration it is equal to 17.5 m.s^{-1} and 35.8 m.s^{-1} after bivariate estimation. The variance of the 10 m wind speed component of the CAL^2 forecast is even greater than the one of the climatology

Method	a_u	b_u	c_u	d_u	a_h	b_h	c_h	d_h
CAL^1	0.09	1.03	0.65	1.85	0.07	0.97	0.00	2.42
CAL^2	-0.49	1.09	3.28	5.20	0.02	1.07	0.00	7.53

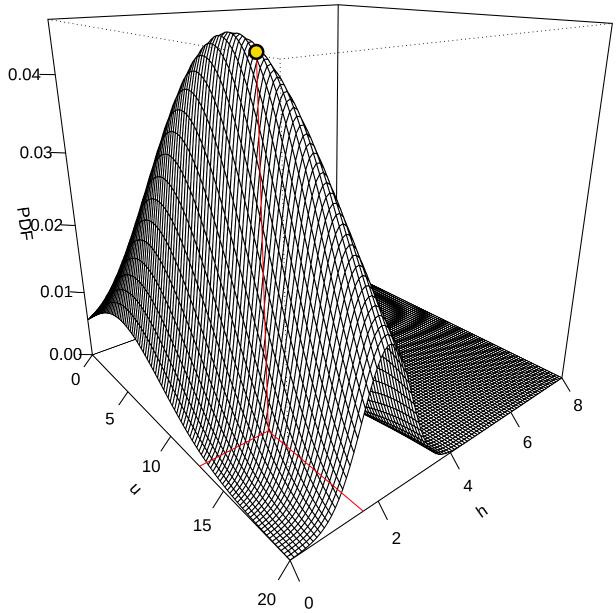
Table 5.6: Example of parameters estimates of the bivariate and univariate calibration method for 48 hours ahead the wind speed (a_u,b_u,c_u,d_u) and the wave height forecasts (a_h,b_h,c_h,d_h) valid on the 18th of June 2011 at 12_{UTC} (issued the 16th of September 2011 at 12_{UTC}).

forecast. In consideration of the previous example, it seems that the maximum likelihood estimation technique is not as efficient for the bivariate approach as for the univariate.

The corresponding EPS and CAL^2 distributions are illustrated in three-dimension in figure 5.18. The spread parameters overestimation can clearly be seen. Indeed, even if the mean correction is correct leading to a very well predicted mean, the distribution is way too wide, tails in the wind speed components goes even beyond 20m.s^{-1}. The maximum of the pdf is approximately 0.04 which is very low compared to the pdf of the EPS (0.2).



(a) EPS



(b) CAL^2

Figure 5.18: Example of three diemnsional 48 hours ahead (a) raw and (b) jointly calibrated forecasts distribution valid on the 18th of June 2011 2011 at 12_{UTC}. The yellow point the corresponding observation

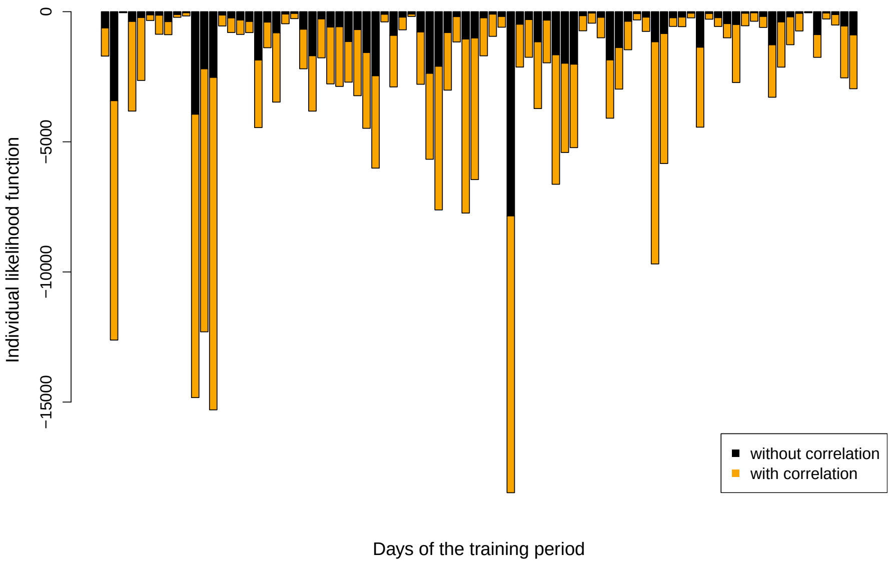


Figure 5.19: Truncated bivariate normal logarithm density values of the different observations introduced in the training period for the bivariate calibration valid on the 18th of June 2011 at 12_{UTC}. The black bars are valid when the correlation is null whereas the orange bars are valid for a correlation equal to the EPS correlation.

Figure 5.19 shows logarithm density values (symbolised by bars) corresponding to the pairs observation - predicted distribution $(x_i, \hat{f}_i)$ present in the training period for the bivariate calibration valid on the 18th of June 2011 at 12_{UTC}. Logarithm density are identical to individual log-likelihood function and therefore inform about the respective contribution of the different pairs for the parameter estimation on the valid date. The black bars are the density values (equation (4.16)) with a correlation equal to zero and the orange bars are density values with a correlation equal to the predicted correlation. It can be noticed that taking into account the correlation strongly impacts the density values. Indeed, the values range is strongly increased with maximum at -20000 instead of approximately -8000. Most of the density values are increased apart from some events like the third from the left or the sixth from the right where it is decreased. It seems that taking into account the predicted correlation gives more weight on outliers resulting in an overestimation of the spread correcting parameters compared to the univariate approach (see Table 5.5).

Figure 5.20 shows more precisely one of the outliers seen in the figure 5.19, corresponding to the 48 hours ahead forecasts valid on the 8th of May 2011 at

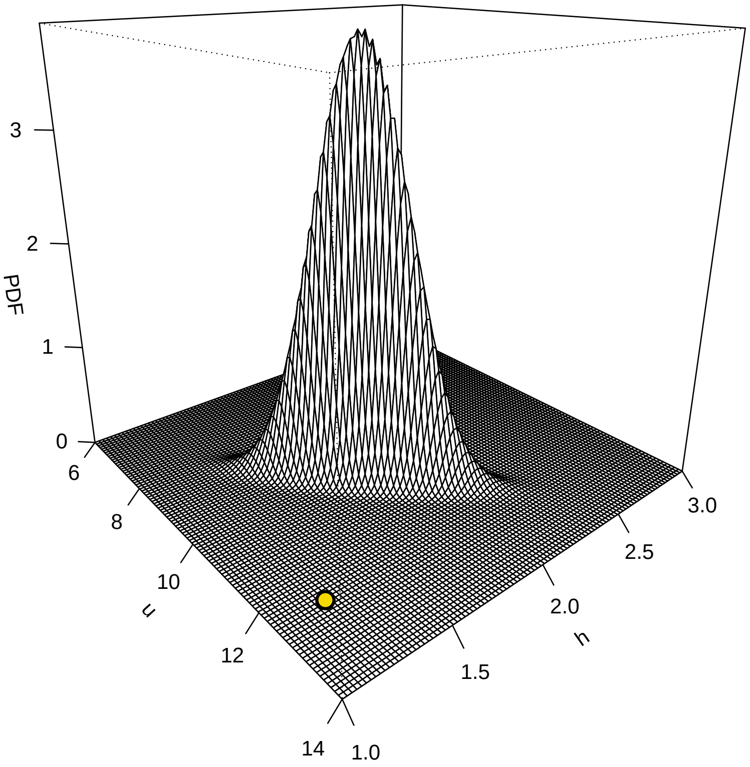


Figure 5.20: Outlier observation of the 8th of May 2011 at 12_{UTC} introduced in the training period for the calibration valid on the 18th of June 2011 at 12_{UTC} with the corresponding raw forecasts with the predicted correlation (black) and with a correlation equal to zero (grey)

12_{UTC}. The corresponding predictive raw bivariate distribution which a non null correlation is represented in three dimensions and the corresponding observation is symbolised by the yellow point. The prediction for that day can be considered as bad, that is the predicted mean is far from the observation and the predictive distribution is way too sharp. Thus, the individual log-likelihood function of this observations is strongly negative. The fact that for this event, the observation lies on the orthogonal direction of the bivariate distribution preferential direction strongly impacts the respective log-likelihood function, that is makes it much lower. Indeed, the figure 5.19 shows that the individual log-likelihood value of this observation (thirteenth from the left of the figure) is approximately -15000 whereas it is around -2500 when the predicted correlation is considered

as zero.
Here we want to point out an issue encountered by the bivariate calibration method. An observation is considered as an outlier when it falls far from the ensemble distribution. However, taking the predicted correlation into account as it is done by the bivariate calibration method proposed here tends to sharper predicted distributions and can amplify the outlier characteristic of an observation and therefore enhance his weight during the parameter estimation process.

Figure 5.21 shows the evolution of the estimated parameters a_u, b_u, c_u, d_u, a_h, b_h, c_h and d_h for the different 48 hours ahead wind speed and the significant wave height forecast calibration methods over the entire period from January 2010 to December 2011. The marginally estimated parameters are represented in black whereas the jointly estimated parameters are represented in red. We can see that the mean correcting parameters a and b evolve anti-symmetrically, that is a becomes larger when b becomes smaller b and vice versa. We can notice that in general, corrections are smaller for the significant wave height forecasts than for the 10m wind speed, a_h amplitudes is approximatively 0.4 whereas a_u take values from -3 to +1. It is even stronger for the variance correcting parameters c_u and c_h. Indeed, c_h's order of magnitude is more than twenty times smaller than for c_u. The parameters a_u and b_u tend to be closer to 0 and 1 during the year 2011. Indeed, as it can be noticed in figure 5.8, the bias of the year 2011 is very close to zero contrary to the year 2010, thus the correction of the location parameter is much less important during the second year data.
Parameters estimated from the bivariate and the univariate calibration method mainly differ from the spread parameters. Parameters a_u,b_u, a_h,and b_h are very similar, their difference might only be due to the difference of the training period. Indeed, we remind that, unlike for the univariate calibration method, both 10 m wind speed and significant wave height have to available to be introduced in the training period of the bivariate calibration method. Parameters c_u,d_u, c_h,and d_h are very different from one method to the other. The one estimated through the bivariate calibration method are always greater than the others and seem to be strongly overestimated at some point. Whatever the parameter, short oscillations can be seen, result of the small length of the training period allowing a faster adaptation to forecast errors variations.

Figure 5.22 compares once again the marginally estimated parameters (represented in black) with parameters estimated through a bivariate approach while fixing the correlation to zero (represented in green). The two method present similar results with slight differences mainly due to the training period used for the estimation. This figure confirms the fact that the overestimation of the spread correcting parameters is not due to the use of a bivariate distribution but essentially to the fact that a non null correlation is provided. Indeed, taking into account the correlation tends to sharper distributions that might have to be strongly dilated to cover observations lying on the orthogonal direction to the one impose by the correlation prescribed by the EPS.

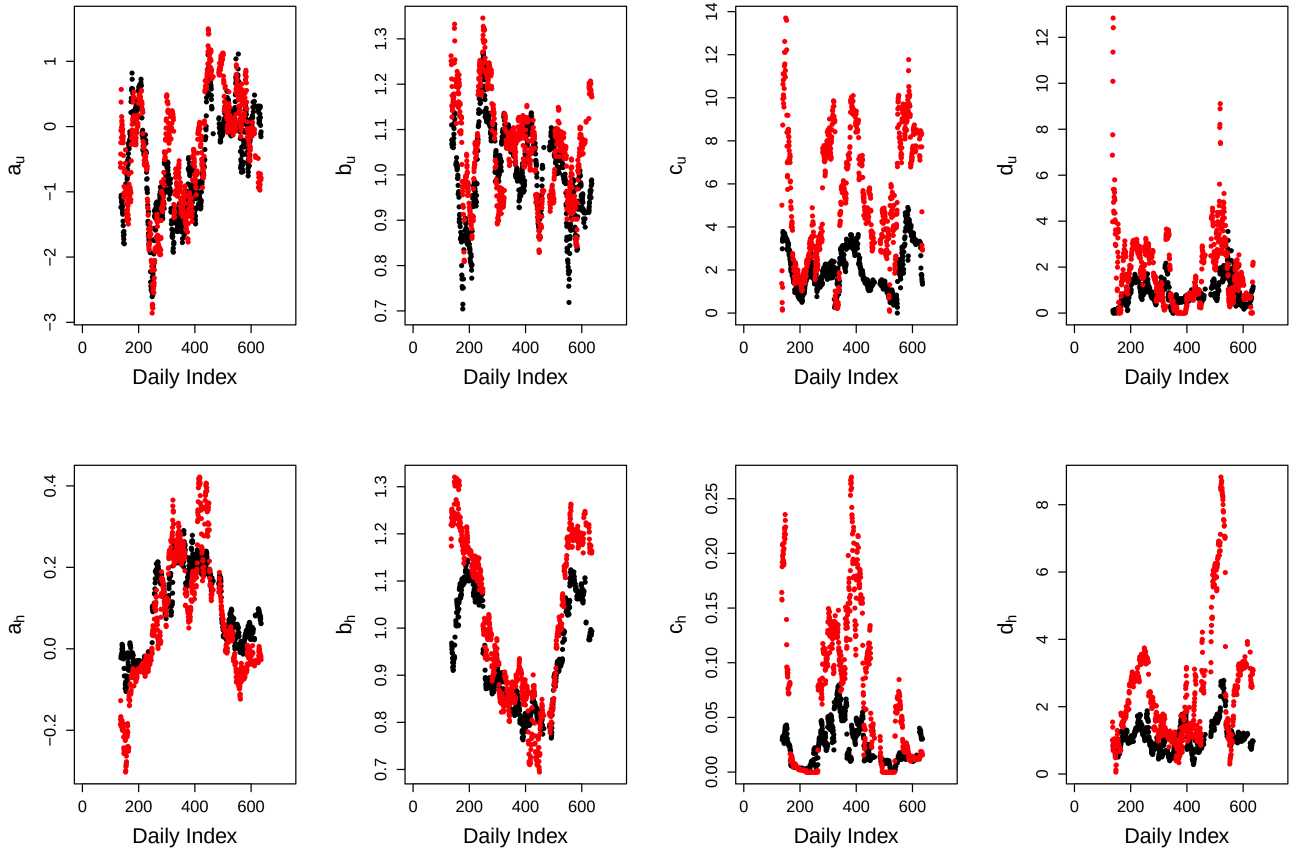


Figure 5.21: Parameter estimates a_u,b_u,c_u and d_u (from left to right) for the 10m wind speed (top) and a_h,b_h,c_h and d_h for the significant wave height (bottom) for the 48 hours ahead univariate calibration CAL^1 (black) and bivariate calibration CAL^2 (red) from January 2010 to December 2011

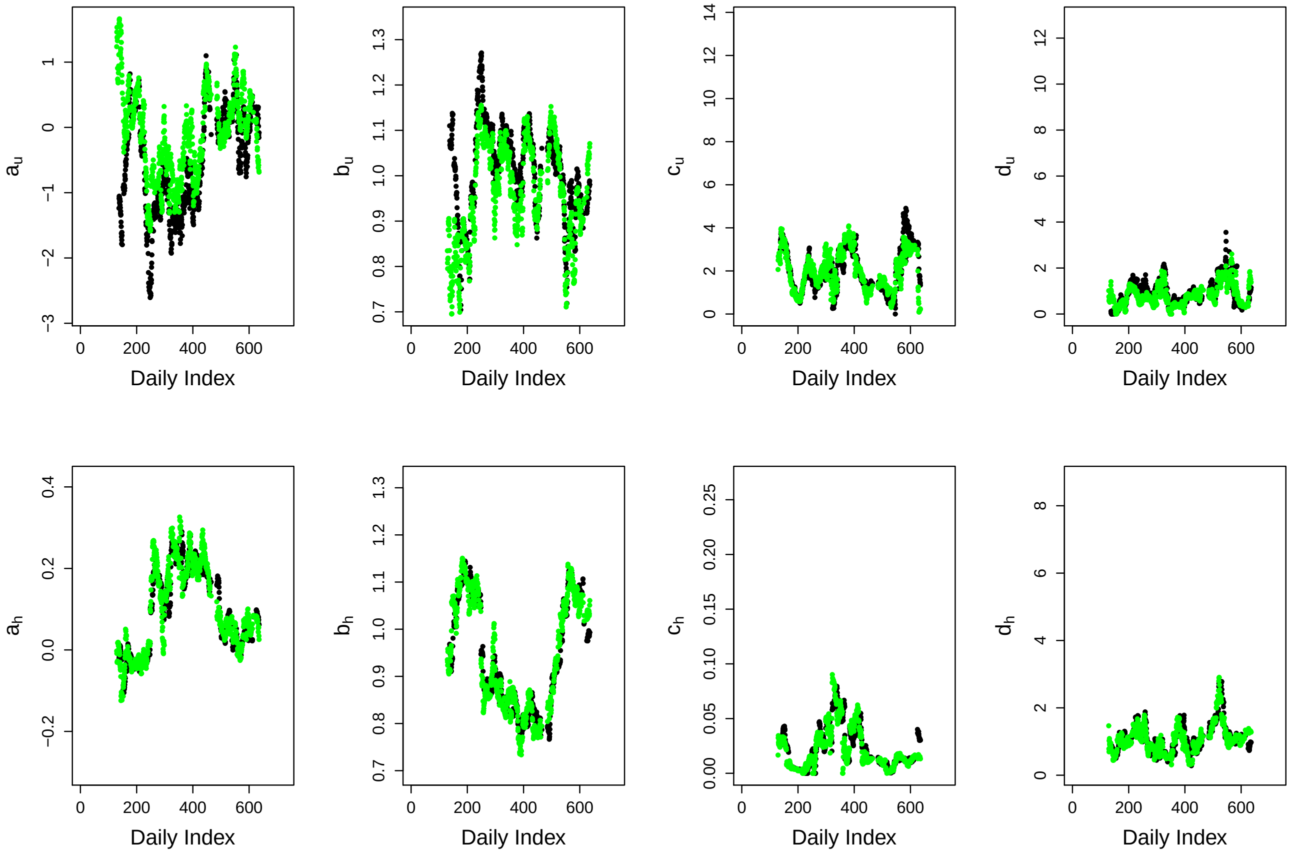


Figure 5.22: Parameter estimates a_u,b_u,c_u and d_u (from left to right) for the 10m wind speed (top) and a_h,b_h,c_h and d_h for the significant wave height (bottom) for the 48 hours ahead univariate calibration (black) and bivariate calibration with a correlation fixed to zero (green) from January 2010 to December 2011

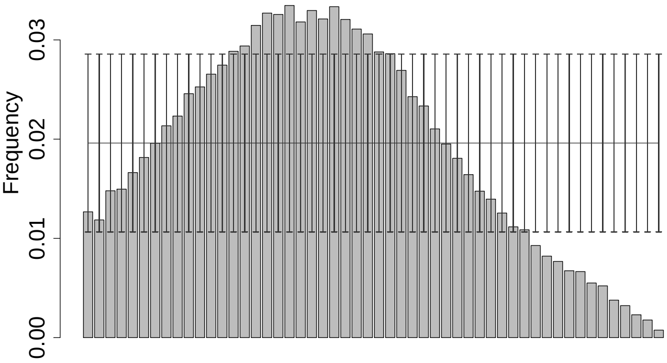


Figure 5.23: Multivariate rank histogram of the 48 hours ahead CAL^2 forecasts of 10m wind speed and significant wave height.

Figures 5.23 shows the multivariate rank histograms of the jointly calibrated forecasts CAL^2. As it has been noticed previously with the example given in the figure 5.15, the jointly calibrated forecasts CAL^2 are clearly overdispersive. This is the result of overestimations of the spread correcting parameters c_u, d_u, c_h and d_h.

Calibration Method	bRMSE	bMAE	es	Δ	DS
EPS	1.998	1.752	1.179	0.481	0.355
Climatology	3.724	3.594	2.078	0.064	1.468
CAL^1	1.971	1.729	1.110	0.094	0.699
CAL^{1+}	1.973	1.733	1.114	0.229	0.51
CAL^2	2.116	1.864	1.212	0.468	0.768

Table 5.7: Comparison of bivariate scores of the different calibration methods for the +48h forecasts over the entire period

Table 5.7 presents bivariate scores of every forecasts types studied in this thesis for the 48 hours prediction horizon. Since results of the first forecast types have already be discussed previously, we only focus here on the bivariate calibration approach CAL^2. Obviously, the bivariate approach jointly calibrating the forecasts CAL^2 presents bad results. Since forecasts sample a truncated bivariate normal distribution, even if the mean correcting parameters are similar to the CAL^1 method, the overestimation of the spread parameters impacts the ensemble mean. Indeed, the mean of the truncated normal distribution is affected by the variance. (see equation (4.4)). Thus, the bRMSE of the CAL^2 forecasts is even worse than for the raw forecasts. It is also true for the bMAE and the energy score. Even if the Reliability index is slightly improve compared to the EPS, the forecasts are strongly overdispersive and present the less sharp calibrated forecasts.

Unfortunately, even if the bivariate approach is promising, the maximum likelihood technique appears to be too sensitive too outliers which is the reason why the method is therefore not efficient.

Chapter 6

Discussion

As explained in the introduction, wind and wave ensemble forecasts are of a great interest for a number of decision-making problems. They inform about the possible future states of the atmosphere and are of substantial economic value. However, ensemble forecasts tend to be uncalibrated, that is biased and underdispersive. Statistical post-processing methods are used to improve predictive performance while providing reliable probabilistic forecasts. So far these calibration methods mainly deal with univariate ensemble forecasts and therefore do not take into account any possible correlation of two-dimensional (or more) forecasts. We proposed a bivariate approach jointly calibrating 10 m wind speed and significant wave height ensemble forecasts so that essential bivariate characteristics can be captured.

Empirical analysis showed that 10 m wind speed and significant wave height could be modelled by a truncated bivariate normal distribution with a cut-off at zero for both variables. The mean vector and the variances of the underlying normal distribution are respectively assumed to be a linear function of the predicted mean vector and variances. The optimal correcting parameters are simultaneously estimated by maximum likelihood estimation on a sliding window of 42 days. A method using the raw predicted correlation to recover the dependence lost during the marginal calibration is also tested. Bivariate distribution provided by this method share the same univariate properties than the marginally calibrated distribution, though they additionally have an informative correlation.

The univariate calibration method CAL^1 has proved to be the most efficient since considerably increasing reliability of raw forecasts while reducing bRMSE of almost 6% and the energy score of approximately 20% for the first prediction horizons. However this technique does not take into account the existing relationship between the two variables and therefore predicts uncorrelated distributions. The EPS-prescribed correlation approach CAL^{1+} does not present better improvements than the univariate approach as well. Even if calibrated forecasts seem more realistic and are sharper, forecasts are less reliable and the energy score is worsened. These results prove that a bivariate approach jointly calibrating forecasts is needed. Yet, the proposed bivariate calibration method CAL^2 does not appear to be efficient. Indeed, the likelihood technique is too sensitive to outliers when the correlation is taken into account and yields an overestimation of the spread correcting parameters. Calibrated forecasts are therefore overdispersive resulting in worse bivariate scores. Outliers falling on the orthogonal direction as the one imposed by the correlation are the cause of the spread parameters overestimation.

We have seen that enlarging the training period is not a sufficient solution to solve the overestimation issue. Therefore, two options can be envisaged: the first one would be to model 10 m wind speed and significant wave height with a distribution having heavier tails than the truncated bivariate normal distribution. The second option would be to use of a robust maximum likelihood technique. These two perspectives could permit to reduce the problem introduced by these so-called outliers.

Furthermore in this thesis, it was always assumed that the predicted correlation of the EPS were exact. Though, the correlation might be incorrect and even show systematic errors. Once the maximum likelihood technique sensitivity issue solved, it would be interesting to investigate a possible correction of the raw predicted correlation. A linear correction had been envisaged,

$$\rho_{t+k} = e + f\hat{\rho}_{t+k|t}$$

The parameters e and f would be estimated during the training period. However, since the correlation parameter $\rho \in \{-1;1\}$, these parameters have then to be constrained so the corrected correlation would also be defined on the same interval. In order to ensure that condition, we could write,

$$\rho_{t+k} = \cos\varepsilon + (1 - |\cos\varepsilon|)\cos\zeta\hat{\rho}_{t+k|t}$$

Appendix A

List of Notations

$\boldsymbol{x}$	Observation vector
x_u	Observed 10 m wind speed
x_h	Observed significant wave height
μ	true mean
σ^2	true variance
ρ	true correlation
$y(j)$	j^{th} ensemble member
$\hat{f}$	predictive probability density function
$\hat{F}$	predictive cumulative density function
$\hat{\boldsymbol{\mu}}$	predictive mean vector
$\hat{\mu}_u$	predictive mean 10 m wind speed
$\hat{\mu}_h$	predictive mean significant wave height
$\bar{y}_u$	ensemble mean 10 m wind speed
$\bar{y}_h$	ensemble mean significant wave height
$\hat{\boldsymbol{\Sigma}}$	predictive variance-covariance matrix
$\hat{\sigma}_u^2$	predictive variance for 10 m wind speed
$\hat{\sigma}_h^2$	predictive variance for significant wave height
$\hat{s}_u^2$	ensemble variance for 10 m wind speed
$\hat{s}_u^2$	ensemble variance for significant wave height
$\tilde{y}_u$	ensemble median for 10 m wind speed
$\tilde{y}_h$	ensemble median for significant wave height
M	number of ensemble members
n	length of the training period
t	issued data of the forecast
$t+k$	valid date of the observation/forecast
k	lead time
$\mathbb{I}(condition)$	Indicator function equal to 1 if condition is true 0 otherwise

Bibliography

Baars, J. (2005). Observations qc summary page. `http://www.atmos.washington.edu/mm5rt/qc_obs/qc_obs_stats.html`.

Bidlot, J. and Holt, M. (1999). Numerical wave modelling at operational weather centres. *Coastal Engineering*, 37:409–429.

Bröcker, J. and Smith, L. (2008). From ensemble forecasts to predicitve distribution functions. *Tellus A*, 60:663–678.

Charnock, H. (1995). Wind stress on a water surface. *Quarterly Journal of the Royal Meteorological Society*, 81:639–640.

Delle Monache, L., Hacker, J., Zhou, Y., Deng, X., and Stull, R. (2006). Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research*, 111:1–15.

Gneiting, T. (2011). Quantiles as otpimal point forecasts. *International Journal of Forecasting*, 27:197–207.

Gneiting, T., Balabdaoui, F., and Raftery, A. (2007). Probabilistic forecasts, calibration and sharpness. *Quarterly Journal of the Royal Meteorological Society*, 69:243–268.

Gneiting, T. and Raftery, A. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102:359–378.

Gneiting, T., Raftery, A., Westveld III, A., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensembe model output statistics and minimum crps estimation. *Monthly Weather Review*, 133:1098–1118.

Gneiting, T., Stanberry, L., Grimit, E., Held, L., and Johnson, N. (2008). Assessing probabilistic forecasts of multivariate quantites, with an application to ensemble predictions of surface winds. Technical Report 537, Department of Statistics, University of Washington.

Hinnenthal, J. (2008). *Robust Pareto – Optimum Routing of Ships utilizing Deterministic and Ensemble Weather Forecasts*. PhD thesis, Technical University Berlin.

Horrace, W. (2005). Some results on the mutlivariate truncated normal distribution. *Journal of Mutlivariate Analysis*, 94:209–201.

Mathiesen, B., Lund, H., and Karlsson, K. (2009). The ida climate plan 2050, background report. Technical report, Aalborg University,RIS∅-DTU.

Möller, A., Lenkoski, A., and Thorarinsdottir, T. (2012). Multivariate probabilistic forecasting using bayesian model averaging and copulas. *arXiv:1202.3956v1*.

Murphy, A. (1993). What is a good forecast? an essay on the nature of godness in weather forecasting. *Monthly Weather Review*, 8:281—-293.

Pawitan, Y. (2001). *In all likelihood : Statistical modelling and inference using likelihood*. Oxford Science Publication.

Pinson, P. (2012). Adaptative calibration of (u,v)- wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138(666):1273–1284.

Pinson, P. and Hagedorn, R. (2012). Verification of the ecmwf ensemble forecasts of wind speed against analyses and observations. *Meteorological Application*, 00:1–20.

Pinson, P., Nielsen, H., Moller, J., and Madsen, H. (2007). Nonparametric probabilistic forecasts of wind power: required properties and evaluation. *Wind Energy*, 10:497–516.

Pinson, P., Reikard, G., and Bidlot, J. (2012). Probabilistic forecasting of the wave energy flux. *Applied Energy*, 93:364–370.

Raftery, A., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174.

Schuhen, N., Thorarinsdottir, T., and Gneiting, T. (2012). Ensemble model output statistics for wind vectors. *Monthly Weather Review*, In Press.

Sloughter, J., Gneiting, T., and Raftery, E. (2010). Probabilistic wind speed forecasting using ensembles and bayestian model averaging. *Journal of the American Statistical Association*, 105:25–35.

Tambke, J., Bye, J., Wolff, J., Tautz, S., Lange, B., Lange, M., and Focken, U. (2004). Modelling offshore wind profiles using inertially coupled wave boundary layers. In *Proceedings of the 2004 European Wind Energy Conference*.

Thorarinsdottir, T. and Gneiting, T. (2008). Probabilistic forecast of wind speed: Ensemble model output statistics unsing hereoskedastic censored regression. Technical Report 546, Deparment of Statistics, University of Washington.

Wilhelm, S. and Manjunath, B. (2010). tmvtnorm: A package for the truncated mutlivariate normal distribution. *The R Journal*, 2(1):25–29.

Wilks, D. (2006). *Statistical Methods in the Atmospheric Sciences, 2nd edition*. Academic Press.

Wilks, D. and Hamill, T. (2007). Comparison of ensemble-mos methods using gfs reforecasts. *Monthly Weather Review*, 135:2379–2390.