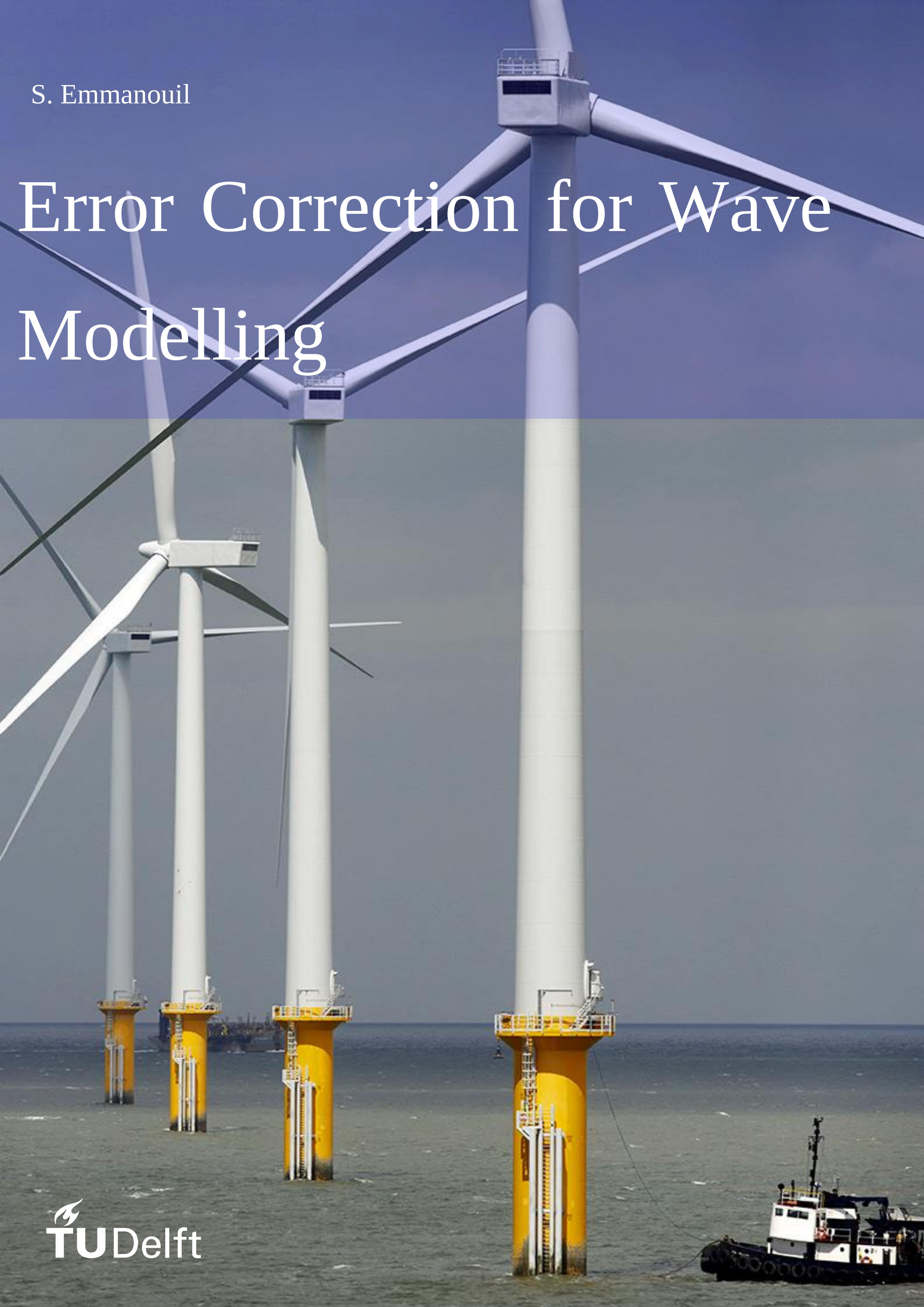
S. Emmanouil

Error Correction for Wave Modelling



TU Delft

TUDelft

Error Correction for Wave Modelling

by

Stergios Emmanouil

Diploma of Civil Engineering (2016)
University of Patras

Submitted to the Department of Civil Engineering
in partial fulfilment of the requirements
for the degree of

Master of Science
in Civil Engineering

at the
Delft University of Technology

September 2018

Supervisor:	Prof. dr. ir. G.F. Nane	TU Delft
Thesis committee:	Prof. dr. ir. M. Kok,	TU Delft
	Ir. G. Leontaris,	TU Delft
Company Supervisors:	Ir. S. Gaytan Aguilar,	Deltares
	Ir. J.-J. Schouten,	Deltares

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Preface

This MSc thesis was conducted in full cooperation with Deltares. Deltares is an independent institute for applied research in the field of water and subsurface. Throughout the world, they work on smart solutions, innovations and applications for people, environment and society. The company's main focus is on deltas, coastal regions and river basins. Managing these densely populated and vulnerable areas is complex, which is why the work is conducted closely with governments, businesses, other research institutes and universities at home and abroad. Their motto is Enabling Delta Life. As an applied research institute, the success of Deltares can be measured in the extent to which the expert knowledge can be used in and for society. For Deltares the quality of the expertise and advice comes first. I personally feel privileged to have been given the chance to work in such an environment, and share knowledge and experience about Civil Engineering, but also life in general.

The "*Error Correction for Wave Modelling*" provides useful tools to support the accurate and efficient conduction of operations in offshore wind farms. The importance of renewable sources of energy in today's world is widely known by now, and certainly the area of offshore wind energy offers many advantages, with wind speeds much higher than those on land, and vast amounts of space. Projections for 2020 estimate an offshore wind farm capacity of 40 GW in European waters, which would provide 4% of the European Union's demand of electricity. The European Wind Energy Association has set a target of 40 GW installed by 2020 and 150 GW by 2030. Offshore wind power capacity is expected to reach a total of 75 GW worldwide by 2020, with significant contributions from China and the United States.

Since the techniques discussed in this research refer to mild (everyday) conditions, the results can form the basis of further research on ocean navigation and safety, but also other fields such as coastal engineering and flood risk management. More importantly this thesis is a proof of the importance of machine learning, and more generally data-based, techniques in civil and environmental engineering, while combined with already developed numerical tools.

S. Emmanouil
Delft, July 2018

Abstract

Installation and maintenance strategies regarding offshore wind farm operations involve extensive logistics, since the main focus is the right temporal and spatial placement of personnel and equipment, while taking into account forecasted meteorological and hydrodynamic conditions. In order for these operations to be successful weather windows characterized by certain permissive wave, wind and current conditions is of enormous importance, whereas unforeseen events result in high cost and risk in terms of safety.

For that purpose, Deltares created Meteo Dashboard, an integrated software system that collects, stores, computes and presents measured and forecasted meteorological and hydrodynamic data for decision making of maintenance or installation activities in an offshore wind farm. The wind speed, as well as the air and water temperatures, result from a meteorological model and serve as an input for the numerical modelling (e.g. SWAN or Delft3D) of waves, water levels and current related parameters. To account for the inherited uncertainty, several error modelling techniques, such as Artificial Neural Networks (ANN), Copulas, Stochastic Interpolation (SI), ARMA models, and Linear Regression (REG), already run operational on Meteo Dashboard and can be implemented in order for the numerical model forecasts to be corrected. A number of the aforementioned techniques require training using historical or present time data, while others can be incorporated forthwith.

In this research, a fully automated ARIMA model and different kinds of Bayesian Network (BN) models are incorporated in order to enhance the accuracy of the significant wave height (H_s) predictions even further. Both techniques are implemented using packages provided by the free software environment of R, namely the `bnlearn` and `forecast`. The implemented BN models differ in terms of training and structure, and provide overall the most satisfying accuracy in comparison to the rest of the error correction techniques, when tested with data retrieved from stations deployed in the Irish Sea (adjacent to the Gwynt-y-Mor and Rhyl Flats offshore wind farms) corresponding to the whole year of 2017 (from January 2017 – to January 2018).

Supplementary, it is also shown that the BN models illustrate even more advantages when compared to the rest of the error correction techniques, since they provide information about the incorporated variables dependence relationship through their structures, while producing estimates for the underlying uncertainty of the phenomenon, by means of 95% confidence intervals extracted by the significant wave height (H_s) conditional distribution.

Finally, all error correction models are tested in operational (online) mode, with real-time data from the aforementioned locations, with the newly implemented BN models producing results of enhanced accuracy, even in the absence of measurements.

Dedicated to
my grandfather

Acknowledgements

Reaching the end of this effort, I would like to thank all the people who helped me and provided the basis for this work to be concluded.

First, I would like to thank all my professors at TU Delft, who provided me the necessary knowledge and tools to be able to confront any challenges that appeared throughout my thesis, but also prepare me sufficiently for whatever it is to follow after my graduation.

I have to give special thanks to the members of my committee, for their contributions in this work and the time they spend providing me continuous support during this year.

I would like to acknowledge and thank Professor Matthijs Kok, who invested time and effort to render his experience and knowledge to the betterment of this work. Also, I want to thank George Leontaris, who apart from giving his insight in various occasions in technical matters, was always there to discuss, encourage me and assist me as a friend.

Further, I would like to thank my professor and friend Tina Nane, who is not only a brilliant researcher, who supplied me endlessly with ingredients to fulfill this thesis, but she is also a most considerate and helpful person. Her day-to-day support is definitely one of the reasons that this work is finalized.

I want to exceptionally thank my research supervisors Sandra Gaytan Aguilar and Jan-Joost Schouten, who created the basis of this work and later provided daily support, both in technical and life aspects, during my stay at Deltares. Sandra and Jan-Joost not only showed me how to interact and conduct myself in a professional environment, but they also made sure that everything will be in my disposal for this thesis to be concluded.

Of course, by no means I can omit thanking my former research supervisor, professor and friend Andreas Langousis, who is the primary reason I became a Hydraulic Engineer. Being an outstanding researcher, he taught me the importance of working hard to achieve my goals, while supplying me with extremely influential knowledge, advice, and motivation, which always follow me. Most importantly, he was always there to support and inspire me in matters far beyond the boundaries of engineering.

Certainly, I cannot forget to thank all the people at Deltares and TU Delft, many of which became my friends, providing daily encouragement and motivation. I would love to name them one by one, but to avoid omitting to mention anyone and degrade their influence I am settling with this reference. My stay in the Netherlands would have been much harder if it wasn’t for them.

Special thanks have to be given to my girlfriend Efi, whose understanding, support and belief on a daily basis helped me go through all the obstacles that came my way. Her patience and care to listen to all my news and ideas, even from distance, was a reassuring pillow which I am grateful to have.

I have to exceptionally thank my cousin Teo Emmanouilidis. Also being a Civil Engineer he formed an example for me from his student days. Even more importantly, he influenced me as a person, since with his humor and advice he enabled me to pull through various situations. A part of his love for programming and computers was imprinted on me and is reflected also in this work.

Finally, I owe a great thanks to my father, my mother and my brother, who always support me and never stop believing in me, while continuously pushing and motivating me to pursue my dreams. All my principles and values originate from them.

Delft, September 2018

Stergios Emmanouil

Contents

Preface v

Abstract.. vii

Acknowledgements... xi

Tables... xvii

Main Text..xvii

Appendix A xix

Appendix C xx

Figures xxi

Main Text..xxi

Appendix A xxiii

Appendix B xxv

Appendix C xxvi

1. Introduction.. 1

1.1. General 1

1.2. Meteo Dashboard.. 1

1.3. Error modelling.. 2

1.4. Objective and Research Questions... 4

1.5. Reader 5

2. Literature Review 7

2.1. Regressive Modelling... 7

2.1.1. Auto-Regressive Model (AR) 7

2.1.2. Moving-Average Model (MA) 7

2.1.3. Auto-Regressive Moving Average Model (ARMA).................................. 8

2.1.4. Auto-Regressive Integrated Moving Average (ARIMA) 8

2.2. Simple Linear Regression Modelling 11

2.3. Artificial Neural Networks (ANNs) 12

2.4. Copula modelling 16

2.5. Bernstein Stochastic Interpolation 19

2.5. Bayesian Networks (BNs) 20

2.5.1. General 20

2.5.2. Hybrid Bayesian Networks 22

2.5.3. Non-Parametric Bayesian Networks 24

2.5.4. Implementation of BNs in Matlab® and R 26

2.6. SWAN 27

2.7. Concluding Remarks 27

3. The Data 29

3.1. Raw Data and Pre-processing 30

3.1.1. Meteorological Data 32

3.2. Data cleaning 33

3.2.1. Methodology and Results 33

3.3. Scatterplots 36

3.3.1. Peak Wave Period (T_p) Case 37

3.4. General Comments 38

3.4.1. Joint Distributions of Data 38

4. The Error Correction Model and its Functionality 41

4.1. General 41

4.2. Operational Techniques in Meteo Dashboard 42

4.2.1. Linear Regression 42

4.2.2. Copulas 42

4.2.3. Artificial Neural Networks 43

4.2.4. Stochastic Interpolation 43

4.3. ARIMA Model 44

4.3.1. Methodology and Training 44

4.3.2. Seasonality 44

4.4. Bayesian Networks (BN) Model 45

4.4.1. Scope 45

4.4.2. Training Methodology 46

4.4.3. Data Fitting in BN Structure 47

4.4.4. Predictions 48

4.5. Non-Operational Tests of Functionality 50

4.5.1. Model settings and training 50

4.5.2. Evaluation Measures/Metrics 51

4.5.3. Preliminary Results 51

4.5.4. BN Structure 53

4.5.5. Further Testing 55

4.6. Preliminary Conclusions 55

5. Method Comparison 58

5.1. Model configuration and settings 58

5.1.1. Training and Fitting 58

5.1.2. Testing and Validation 59

5.2. Technique Comparison – Yearly Tests 59

5.2.1. Timeseries and Evaluation of 48-hr Forecasts 59

5.2.2. The potential of Bayesian Networks 64

5.2.3. Summative (Yearly) Evaluation Metrics 67

5.2.4. Error Distribution 70

5.2.5. Taylor Diagrams for the yearly predictions 72

5.2.6. Scatterplots of EC Model Vs Observations 73
5.3. Uncertainty Evaluation 76
5.3.1. Normal and Log-Normal Uncertainty Bounds 77
5.3.2. Comparison of the provided Uncertainty Bounds 79
5.4. BN Structures and Configurations 81
5.4.1. BN Structures with 6 Variables 81
5.4.2. Comparison of Alternative BN Structures 85
5.5. Remarks – Comments 94
6. Conclusions 96
6.1. Overall Model Performance 96
6.2. Analysis of the BN Structures 97
6.3. Evaluation of the Uncertainty Estimates 97
6.4. Operational Functionality 98
6.5. Research Directions 99
Appendix A 100
A.1. Preliminary Analysis BN Structures 100
A.2. Periodograms 108
A.3. Functionality Test Results 113
Appendix B 119
B.1. Numerical Model (SWAN) Timeseries 119
B.2. Raw and Clean Observational Data Timeseries 123
B.3. Numerical Model Data (SWAN) Vs Measurements 130
B.4. Density scatterplot for GyM 132
B.5. Data Cleansing Filters 132
B.6. Peak Wave Period (T_p) Cleaning Procedure 133
B.7. Joint Occurrence of Wind and Waves 134

Appendix C 135

C.1. Error Timeseries for the Gwynt-y-Mor wind farm.. 135

C.2. Scatterplots for the Rhyl Flats wind farm.. 136

C.3. Taylor diagram for the Rhyl Flats wind farm... 138

C.4. Uncertainty estimates for the Rhyl Flats wind farm 139

C.5. BN structures for the Rhyl Flats wind farm.. 140

C.6. Comparison of data-driven and imposed structures.................................... 143

Appendix D 144

D.1. Copula Families 144

D.2. Cramer-von Mises.. 145

D.3. Artificial Neural Networks Basics.. 145

D.4. Applications of ARMA 146

Bibliography 147

Tables

Main Text

Table 1. Preliminary Evaluation Metrics (RMSE - MAE table) for 28-30 March 2015, at Rhyl Flats (Irish Sea)... 53

Table 2. Correlation matrix for short-trained BN, 28-30 March 2015 - Rhyl Flats..... 54

Table 3. Correlation matrix for long-trained BN, 28-30 March 2015 - Rhyl Flats...... 55

Table 4. General model accuracy evaluation metrics for Gwynt-y-Mor at 18:00 on 24-02-2017.. 62

Table 5. General model accuracy evaluation metrics for Gwynt-y-Mor at 00:00 on 25-02-2017.. 62

Table 6. Yearly (2017) Evaluation Metrics for Gwynt-y-Mor 67

Table 7. Application-specific Evaluation Metrics for the Gwynt-y-Mor case study.... 68

Table 8. Yearly (2017) Evaluation Metrics for Rhyl Flats. 69

Table 9. Application-specific Evaluation Metrics for the Rhyl Flats case study. 69

Table 10. Uncertainty Bounds comparison for the Gwynt-y-Mor case study. 80

Table 11. Uncertainty Bounds comparison for the Gwynt-y-Mor case study, if negative values are taken into account. 80

Table 12. Correlation matrix for the long-trained BN model structure (Gwynt-y-Mor). 84

Table 13. Correlation matrix for the short-trained BN model structure on 29-12-2017 at 18:00 (Gwynt-y-Mor). 85

Table 14. Evaluation metrics for the case of the 4-variable BN models (Gwynt-y-Mor). 87

Table 15. Application specific evaluation metrics for the case of the 4-variable BN models (Gwynt-y-Mor). 87

Table 16. Uncertainty estimates' performance for the case of the 4-variable BN structures (Gwynt-y-Mor). 88

Table 17. Comparison of evaluation metrics for the 4- and 6-variable BN models (Rhyl Flats). 88

Table 18. Comparison of application specific evaluation metrics for the 4- and 6-variable BN models (Rhyl Flats). 89

Table 19. Evaluation metrics for the case of the 5-variable BN models (Gwynt-y-Mor). 90

Table 20. Application specific evaluation metrics for the case of the 5-variable BN models (Gwynt-y-Mor). 91

Table 21. Uncertainty estimates' performance for the case of the 5-variable BN structures (Gwynt-y-Mor). 91

Table 22. Evaluation metrics for the case of the 5-variable BN models (Rhyl Flats). 92

Table 23. Application specific evaluation metrics for the case of the 5-variable BN models (Rhyl Flats). 93

Table 24. Uncertainty estimates' performance for the case of a 5-variable BN structure (Rhyl Flats). 93

Appendix A

Table A. 1. Long-trained BN correlation matrix on 28-30 May 2015 (Rhyl Flats - Irish Sea).. 100

Table A. 2. Short-trained BN correlation matrix on 28-30 May 2015 (Rhyl Flats - Irish Sea).. 101

Table A. 3. Long-trained BN correlation matrix on 28-30 May 2015 (Gwynt-y-Mor - Irish Sea)... 102

Table A. 4. Short-trained BN correlation matrix on 28-30 May 2015 (Gwynt-y-Mor - Irish Sea)... 102

Table A. 5. Long-trained BN correlation matrix on 29-31 July 2015 (Rhyl Flats - Irish Sea).. 103

Table A. 6. Short-trained BN correlation matrix on 29-31 July 2015 (Rhyl Flats - Irish Sea).. 103

Table A. 7. Long-trained BN correlation matrix on 29-31 July 2015 (Gwynt-y-Mor - Irish Sea)... 104

Table A. 8. Short-trained BN correlation matrix on 29-31 July 2015 (Gwynt-y-Mor - Irish Sea)... 105

Table A. 9. Long-trained BN correlation matrix on 28-30 August 2015 (Rhyl Flats - Irish Sea)... 105

Table A. 10. Short-trained BN correlation matrix on 28-30 August 2015 (Rhyl Flats - Irish Sea)... 106

Table A. 11. Long-trained BN correlation matrix on 28-30 August 2015 (Gwynt-y-Mor - Irish Sea)... 107

Table A. 12. Short-trained BN correlation matrix on 28-30 August 2015 (Gwynt-y-Mor - Irish Sea)... 107

Table A. 13. Preliminary Evaluation Metrics (RMSE - MAE table) for 28-30 May 2015, at Rhyl Flats (Irish Sea)... 113

Table A. 14. Preliminary Evaluation Metrics (RMSE - MAE table) for 28-30 May 2015, at Gwynt-y-Mor (Irish Sea). 114

Table A. 15. Preliminary Evaluation Metrics (RMSE - MAE table) for 29-31 July 2015, at Rhyl Flats (Irish Sea)... 115

Table A. 16. Preliminary Evaluation Metrics (RMSE - MAE table) for 29-31 July 2015, at Gwynt-y-Mor (Irish Sea). 116

Table A. 17. Preliminary Evaluation Metrics (RMSE - MAE table) for 28-30 August 2015, at Rhyl Flats (Irish Sea). 117

Table A. 18. Preliminary Evaluation Metrics (RMSE - MAE table) for 28-30 August 2015, at Gwynt-y-Mor (Irish Sea). 118

Appendix C

Table C. 1. Uncertainty estimates' performance for the case of the 4-variable BN structures (Rhyl Flats). 139

Table C. 2. Uncertainty estimates' performance for the case of the 6-variable BN structures (Rhyl Flats). 140

Table C. 4. Correlation matrix for the short-trained BN model 4-variable structure (Rhyl Flats). 142

Table C. 5. Correlation matrix for the short-trained BN model 5-variable structure (Rhyl Flats). 142

Table C. 6. Correlation matrix for the long-trained BN models (Rhyl Flats). 143

Figures

Main Text

Fig. 1. Example of a 3-layered ANN (Source: Google Images for Artificial Neural Networks). 13

Fig. 2. Relations between nodes in a Bayesian Network. (source Google Images) 21

Fig. 3. Map of the Irish Sea wind farms (underlined with red) under consideration (measurement stations visible as blue diamonds). 29

Fig. 4.Gwynt-y-Mor wind farm (left) and wave rider buoy deployed in the area (left) (Source: Google Images). 30

Fig. 5. Significant wave height (H_s) raw data at Gwynt-y-Mor (01-09-2012 to 31-01-2018). 31

Fig. 6. Wind velocity profile above sea level (Source: Google Images). 32

Fig. 7. Wind velocity (U_{10}) timeseries at Gwynt-y-Mor (01-09-2012 to 31-01-2018). 33

Fig. 8. Clean and raw significant wave height (H_s) datasets, as resulted from the ORCA cleaning procedure (Gwynt-y-Mor). 34

Fig. 9. Clean and raw zero-crossing wave period (T_z) datasets, as resulted from the ORCA cleaning procedure (Gwynt-y-Mor). 35

Fig. 10. Clean and raw wind velocity (U_{10}) datasets, as resulted from the ORCA cleaning procedure (Gwynt-y-Mor). 35

Fig. 11. Numerical modelled versus Observed significant wave height (H_s) data (Gwynt-y-Mor). 36

Fig. 12. Numerical modelled versus Observed zero-crossing wave period (T_z) data (Gwynt-y-Mor). 36

Fig. 13. Numerical modelled versus Observed wind velocity (U_{10}) data (Gwynt-y-Mor). 37

Fig. 14. Numerical modelled versus Observed peak wave period (T_p) data (Gwynt-y-Mor). 38

Fig. 15. Joint distributions for the available hydrodynamic and meteorological variables at Gwynt-y-Mor. 39

Fig. 16. Results of the parametric distribution fitting procedure to the significant wave height (H_s) data of Gwynt-y-Mor. 40

Fig. 17. Flow Chart presenting the basic function of the BN model. 49
Fig.18. Comparison between the newly implemented correction techniques for 28-30 March 2015, at Rhyl Flats (Irish Sea). 51
Fig. 19. Periodogram displaying the major frequencies extracted by the original dataset (28-30 March 2015 – Rhyl Flats). 52
Fig. 20. BN structure for 28-30 March 2015, at Rhyl Flats (Irish Sea). 54
Fig. 21. Incorporation of an ARIMA model, with a 60 min seasonality setting applied, in 28-30 March 2015 (Rhyl Flats - Irish Sea). 56
Fig. 22. Significant wave height (H_s) predictions produced for Gwynt-y-Mor at 18:00 on 24-02-2017. 60
Fig. 23. Significant wave height (H_s) predictions produced for Gwynt-y-Mor at 00:00 on 25-02-2017. 61
Fig. 24. Example of the SI model's predictive accuracy, in comparison to the rest of the methods (Gwynt-y-Mor). 63
Fig. 25. Significant wave height correction (H_s) attempts, in which the BN methods displayed satisfying performance (GyM). 64
Fig. 26. Examples of the BN methods' predictive accuracy, with emphasis on the fixed structure BN (Gwynt-y-Mor). 65
Fig. 27. Occasions in which the BN methods provided satisfying predictive accuracy (Gwynt-y-Mor). 66
Fig. 28. Distribution of errors for the significant wave height (H_s) predictions at Gwynt-y-Mor (2017). 71
Fig. 29. Distribution of errors for the significant wave height (H_s) predictions at Rhyl Flats (2017). 71
Fig. 30. Evolution of errors for the Rhyl Flats case study (2017). 72
Fig. 31. Taylor diagram for significant wave height (H_s) predictions (Gwynt-y-Mor in 2017). 73
In terms of the BN error correction models (Figure 32), the close relation between the predictions produced by the long-trained BN (top right) and the numerical model (SWAN) is evident. Regarding the uncertainty, the scatterplots of the aforementioned models are similar, while the short-trained BN model (bottom left) introduces a generally larger scatter under the diagonal. Thus it can be concluded by the graph that the short-trained BN has a tendency to under-predict the measurements. 73

Fig. 32. Scatterplots of SWAN (upper left) and BN models in relation to the measurements (Gwynt-y-Mor). 74
Fig. 33. Scatterplots of implemented error correction models in relation to the measurements (Gwynt-y-Mor). 75
Fig. 34. Log-Normal (left) and Normal (right) 95% Uncertainty Bounds (fixed BN method for Gwynt-y-Mor on 09-03-2017). 77
Fig. 35. Log-Normal (left) and Normal (right) 95% Uncertainty Bounds (long-trained BN for Gwynt-y-Mor on 09-03-2017). 78
Fig. 36. Log-Normal (left) and Normal (right) 95% Uncertainty Bounds (short-trained BN for Gwynt-y-Mor on 09-03-2017). 79
create a structure using general knowledge on the incorporated variables (i.e. knowledge on the underlying relations procured by the literature or by experts) only hindered the prediction/correction procedure instead of enhancing its accuracy. 81
Fig. 37. Structure for the long-trained and fixed BN models, incorporating 6 variables (Gwynt-y-Mor). 82
Fig. 38. Structures for the short-trained BN model incorporating 6 variables for the case of Gwynt-y-Mor. (29-12-2017 at 18:00 and 17-11-2017 at 06:00) 82
Fig. 39. Long- (left) and short-trained (right) BN models' structures for the case Gwynt-y-Mor incorporating 4 variables. (27-10-2017 at 18:00) 86
Fig. 40. Long- (left) and short-trained (right) BN models' structures for the case Gwynt-y-Mor incorporating 5 variables. (27-10-2017 at 18:00) 90

Appendix A

Fig. A. 1. BN structure for short and long training on 28-30 May 2015 (Rhyl Flats - Irish Sea). 100
Fig. A. 2. BN structure for long training on 28-30 May 2015 (Gwynt-y-Mor - Irish Sea). 101
Fig. A. 3. BN structure for short training on 28-30 May 2015 (Gwynt-y-Mor - Irish Sea). 101
Fig. A. 4. BN structure for long training on 29-31 July 2015 (Rhyl Flats - Irish Sea). 102
Fig. A. 5. BN structure for short on 29-31 July 2015 (Rhyl Flats - Irish Sea). 103

Fig. A. 6. BN structure for long training on 29-31 July 2015 (Gwynt-y-Mor - Irish Sea). 104
Fig. A. 7. BN structure for short on 29-31 July 2015 (Gwynt-y-Mor - Irish Sea). 104
Fig. A. 8. BN structure for long training on 28-30 August 2015 (Rhyl Flats - Irish Sea). 105
Fig. A. 9. BN structure for short on 28-30 August 2015 (Rhyl Flats - Irish Sea). 106
Fig. A. 10. BN structure for long training on 28-30 August 2015 (Gwynt-y-Mor - Irish Sea).......... 106
Fig. A. 11. BN structure for short on 28-30 August 2015 (Gwynt-y-Mor - Irish Sea). 107
Fig. A. 12. Periodogram corresponding to the 28-30 May 2015 data (Rhyl Flats - Irish Sea).......... 109
Fig. A. 13. Periodogram corresponding to the 28-30 May 2015 data (Gwynt-y-Mor - Irish Sea). 109
Fig. A. 14. Periodogram corresponding to the 29-31 July 2015 data (Rhyl Flats - Irish Sea).......... 110
Fig. A. 15. Periodogram corresponding to the 29-31 July 2015 data (Gwynt-y-Mor - Irish Sea). 110
Fig. A. 16. Periodogram corresponding to the 28-30 August 2015 data (Rhyl Flats - Irish Sea). 111
Fig. A. 17. Periodogram corresponding to the 28-30 August 2015 data (Gwynt-y-Mor - Irish Sea).......... 111
Fig. A. 30. Comparison between the BN and ARIMA techniques for 28-30 May 2015 (Rhyl Flats - Irish Sea).......... 113
Fig. A. 31. Comparison between the BN and ARIMA techniques for 28-30 May 2015 (Gwynt-y-Mor - Irish Sea). 114
Fig. A. 32. Comparison between the BN and ARIMA techniques for 29-31 July 2015 (Rhyl Flats - Irish Sea).......... 115
Fig. A. 33. Comparison between the BN and ARIMA techniques for 29-31 July 2015 (Gwynt-y-Mor - Irish Sea). 116
Fig. A. 34. Comparison between the BN and ARIMA techniques for 28-30 August 2015 (Rhyl Flats - Irish Sea).......... 117

Fig. A. 35. Comparison between the BN and ARIMA techniques for 28-30 August 2015 (Gwynt-y-Mor - Irish Sea). 118

Appendix B

Fig. B. 1. Significant wave height (top left), zero-crossing wave period (top right), wave direction (bottom left), and peak wave period (bottom right) produced by SWAN (Gwynt-y-Mor). 119

Fig. B. 2. Numerical model (HIRLAM) wind velocity data, which served as input data to SWAN (Gwynt-y-Mor). 120

Fig. B. 3. Numerical model (HIRLAM) wind direction data, which served as input data to SWAN (Gwynt-y-Mor). 120

Fig. B. 4. Timeseries of the swell components, as produced by SWAN (Gwynt-y-Mor). 120

Fig. B. 5. Hydrodynamic data timeseries as produced by SWAN for the case of Rhyl Flats. 121

Fig. B. 6. Meteorological data timeseries as used in SWAN for the case of Rhyl Flats (produced by HIRLAM). 121

Fig. B. 7. Timeseries of the swell components, as produced by SWAN (Rhyl Flats). 122

Fig. B. 8. Clean (blue circles) and raw (red dots) significant wave height (H_s) timeseries for the Rhyl Flats wind farm. 123

Fig. B. 9. Clean (blue circles) and raw (red dots) zero-crossing wave period (T_z) timeseries for the Rhyl Flats wind farm. 123

Fig. B. 10. Clean (blue circles) and raw (red dots) wind velocity (U_{10}) timeseries for the Rhyl Flats wind farm. 124

Fig. B. 11. Clean (blue circles) and raw (red dots) peak wave period (T_p) timeseries for the Rhyl Flats wind farm. 124

Fig. B. 12. Clean (blue circles) and raw (red dots) wave direction (D_{irp}) timeseries for the Rhyl Flats wind farm. 125

Fig. B. 13. Clean (blue circles) and raw (red dots) wind direction (U_{dir}) timeseries for the Rhyl Flats wind farm. 125

Fig. B. 14. Clean (blue circles) and raw (red dots) wave direction (D_{irp}) timeseries for the Gwynt-y-Mor wind farm. 126
Fig. B. 15. Clean (blue circles) and raw (red dots) wind direction (U_{dir}) timeseries for the Gwynt-y-Mor wind farm. 126
Fig. B. 16. Wind rose for the case of Gwynt-y-Mor. 127
Fig. B. 17. Wind rose for the case of Rhyl Flats 128
Fig. B. 18. Joint and individual distributions of the hydrodynamic and meteorological variables at Rhyl Flats. 129
Fig. B. 19. Joint occurrence of wind and waves at Rhyl Flats. 129
Fig. B. 20. Numerical data vs Measurements for the Rhyl Flats' meteorological and hydrodynamic datasets. 130
Fig. B. 21. Numerical data vs Measurements of wind (U_{dir}) and wave (D_{irp}) direction at Rhyl Flats and Gwynt-y-Mor. 131
Fig. B. 22. Density scatterplot illustrating the relation between wind velocities at different levels (Gwynt-y-Mor). 132
Fig. B. 23. Clean and raw peak wave period (Tp) datasets, as resulted from the ORCA cleaning procedure (Gwynt-y-Mor). 133
Fig. B. 24. Joint-occurrence of wind and waves at Gwynt-y-Mor. 134

Appendix C

Fig.C. 1. Evolution of errors in relation with time for the Gwynt-y-Mor case study. 135
Fig.C. 2. Scatterplots of implemented BN models compared with SWAN results in relation to the observations in Rhyl Flats. (Jan 2017 – Jan 2018) 136
Fig.C. 3. Scatterplots of already operation error correction techniques in relation to Rhyl Flats' observations. (Jan 2017 – Jan 2018) 137
The Taylor diagrams of the 5-variable and 6-variable structures for the case of Rhyl Flats (Figures C.4 and C.5) display clearly the suitability of the first one for the application. 138
Fig.C. 4. Taylor diagram for the case where 5-variable BN models are incorporated (Rhyl Flats). 138
Fig.C. 5. Taylor diagram for the case where 6-variable BN models are incorporated (Rhyl Flats). 138

Fig.C. 6. Long-trained (left) and short-trained (right) BN model 4-variable structures for the Rhyl Flats case. (21-02-2017 at 06:00) 141

Fig.C. 7. Long-trained (left) and short-trained (right) BN model 5-variable structures for the Rhyl Flats case. (21-02-2017 at 06:00) 141

Fig.C. 8. Long-trained (left) and short-trained (right) BN model 6-variable structures for the Rhyl Flats case. (21-02-2017 at 06:00) 141

Fig.C. 9. Example of differences between a data-driven and an imposed structure results for Gwynt-y-Mor. 144

1. Introduction

1.1. General

Marine structures like offshore wind turbines can ensure safety and serve their main function adequately, in both reliability and economy terms, when most – if not all – of the parameters involved in their design are modelled as accurately as possible. The specification of the uncertainties related to the environmental parameters describing the ocean conditions is continuously gaining importance and interest by the offshore, coastal, and the emerging renewable energy industries.

Several studies have been conducted in order to describe, classify, or quantify the uncertainties and errors related to meteorological and ocean climate variables (see e.g. Bitner – Gregersen et al., 2014; Haver and Moan, 1983; Bitner – Gregersen and Hagen, 1990). Simplistically, the uncertainty can be classified as:

a. Phenomenon related uncertainty, which is a product of the natural randomness and stochastic nature of the variables incorporated and cannot be reduced.
b. Data related uncertainty, which surfaces either from the measuring devices' accuracy, or the insufficient number or quality of the observations.
c. Model related uncertainty, which constitutes a product of inaccurate idealisations, crude assumptions, or even insufficient use of either the meteorological or the hydrodynamic model. It is obvious that the true nature of any phenomenon cannot be modelled exactly and that even if the probability distributions of some variables are known a priori, the extreme complexity of the met-ocean environment makes the distributions of the rest completely unknown.

The estimation of the bias, or systematic error, and the random error evaluation are the first steps to quantify the uncertainty of any variable.

1.2. Meteo Dashboard

In the case of offshore wind farms, the installation and maintenance strategies involve extensive logistics. The main focus is the right placement, in time and space, of both the personnel and the equipment, while taking into account forecasted meteorological and hydrodynamic conditions. In order for the aforementioned procedures to be carried out successfully, weather windows, interwoven with certain permissive wave, wind and current conditions, are of major importance, while unforeseen weather or sea climate events result in high cost and risk, primarily in terms of safety. Subsequently, successful operations require accurate and representative data for the wind farm sites, which unfortunately are inadequately - if at all - provided by surrounding stations.

For that purpose, Deltares created Meteo Dashboard; an integrated software system that collects, stores, computes and presents measured and forecasted meteorological and hydrodynamic data for decision making of maintenance or installation activities in an offshore wind farm. These forecasts can be used for other instances as well, such as flood warning, workability in open seas and nautical safety. The wind speed, as well as the air and water temperatures, result from a meteorological model and serve as an input for the numerical modelling (e.g. SWAN or Delft3D) of waves, water levels and current related parameters.

To further improve the accuracy and usability of the forecasts and to quantify the model's uncertainty, various techniques have been implemented in the Meteo Dashboard. In general, such techniques are referred to as data-model integration (DMI) or data assimilation (see e.g. Bidlot and Holt, 1999; De Las Heras et al., 1994; Anderson et al., 1996; Lefevre and Aouf, 2012). Based on the modelled error[1] exported by the aforementioned techniques, the original (uncorrected) forecast can be corrected, thus improving the accuracy of the output, which is provided by Meteo Dashboard every 6 hours.

Comparison of the wave model forecasts with observations is essential for characterizing the model deficiencies, identifying systematic and random model errors, thus providing areas for improvement. The error correction models are either trained offline, using a substantial amount of numerical model hindcast[2] data and measurements from the same period, or in operational (online) mode using a smaller amount of data, containing only the most recent hindcast and measured values of interest.

1.3. Error modelling

Several error modelling techniques exist and can be implemented in order for the numerical model forecasts to be corrected. All of them constitute soft computing methods and ensure a reasonable computational load. A number of the aforementioned techniques require training using historical or present time data, while others can be incorporated forthwith. Numerous studies have tried to produce valid met-ocean climate forecasts using coupled (hybrid)

[1] The term "*error*" refers to the difference between simulation model output and observations, while the "*model uncertainty*" term corresponds to the variability in the output of the simulation model resulting from (minor) differences in the input.

[2] Hindcast is the exact opposite of forecast; numerical model results for a past time, where observations exist, making the calculation of the errors possible.

methods[3], as the one discussed in this thesis, or incorporate one of the techniques discussed below to predict the environmental conditions therewithal. Certainly the use of a single soft computing method for prediction reduces the computational time significantly, but often at the expense of accuracy.

Special attention is given in the implementation of the Bayesian Networks (BNs), graphical models which allow the representation of a probability distribution over more than one variables and whose use has not been that widespread in offshore applications (an example can be found in Malekmohamadi et al., 2011), but has been tested effectively in other engineering problems, such as coastal morphology (see e.g. Poelhekke et al., 2016; Kroon et al., 2017; Wilson et al., 2015; Plant and Holland, 2011), environmental modelling (see Chen and Pollino, 2012; Aguilera et al., 2011), construction reliability (Morales-Napoles and Steenbergen, 2014), or flood risk analysis, for which the reader is referred to Sebastian et al. (2017). An overview of many practical BN applications can be found in the work of Hanea et al. (2015).

One of the techniques commonly practiced in a variety of time series forecasting applications is the Auto-Regressive Moving Average modelling or ARMA, which is a stationary stochastic process consisting of sums of auto-regressive and moving average components. Auto-regressive (AR) models are basically recurrence relations with linear terms of past states of the variable itself, plus a noise term. Moving average (MA) models contain linear terms of the past error values, plus an expected value term. The main task when employing the ARMA model is to estimate the model's parameters. ARMA assumes that the time series is stationary, that is the average and variance of observations do not vary in time. Moreover, the errors have to be independent and normally distributed, or in other words the variance and the average must be assumed constant in time (data stationarity). When the data is non-stationary, the Auto-Regressive Integrated Moving Average (ARIMA) model, a generalization of the ARMA model presented by Box and Jenkins (1976), is employed. This is, in fact, the case with the variables of interest in met-ocean environments (e.g. wave heights, wind speed, current direction, etc.). Incorporating the aforementioned stochastic models in offshore engineering applications has been done extensively in the works of Manouchehr (1997); Li and Kareem (1993); Sobey (1996); Martzikos and Soukissian (2017); Zhang (2003); Khashei and Bijari (2010); Spanos (1983); Pena-Sanchez and Ringwood (2017); Delicado and Justel (1999).

[3] By "*coupled*" or "*hybrid*" methods the use of more than one error modelling techniques, or a combination of a soft computing method and a numerical model, is implied.

Another approach is the Artificial Neural Networks (ANN), which are information processing paradigms composed of large number of highly interconnected processing elements (neurons) working together. Similarly, ANN have been used extensively in offshore and coastal applications (see e.g. Deo, 2010; Deo et al., 2001; Tsai et al., 2002; Makarynskyy, 2004; Malekmohamadi et al., 2008; Kumar et al., 2017; Deo and Sridhar Naidu, 1999; Makarynskyy et al., 2005; Londhe et al., 2016; Agrawal and Deo, 2002; Mandal et al., 2005; Londhe and Panchang, 2005; Zhang et al., 2006; Deshmukh et al., 2016; Makarynskyy, 2007; Londhe and Panchang, 2006; Makarynskyy, 2005). There are different topologies and learning processes to be chosen when constructing an ANN, and after sufficient training with historical data, it can be used operationally every time that the model data need correction.

Useful tools for the quantification of the uncertainty in forecasts while accounting for the dependence between random variables are the copulas. Copulas are multivariate probability distributions, for which the marginal distribution of each variable is uniform (see e.g. Genest and Favre, 2007; Embrechts et al., 2001; Nelsen, 2006; Schmidt, 2006). In offshore applications, copulas have been used in various occasions to model the dependency of ocean related variables and predict their behavior, as it has been done in the works of Leontaris et al. (2016) and Jane et al. (2016).

More straightforward and simple methods, but equally effective in numerous occasions, are the linear regression and the stochastic interpolation. Both of these techniques have been used extensively in a variety of engineering applications, including offshore and coastal related (see e.g. Asma et al., 2012; Scotto and Guedes Soares, 2007), do not require training and pose serious advantages in terms of the computational time and load.

This thesis studies the effectiveness of the aforementioned methods in error modelling, focused on the minimization of the uncertainty in met-ocean climate forecasts. More specifically, a sufficiently large amount of met-ocean data, observed by stations deployed in the Irish and North Seas, in combination with produced hindcast and forecast numerical model data, is inserted in a hybrid error correction model, which incorporates all of the aforementioned statistical and stochastic methods. First, the validation of the techniques is done in non-operational mode (offline), with the ultimate goal being their implementation on the operational Meteo Dashboard platform.

1.4. Objective and Research Questions

This research aims to address the accuracy and quantify the possible errors present in the numerical model (e.g. SWAN) significant wave height (H_s) forecasts used in Meteo Dashboard, and provide corrections for these errors using automated statistical and stochastic models. To grant these corrections several BN models, that differ in terms of their training, their structure, and the incorporated variables, as well as a fully automated ARIMA model, are created, tested and validated with data retrieved from stations deployed in the Irish Sea, adjacent to the Gwynt-y-Mor and the Rhyl Flats offshore wind farms. A comparison of the performance of all the implemented statistical and stochastic techniques is also taking place, to ascertain which one performs better, using widely used evaluation metrics, such as the Root-Mean-Square-Error (RMSE), and more case specific indicators created for the purposes of the application under consideration.

Supplementary, uncertainty estimates are provided by the conditional distribution given by the BN models, and then compared with the confidence intervals derived by the already incorporated Gumbel Copula. Finally, the ability of the error correction techniques to perform in operational (real-time) conditions is investigated, to evaluate their performance even with the possible absence of measurements.

1.5. Reader

Chapter 2 reviews the literature on the methodological background of the employed techniques and examples of the error correction techniques in engineering practice, presenting advantages and disadvantages of each method. Hybrid and autonomous versions of those methods are both examined and analysed, so that the basis of the project can be set. The main assumptions governing the error correction model, primarily focused on the spatial correlation of the met-ocean variables, as well as the Matlab® and R toolboxes' possibilities and validity, are also discussed.

Chapter 3 includes the techniques incorporated to manipulate new data for testing, presenting simultaneously the observational and numerical model datasets to be used during the simulations. A critical evaluation of the variables' availability and appropriateness for use is also made, critically taking into account or discarding variables in order to make the models as robust and accurate as possible.

Chapter 4 provides an overview of the methodology used to incorporate each error correction technique, while assessing the functionality of the newly developed methods (Bayesian Networks (BNs) and ARIMA error correction models) with a preliminary analysis. Some preliminary conclusions on the functional performance follow the aforementioned analysis.

In Chapter 5 the overall performance of the error correction techniques is evaluated, with general and application-specific metrics, including also a critical comparison between various BN structure configurations, and the influence of certain variables on the models' accuracy. Supplementary, a discussion is made on the advantages that the BN models' uncertainty estimates provide, while analysing two different kinds of confidence intervals, produced by different assumptions and approaches.

Finally, in Chapter 6 the conclusions are present, combined with possible future research directions.

2. Literature Review

2.1. Regressive Modelling

2.1.1. Auto-Regressive Model (AR)

An order p autoregressive model (AR) allows the simulation of a stochastic process at a certain time, specifying that the output value depends linearly on its past values, as well as on a stochastic term. If X_t denotes a time series then the model AR(p) takes the following form:

$$X_t = c + \alpha_1 \cdot X_{t-1} + \alpha_2 \cdot X_{t-2} + \cdots + \alpha_1 \cdot X_{t-p} + \varepsilon_t \tag{2.1}$$

Where c is a constant, α_1, α_2, …, α_p, are the model parameters that have to be estimated, while ε_t constitutes a zero-mean white noise. Evidently, the variable of interest X_t is a linear combination of its own past values. Equivalently, Eq. (2.1) can be rewritten in the form:

$$X_t = c + \sum_{i=1}^{p} \alpha_i \cdot X_{t-i} + \varepsilon_t \tag{2.2}$$

Using the backshift operator B, which operates on a time series element in order to produce the previous element, the above equation can be written as:

$$X_t = c + \sum_{i=1}^{p} \alpha_i \cdot B^i \cdot X_t + \varepsilon_t \tag{2.3}$$

In order for the model to remain wide-sense stationary[4] the (complex) roots of the polynomial $z^p - \sum_{i=1}^{p} \alpha_i \cdot z^{p-i}$ must satisfy that $|z_i| < 1$.

2.1.2. Moving-Average Model (MA)

An order q moving average model (MA) is a filter allowing the simulation of a stochastic process at a specific time, based on the past and present white noise processes, which serve as an input. It specifies that the variable of interest depends linearly on the present and past values of a stochastic term, and it is commonly used to model univariate time series. Contrary to the AR model, the MA model is always stationary. The model MA(q) is defined as:

$$X_t = \mu + \varepsilon_t + \beta_1 \cdot \varepsilon_{t-1} + \cdots + \beta_q \cdot \varepsilon_{t-q} \tag{2.4}$$

[4] Weak or wide-sense stationarity is a weaker form of stationarity employed in signal processing, only requiring that the mean and the auto-covariance do not vary in time.

Where μ is the mean of the time series, β_1, …, β_q are the model parameters that need to be estimated, and ε_t, ε_{t-1}, …, ε_{t-q} are the white noise terms. Again, using the backshift operator (B), Eq. (2.4) can be rewritten as:

$$X_t = \mu + \left(1 + \beta_1 \cdot B + \cdots + \beta_q \cdot B^q\right) \cdot \varepsilon_t \tag{2.5}$$

It can be seen that the MA model is actually a linear regression of the present timeseries values against present and past white noise error terms, which are assumed to be mutually independent and to originate from the same distribution, most often a normal distribution, with location at 0 and constant scale, i.e. $\varepsilon \sim N(0,\sigma^2)$. For more details and information the reader is referred to the work of Shumway and Stoffer (2017).

2.1.3. Auto-Regressive Moving Average Model (ARMA)

An order p, q autoregressive moving average model (ARMA) constitutes a filter that allows the simulation of a vector y at a certain time by its past time histories and the past and present white noise processes. It is a mixed form, which combines the AR(p) and MA(q) models and is formed as follows:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i \cdot X_{t-i} + \sum_{i=1}^{q} \beta_i \cdot \varepsilon_{t-i} \tag{2.6}$$

Where $\{\varepsilon_t\}$ is a Gaussian white noise series with a zero mean and variance σ_z^2. The ARMA model is described in the book of Box and Jenkins (1976), where a method is derived, most commonly known as the Box-Jenkins method, to estimate the parameters of the model. An important notice is that in case the assumption that the error terms are independent identically distributed random variables sampled from a Gaussian distribution with zero mean is weakened, the properties of the model might change significantly. Applications of ARMA in offshore applications can be found in Appendix D.

2.1.4. Auto-Regressive Integrated Moving Average (ARIMA)

The autoregressive integrated[5] moving average model (ARIMA), and the method for assessing its parameters, were primarily developed by Box and Jenkins (1976), as a tool which allowed the prediction and control of time series. The ARIMA model is a generalization of the ARMA model described previously, and it is applied often in cases where the data show non-stationarity. In general, non-seasonal models are denoted as ARIMA(p,d,q), where p is the order of the AR model, d is the degree of differencing, and q the order of the MA model. Seasonal ARIMA models are denoted as $\text{ARIMA}(p,d,q)(P,D,Q)_m$, where m is the

[5] The "*integrated*" (I) part of the ARIMA model indicates that the current data have been replaced with the difference between their present and past values.

number of periods in each season, and (P,D,Q) refers to the autoregressive, differencing, and moving average terms of the seasonal part. It is important to notice that in an ARIMA model, the future value of a variable is assumed to be a linear function of several past observations and random errors. That said, the model's general form is given by:

$$\left(1-\sum_{i=1}^{p}\alpha_i\cdot L^i\right)\cdot(1-L)^d\cdot X_t=\delta+\left(1+\sum_{i=1}^{q}\beta_i\cdot L^i\right)\cdot\varepsilon_t \tag{2.7}$$

Where L is the lag operator[6], and d is the number of unit roots of the autoregressive polynomial. Generally, an ARIMA model can be thought as a particular case of an ARMA(p+d,q), and for that reason it can never be wide sense stationary when d>0. It has to be noted that the error terms ε_t are assumed to be independent, identically distributed variables sampled by a normal distribution with zero mean.

For forecasting purposes, the ARIMA model can be viewed as a combination of two models. One is a non-stationary:

$$Y_t=(1-L)^d\cdot X_t \tag{2.8}$$

While the second one is wide-sense stationary:

$$\left(1-\sum_{i=1}^{p}\alpha_i\cdot L^i\right)\cdot Y_t=\left(1+\sum_{i=1}^{q}\beta_i\cdot L^i\right)\cdot\varepsilon_t \tag{2.9}$$

More in line with Eq. (2.6), the general forecasting equation for an ARIMA model is presented below:

$$X_t=\mu+\sum_{i=1}^{p}\alpha_i\cdot X_{t-i}-\sum_{i=1}^{q}\beta_i\cdot\varepsilon_{t-q} \tag{2.10}$$

where the constant term μ is the average difference in X. Here the moving average parameters (β) are defined so that their signs are negative in the equation, following the convention introduced by Box and Jenkins (1976). Some authors and software (including the R programming language) define them so that they have plus signs instead. When actual numbers are plugged into the equation, there is no ambiguity, but it's important to know which convention each software uses when an output is produced.

Even though ARIMA models can be very flexible and are able to represent several different types of time series, the pre-assumed linear form of the model poses a major limitation. Since a linear correlation structure is assumed for the time series values, it is impossible for the ARIMA model to capture non-linear patterns. Usually, linear model approximations to the

[6] The lag operator (L) has the same property as the backshift operator (B); it operates on an element of a time series in order to produce the previous element.

complex real world phenomena are far from satisfactory. Zhang (2003) suggested that since real world time series most often contain both linear and non-linear patterns, the ARIMA model cannot deal with non-linear relationships by itself. It is a fact that the universal forecasting literature supports that more than one method (or simply hybrid methods) can model complex structures more accurately and that no method can be considered universally best in every situation (see e.g. Jenkins, 1982; Makridakis et al., 1982; Chatfield, 1996). Lastly, Khashei and Bijari (2010) proposed a hybrid method for time series forecasting, using ARIMA modelling and an Artificial Neural Network (ANN) to produce more accurate results, than using the methods individually.

In offshore engineering applications, the ARIMA model was tested and compared to an Artificial Neural Network, for the short-term prediction of the Caspian Sea surface water level (see Vaziri, 1997). It was concluded that the ARIMA is a useful tool for the forecasting of sea data time series, producing quite reasonable predictions in winter period. Supplementary it was suggested in the same work that the ARIMA short-term predictions could be incorporated in longer-term computer simulation model forecasts. Seasonal ARIMA models, as well as models including non-linear seasonal components, were used for short-term forecasting of real data series with missing data significant wave height) in the work of Delicado and Justel (1999).

For wind speed predictions and modelling, the ARMA and ARIMA models have been also widely used (see e.g. Torres et al., 2005; Kamal and Jafri, 1997; Erdem and Shi, 2011; Chen et al., 2010). Kavasseri and Seetharaman (2009) used an f-ARIMA[7] model, to make 24-hour forecasts, and proposed models whose results indicated significant improvements in forecasting accuracy. A hybrid ARIMA-ANN model was used for wind speed forecasting in three different regions of Mexico, by Cadenas and Rivera (2010), showing higher accuracy than the individual ARIMA and ANN models in the examined sites. A review on the forecasting methods for wind speed is given in the work of Lei et al. (2009).

In Matlab®, the Econometrics toolbox provides useful functions to create and employ an ARMA or ARIMA model. There are also functions to fit a model to current data and generate simulations. Extended information can be found in Matlab® online documentation[8]. In R, the `forecast` package provides useful tools for automatic ARIMA modelling.

[7] *Fractional*-ARIMA or f-ARIMA models are generalizations of the simple ARIMA model, which allow non-integer values for the differencing parameter.

[8] For ARIMA modeling see: https://nl.mathworks.com/help/econ/specify-arimap-d-q-models.html, while for ARMA modeling see: https://nl.mathworks.com/help/econ/arma-models.html.

2.2. Simple Linear Regression Modelling

In statistics, the simple linear regression is a linear regression model with a single explanatory variable, which is also called the independent variable. It states that the true mean of the dependent variable changes at a constant rate as the value of the independent variable increases or decreases. The simple linear regression model is formulated as follows:

$$Y = \beta_o + \beta_1 \cdot X + \varepsilon \tag{2.11}$$

where Y is the dependent variable, X the independent variable, ε the random error, β_0 is the y-intercept of the line y = $\beta_0 + \beta_1 \cdot x$, and β_1 is the slope of the same line. In case Y and X are assumed linearly correlated, Eq. (2.11) takes the form of a line:

$$Y = \beta_o + \beta_1 \cdot X \tag{2.12}$$

In offshore practice, the Y variable is usually chosen to be the observed significant wave parameter (wave height, wave period, etc.), and the X variable the numerically modelled wave parameter. The choice of an axis for the observed and modelled (predicted) data is actually of great importance, as it has been demonstrated in Pineiro et al. (2008). The fact that the slope and intercept of the model can be significantly different, when changing the position of the data on the axis, can be easily overlooked due to the similar r^2, i.e. the percentage of the response variable variation that is explained by a linear model.. Nevertheless, in the aforementioned work, it is shown that the observed data should be on the y-axis (dependent variable), while the numerically modelled data should take place in the x-axis (independent variable).

In case random errors are included, the following assumptions have to be made (see also Soukissian and Kechris, 2007; Altunkaynak and Ozger, 2004):

i. The random variables ε_i have the same variance Var[ε_i] = σ_ε^2, i = 1,2, …, n. In case this condition is not met, the regression parameters have to be estimated with the weighted least squares method.

ii. The error terms are uncorrelated to each other and uncorrelated with the independent variable.

iii. The random variables ε_i are normally distributed with zero mean value. This assumption implies that the errors are independent and that the dependent variable Y is also normal, which leads to the conclusion that Y_i and Y_j are uncorrelated and independent as well.

The observations of the independent variable X are generally assumed to be measured without error (see Rawlings et al., 1998; Webster, 1997), which is in reality met very rarely. In order to account for an error in both variables, Soukissian and Kechris (2007) proposed a more general methodology, the so-called Errors in Variables (EIV), achieving a minor correction of the data prediction, compared to the simple model. An example of the implementation of linear regression for wave height forecasting, using wind speed as the independent variable, is given by Altunkaynak and Ozger (2004), while Rusu and Guedes Soares (2016) applied a linear regression model to an ensemble of measurements and the corresponding numerical model forecast (SWAN). In the latter work, the regression parameters are updated in every time step, only to be used to correct/improve the produced forecast for the next time step at a local scale. Finally, linear regression has been used as a data assimilation technique, to perform corrections on the results of a numerical wave model (SWAN) in the Romanian nearshore using satellite data, by Raileanu et al. (2015), showing promising results.

2.3. Artificial Neural Networks (ANNs)

The function of an Artificial Neural Network (ANNs) is similar to a biological neural network, encountered for instance in the human brain. ANNs constitute soft computing tools, which roughly mimic the ability of the human mind to decisively employ modes of reasoning and recognize certain patterns. The basis of the network is the neuron, which receives an argument formed as a sum of a weighted input and bias, and produces an output using a transfer function. Several neurons can be combined in a layer, while a network can be composited by more than one interconnected layers. The concept of ANN was first introduced by W.S. McCulloch and W. Pitts (1943), and since then, many models were developed based on them (e.g. see Rumelhart et al., 1986; Peterson and Anderson, 1987; Psaltis et al., 1988).

A commonly used type of neural network in engineering applications (see e.g. French et al., 1992; Yeh et al., 1993; Kasperkiewiecz et al., 1995; Grubert, 1995; Thirumalaiah and Deo, 1998; Thirumalaiah and Deo, 2000; Deo and Kiran Kumar, 1999) is the 3-layered, feed forward network, given in Figure 1. Its basic advantage is the ability to approximate any non-linear mathematical dependency structure. For more information on the function and theoretical basis of ANN, the reader is referred to Appendix D.

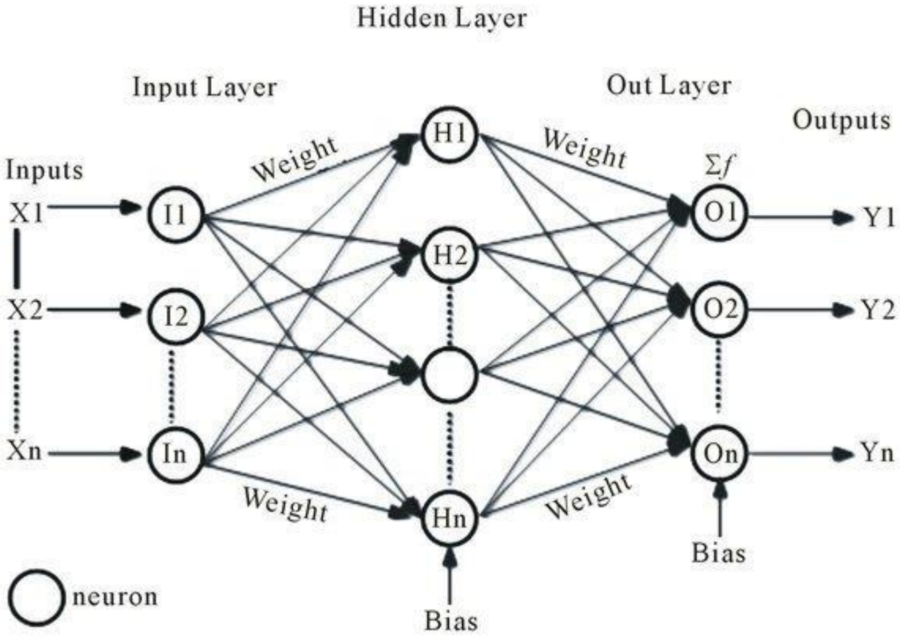


Fig. 1. Example of a 3-layered ANN (Source: Google Images for Artificial Neural Networks).

In ocean engineering, ANNs have been used in order to predict accurately the values of met-ocean parameters (see e.g. Deo, 2010). Deo and Sridhar Naidu (1999) achieved real time wave height forecasting (lead times 3-24 hrs) using neural networks, while comparing the results of different training algorithms. All of the training algorithms appeared to have similar correlation coefficients, but the cascade correlation algorithm was the fastest in terms of training. They also concluded that neural networks have many practical advantages compared to statistical models, like the AR, since they are more flexible and adaptable, but unlike numerical models, they are site specific and applicable only when the wave data are gathered on site. Simple 3-layered neural networks, incorporating wind speed data in order to produce significant wave height and average wave period predictions, were used satisfactory by Deo et al. (2001). They came up with the conclusion the network can represent successfully the relation between wind and waves in deep water for large prediction intervals, such as a week. For shorter time intervals, the rapid variations in wind measurements make the training procedure difficult, so a separate training for extreme and mild weather seasons is beneficial. Tsai et al. (2002) used a neural network, with a back propagation training algorithm, to conduct wave height and period forecasting or data supplement between neighbouring stations, concluding that the ANN model performs well for both procedures when using short-term wave data (see also Mandal et al., 2005). Stochastic (ARMA and ARIMA) and neural approaches were compared in terms of their operational forecasting ability in the work of Agrawal and Deo (2002). For short intervals (3-6 hrs), the neural networks proved to have a distinct superiority, while both models provided satisfactory results in long range predictions (12-24 hrs). The aforementioned conclusion is verified by the work of Londhe and Panchang (2005), which provided one day wave forecasts, both in online and offline modes, using buoy data and an ANN. Additional examples of the use of ANN’s in met-ocean parameter estimation, with the incorporation of field observations, can be found in the works of Makarynskyy (2006) and Makarynskyy et al. (2005), where 3-layered feed forward networks with a non-linear differentiable log-sigmoid transfer function in the hidden layer and a linear transfer function in the output layer perform well in ocean wave parameter simulation. Regional wave height predictions using sequential learning neural networks was conducted by Krishna Kumar et al. (2017), demonstrating the performance advantages of neural networks as opposed to other soft-computing techniques.

A different approach, than the ones followed in the aforementioned references, is the potential coupling of an ANN and a numerical model (e.g. SWAN, Delft3D, or WAM). Krasnopolsky et al. (2002) proposed a neural network technique, intending to improve the efficiency and accuracy of numerical ocean models. They showed that the use of neural networks is a fast, computationally beneficial, and accurate approximation of the continuous mappings[9], which can essentially represent the parameterization[10] of several types of physical processes, thus accelerating and simplifying time-consuming calculations in environmental numerical models. Makarynskyy (2004) indicated that wave predictions (wave heights and periods) can be improved using ANNs. He proposed two procedures for an oceanic and a sea site; one using only the initial numerical model results (leading times from 1 to 24 hrs), and then another one merging site measurements and forecast data. The forecasts were improved in the case of the oceanic site, but their accuracy continuously reduced with larger time intervals for the sea site, where the waves are temporally inconstant. Improved numerical hindcast outputs can be produced for various ocean parameters, while incorporating neural networks, as presented by Makarinskyy (2005). The techniques displayed in the aforementioned work can be embedded into operational procedures of model output corrections, proving the applicability of the neural methodology for an area close to the position where the online observations are made. A proper combination of numerical model results and ANNs can reduce the cost and time, accompanied with the desired accuracy. The literature points out occasions where numerical wave results are produced, using wind data as input, and then channelled to an ANN, which is designed and trained according to them, in order to accurately and sufficiently provide location specific ocean wave predictions (see e.g. Malekmohamadi et al., 2008).

[9] Continuous mappings are synonymous to continuous dependencies between two vectors.

[10] The parameterization of the physical processes governing the ocean environment generally requires complex, non-linear mathematical expressions, based on statistical or physical models. Initially, Krasnopolsky et al. (2002) indicated that those parameterizations can be considered continuous mappings.

The incorporation of ANNs, for the prediction of the forecast errors, and data assimilation techniques into a wind-wave model for wave forecasting was conducted by Zhang et al. (2006). It was shown that the accuracy of the forecasting results is enhanced, always taking into account the seasonal variation for the wind data, which is significant. A proposed solution is the training of the neural network in specific seasons, in case highly different (in nature) seasonal data need to be examined[11]. Londhe et al. (2016) predicted the forecast errors of a numerical model using a neural network, and then added or subtracted them to the numerical forecast, increasing the accuracy of 24-hr wave predictions considerably. As input to calculate the error between forecasted and observed data, the wave height from the present and the previous time steps was used. Evidence provided by the literature (see Deshmukh et al., 2016) points out that a sufficiently trained neural network that predicts the errors for future time steps, coupled with a numerical model, offers more sustained prediction performance for wave parameters, than a standalone neural network model, for time intervals varying from 3 to 24 hrs in all seasons. In such an occasion, the neural network is trained using buoy (or generally observed) and numerical data from past time steps.

In terms of real-time forecasting, Jain and Deo (2007) showed that really accurate wave forecasts can be achieved by neural networks for short (3 hr) and long (24 hr) intervals, if the amount of annual data gaps is around 2%. Yet, they noticed that the prediction accuracy decreases in harsh environments. Missing wave heights were estimated in real-time basis, using wave height data from different locations, ANNs, and Genetic Programming (GP), in the work of Londhe (2008). It was found that the soft computing methods performed reasonably well, providing in some cases more accurate results than a numerical model. The GP approach yielded better results in extreme events, but both methods offered satisfactory results. For more information on ANNs' efficiency and accuracy advantages over other soft-computing methods (e.g. GP and model trees) in real-time wave forecasting, when using wind time history and a numerical model, the reader is referred to Jain et al. (2011). There, it is stressed out that due to the very large number of available options in terms of training, architecture, and flexibility, ANN's are a safe choice when it comes to wave parameter prediction, even if all of the presented soft-computing methods provide sufficient results.

[11] It is trivial that different seasons produce events of highly different intensity. For instance, often winter events (storms) are more violent than summer or spring events, and as a result inter-seasonal input data will most probably affect the design and function of the ANN at hand, causing inaccuracies.

2.4. Copula modelling

Multidimensional phenomena, like the ones included in the ocean environment, require the joint modelling of several random variables. The main limitation of traditional approaches is that the individual behaviour of two variables must be characterized by the same parametric family of univariate distributions. This restriction can be avoided with the incorporation of copula models (see Genest and Favre, 2007).

Copulas are multivariate distribution functions defined on the unit cube $[0,1]^n$, with uniformly distributed marginals. For more information on the mathematical background and definition of copulas the reader is referred to Joe (1997), Nelsen (1999), or Embrechts et al. (2001). The roots of the copula approach in dependence modelling are in the representation theorem derived by Sklar (1959), which states that the joint cumulative distribution function (CDF) H (x, y) of any pair of continuous random variables (X, Y) may be written in the following form:

$$H(x,y) = C\{F(x), G(y)\},\ \ x,y\ \in R \tag{2.13}$$

where F(x) and G(y) are the marginal distributions, and C: $[0, 1]^2 \rightarrow [0, 1]$ is the copula. A valid model can arise for the above equation for (X, Y) whenever C, F and G are chosen from given parametric families of distributions:

$$F\ \in\ (F_\delta), \qquad G \in \left(G_\eta\right), \qquad C\ \in\ (C_\theta)$$

A large variety of copulas can be used to model joint distributions with different characteristics. The Gaussian, the Gumbel and the Clayton copulas are the most common families (see Appendix D), since they can model different tail asymmetries of the joint distributions and have been used in many applications, such as in finance (see e.g. Aas et al., 2009).

According to the theory described previously, the copula representation of a joint distribution function H (x_1, x_2, …, x_d) is C ($F_1(x_1)$, …, $F_d(x_d)$), where C constitutes a unique cumulative distribution function with uniform margins on (0,1). A dependence structure of a multivariate distribution is well represented by a specific parametric family C_o of copulas, when the following hypothesis is valid:

$$H_o: C \in C_o$$

In order to fit a copula to a structure there are various goodness-of-fit tests, like the Cramer-von Mises statistic, or the Kolmogorov-Smirnov statistic (for more information see Genest et al., 2009). As stated by Genest et al. (2009), because the marginal distributions of random variables are often unknown, the only reasonable choice is to base the tests on a sample from the underlying copula C, a collection of pseudo-observations U_1 = (U_{11}, …, U_{1d}), …, U_n = (U_{n1}, …, U_{nd}), deduced from the ranks:

$$U_{ij} = \frac{R_{ij}}{n+1} \tag{2.14}$$

where $R_{i,j}$ are the ranks of the observations.

The most well-known goodness-of-fit tests are “rank based”. The Cramer – von Mises statistic is based on the empirical copula, and focuses on comparing its differences with a parametric one (Gaussian, Gumbel, Clayton, t-student, etc.). Information on how its is formulated can be found in Appendix D.

Large values of the aforementioned statistic lead to rejection of the H_o hypothesis. The most appropriate parametric copula to represent the dependence between two variables is the one with the smallest sum of the square differences.

Leontaris et al. (2016) adopted the technique of calculating the Pearson correlation for the upper and lower quadrant of the actual observations, transformed to standard normal N(0,1) margins, in order to model the dependence between significant wave height and wind speed. As described by Joe (2014), $Z_j = \Phi^{-1}(U_j)$, for j = 1, …, d, are the standard normal transforms of the pseudo-observations. After dividing the standard normal transforms of observations into four quadrants, for positive correlation, the semi-correlations of the upper (NE) and lower (SW) quadrants are respectively formulated as follows:

$$\rho_{ne} = \rho(Z_1, Z_2) \cdot Z_1 > 0, Z_2 > 0 \tag{2.15}$$

$$\rho_{sw} = \rho(Z_1, Z_2) \cdot Z_1 < 0, Z_2 < 0 \tag{2.16}$$

The aforementioned correlations indicate the existence of tail asymmetry, which is present in case of significant difference between them (Joe, 2014). Also, if the values of the semi-correlation coefficients are larger than the overall Pearson correlation coefficient, an indication of tail dependence in given; for more information on tail dependence the reader is referred to Embrechts et al. (2003). An example of this procedure, in the context of traffic load measurements, can be found in Morales-Napoles and Steenbergen (2014).

Numerous studies exist, focusing on the dependence modeling of various parameters in engineering and financial practice (see e.g. Vaz de Melo Mendes and Martins de Souza, 2004; Zhang et al., 2015; Accioly and Chiyoshi, 2004). Salvatori et al. (2013) used copulas to estimate the joint distributions of wave characteristics, such as the wave height and period, while Yang and Zhang (2013) used copulas to estimate the joint distribution of wind speed and significant wave height without taking into account the autocorrelation, which is essential when time series are required. A similar approach was followed by Leontaris et al. (2016), where the authors used copula-based environmental time series to include environmental uncertainties in offshore wind farm operations and to produce realistic time series concerning met-ocean variables. It was also found, that dependently constructed synthetic[12] time series provide better results compared to the case where only observed time series are used, while the importance of including the dependence between wind speed and wave height, in order to construct independent time series, is stressed out. Jane et al. (2016) proposed a copula-based approach for the estimation of wave height records through spatial correlation, in coastal environments. The predictions provided in this study were equally accurate to numerical model results (SWAN)[13].

In Leontaris et al. (2016) and Jane et al. (2016), it is suggested that Vine copulas (see also Brechmann and Czado, 2015; Smith, 2015; Bedford and Cooke, 2001, 2002; Joe, 1996; Kurowicka and Cooke, 2006), which capture the dependence structure between a set of variables by arranging a series of bivariate copulas in a tree structure, can provide larger modeling flexibility and describe better the existent multivariate distributions. Nevertheless, it has been proven that in a trivariate setting, the simpler elliptical copulas, and especially the student-t, are more than capable of capturing the dependence structure between a set of random variables (see Ma et al., 2013; Poulomi and Reddy, 2013; Wang et al., 2010; Wong et al., 2010), thus they can also be explored.

Supplementary, a Pair-Copula approach has been introduced to construct a multivariate model, which can fully consider the dependence characteristics of wind power forecasting errors (Hu et al., 2017). Kazianka (2012) developed a set of open source Matlab® functions, which provide copula-based spatial analysis for non-Gaussian and extreme value data. Joint models for significant wave height and wave period were investigated by Vanem (2016), concluding that the model selection remains a challenge and it is difficult to unambiguously identify the one that better describes dependence. Finally, the author also suggests that the correlation in the data between extreme significant wave heights and wave period increases in a future climate, and might influence the risk related to ocean environments.

[12] *Synthetic time series* are created by a combination of observed and model-developed time series.
[13] For more information on SWAN the reader is referred to: http://swanmodel.sourceforge.net/.

2.5. Bernstein Stochastic Interpolation

The stochastic Bernstein method, which should not be confused with the Bernstein polynomials, is a novel and significantly improved non-polynomial global method of signal processing, suitable for erratically distributed input data interpolation and approximation (see Kolibal and Howard, 2006). Generally, stochastic interpolation represents a family of methods, posing clear advantages over polynomial interpolation methods.

The theoretical base of the method can be found in Kolibal and Howard (2006), while for its implementation in practical applications the reader is referred to Kolibal and Howard (2006, 2008). In a nutshell, considering the function $f(x)$ sampled at points $x_k \in [0,1]$: $f(x_k) = y_k$. The natural continuum extension of the Bayesian polynomials on the set of data $\{(x_k, y_k)\}$ is expressed by the following sum:

$$K_n(x) = \sum_{k=0}^{n} \frac{y_k}{2} \cdot \left[erf\left(\frac{z_{k+1}-x}{\sqrt{\sigma(x)}}\right) + erf\left(\frac{x-z_k}{\sqrt{\sigma(x)}}\right) \right] \tag{2.17}$$

$$\sigma(x) = \frac{2}{n} \cdot x \cdot (1-x) \tag{2.18}$$

$$z_k = \begin{cases} -\infty & k = 0 \\ (x_{k+1} + x_k)/2 & k = 1,2,\dots,n-1 \\ \infty & k = n \end{cases} \tag{2.19}$$

where f is assumed piecewise constant and equal to y_k in (z_{k-1}, z_k). In most cases it is convenient to choose a constant $\sigma(x)$ throughout the interval, which makes the method non polynomial. The approach offers freedom in terms of scaling parameters, due to its generality. The natural continuum extension (K_n) consists of matrix vector multiply, where the $n \times n$ matrix is denoted as $A_{nn} = (\alpha_{jk})$. Hence, for a constant σ:

$$\alpha_{jk} = \frac{1}{2} \cdot \left[erf\left(\frac{z_{k+1}-x_j}{\sqrt{\sigma}}\right) + erf\left(\frac{x_j - z_k}{\sqrt{\sigma}}\right) \right] \tag{2.20}$$

As a result, $K_n(x_k) = A_{mn} \cdot y$, where $y = (y_1, \dots, y_n)$ and A_{mn} is a row-stochastic matrix whose kth row is generated using Eq. (2.20). A deconvolution operator on the data is A_{nn}^{-1}, thus an elegant solution for the interpolation of data is provided by $A_{mn} \cdot A_{nn}^{-1} \cdot y$. Different choices of σ in A_{nn} and A_{mn} would yield a range of data representation forms, ranging from pure smoothing, interpolation, and deconvolution.

Kolibar and Howard (2006) concluded that stochastic interpolation methods built around Bernstein functions will fit complex two dimensional surface data, and as a result they could find application in engineering practice. In contrary to polynomial methods, the stochastic

Bernstein interpolation allows the use of any number of points, in case computational stencils are used, allowing for infinitely differentiable surfaces, where necessary.

In general, to the writer's best knowledge, the Bernstein stochastic interpolation has not been used in civil or hydraulic engineering applications. Examples of its application can be found, however, in other fields, such as visual surveillance systems (see e.g. Kim and Ko, 2011), and computer graphics (see e.g. Seyfarth et al., 2006). Nonetheless, stochastic interpolation (Gaussian process regression) as a technique has been used for ocean parameters. For instance, Scotto and Guedes Soares (2007) used Bayesian inference to make long term predictions of the significant wave height, concluding that the procedure provides adequate flexibility, as well as consistent forecasting results.

Finally, it has to be noted that the Bernstein stochastic model is not available in a Matlab® toolbox, and as a result any effort to implement has to be improvised by the modeller. Matlab's Statistical toolbox includes, however, Gaussian Process Regression Models (see Rasmussen and Williams, 2006), able to provide forecasts with uncertainty intervals, and compute the regression error. For more information the reader is referred to the Matlab® online documentation[14].

2.5. Bayesian Networks (BNs)

2.5.1. General

Bayesian Networks (BNs) are graphical models, which allow the representation of a probability distribution over a set of random variables (see Jensen and Nielsen, 2007; Morales-Napoles et al., 2013; Hanea et al., 2015; Weber et al., 2012). They consist of a directed acyclic graph (DAG) built on discrete (discrete networks), or continuous (continuous networks), or both kinds (hybrid networks) of random variables (X_1, X_2, …, X_n), and a set of (conditional) distributions. A DAG is constituted by a set of nodes, that represent random variables, and a set of arcs, in a way that a directed cycle cannot be created. Within the graph, an ordering of the variables can be established, given the directionality, which provides information on the sampling order, i.e. the order which has to be followed so that a sample can be taken from this joint distribution. As a result, some of the nodes are characterized as "parents" and others as "children", depending on whether they precede or success the node of interest (see Figure 2). A marginal distribution is assigned to each node with no parent, and a conditional distribution is associated with each child node, which

[14] See https://nl.mathworks.com/help/stats/gaussian-process-regression-models.html for more information.

provides quantitative information[15] about the dependences between the variables. An important notice is that each variable X_i is conditionally independent of its non-descendants, given its parents in the DAG. This constitutes the Local Markov property:

$$P(X_v = x_v \mid X_i = x_i) = P\big(X_v = x_v \mid X_j = x_j\big) \qquad (2.21)$$

where X_i is not a descendant of X_v, and X_j is a parent of X_v. It has to be noted that the set of parents is a subset of the set of non-descendants because the graph is acyclic.

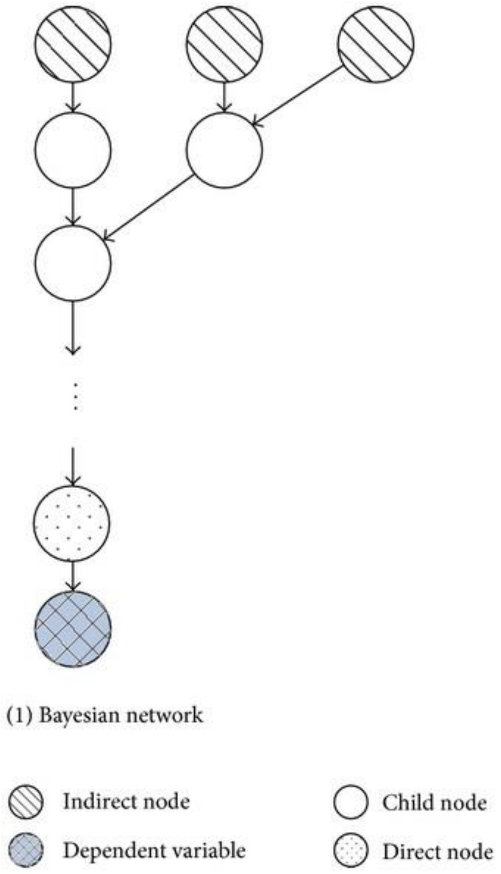


Fig. 2. Relations between nodes in a Bayesian Network.
(source Google Images)

Denoting the parent nodes as $Pa(i)$, the conditional probability function (i.e. the joint density of X_1, X_2,…, X_n) of a variable given its parents is formulated as follows:

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^{n} f_{X_i|X_{Pa(i)}}\big(x_i \big| x_{Pa(i)}\big) \qquad (2.22)$$

[15] The quantitative information can be either retrieved from data, or from expert judgment.

where $f_{X_1,\dots,X_n}$ is the joint density of the n variables, f_{X_i} denotes their marginal densities, and $f_{X_i|X_j}$ are the conditional densities.

BNs are quantitative tools, able to evaluate conditional probabilities between variables, and at the same time constitute valuable conceptual models, since they visually represent independent and dependent variables in causation relationships (see Chen and Pollino, 2012; Palmsten et al., 2014; Stewart-Koster et al., 2010). The principles of BNs as a modelling tool are described thoroughly in Pearl (1988) and Jensen (1996). The main property of the BNs is inference, which constitutes their ability to provide updated distributions, given observations, but also characterization of the relationship between the variables. Generally, the simple visualization of the complicated relationships between the random variables, as well as their polyvalence, i.e. the ability to deal with issues such as prediction, diagnosis, optimization, data analysis of feedback experience, and model updating, makes the use of BNs appealing.

Many applications of the BNs on dependability, risk analysis and maintenance can be found in Weber et al. (2012) and Medina Oliva et al. (2009). Most of the applications, however, use networks consisting of nodes that represent discrete random variables. Those networks are characterized as discrete BNs and suffer for serious limitations, since the provided discrete representation of variables for many important problems is inadequate.

2.5.2. Hybrid Bayesian Networks

Many domains require information about the joint behaviour of both discrete and continuous variables, hence they are called hybrid (see Langseth et al., 2009). As a result, the existence of continuous and discrete variables makes a BN hybrid as well. One way of dealing with Hybrid BNs, or HBNs hereon, is the use of the conditional Gaussian model, as described in Shachter and Kenley (1989), Phillips (1998), and Lauritzen (1992). This form of discrete-normal HBNs suffers from the restriction to the joint normal distribution. Exact inference algorithms for discrete BNs (see e.g. Pearl, 1988; Zhang and Poole, 1994) have been extended to discrete-normal HBNs, while approximation algorithms are also available (see e.g. Lerner, 2002).

When the joint normality assumption is not appropriate, the most common method to deal with continuous variables is discretisation. When discretizing a continuous variable, a large number of partitions should be used to obtain a reasonable approximation. However, in complex structures this approach leads to extremely large conditional probability tables (CPTs) that have to be quantified in a defendable way. The large amount of required data to achieve this quantification is rarely available, hence a small number of partitions is used in order to approximate continuous variables. Additionally, even when the quantification can be done, exact inference might prove unfeasible due to the large number of the required calculations (see Murphy, 2002). Further, many discretisation techniques have only local application, thus they often fail to account for the entire dependence structure of the variables, leading to poor control of the global error in the model (see Kozlov and Koller, 1997; Langseth et al., 2009).

Other methods that deal with HBNs include Markov Chain Monte Carlo simulations, variational methods, enhanced BNs (see Straub and Der Kiureghian, 2010), and mixtures of truncated exponentials (MTEs) or polynomials (see e.g. Shenoy and West, 2011). An extensive review of some of these methods is given by Langseth et al. (2009, 2012), while Hanea et al. (2015) briefly presented some advantages and disadvantages of each one.

The methods mentioned so far demonstrate the same pathogenicity in terms of their quantification, since in most of the cases this process is not transparent, reliable, and defensible. Besides these setbacks, the aforementioned “classic” approaches are used extensively in real life problems. Plant and Holland (2011) demonstrated how a discretized BN model can be used to provide accurate predictions of wave-height evolution in the surf zone of a coastal region, given very sparse or inaccurate boundary-condition data. The accuracy of the predictions was similar to this of a numerical model, while it was noticed that more consistent forecasts and uncertainties were obtained if the model parameter errors were included, as a source of input uncertainty. The authors, in a companion work (see Plant and Holland, 2011), prove that the BN model can be used effectively for predicting offshore wave heights, given limited wave height observations from an onshore location (inverse procedure). It is stressed out that a major advantage of the Bayesian Networks is that they are simpler than a detailed numerical model, providing accurate wave height forecasts, accompanied with uncertainty estimates for all predictions, while simultaneously estimating model parameters; for more successful applications of BNs in coastal engineering practice the reader is referred to Pelhekke et al. (2016), Kroon et al. (2017), Jager et al. (2017), Wilson et al. (2015). Supplementary, Palmsten et al. (2013) showed that BNs can be transferred to new settings, if the observations between study sites show adequate similarity.

An extended review of the use of Bayesian Networks in environmental modelling is given in Aguilera et al. (2011), where the authors intone that BNs are recommended for studies with missing values, which is common in ocean environments. Generally, there is no extensive literature on applications regarding specifically met-ocean variables, at an offshore site.

Nevertheless, the extended number of applications where the BNs have been used, some of which involve parameters found in open seas (such as the wave height, wave period, current direction, etc.), provide a stable base so that a model for ocean variables can be established. An example of the use of BNs in ocean wave height prediction can be found in the work of Malekmohamadi et al. (2011), where a useful comparison between BNs and ANNs is also given. In the same work, it is noted that when the probability mass function of wave parameters and the confidence intervals of the forecast are important, BNs can be used efficiently.

A concluding remark regarding the nature of the prediction errors in time is that they can be either homoscedastic or heteroscedastic. This, however, makes a significant difference in the modelling strategy. In similar fashion, the variance of the errors can be constant or variable in space. In this case, there is variation in the parameters that link the predicted error at a single location to the preceding error values at that same location. A Bayesian modelling approach can allow spatial heterogeneity of the error. Furthermore, the fact that the spatial spreading problem is characterized usually by a small set of available observation data points in space, could favour a BN modelling method over others.

2.5.3. Non-Parametric Bayesian Networks

A method that handles HBNs was introduced by Kurowicka and Cooke (2004), and extended by Hanea et al. (2006, 2010, 2015), is the Non-Parametric Bayesian Networks, or simply NPBNs. This methodology was initially developed purely for continuous BNs. The NPBNs associate nodes with random variables for which marginal distributions were assumed, and arcs with one parameter conditional copulas (see Joe, 1997). The conditional copulas, alongside with the one-dimensional marginal distributions and the conditional independence statements implied by the DAG, uniquely determine the joint distribution, making the NPBN specification consistent (Hanea et al., 2006). The marginal distributions can either be estimated from data or elicited by experts (see Cooke, 1991). In most cases, the empirical marginal distributions are used, but parametric forms can also be fitted. The conditional copulas incorporated are parameterized by constant conditional rank correlations (Spearman's), which can be calculated from data, as well as obtained from expert judgment (see Morales et al., 2008). The rank correlation has numerous attractive properties, the most important of which are the ability to measure monotone dependence, rather than just linear, and the fact that it is independent of the marginal distributions. Associating the arcs $i_{p(i)-k} \rightarrow i$ with the conditional rank correlations, and assuming $Pa(i) = \{i_1 \dots i_{p(i)}\}$:

$$\begin{cases} r_{i,i_{p(i)}} & if\ k=0 \\ r_{i,i_{p(i)-k}|i_{p(i)},\dots,i_{p(i)-k+1}} & if\ 1 \leq k \leq p(i)-1 \end{cases} \tag{2.23}$$

The assignment is vacuous in case $\{i_1 \dots i_{p(i)}\} = \emptyset$. Hanea et al. (2015) presented a theorem, showing that these assignments are algebraically independent and uniquely determine the joint distribution for a particular choice of copula. The authors also proposed a modification of the aforementioned theorem, allowing for various types of copula, after specifying the conditional independence statements as independent copula, instead of zero rank correlations. In this case, the conditional rank correlations associated with the arcs could be realized by any copula which realizes all correlations $[-1,1]$. That way, an NPBN could be quantified with a mixture of conditional independent copula and t-copula, with different tail dependence for each pair of variables, allowing the modeller to capture phenomena involving dependent extreme values.

The quantification of NPBNs is actually brought to the quantification of marginal distributions, the number of which is equal to the number of variables, and of conditional dependence parameters, equal to the number of arcs existing in the NPBN. Assuming that the DAG of a NPBN is known, and that data is available, all of the above can be estimated. In case of data scarcity, expert judgment is a necessity. Assuming, now, that the DAG is unknown as well, and that is form is to be determined by data, Hanea et al. (2010) proposed a structure learning algorithm from an ordinal multivariate data set, which may contain a large number of variables. The major assumption of this algorithm is that the joint distribution has to be a normal copula, thus the rank dependence structure of the variables is that of a joint normal distribution. The one dimensional marginal distributions are retrieved directly from the data.

Validating the learned NPBN model involves two steps; validating that the joint normal copula adequately represents the multivariate data, and verifying that the NPBN is an adequate model of the saturated graph. The validation procedure requires an overall measure of multivariate dependence, which could be provided by the determinant of the rank correlation matrix (see Hanea et al., 2015). Hanea et al. (2010) stated that the major reason for choosing the aforementioned determinant as s a multivariate dependence measure is that an approximation of it factorizes on the arcs of the NPBN.

After the quantification, the joint distribution becomes attainable, but the only way to stipulate it is by sampling (see Hanea et al., 2006). The independent copula has to be used to represent the absence of an arc, and any one-parameter copula could be used to realize the rank correlations associated with the arcs. However, when arbitrary copulas are used, the sampling procedure involves numerical evaluations of multiple integrals, which constitutes a large disadvantage of NPBNs in real-time decision support problems, as the one elaborated in this thesis. Nonetheless, this problem can be solved by assuming a normal copula and transforming each marginal distribution to a standard normal. That way, a joint normal distribution with the same rank correlation structure as the original one can be obtained, thus providing a fast and efficient procedure; examples and comparisons are given in Hanea et al., 2006.

Different ways of performing inference in NPBNs are described in Hanea et al. (2015). The only way that is going to be described here is the use of normal copula to realize the rank correlations, which is the fastest and can be seen as an exact propagation method. Briefly, the transformation of the marginals to standard normal essentially leads to a joint normal distribution, with any conditional distribution normally distributed with known mean and variance (see Whittaker, 2009), thus enabling inference to be performed analytically. Transforming backwards, using the inverse distribution function of the variables and the standard normal distribution function, provides the conditional distributions of the original variables (see also Hanea et al., 2006).

Applications of NPBNs in engineering practice, and more precisely in risk analysis applications, reliability of structures, material properties, traffic predictions, and flood protection can be found in Hanea et al. (2015), Worm et al. (2011), Morales - Napoles et al. (2013), and Morales-Napoles and Steenbergen (2014). In conclusion, the biggest advantage of the NPBN is that it can handle a large amount of mixed (discrete and continuous) variables in a short time, while the quantification procedure requires a small number of parameters. This makes the use of NPBNs in an offshore ocean environment quite possible. It should be emphasized, however, that the normal copula, which accelerates the whole procedure significantly, may not be appropriate for all real-life applications.

2.5.4. Implementation of BNs in Matlab® and R

In order to use Bayesian networks in Matlab®, Kevin Murphy (2001) created the Bayes Net Toolbox[16], an open source package able to support many types of conditional probability distributions for both static and dynamic BN modelling, with exact or approximate inference. Some setbacks in the use of the aforementioned toolbox are that it is relatively slow, it has little support for undirected models, and it does not support online inference or learning. In R, the `bnlearn` package[17] provides tools for learning the graphical structure of Bayesian Networks, estimate their parameters and perform some useful inference. Also, the package

[16] The Bayes Net Toolbox can be found in: https://github.com/bayesnet/bnt.

[17] See also: http://www.bnlearn.com/.

supports both discrete and continuous variables, while assuming that the latter ones are jointly normally distributed.

An alternative to the aforementioned tools could be the use of UniNet®, a standalone uncertainty analysis software package developed by TU Delft (see Morales-Napoles et al., 2013), which implements NPBNs under the normal copula assumption. Through the UninetEngine classes and functions[18] the user is able to call UniNet's engine from Matlab® in order to create and use a model for a specific engineering problem (such as the problems mentioned above).

2.6. SWAN

SWAN is a third-generation wave model, developed at Delft University of Technology, which computes random, short-crested wind-generated wave in coastal regions and inland waters. The model accounts for a variety of physics that include (1) wave propagation in time and space, shoaling, refraction due to current and depth, frequency shifting due to currents and non-stationary depth, (2) wave generation by wind, (3) whitecapping, bottom friction and depth-induced breaking, (4) three- and four-wave interactions, (5) dissipation due to aquatic vegetation, turbulent flow and viscous fluid mud, (6) wave-induced set-up, (7) propagation from laboratory up to global scales, (8) transmission through and reflection against obstacles, and (9) diffraction.

SWAN provides output quantities in numerical files containing tables, maps and timeseries. Those quantities include the (1) one- and two-dimensional spectra, (2) significant wave height and wave periods, (3) average wave direction and directional spreading, (4) one- and two-dimensional spectral source terms, (5) root-mean-square of the orbital near-bottom motion, (6) dissipation, (7) wave-induced force, (8) set-up, (9) diffraction parameter, etc. A limitation of the SWAN numerical model is that it does not account for Bragg-scattering and wave tunnelling.

2.7. Concluding Remarks

All of the previously described statistical and stochastic techniques will be compared in terms of performance, while making predictions for hydrodynamic variables in offshore environments. The main focus is the Bayesian Network models, which have not been used extensively in similar applications, but can incorporate a variety of variables, granting also estimates for the underlying uncertainty. The free availability and open-source nature of the tools provided by the `bnlearn` package in R, which facilitates the creation and use of BN

[18] Full documentation can be found at www.lighttwist.net.

models in automated procedures, make them attractive for this research. The same holds for the case of the ARIMA models to be employed, which have to be completely automated, due to the real-time nature of the application (`forecast` package in R). Certainly, the assumption of the multivariate normality included in the `bnlearn` package for the marginals and the conditional distribution of the variable of interest, is a matter of discussion in the chapters to follow, where the level of sufficiency of the underlying assumptions will become evident.

The complete methodology, on how the different error correction models, already existent in Meteo Dashboard (ANN, Linear Regression, Bernstein Stochastic Interpolation, and Copula) or newly created (Bayesian Networks and ARIMA), can be found in the following chapters (Chapter 4).

3. The Data

To move a step further, and evaluate the performance of the error correction model as a whole, data were retrieved from stations deployed in the Irish Sea (see Figures 3 and 4). The measurement stations, which are actually wave rider buoys and meteorological masts, are adjacent to the wind farms of Gwynt-y-Mor[19] and Rhyl Flats[20], located within the Liverpool Bay. The received datasets consist of measurements of hydrodynamic and meteorological data, obtained between 01-09-2012 to 31-01-2018. The aforementioned area and time interval will be the focus of this research.

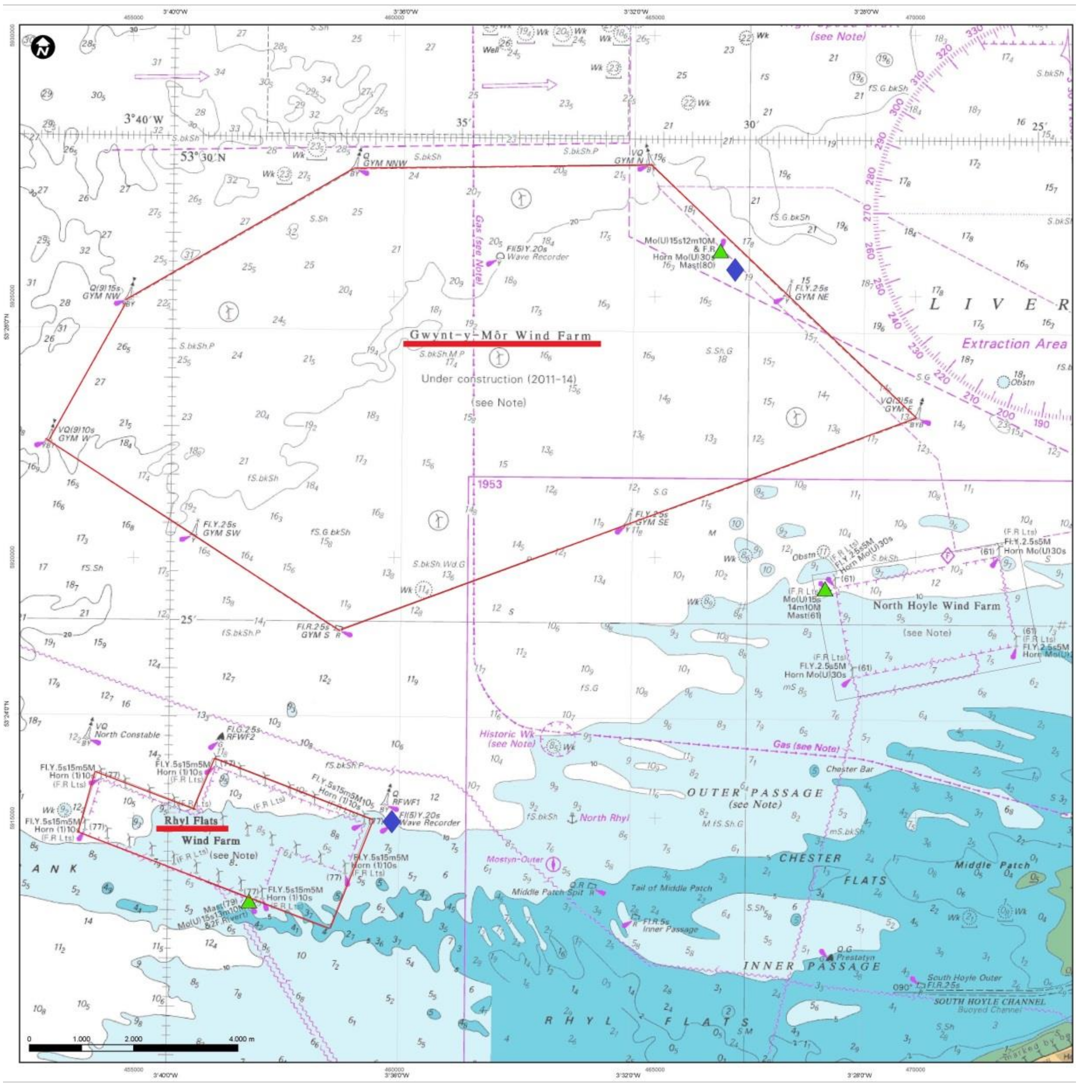


Fig. 3. Map of the Irish Sea wind farms (underlined with red) under consideration (measurement stations visible as blue diamonds).

[19] Gwynt-y-Mor Offshore Wind Farm (53°27′N 03°35′W) is located off the coast of North Wales and is the 2nd largest operating wind farm in the world (160 wind turbines).

[20] Rhyl Flats Offshore Wind Farm (53°22′N 03°39′W) is a 25 turbine wind farm, located approximately 8 km north east of Llandudno in North Wales.

Certainly in order to be able to use the data to perform training and testing of the error correction model, certain procedures have to be followed to transform the raw datasets to usable timeseries. This chapter includes all the data preparation and manipulation techniques and results, since their importance in obtaining satisfying and robust results is noteworthy. In Section 3.1 the raw data, alongside with the corresponding results of a pre-processing procedure, are presented. Section 3.2 includes a presentation of the "clean" data, accompanied by an explanation of the analysis that took place, while in Section 3.3 the scatterplots showing the relations between the clean observational data and the results of the numerical model (SWAN) for the same time period can be found. Finally, some general comments on the data, as well as a presentation of the behaviour of the available variables after their preparation, are presented in Section 3.4.

It has to be stressed that the error correction techniques are suitable for any offshore environment, given the required training, and are not limited in the area of the Irish Sea. The case presented here serves as an example of the applicability of the models in real life applications. The same procedures and techniques would have to be followed in any similar case, aiming to accurately predict the variables' behaviour in offshore (mild) environments.

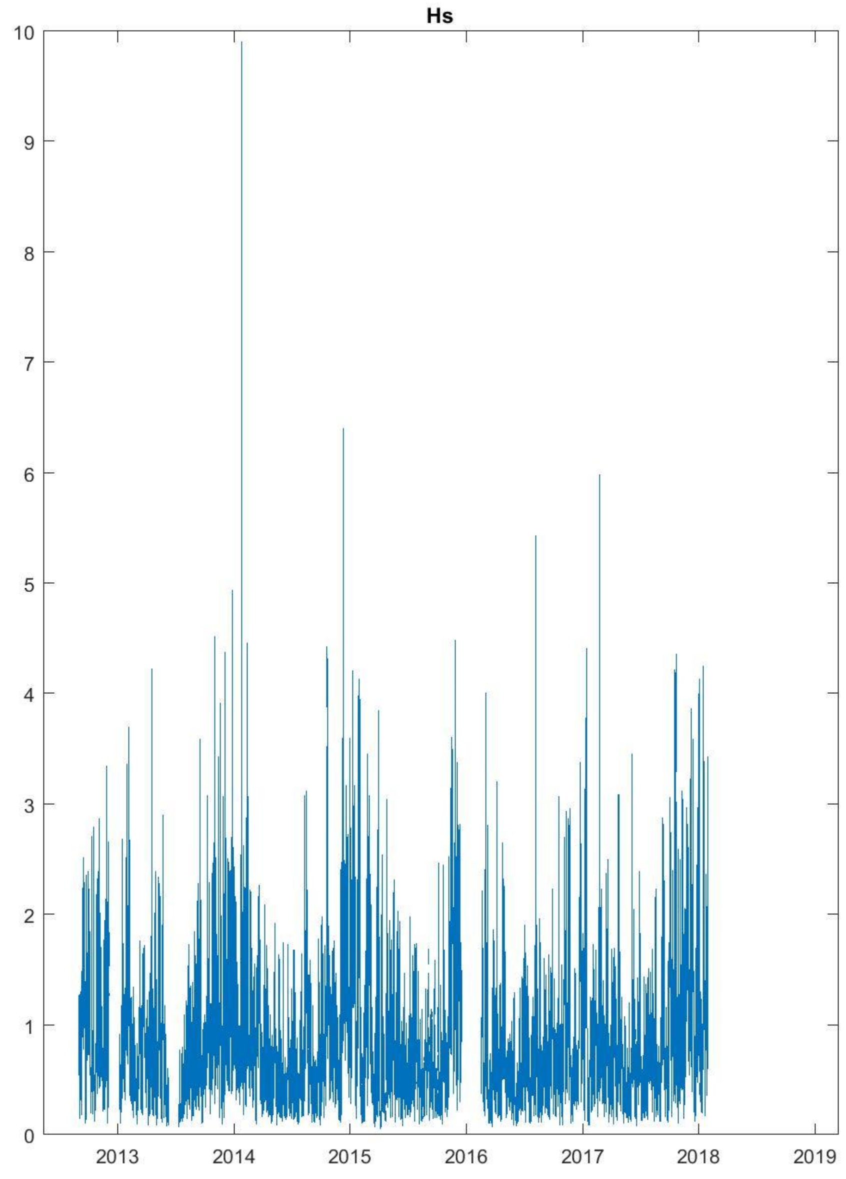
Fig. 4.Gwynt-y-Mor wind farm (left) and wave rider buoy deployed in the area (left) (Source: Google Images).

3.1. Raw Data and Pre-processing

Initially, it is important to have an overview of the raw data, before any processing has been done. In general, the pre-processing of hydrodynamic raw data is more straight-forward than the procedure followed for the meteorological dataset. Hydrodynamic raw data have to undergo a procedure of checking for continuously repeated measurements (which indicates a device malfunction), which are then replaced by custom fixed values (e.g. NaN[21] values) in order for the timeseries to be presentable. An example of such a timeseries is illustrated below (Figure 5) for the case of the significant wave height (H_s) at the Gwynt-y-Mor station.

[21] "NaN" is the abbreviation for" *Not a Number*", which is extensively used in programming.

The rest of the timeseries, concerning other variables included in the analysis, are moved to Appendix B.



Innogy 2018 - Raw hourly measurements Top panel: H_s (m)	Innogy 2018	Measurements
	Sep2012 - Jan2018	
Deltares	Hxxxx	Fig. 2.1

Fig. 5. Significant wave height (H_s) raw data at Gwynt-y-Mor (01-09-2012 to 31-01-2018).

3.1.1. Meteorological Data

Before the wind velocity timeseries could be presented or manipulated, it had to be transformed to the wind velocity corresponding to 10 m above sea level (Figure 6). This reference wind speed is used since it constitutes also an input to the numerical model (most theoretical relations concerning ocean waves include the 10 m wind velocity). This was achieved using the logarithmic wind velocity profile, by means of the ORCA tool[22] (met-**O**cean data t**R**ansformation, **C**lassification and **A**nalysis), provided by Deltares. ORCA integrates the main aspects of analysing met-ocean data, having four basic functionalities: (1) Data validation, (2) Normal conditions, (3) Extreme conditions, and (4) Sea State analysis.

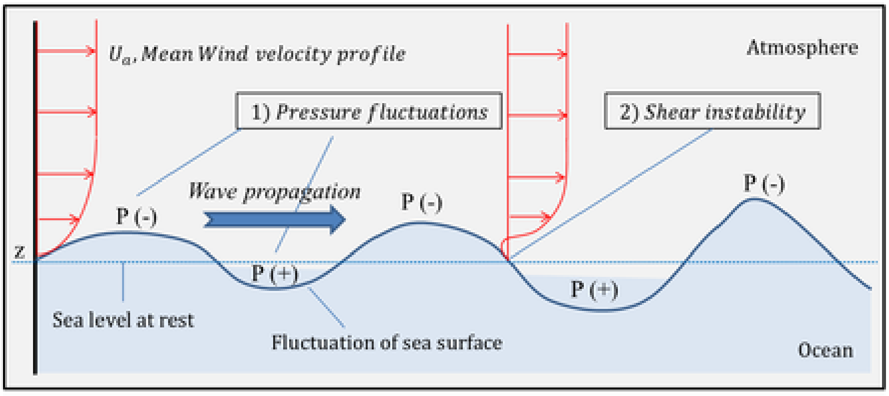


Fig. 6. Wind velocity profile above sea level (Source: Google Images).

A density scatterplot, showing the relation between the 70 m wind velocity[23] at Gwynt-y-Mor and the required 10 m wind velocity can be seen in Appendix B. A similar analysis was followed for the case of the observations corresponding to the meteorological mast at Rhyl Flats, where the wind velocity was measured at 58 m above sea level. The results of the meteorological data analysis for Rhyl Flats can also be found in Appendix B.

After the 10 m wind velocity (U_{10}) was calculated, the corresponding timeseries could be presented (Figure 7). Again for the meteorological dataset, a procedure of removing repeated values was carried out, only to be later replaced by NaN values (a procedure also known as gap filling).

[22] Deltares developed a method to standardize the execution of met-ocean studies, by developing guidelines along with a Matlab® toolbox, called ORCA (see also: https://www.deltares.nl/en/software/orca/).
[23] Wind velocity is obviously measured above sea level, so the reference levels correspond to distances measured from the ocean surface.

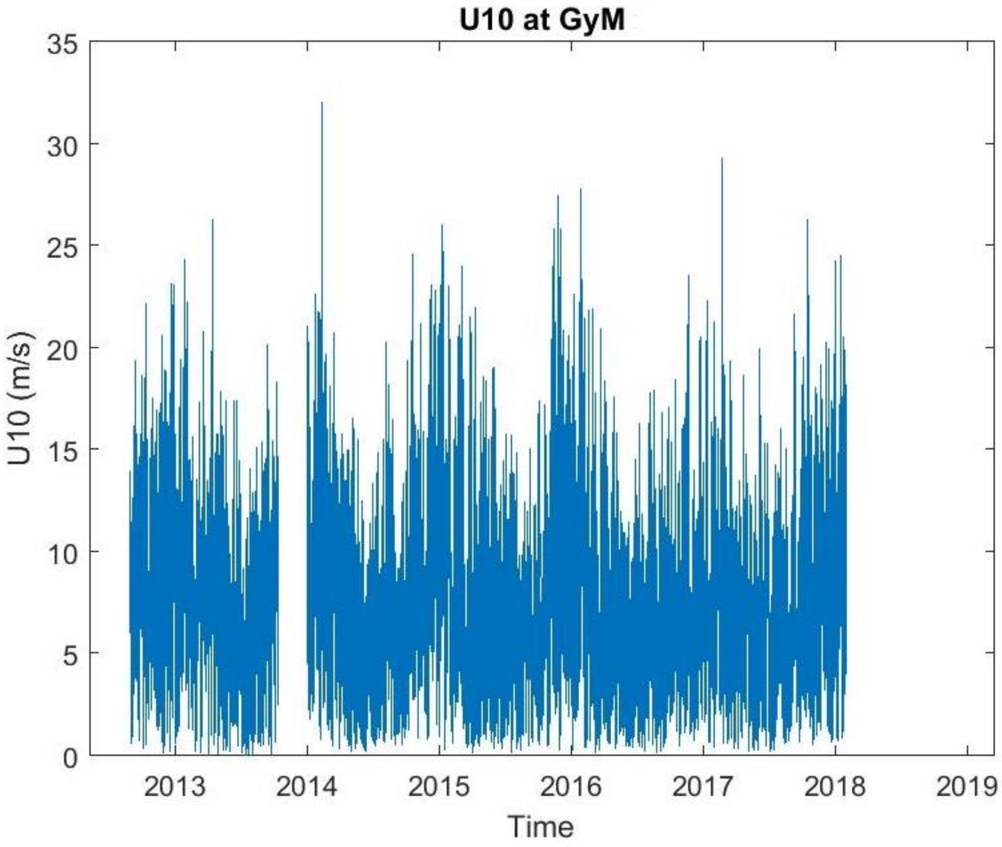


Fig. 7. Wind velocity (U_{10}) timeseries at Gwynt-y-Mor (01-09-2012 to 31-01-2018).

3.2. Data cleaning

3.2.1. Methodology and Results

A procedure of significant importance in terms of robustness in the final result of data-driven methods, such as the statistical error correction techniques under study in this thesis, is the data cleaning or data cleansing. Data cleaning is the process of detecting and correcting or removing inaccurate records (outliers) from a dataset. Incorrect or inconsistent data can lead to false conclusions and misdirected actions, which might have a large impact, depending on the application.

In general, to be able to perform a consistent cleaning of the dataset in hand, certain filters have to be implemented. In this research, again, the ORCA tool was used for the cleaning procedure. The filters incorporated were two: (1) a moving average filter, and (2) a strict moving standard deviation filter. For more information on the aforementioned filters the reader is referred to Appendix B.

The cleaning procedure of the whole set of variables had to be done consistently according to their dependence. That means that a removed significant wave height value has to lead to a removal of the corresponding zero-crossing wave period value (same for wave direction, etc.). As a result the cleaning process of hydrodynamic variables was done according to the significant wave height (H_s), while for the meteorological variables the wind velocity (U_{10}) was considered the point of focus. The results of the aforementioned analysis are presented below (Figures 8 to 10). Because of the nature of the timeseries produced for the wind and wave direction (not actual visible points of removal), the comparison plots for those two variables are not presented here, but are included in Appendix B for the sake of completeness. In Appendix B, the reader can also find the results produced for the measurements retrieved from Rhyl Flats.

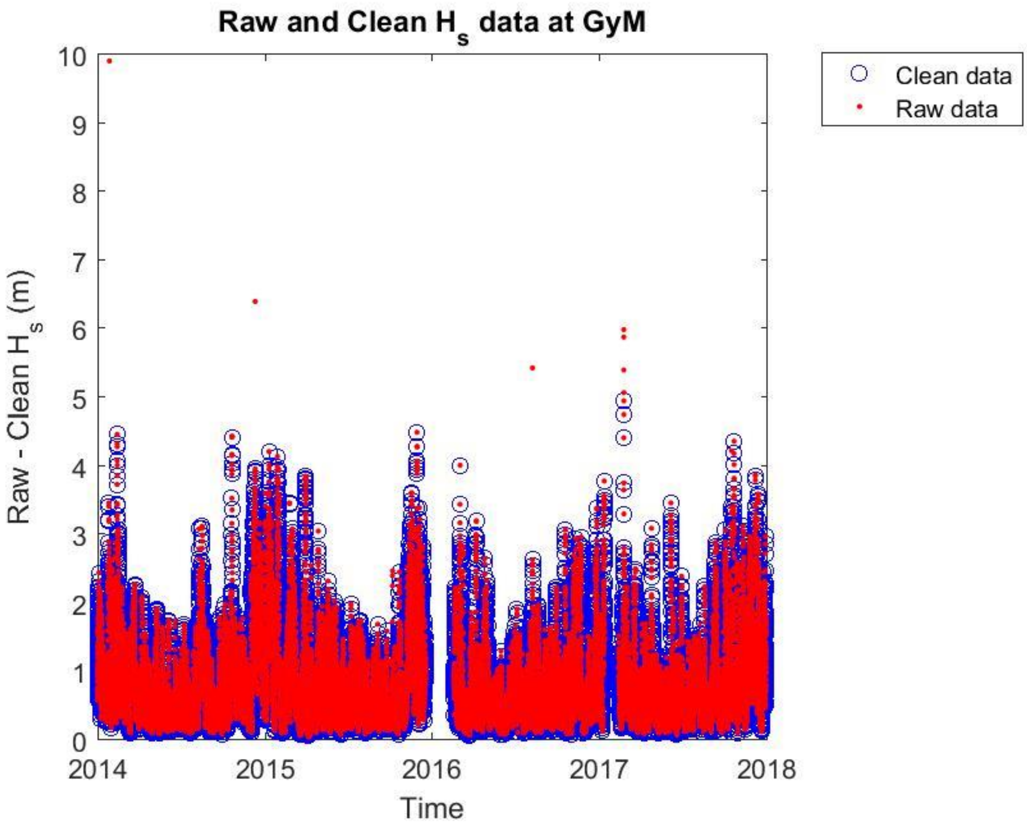


Fig. 8. Clean and raw significant wave height (H_s) datasets, as resulted from the ORCA cleaning procedure (Gwynt-y-Mor).

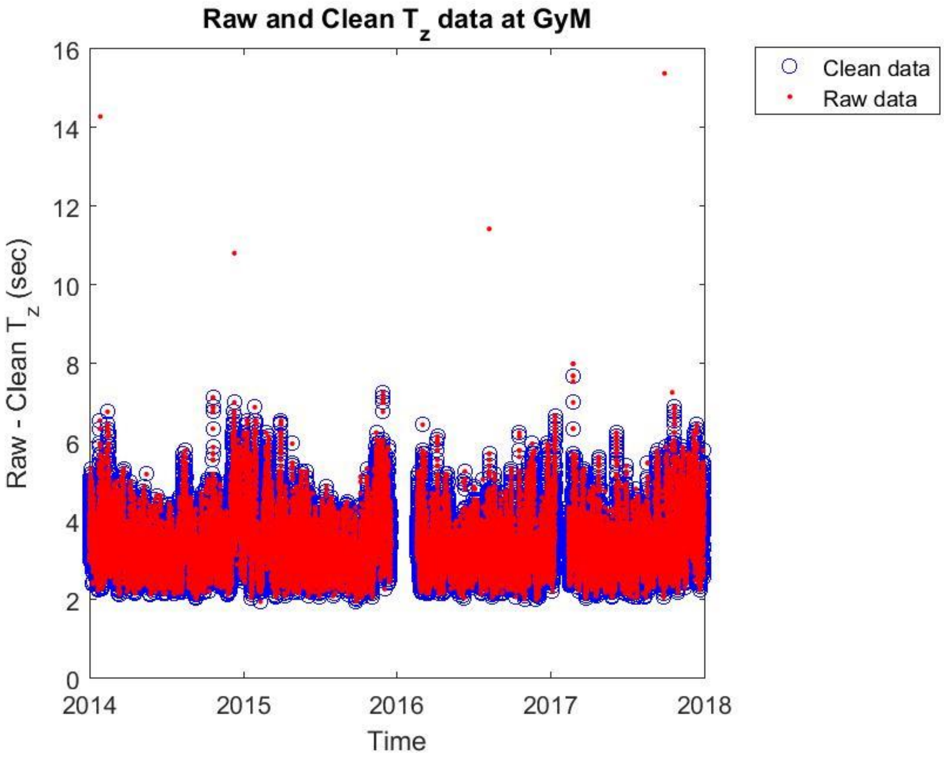


Fig. 9. Clean and raw zero-crossing wave period (T_z) datasets, as resulted from the ORCA cleaning procedure (Gwynt-y-Mor).

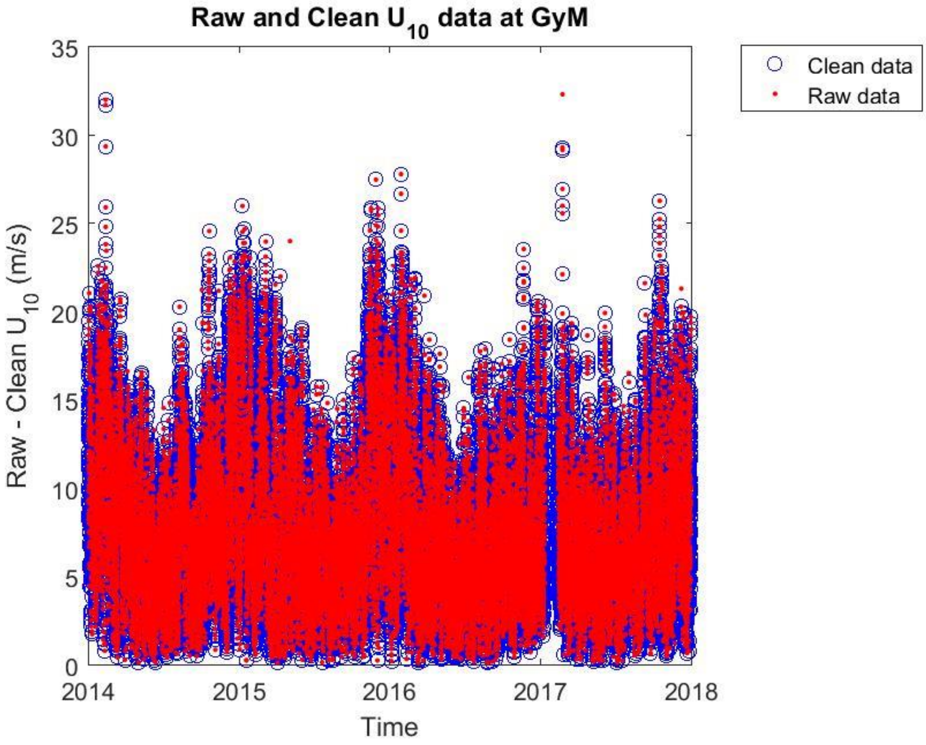


Fig. 10. Clean and raw wind velocity (U_{10}) datasets, as resulted from the ORCA cleaning procedure (Gwynt-y-Mor).

3.3. Scatterplots

A robust way to check whether the analysis followed is satisfying is to examine the scatterplots between the clean observational data and the numerical model outputs (Figures 11 to 13). The results from the measurement stations adjacent to Gwynt-y-Mor are going to be presented and evaluated in this section. The respecting Rhyl Flats results can be found in Appendix B.

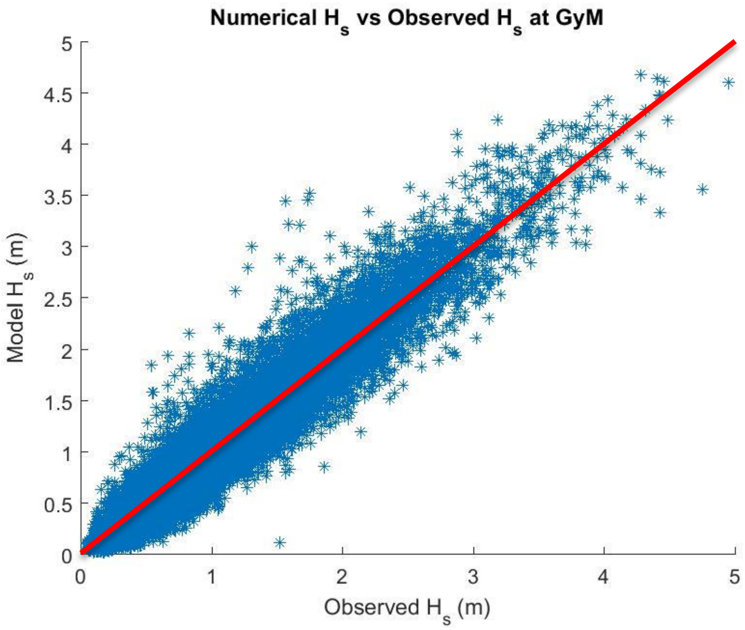


Fig. 11. Numerical modelled versus Observed significant wave height (H_s) data (Gwynt-y-Mor).

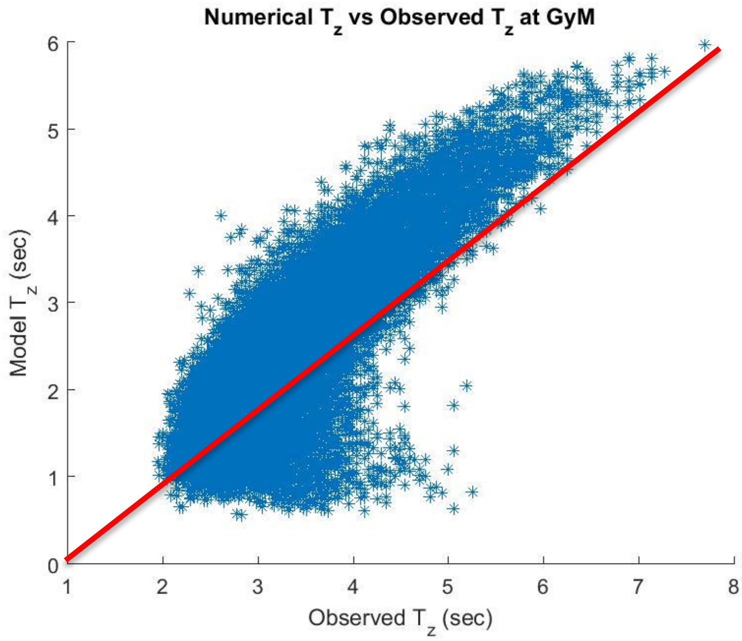


Fig. 12. Numerical modelled versus Observed zero-crossing wave period (T_z) data (Gwynt-y-Mor).

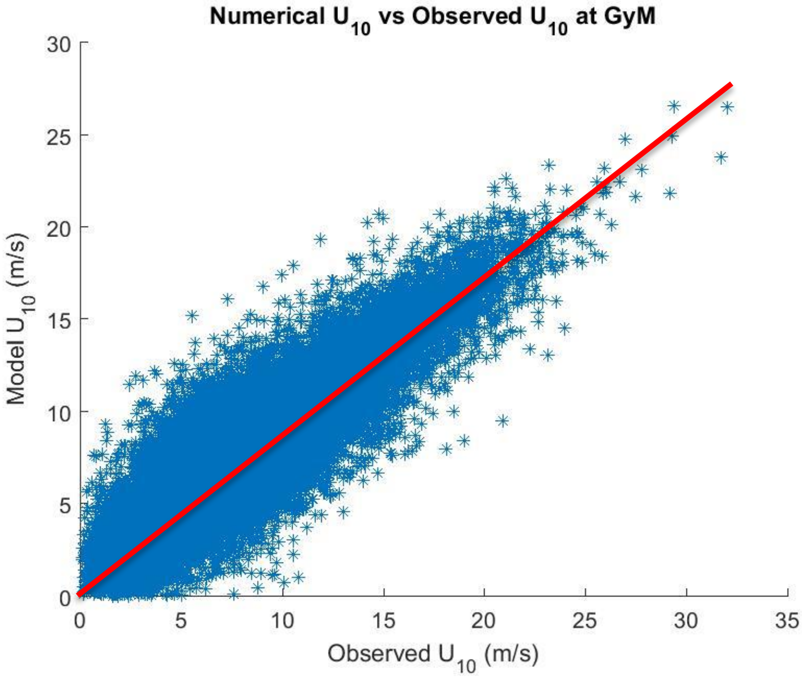


Fig. 13. Numerical modelled versus Observed wind velocity (U_{10}) data (Gwynt-y-Mor).

It is evident that there is a clear and satisfying relation between the modelled and measured datasets, close to the diagonal. Some values of the zero-crossing wave period seem to differ (lower area in Figure 12), but the behaviour was considered reasonable, since the wave period cleaning was conducted according to the wave height procedure. As a result, the cleaning procedure was considered successful for those variables, and their use in testing was justified.

3.3.1. Peak Wave Period (T_p) Case

Examining the modelled and observed peak wave period scatterplot, it becomes even more apparent that the procedure followed to calculate the variable's values with the numerical model (SWAN) produces significantly different outputs, in comparison to the measurements (see Figure 14). While the datasets seem to present a relation close to the requested in one part, there is a distinctive portion of the values differing significantly. This difference, results most probably from the inclusion of swell in the observed data, while the numerical simulations separated the swell components and produced peak wave periods referring only to the corresponding wind waves.

Also, the peculiar behaviour of the data, i.e. forming lines perpendicular to the observed data axis (x-axis), is even more visible here. Therefore, since the peak wave period develops a behaviour which cannot be handled consistently, it was decided not to be included as a variable, during the error correction simulations. Certainly, a more focused and separate cleaning, or even calculation, of the peak wave period could be made, but this kind of analysis would deviate from the goals of this research and the real-time nature of the application.

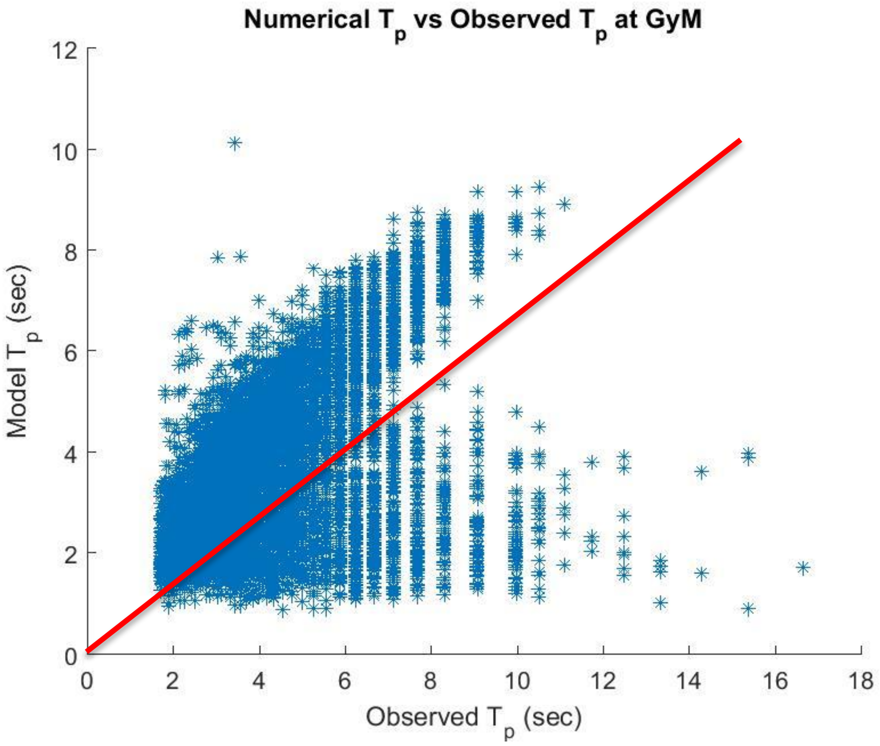


Fig. 14. Numerical modelled versus Observed peak wave period (T_p) data (Gwynt-y-Mor).

3.4. General Comments

3.4.1. Joint Distributions of Data

Another way to evaluate the relations governing this application is by looking at the bivariate distributions of the variables in general (Figure 15). The distributions of the individual variables are visible at the diagonal of Figure 15, where it is shown that the wind velocity (U_{10}) and the significant wave height (H_s) seem to follow a log-normal or Rayleigh distribution as described many times in the literature (see also Tayfun, 1980; McWilliams, Newmann and Sprevak, 1979; Li et al., 2016; J. Mathisen and E. Bitner-Gregersen, 1990). The fit test was carried out by means of the FDB[24] tool in Matlab®, which incorporates certain criteria (AIC, BIC, etc.) to define the best parametric distribution for the data in hand. As can be seen in

[24] For extended description and information see: https://nl.mathworks.com/matlabcentral/fileexchange/36000-fbd--find-the-best-distribution--tool?focused=5245793&tab=function.

Figure 16, the lognormal distribution provides a good fit for the significant wave height data (H_s), which will be proved really useful in the simulations to follow (see Chapter 5).

The variables, whose joint distribution (scatter) is concentrated around the diagonal, have a clear and almost linear relation. Such relations can be seen between (1) the significant wave height (H_s) and the zero-crossing wave period (T_z), (2) the wind velocity (U_{10}) and the zero-crossing wave period (T_z), as well as between (3) the significant wave height (H_s) and the wind velocity (U_{10}).

For the rest of the relations it is difficult to conclude anything a priori, but for the aforementioned variables it has to be expected that the correlation in the Bayesian Network (BN) analysis will be large. Having an indication on the dependence between the variables can also play an important role on the decision process to establish which relations can be fixed in a custom-built BN, which will be the case in the remaining chapters and will help examine and validate the performance of the BN models more consistently.

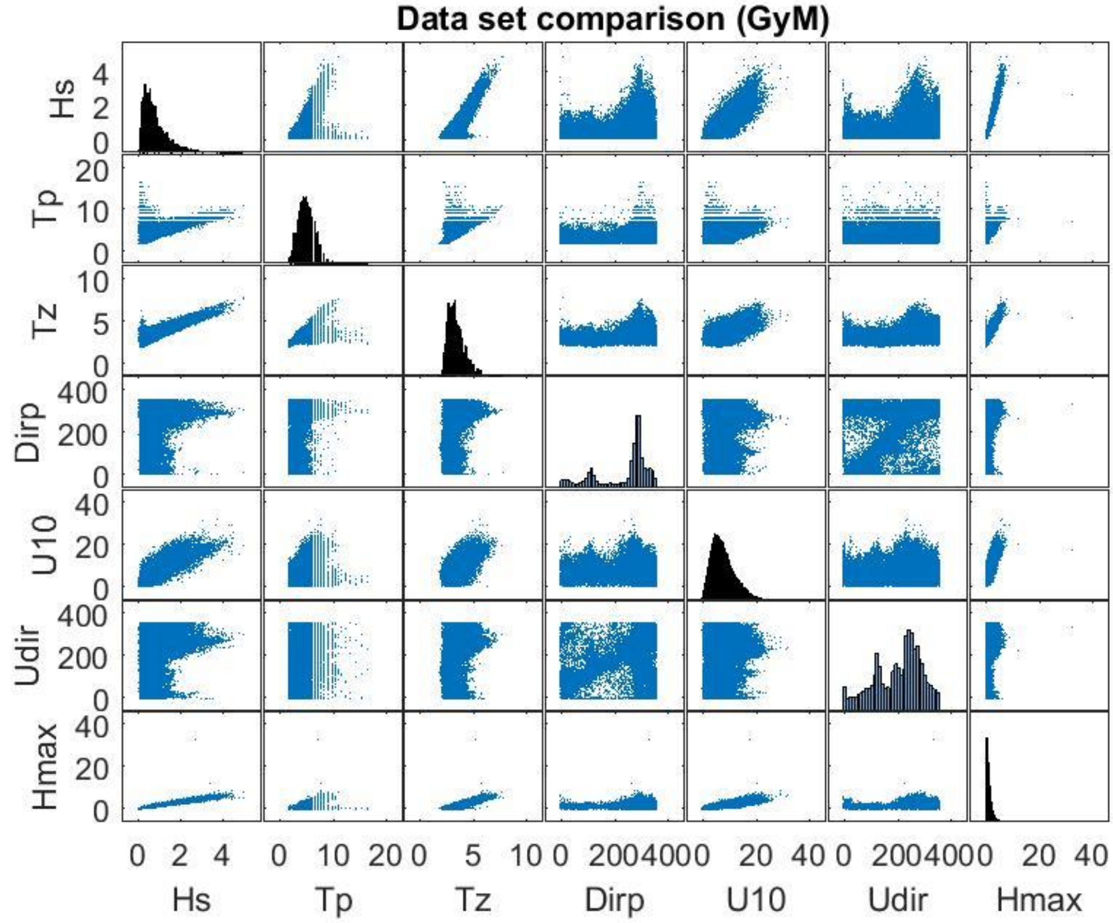
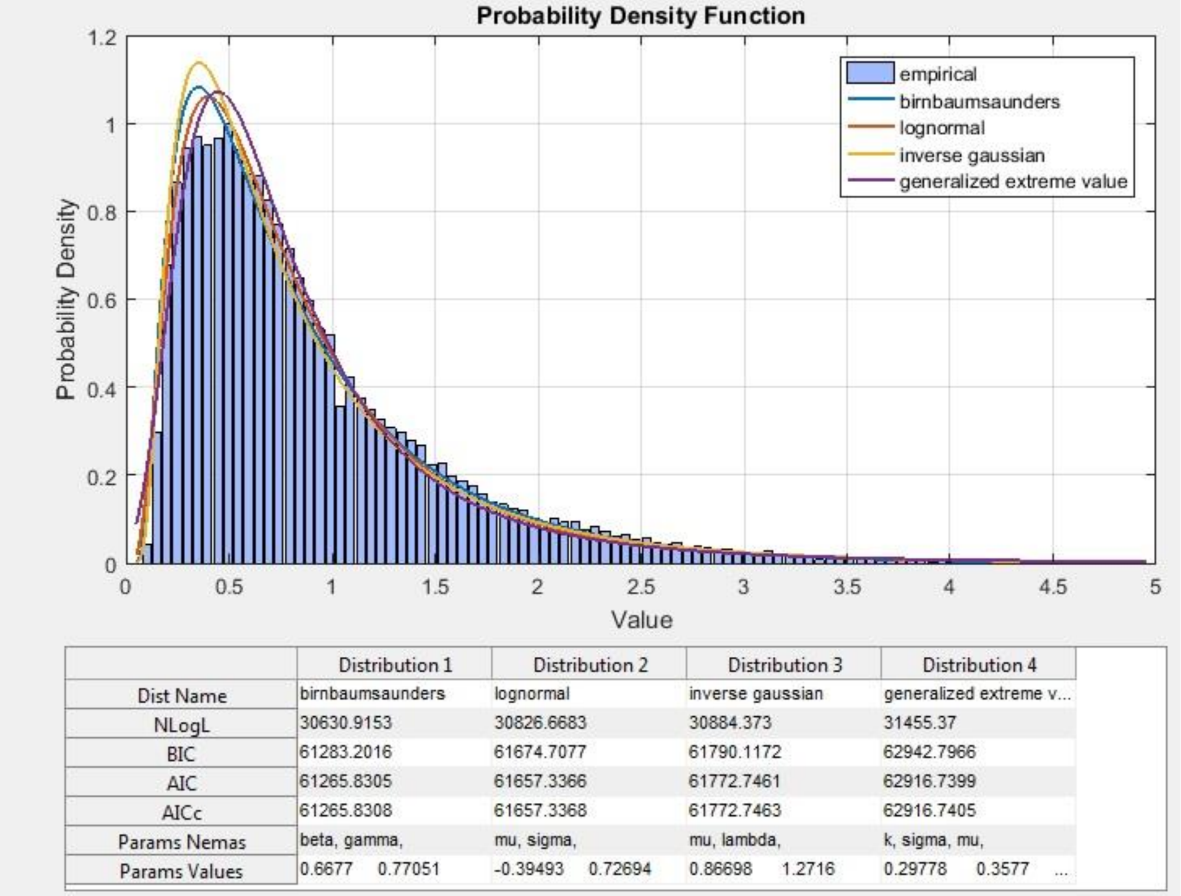


Fig. 15. Joint distributions for the available hydrodynamic and meteorological variables at Gwynt-y-Mor.

One relation that would also seem obvious is the one between the wind (U_{dir}) and wave (D_{irp}) directions. Nevertheless, this is not the case at all, as it will also become evident in the following chapters. The wave direction differs significantly in many occasions from the wind direction, due to the fact that its measured values might correspond to waves generated far from the area of interest. As described earlier, waves might be generated by storms which occurred may kilometres away from the measurement area. As a result the wave direction includes the swell wave direction as well.

Both directions are going to be incorporated as variables in the BN model, but for the case of a custom-built network, it is better to implement the data driven relation between those variables, due to the uncertainty of their origins. As a general comment it has to be stressed that it is more beneficial, when sufficient data are available, to incorporate the data driven relations in order to perform predictions. The reasons why this option is usually less prone to mistakes will be clearly displayed in the remaining of this thesis. The common reason would say that the less the human intervention, the more the natural behaviour is encapsulated, given a vast amount of information and case specific measured data.



	Distribution 1	Distribution 2	Distribution 3	Distribution 4
Dist Name	birnbaumsaunders	lognormal	inverse gaussian	generalized extreme v...
NLogL	30630.9153	30826.6683	30884.373	31455.37
BIC	61283.2016	61674.7077	61790.1172	62942.7966
AIC	61265.8305	61657.3366	61772.7461	62916.7399
AICc	61265.8308	61657.3368	61772.7463	62916.7405
Params Nemas	beta, gamma,	mu, sigma,	mu, lambda,	k, sigma, mu,
Params Values	0.6677 0.77051	-0.39493 0.72694	0.86698 1.2716	0.29778 0.3577 ...

Fig. 16. Results of the parametric distribution fitting procedure to the significant wave height (H_s) data of Gwynt-y-Mor.

4. The Error Correction Model and its Functionality

4.1. General

The error correction model, described in this thesis, is essentially a forecasting tool, which attempts to predict the hydrodynamic conditions in open seas more accurately than a numerical model (in this case SWAN[25]). Hence, it is referred to as "*error correction*" model, since its nature and behaviour deviates slightly from a pure prediction tool.

In general, the model is able to perform both in non-operational (offline) and operational (online) situations. By operational situation, the continuous flow of the required data in real time is implied, while in non-operational mode, the model interacts with data stored in the computer memory. Nevertheless, in both cases the nature of the data, and the number of variables included in each simulation, is the same. The error correction model requires three types of data:

1) On site measurements (observations), which are manipulated and processed, before used (i.e. filling of gaps by interpolation and cleaning of outliers).
2) Numerical model hindcast[26] data for a time interval prior to the one under consideration. Instead of using hindcast data for the analysis, one could alternatively use past forecast data of the numerical model, which of course will be lees accurate, due to the input of wind data produced by a numerical model (in this case HIRLAM[27]).
3) Numerical model forecast data for the time interval under consideration (48 hours ahead of current time). The numerical model forecasts are produced every 6 hours, so there are 4 forecasts per day, each one for 48 hours ahead.

Depending on the error correction method some of the above data may or may not be used. The functionality of each error correction technique is described in the following paragraphs. We start in Section 4.3 with the methods implemented also in the past, i.e. the Linear Regression, the Copulas, the Artificial Neural Networks and the Stochastic Interpolation, continuing in Section 4.3 with the function and methods incorporated into the ARIMA model.

[25] For more information on the SWAN model the reader is referred to Chapter 2 and the official SWAN website: *http://swanmodel.sourceforge.net/online_doc/swanuse/swanuse.html*.

[26] The numerical model hindcast data are produced by incorporation of observational wind data as input to the model and a reverse procedure to obtain the results (i.e. the opposite of a forecast procedure).

[27] For more information on HIRLAM the reader is referred to: *http://hirlam.org/index.php/hirlam-documentation*.

In Section 4.4 the incorporation of Bayesian Networks as an error correction technique is discussed. Finally, some initial tests and results on the newly implemented models' functionality, alongside with some preliminary conclusions, are presented in Sections 4.5 and 4.6 respectively.

4.2. Operational Techniques in Meteo Dashboard

In this section, a short description is given, about the techniques already implemented in Meteo Dashboard. The operational functionality of these methods has already been tested in the past, for a specific area in the Irish Sea, but their performance and ability to predict in comparison to each other, but also to newly incorporated techniques (i.e. ARIMA and the Bayesian Networks) is still a matter of research. Thus, it is important for the reader to get acquainted with their functionality in general terms, since their performance will be discussed extensively in the following chapters. The theoretical basis of the discussed methods has been presented in preceding chapters, hence only their operation as part of the error correction model will be addressed.

4.2.1. Linear Regression

For the Simple Linear Regression incorporated in the error correction model, the predicted significant wave height ($\widehat{H_s}$) is given by the following equation:

$$H_s = \beta_0 + \beta_1 \cdot H_{s,num} + \varepsilon \tag{4.24}$$

where, $H_{s,num}$ is the modelled (numerically) significant wave height, ε is a random error variable, β_0 is the y-intercept of the line y = β_0 + β_1 · x, and β_1 is the slope of the aforementioned line.

In the previously presented model, $\widehat{H_s}$ and $H_{s,num}$ are assumed to be correlated, i.e. linearly related. Thus, the model function takes the form of a line:

$$\widehat{H_s} = \widehat{\beta_0} + \widehat{\beta_1} \cdot H_{s,num} \tag{4.25}$$

The parameters β_0 and β_1 are computed using the data available in the last 48 hours at t_0, assuming that they are constant for the next 48 hours, after t_0 (prediction interval).

4.2.2. Copulas

Having six months of simulated significant wave height data by SWAN, the modelled data and the observed data are used to construct a copula. The multivariate densities are difficult to be estimated due to possibly complicated forms of the data distribution and the curse of dimensionality. The use of copulas is simplified, by separating the learning of the marginal distributions from the learning of the multivariate dependence structure.

For this particular application, the use of the Gumbel copula was chosen. For the determination of the degree of fit of the chosen copula, a simple fitting test was carried out successfully. Two variables were used for the bivariate copula; (1) the modelled, and (2) the observed significant wave height, which are linked together into a density model.

The fitted copula is saved and used in operational (online) mode. The density model can be conditionalized according to one of the variables, in this case, the modelled significant wave height, giving back the marginal distribution of the observed significant wave height. This way, the confidence intervals for the expected conditionalized value are provided.

4.2.3. Artificial Neural Networks

As described previously (see Chapter 2), there are different network topologies and learning processes to be chosen, when building an Artificial Neural Network (ANN). For simplicity, in this application, a three (3) layered architecture is used, i.e. an input layer, a hidden layer, and an output layer. The back propagation algorithm is used and the network's input consists of the modelled significant wave height at the measurement stations (numerical model forecast data), with the output being the corrected predictions at the same stations. The ANN has an autoregressive component, as significant wave height data for the last 3 hours (prior to the forecast) have been used as an input.

Initially, the network had to be trained, tested, and validated. After the ANN had been optimally trained, it was stored and used operationally, every time the model is in use. For the training phase, four (4) months of data were used, while for testing 1 month was incorporated to avoid over-fitting. The last 1 month, of the total of 6 months (same with the ones used for the copula construction), was used to validate the network.

4.2.4. Stochastic Interpolation

For the case of Stochastic Interpolation, a simple regularization routine was used, which is able to combine past significant wave height observations and SWAN forecast data, incorporating the Bernstein Stochastic Interpolation (see also Chapter 2) method, presented by Howard and Kolibal. Here, three basic components are required, for the model to produce results; (1) past (48 hours prior to the time the forecast started) observed significant wave height data, (2) past numerical model significant wave height data (from the same time interval previously mentioned), and (3) the numerical model forecast for the significant wave height (48 hours ahead).

An important comment has to be made here; the Bernstein Stochastic Interpolation is not a unique forecasting/predicting tool, but it was created primarily for spatial modelling. As a result its forecasting capabilities are limited, compared to its ability in spreading the forecast in space. Further results and conclusions on the method's functionality and performance, will be found in the following chapters.

4.3. ARIMA Model

4.3.1. Methodology and Training

The ARIMA model was built around the `forecast`[28] package provided by R, which needs only past observational data of the variable of interest (in this case the significant wave height or simply H_S) to be able to function, and follows the previously described Box – Jenkins approach (see also Chapter 2). By means of these past measurements, the model can be trained sufficiently in order to retrieve its seasonal and non-seasonal order/component. The `auto.arima()` function in R uses a combination of unit root tests, minimization of the Akaike Information Criterion (AIC) and the maximum likelihood estimation (MLE) to obtain (fit) an ARIMA model.

Here arises the important question of how much training is needed in order for results of satisfying accuracy to be produced. Normally, the prediction interval represents 20-30 % of the total time under consideration; hence the training data can include 70-80 % of the data[29]. That amount of training, in many occasions, (such as this one) makes the computational load big enough for the model to become exceptionally slow. That, of course, is an undesirable effect in real time simulations, and deviates from the goals of a soft computing technique. As a result, the amount of training has to be reduced to the minimum possible, which would provide under certain circumstances reasonable results. Various tests were carried out, but in general the phenomena involved, as well as the real-time nature of the application, limit the ARIMA model's ability to perform.

4.3.2. Seasonality

Certainly, the aforementioned drawback could be set aside in case the results were exceptionally satisfying. As it can be seen in the following sections, (Section 4.4 on Preliminary Results) the accuracy of the model can deviate a lot from the desired one, even with the amount of training mentioned above. The possible reasons behind this behaviour are extensively discussed in following sections, but the way the results are produced can partly give an explanation. In order for the data to be inserted into the ARIMA model and be

[28] See also: *https://cran.r-project.org/web/packages/forecast/forecast.pdf*.
[29] Similar examples of training and test sets can be found in Chapter 2.

analysed, they have to be transformed into an artificial timeseries, compatible to the model. For this to happen, the seasonality of the dataset has to be known.

To determine the seasonality, a Fourier Transform has to be performed, from which the two top frequencies are extracted. In case the top frequency indicates a time larger than the time interval in hand, (false estimation) it is replaced by the second most dominant frequency, given by the aforementioned analysis. Having the frequency, i.e. the time period, a timeseries can be inserted into the ARIMA model in order to produce prediction of the variable of interest (i.e. the significant wave height or simply H_s).

Also, to make things even more complex, the ability of the model to produce accurate forecasts is further reduced when the seasonal component is omitted. Generally, a simple training would probably provide only the non-seasonal components (p, d, q), since the seasonal components (P, D, Q) are in most occasions zero. As a result, the aforementioned order has to be artificially inserted into the model, i.e. by setting the degree of differencing (D) of the seasonal component equal to 1. That way, the ARIMA model is of the order (p, d, q)(P,1,Q), with the seasonality inserted to the timeseries. As it can be easily understood, a large part of the procedure is artificial, which is reflected in the final result.

4.4. Bayesian Networks (BN) Model

4.4.1. Scope

The Bayesian Networks model has an extended multivariate input in comparison to the rest of the error correction techniques. As it was described in the preceding paragraphs, the error correction model, without the addition of Bayesian Networks, would just need past measurements and numerical model data, as well as the numerical model forecast of the significant wave height, to produce a possible correction. The nature of Bayesian Networks imposes the use of more variables, whose dependency with the variable of interest can produce a forecast of reasonable accuracy.

As stated previously, the perspective of this thesis deviates from providing just a forecast, accompanied with a desired level of accuracy. The goal is to learn from the errors in the numerical model due to certain phenomena, understand and quantify those relations to eventually correct/improve the prediction of the numerical model, which is solely based on empirically and theoretically derived formulas. The consideration of Bayesian Networks aims to the description and representation of the underlying uncertainty in nature's behaviour, as accurately as possible, while combining different dependent components, e.g. the wave period, the wave direction, the wind velocity, etc.

Here an important notation has to be made, that is, the procedure is data driven, i.e. the imputed data sets determine the dependencies between the variables. Efforts were made to also impute some relations, such as the relation between wind velocity and significant wave height, but as it will become obvious by the results displayed in following chapters, these relations are not well reflected by the BN dependence structure implied by the data.

4.4.2. Training Methodology

The Bayesian Networks, as most of the data driven techniques, need a specific amount of data in order to be trained sufficiently and be able to represent the desired relations. When the BN structure is acquired through the data, then a significant amount of data is needed. In every application the characterization of a training procedure as "sufficient" depends largely on the type and behaviour of the data. A sensitivity analysis would be in place to determine what "sufficient amount" actually means for the application. The significant wave height, for instance, is a variable whose behaviour is highly dynamic, i.e. it can change radically in short time intervals (e.g. hours). As a result, the more training the model has the better, since it can assimilate, and later reflect a larger range of behaviours.

In this thesis, the training techniques are divided into two major categories; (1) the long training, which involves past observational and numerical data, even from 3 years prior to the current date[30], and (2) the short training, which only involves measurements and numerical model data from 48 hours prior to the start of the forecast.

In order to obtain the structure of the Bayesian Network, the `bnlearn`[31] package of R is used. In general, there are two broad categories to learn the structure of a BN, the score-based and the constraint-based. The constraint-based case employs the independence test to identify a set of edge constraints for the graph and then finds the best DAG that satisfies the constraints. This approach works well with some other prior (expert) knowledge of structure, but requires lots of data samples to guarantee testing power. So it is less reliable when the number of samples is small. The score-based approach first defines a criterion to evaluate how well the BN fits the data, and then searches over the space of DAGs for a structure with maximal score. In this way, the score-based is essentially a search problem which consists of two parts: (1) the definition of a score metric, and (2) the search algorithm. A hill climbing (HC) score-based structure learning algorithm was used to train the network, by means of a pre-specified dataset. For the score based analysis, the package incorporates four criterions; (1) the multivariate Gaussian log-likelihood, (2) the corresponding Akaike Information

[30] By "*current date*" the date where the forecast takes place is implied.

[31] See also: *https://cran.r-project.org/web/packages/bnlearn/bnlearn.pdf*.

Criterion (AIC), (3) the corresponding Bayesian Information Criterion (BIC), and (4) a score equivalent Gaussian posterior density (BGe). The package also assumes a multivariate normal distribution for continuous variables, such as the hydrodynamic variables in hand, for the marginals and the resulting conditional distribution of the variable of interest. This assumption can be considered restricting in many occasions, but as it will become obvious, the results of such an analysis are quite reasonable. In case the assumption of multivariate normality is violated, the non-parametric Bayesian Networks could produce a more accurate conditional distribution and possibly (that is completely uncertain) in some occasions more accurate forecasting results. Nevertheless, the assumption of multivariate normality was considered sufficient to test the BN behaviour and performance, and the `bnlearn` package, which is provided for free (open source), as the most suitable one for this particular application.

In the case of the long training, the training dataset is continuously enriched with new measurements, encapsulating even more relations and behaviours. Certainly, this requires a relatively large part of the computer's memory. This effect can be impugned by incorporation of new variables and deletion of older, or with smaller training sets, i.e. in the order of months instead of years.

4.4.3. Data Fitting in BN Structure

In succession to the BN structure training, the fitting (feeding) of the data in the network has to take place. That way, according to the data-driven structure retrieved by the aforementioned procedure, the relations of the variables are inserted, in terms of correlations. The fitted dataset involves past observational data for all variables, as well as past numerical model data for the variable of interest only.

In this part, the user can impute his/her own structure, by whitelisting or blacklisting certain relations, i.e. providing a custom and not data-driven fit. This, certainly, creates large differences in the results, since in many occasions the whitelisted arc is not supported by the BN structure in representing the joint density. Thus, it is suggested by the writer that the procedure should be carried out using data-driven structure learning and fitting techniques, even if a given/produced relation seems completely unrealistic. Some examples will be provided in following chapters, where a seemingly realistic connection provides completely unrealistic results, due to large uncertainties in data retrieving procedures and in the natural phenomena themselves (for instance in offshore environments the incorporated swell components can make a huge difference to what is realistic and what is not).

4.4.4. Predictions

The predictions provided by the Bayesian Networks (BN) model, are retrieved from the conditional distribution of the variable of interest, which is dependent on the rest of the incorporated variables. The method to retrieve a forecast from the BN is based on the parent-descendant (child) relations, previously established by training and fitting.

Since it is impossible to have future measurements for the variables incorporated, forecasted numerical model data for these variables are used to construct the conditional distribution for every point prediction. In other words the network is trained and fitted with past observational data, as well as numerical model data for the variable of interest, and then provides a forecast based on forecasted numerical model data (essentially we are conditionalizing on forecast numerical model data). The point prediction is the expected value (E[x]) of the conditional distribution, which is in our case normal (Gaussian), due to the multivariate normality assumption made in the structure learning procedure. Unfortunately, the aforementioned assumption prevents us from retrieving realistic uncertainty bound for the variable (i.e. significant wave height). The significant wave height (H_s) is not in any case normally distributed (see Section 3.4); therefore the assumption is in certain occasions[32] not appropriate. Nevertheless, symmetrical (normal) uncertainty intervals can be extracted, sometimes unrealistic due to negative values, which provide a fairly good coverage of the observations (more information and examples can be found in the following chapters). A flow chart, showing the function of the BN model is shown in Figure 17.

In terms of the variable, one could say that the assumption of normality is not appropriate, but referring to the prediction itself and its uncertainty, i.e. predicting the conditional expectation, along with the confidence bounds, the normal assumption is the most correct and reasonable one, due to the Central Limit Theorem (CLT). Since the predictions are the expected values of the conditional distributions, and simultaneously they are independent, their distribution tends towards a normal distribution, even if the original variables are not normally distributed. For more information on the classical CLT, the reader is referred to Lindeberg (1922); Abramowitz and Stegun (1972); Feller (1945, 1968, 1971); Kallenberg (1997); Spiegel (2003); Trotter (1959); Weisstein; and Zabell (1995). It should be emphasized that the focus of this research is to account for the uncertainty in the significant wave height (H_s) distribution and not in the estimates, and as a result the conditional distribution, in most occasions, cannot be expected to be normal.

[32] The conditional distribution can be in certain cases normal, even if the marginals are not represented by a Gaussian distribution. Thus, the assumption of the normality of the conditional distribution is partly rough.

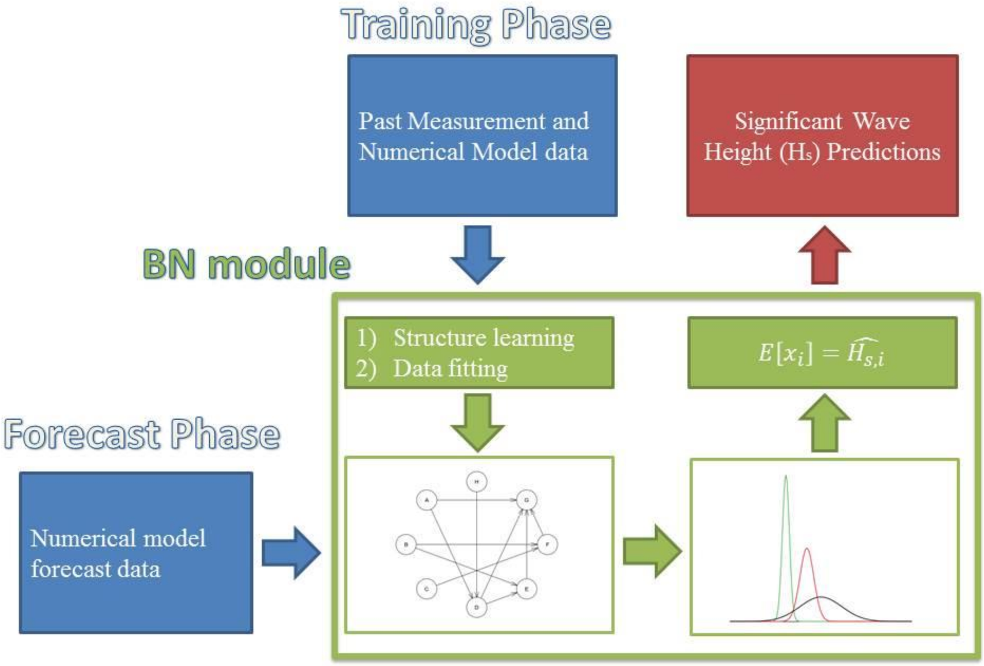

Training Phase
Past Measurement and Numerical Model data
Significant Wave Height (H_s) Predictions
BN module
1) Structure learning
2) Data fitting
$E[x_i] = \widehat{H_{s,i}}$
Forecast Phase
Numerical model forecast data

Fig. 17. Flow Chart presenting the basic function of the BN model.

4.5. Non-Operational Tests of Functionality

In order to test the functionality of the BN error correction model, past data retrieved from two stations (buoys) deployed in the Irish Sea, were used. Each of the stations is adjacent to an offshore wind farm in the same area; (1) the Gwynt-y-Mor wind farm, and (2) the Rhyl Flats wind farm.

The data collected by the aforementioned wave rider buoys correspond to a time interval starting on 1st of March 2015, and ending on the 30th of September 2015. The months were selected randomly, just to evaluate the functionality of the newly implemented methods (ARIMA and BN models). Since this analysis has just a testing nature, the focus is on specific months, rather than larger time intervals (as done later in Chapter 5). Given that, the corrections provided refer to the 2 or 3 last days of each available month, so that sufficient training could take place for the cases of the ARIMA and the long-trained BN models.

4.5.1. Model settings and training

The simulations were executed for both of the two previously referenced stations (GyM and RF). Due to the inability (direct or indirect) of the previous models incorporated in the Meteo Dashboard platform to implement more variables in the analysis, the observations received up until that point included only three variables: (1) the significant wave height (H_s), (2) the zero-crossing wave period (T_z), and (3) the wave direction (D_{irp}). As a result the networks created for functionality testing are relatively small (4 nodes). Nevertheless, they demonstrate the functionality and correction/prediction abilities of the created BN models, sufficiently to make a first performance evaluation and draw some important preliminary conclusions.

As stated before, two BN models are used: (1) a long-trained one, whose learning/training set in this occasion consists of 90% of the variables' values (27 days for a month of data), while the test set consists of 48 hours of data (2 days), and (2) a short-trained BN, whose training set includes just 48 hours of data (i.e. 48 values of each variable) prior to the forecast, and a test set composed of 48 values[33]. This preliminary analysis took place in order to evaluate the performance of the newly implemented methods in a short time interval, hence the rest of the error correction techniques are not presented yet. The numerical model data, as well as the measurements, are hourly[34].

[33] 48 hours ahead is the forecast time interval.

[34] By hourly data it is implied that one measurement exists every hour.

4.5.2. Evaluation Measures/Metrics

Since this is only a preliminary stage in comparison to the consequent veritable tests, presented in the following chapters (Chapter 5), the only evaluation metric calculated and presented here is the root-mean-square-error (RMSE). Certainly, it is extremely difficult to evaluate a forecast consistently using simple metrics, let alone just one of them, since a "good" forecast is a matter of subject and application. In later stages of this research, more measures are incorporated to evaluate the accuracy and performance of the various error correction techniques, focusing more into specific aspects of the application in hand (i.e. maintenance operations in offshore wind farms). In this part we are confined into commenting and reasoning on the applicability of the BN and ARIMA models, without trying to take a final decision or draw any ultimate conclusions on their accuracy.

4.5.3. Preliminary Results

Below, the results of the preliminary analysis (limited to some days) are presented, demonstrating the performance of the BN and ARIMA models individually and in comparison plots. The individual performance graphs of just one month are attached in Appendix A, since the overall performance can also be seen in the corresponding comparison/summative plots (see Figure 18). The rest of the graphs, corresponding to the remaining months, can be found in Appendix A.

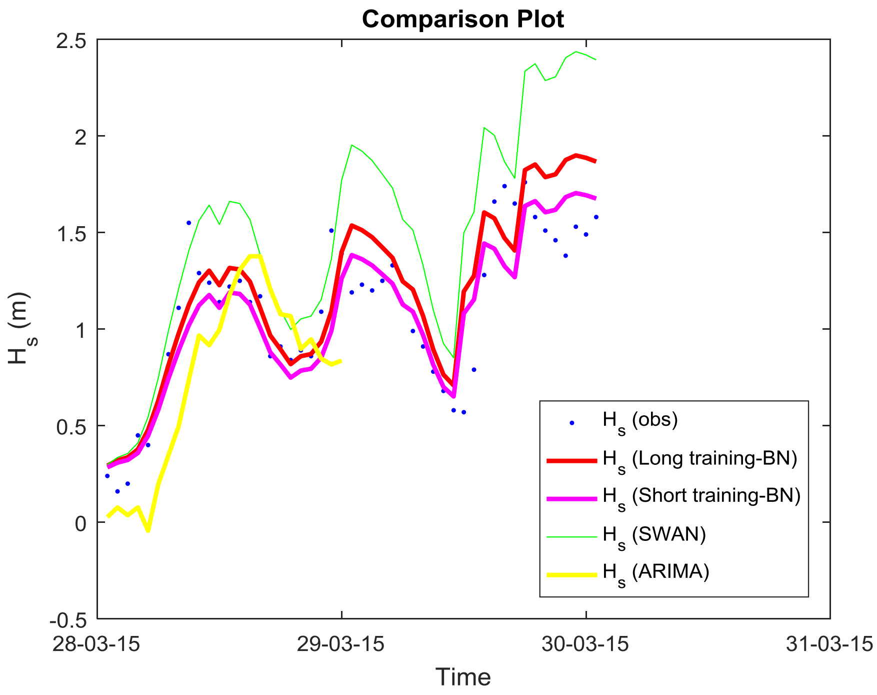


Fig.18. Comparison between the newly implemented correction techniques for 28-30 March 2015, at Rhyl Flats (Irish Sea).

It becomes obvious, by looking at the graphs, that the ARIMA forecast displays a serious time lag. Also, it is virtually impossible to produce results after the 24 hour mark, since the training the model has is not sufficient. In Figure 18, the training time of the ARIMA model is 8 days prior to the forecast, and the fitted model has an order equal to (0,1,1)(0,1,1) with a seasonality of 48 hours. The periodogram, produced by the Fourier Transform, from which the seasonality was computed, can be seen in Figure 19. Unfortunately, even with more training (10 days or even 15 days prior to the forecast) the model still displays an undesirable behaviour.

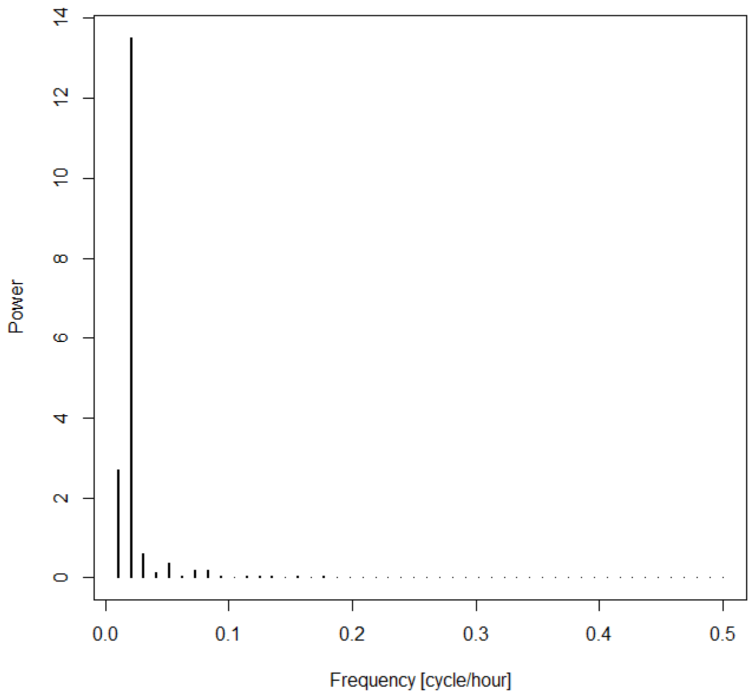


Fig. 19. Periodogram displaying the major frequencies extracted by the original dataset (28-30 March 2015 – Rhyl Flats).

Regarding the overall performance of the model, in terms of the RMSE, Table 1 presents the calculated values for each technique over the period of 28-30 of March 2015, so that a comparison with the numerical model can be made. As an addition, the maximum absolute error of each method was added, to give an indication of the biggest variation from the observations.

Table 1. Preliminary Evaluation Metrics (RMSE - MAE table) for 28-30 March 2015, at Rhyl Flats (Irish Sea).

Method	SWAN	Long-trained BN	Short-trained BN	ARIMA
RMSE (48 hours)	0.5015 m	0.2349 m	0.2002 m	-
RMSE (24 hours)	0.2528 m	0.1461 m	0.1867 m	0.3707 m
Maximum Absolute Error	1.0249 m	0.6246 m	0.5301 m	0.8230 m

As it can be seen in Table 1, but also in the presented graphs, the performance of the BN model is quite satisfying for the given days. In terms of both the RMSE and the MAE, the short-trained, as well as the long-trained BN outperform the numerical model (SWAN) and the ARIMA model. Another important aspect and highly positive aspect of the BN performance, is the lower value of the errors as we come closer to current time, i.e. the forecast the first 24 hours is even more accurate. As shown in Table 1, both BN methods display much lower RMSE values for the 24-hour than for the 48-hour forecasts. Nevertheless, the BN predictions/corrections fail to include the high value ($H_s \cong 1.6$ m) on the 28th of March (see Figure 18), which brings up the question on what do we truly consider a good forecast.

Certainly, encapsulating most of the behaviour of the observations is more important than simulating one or two really high values, and here is the point where application specific evaluation metrics have to come into play. Since this discussion diverges from the goal of this section, it will be addressed on a later chapter, when all the error correction techniques can be compared. For now, the RMSE is considered enough to present the general tendency of the methods' performance and accuracy.

4.5.4. BN Structure

Another component of the BNs, that has to be addressed, is their structure. As described previously the structure determines all the underlying relations between the variables, and largely affects the outcome of the prediction procedure. For these examples, due to the similar behaviour of the variables within a month (which is a relatively short interval), the structure is the same in terms of the arc direction for both the long – and short – trained BN. The components that changes and lead to slightly different results are the correlations between the variables. Figure 20 shows the BN structure, for March 2015, at the Rhyl Flats (RF) wind farm, while Tables 2 and 3 present the correlations between the variables. The correlations between the two alternative methods are close, leading to similar results. The whole procedure was completely data-driven, i.e. no relation was imputed a priori, and the predictions were produced according to the methodology described in Sub-Section 4.4.4.

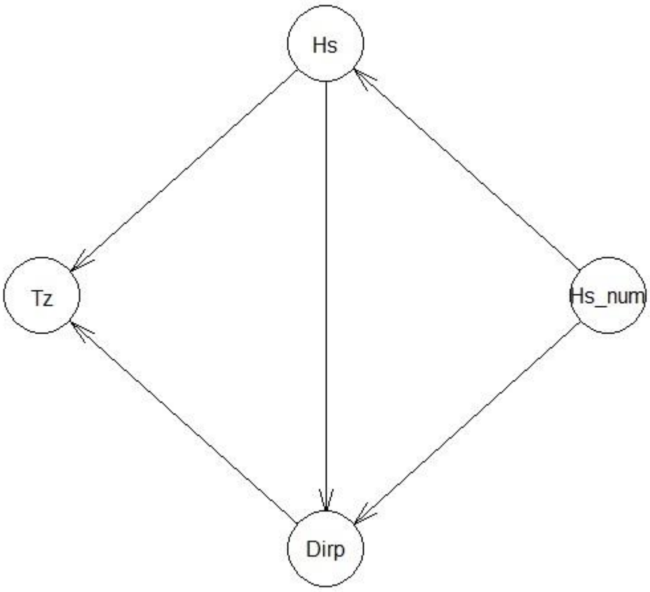


Fig. 20. BN structure for 28-30 March 2015, at Rhyl Flats (Irish Sea).

Table 2. Correlation matrix for short-trained BN, 28-30 March 2015 - Rhyl Flats.

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.6198	0.8028	0.8325
D_{irp}	0.6198	1	0.2967	0.9795
$H_{s,num}$	0.8028	0.2967	1	0.9682
H_s	0.8325	0.2795	0.9682	1

Table 3. Correlation matrix for long-trained BN, 28-30 March 2015 - Rhyl Flats.

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.5045	0.7649	0.7768
D_{irp}	0.5045	1	0.3260	0.2892
$H_{s,num}$	0.7649	0.3260	1	0.9671
H_s	0.7768	0.2892	0.9671	1

4.5.5. Further Testing

In order to have a more clear view, and to be able to draw some preliminary conclusions on the overall functionality of the newly implemented models, the comparison plots, and the evaluation metrics tables for chosen days on May, July, and August 2015 are presented in Appendix A. For the corresponding BN structures, as well as the graphs displaying the individual performances of each model, and the periodograms used in the ARIMA analysis, the reader is also referred to Appendix A.

4.6. Preliminary Conclusions

As expected, the partly artificial nature of the ARIMA model hinders the production of satisfying and robust results. Another important cause of this behaviour is probably the amount of training, which is insufficient for this particular application. Unfortunately, larger training sets elongate the time of training significantly, and as a result amplify the computational load extremely (waiting times up to 30-40 minutes for training with 70-80 % of the data), without actually providing significant accuracy enhancement. It is clear by the preceding graphs, as well as the ones displayed in Appendix A, that the ARIMA predictions are completely inaccurate, which is also the case in terms of the RMSE and MAE. With further tests and preparation of the input data, probably a more robust ARIMA model could be established per case, but this deviates from the goals of this particular research, since the real-time tenor of the particular application makes human intervention and judgement useless when the model runs. Consequently, the created ARIMA model was considered unsuitable for this application, and as a result its predictions will not be discussed or presented further in this thesis.

It is reasonable to suggest, since the data inserted in the ARIMA model are hourly, the seasonality of the timeseries data frame created has to be equal to 60 minutes. Just for demonstrational purposes, to exhibit the behaviour of the ARIMA with an imputed seasonality, the results of 28-30th of March 2015 are shown below (Figure 21). It is visible that this intervention not only did not enhance the model's accuracy, but clearly diminished it. A general conclusion is that it is really difficult to automatically create and ARIMA model, which encapsulates the natural behaviour of offshore hydrodynamic variables. As stated before, the construction of the model has to be more consistent and case specific, with an analysis taking into account the auto-correlation function (ACF), as well as smoothing procedures.

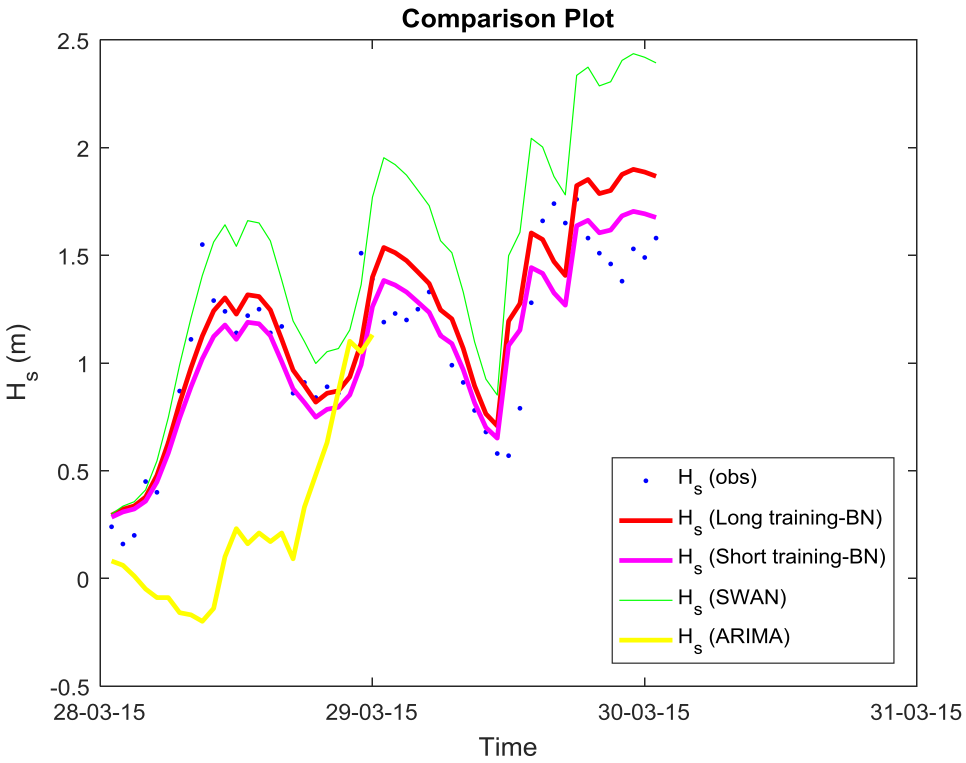


Fig. 21. Incorporation of an ARIMA model, with a 60 min seasonality setting applied, in 28-30 March 2015 (Rhyl Flats - Irish Sea).

The graphs and the metrics display the general tendency, of both BN methods, to enhance the forecasting accuracy in terms of the RMSE, and in some occasions of the MAE. Certainly, it is extremely difficult to judge the forecast's/correction's quality, just by looking at the presented numbers. In the remaining chapters of this thesis, a more concentrated evaluation of the predicting ability of the models will be made, by implementation of case specific metrics and application specific boundaries, connected to the time and wave height boundaries between which the prediction is of higher importance[35]. Especially for values close to the upper boundary of interest (usually H_s = 1.5 m), the results have to be as accurate as possible, since a wrong prediction on either side of the boundary might have a great impact in terms of safety and cost.

Nevertheless, the BN model preliminary results are quite promising with reference to the overall (whole period) and individual (separate months) performances. As a result, further analysis with extended time intervals of training and testing for the case of the Irish Sea, as well as comparison with the rest of the error correction techniques incorporated in Meteo Dashboard, were fully justified and can be found in the remaining of this research. An analysis, accompanied with comments, of the data including the variables to be used in the extensive and more robust method validation, can be found in the following chapter (Chapter 5).

[35] In offshore wind farm maintenance operations the critical wave heights are in the interval $0.5 \leq H_s \leq 1.5$ m.

5. Method Comparison

In Chapter 3 the preparation and manipulation of observational and numerical model data for multiple consecutive years was displayed. In this chapter the use of these data in order to train, test and compare the results of different error correction techniques (presented in Sections 4.3 and 4.4) is going to be discussed thoroughly. The time interval chosen for this particular analysis extends from 01-01-2014 to 31-12-2017 (4 consecutive years), for the measurement stations situated near the Gwynt-y-Mor (GyM) and Rhyl Flats (RF) offshore wind farms. In Section 5.1 a brief description of the models' configuration and the use of data sets in the training and testing procedures are shown. Section 5.2 includes a presentation of some representative examples of the models' behaviour, followed by a technique comparison based on general and case specific metrics, while Section 5.3 provides an overview and comparison of the prediction uncertainty provide by the BN and Copula models. In Section 5.4 the reader can find an analysis comparing solely different configurations and structures of the newly developed BN models, and their ability to produce accurate predictions for the case in hand. Finally, in Section 5.5 some remarks and comments on the accuracy and robustness of each error correction technique can be found, by means of which the most suiting model for this application can be chosen.

5.1. Model configuration and settings

5.1.1. Training and Fitting

As described in previous sections (Sections 4.3 and 4.4), each error correction technique incorporates different sets for training, while some of them do not need training at all. To be more exact, the simple linear regression and the stochastic interpolation take as an input only numerical data and measurements corresponding to a time interval just 48 hours prior to the forecast. The artificial neural network (ANN) and the copula (Gumbel) are trained with 6 months of data, corresponding to the months of March – September 2015, and then used implementing the same input delineated for the aforementioned techniques.

The newly developed Bayesian Network (BN) models incorporate three different types of training; (1) long-training with data from 01-01-2014 to 31-12-2016, i.e. 3 years of training, (2) short-training with hourly data corresponding to 48 hours prior to the forecast, i.e. 2 days of training, and (3) a fixed structure, produced by 3 years of training (2014 - 2016), and fitted with data tallying to 48 hours prior to the respective 48-hr forecast, i.e. 3 years for training and 48 hours for fitting and retrieving the required variable relations, necessary to produce a prediction.

When producing a prediction with the BN model, there should be an input of the variables on the basis of which the conditional distribution is being produced (*conditionalization*). As described in Section 4.4 the numerical model forecast data (48 hours ahead) for the rest of the variables (T_z, U_{dir}, U_{10}, D_{irp}, $H_{s,num}$) are used as conditionalizing values to generate an accurate forecast for the variable of interest, namely the significant wave height (H_s).

5.1.2. Testing and Validation

For testing, validation, and comparison, data retrieved for the year of 2017 were used (01-01-2017 to 31-12-2017). In order to simulate effectively the real-time nature of the application, a forecast was corrected every six hours of each day. Because SWAN produced 4 forecasts per day, one every 6 hours, each one of the error correction techniques, generated a potential corrected (potentially more accurate) prediction an equal number of times. It can be easily realized that the extremely large amount of information makes it absolutely impossible for all the results to be presented. Thus, a representative set, displaying different types of behaviours, is going to be exhibited in the sections to follow. Supplementary, the individual forecast metrics, as well as the ones encapsulating the performance over the whole year, were calculated, in order to show and evaluate the application-focused and more general model behaviours respectively.

5.2. Technique Comparison – Yearly Tests

5.2.1. Timeseries and Evaluation of 48-hr Forecasts

In order to establish a consistent evaluation of the error correction techniques, and conclude which one serves the purposes of the offshore significant wave height predictions better, yearly tests were conducted, involving numerical model data and measurements from 01-01-2017 to 31-12-2017. To start with, the timeseries incorporating all the available techniques can be seen below. For the reader to be able to distinguish the predictive performance of the newly developed BN error correction model, plots comparing solely the three (3) implemented BN techniques are also displayed.

In Figures 22 and 23 two crucial cases for the maintenance operations in offshore wind farms are introduced. The crucial nature of these cases, as well as other similar ones presented later, stems from the level over which nautical operations within the wind farm seize to take place, i.e. the significant wave height (H_s) upper boundary of 1.5 m. It is shown in the following graphs that SWAN over-predicts the wave height to a level where no operations could take place, while the observations, as well as the long-trained BN predictive model, providing a value smaller than 1.5 m. Exactly the same behaviour can be seen 6 hours later, for the next 48-hr SWAN forecast (see Figure 23).

In terms of the difference in the RMSE or MAE, this error might not indicate a behaviour that arises any worry for the models accuracy, even if the long-trained BN would still be more accurate. In a real-life situation such a wrong prediction may encapsulate a large risk for the maintenance operation, since in monetary terms this miss-prediction might cost thousands of euros (€).

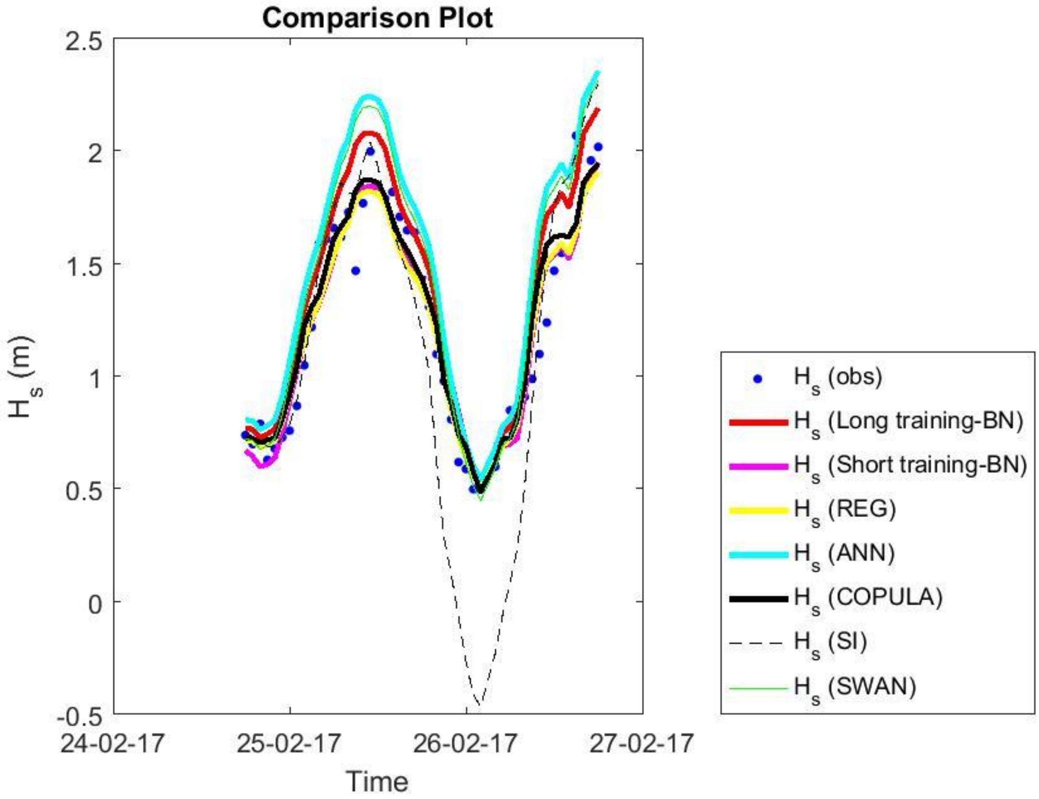


Fig. 22. Significant wave height (H_s) predictions produced for Gwynt-y-Mor at 18:00 on 24-02-2017.

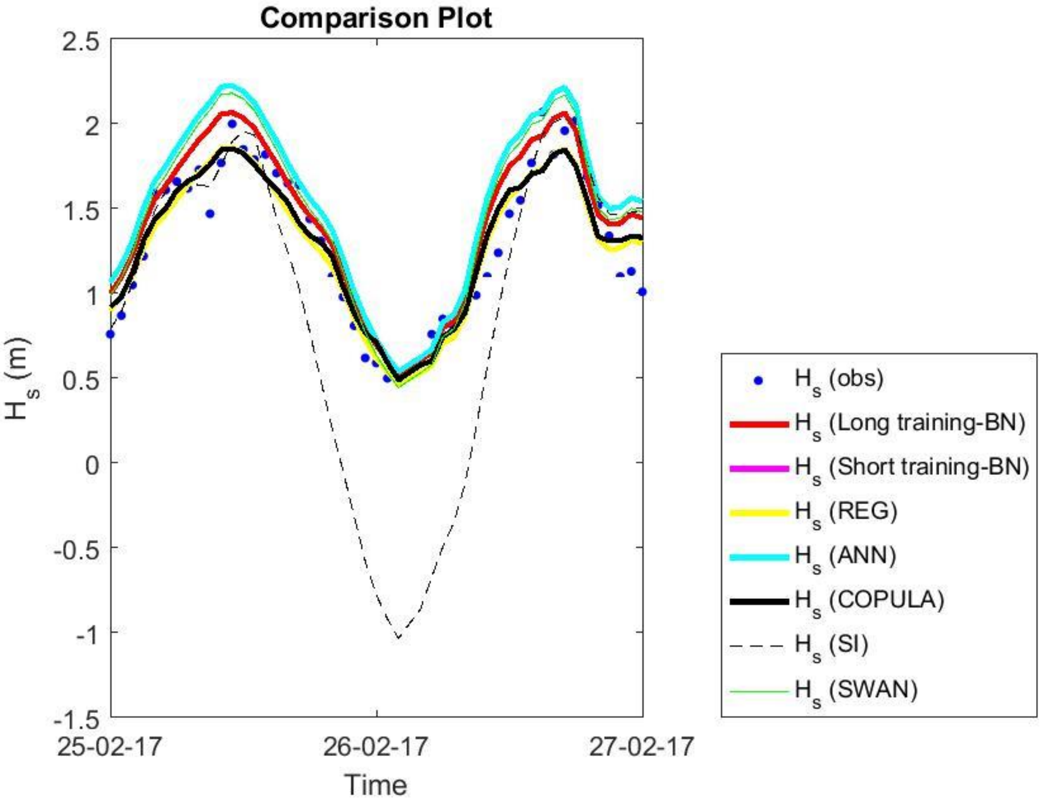


Fig. 23. Significant wave height (H_S) predictions produced for Gwynt-y-Mor at 00:00 on 25-02-2017.

The aforementioned issue regarding the robust and consistent validation of each prediction can be resolved with the use of case specific metrics, i.e. indicators displaying the models' accuracy within and around the significant wave height boundaries of this specific application, i.e. $0.5 \leq H_s \leq 1.5$ m. Particular interest is focused around the upper boundary of 1.5 m, which is certainly the most crucial for the offshore maintenance operations. In each individual forecast there is a no point of displaying the performance in terms of a number, since the reader can distinguish easily whether a forecast is accurate in high values or not. Nevertheless, the individual case general evaluation metrics (RMSE, MAE, BIAS, and Unbiased RMSE) can certainly give an indication of how "good" a forecast is. For the cases of Figures 22 and 23, Tables 4 and 5 give the corresponding accuracy metrics.

Table 4. General model accuracy evaluation metrics for Gwynt-y-Mor at 18:00 on 24-02-2017.

Method	SWAN	BN Long Training	BN Short Training	REG	ANN	Copula	SI
RMSE (m)	0.179	0.146	0.149	0.142	0.169	0.142	0.373
MAE (m)	0.671	0.449	0.560	0.426	0.716	0.390	0.966
BIAS (m)	0.171	0.143	-0.006	-0.008	0.237	0.027	-0.146
URMSE (m)	0.179	0.146	0.149	0.142	0.169	0.142	0.373

Table 5. General model accuracy evaluation metrics for Gwynt-y-Mor at 00:00 on 25-02-2017.

Method	SWAN	BN Long Training	BN Short Training	REG	ANN	Copula	SI
RMSE (m)	0.236	0.144	0.196	0.144	0.281	0.147	0.687
MAE (m)	0.611	0.336	0.506	0.337	0.658	0.346	1.533
BIAS (m)	0.169	0.132	-0.016	-0.0165	0.232	0.007	-0.378
URMSE (m)	0.165	0.144	0.143	0.143	0.158	0.146	0.573

According to Tables 4 and 5 a universally and clearly better model is difficult to be chosen. The linear regression model (REG) demonstrates a generally satisfying behaviour, and so does the long-trained BN. Considering its accuracy in “high” value predictions, the long-trained BN can be distinguished as the most suitable one for this occasion. Certainly, a correction of the SWAN forecasts is achieved in the presented dates, which of course is the ultimate goal.

An interesting fact, displayed in the foregoing figures and tables, is the behaviour and accuracy of the Bernstein Stochastic Interpolation (SI). While the model provides excellent predictions in high and medium values, its lower value predictions are often really inaccurate and completely unrealistic[36]. This behaviour is noticed generally in the yearly predictions provided by this model, as for instance in the case of Figure 24, where the accuracy of the forecast is crucial (close to the 1.5 m upper boundary) and the rest of the methods were unable to predict the respective wave height values. That of course has a huge impact on the general metrics presented in the preceding tables. So, a general comment one could make is that the Bernstein Stochastic Interpolation can certainly simulate the high values, even if the evaluation metrics calculated for this technique are unsatisfying and do not seem to provide any correction.

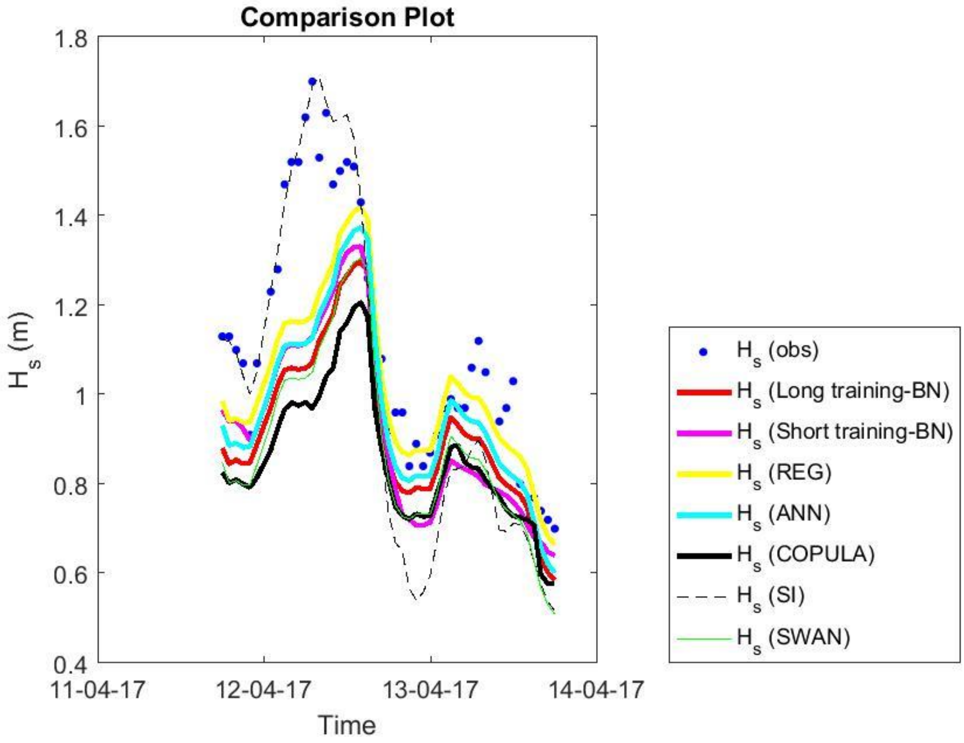


Fig. 24. Example of the SI model’s predictive accuracy, in comparison to the rest of the methods (Gwynt-y-Mor).

[36] It is impossible for the significant wave height (H_S) to display negative values.

5.2.2. The potential of Bayesian Networks

As displayed before, the long-trained BN error correction model can in many circumstances provide a robust and consistent prediction of significant wave height values. Nevertheless, such a performance is not limited only to this kind of BN models, but also to the short-trained as well as the ones whose structure is fixed. To be more exact there are various occasions which illustrate clearly the need and potential of each BN method. Especially in occasions similar to the ones presented in Figure 25, the BN techniques produced forecasts, which would be crucial in a real-life situation.

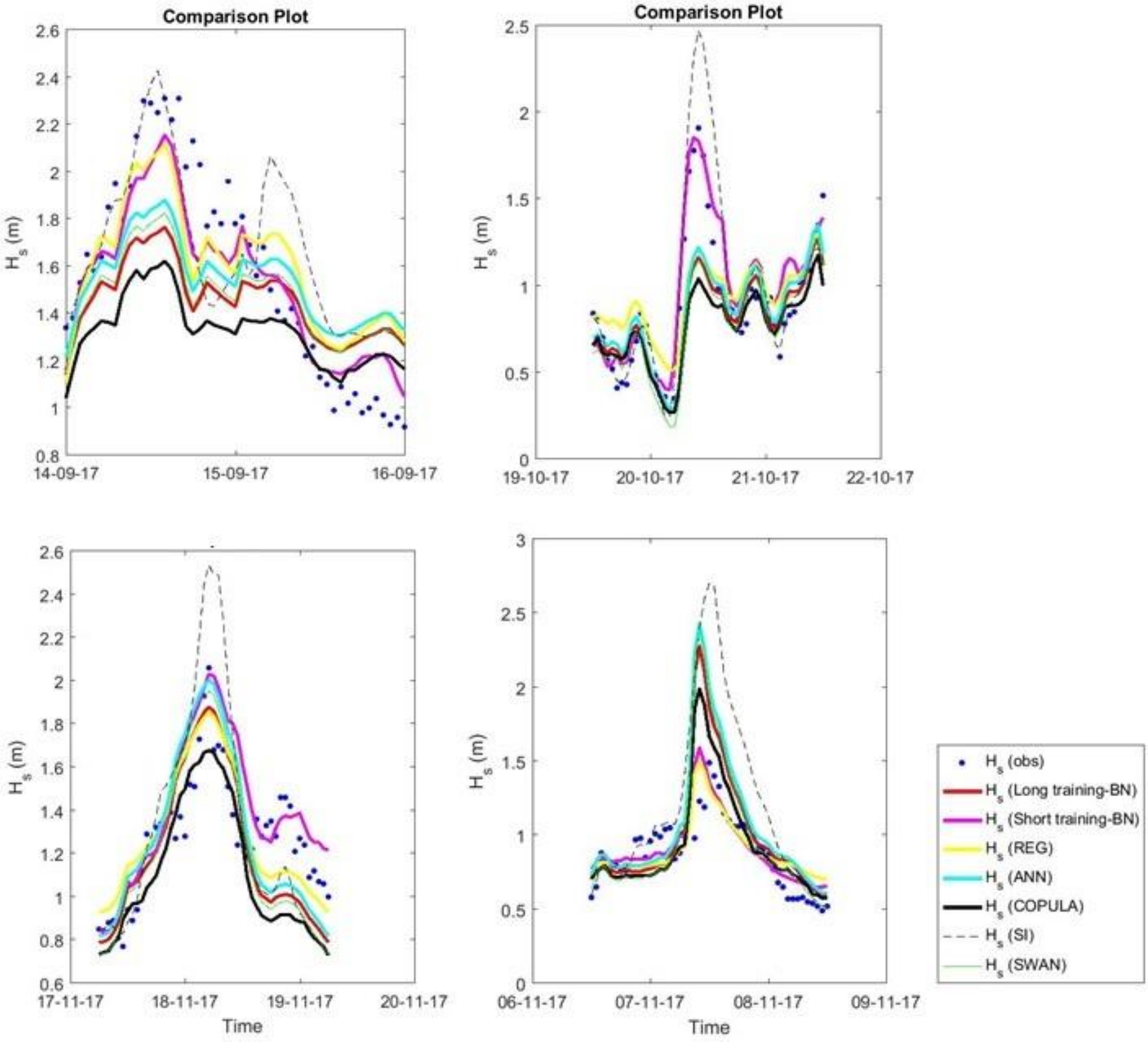


Fig. 25. Significant wave height correction (H_S) attempts, in which the BN methods displayed satisfying performance (GyM).

For instance in the top and lower left graphs of the above figure (Figure 25), the numerical model (SWAN) as well as some of the error correction techniques forecasted wave height values smaller than 1.5 m, while the reality was quite different. That would have a huge impact on the maintenance operation, which is translated in monetary terms, and potentially put human lives at risk. A similar example, which would not have an impact on the conduction of the operation, but might have a serious effect on decision making, is presented in the top right case, where the short-trained BN was the only model capable to simulate nature's behaviour.

The opposite would take place in the lower right graph's case, in which the numerical model, as well as the majority of the error correction techniques, predicted wave heights larger than 1.5 m. Again in this case, the short-trained BN method (and the linear regression) predicted realistic values, which would change the fate of a possible operation, with a correction reaching nearly 1 m in terms of the significant wave height (H_s).

Up until now, for display purposes, the fixed BN method's results have not been presented. But, this technique's performance is also of particular interest, since it provides useful and crucial corrections/predictions in various occasions (see e.g. Figures 26 and 27). A positive characteristic of the BN methods as a group is that when one of the techniques predicts poorly, one or both of the others display satisfying results, as for example in the case of Figure 26. Another benefit of the fixed structure in particular is that it provides satisfying results consistently, even in the absence of measurements from 48 hours prior to the forecast, while the short-trained BN is more dependent on the recent offshore climate, and as a result its behaviour can be erratic (left graph of Figure 26).

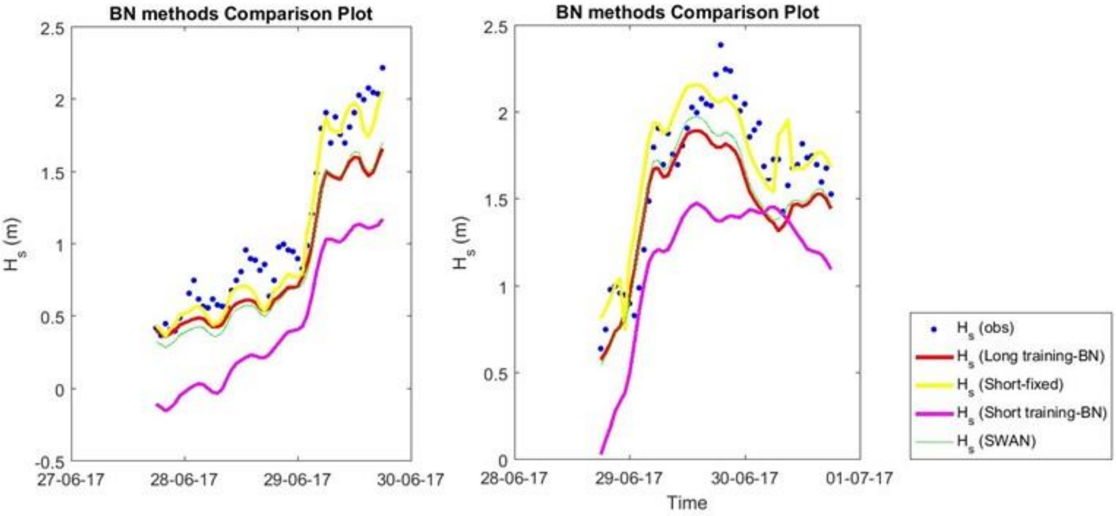


Fig. 26. Examples of the BN methods' predictive accuracy, with emphasis on the fixed structure BN (Gwynt-y-Mor).

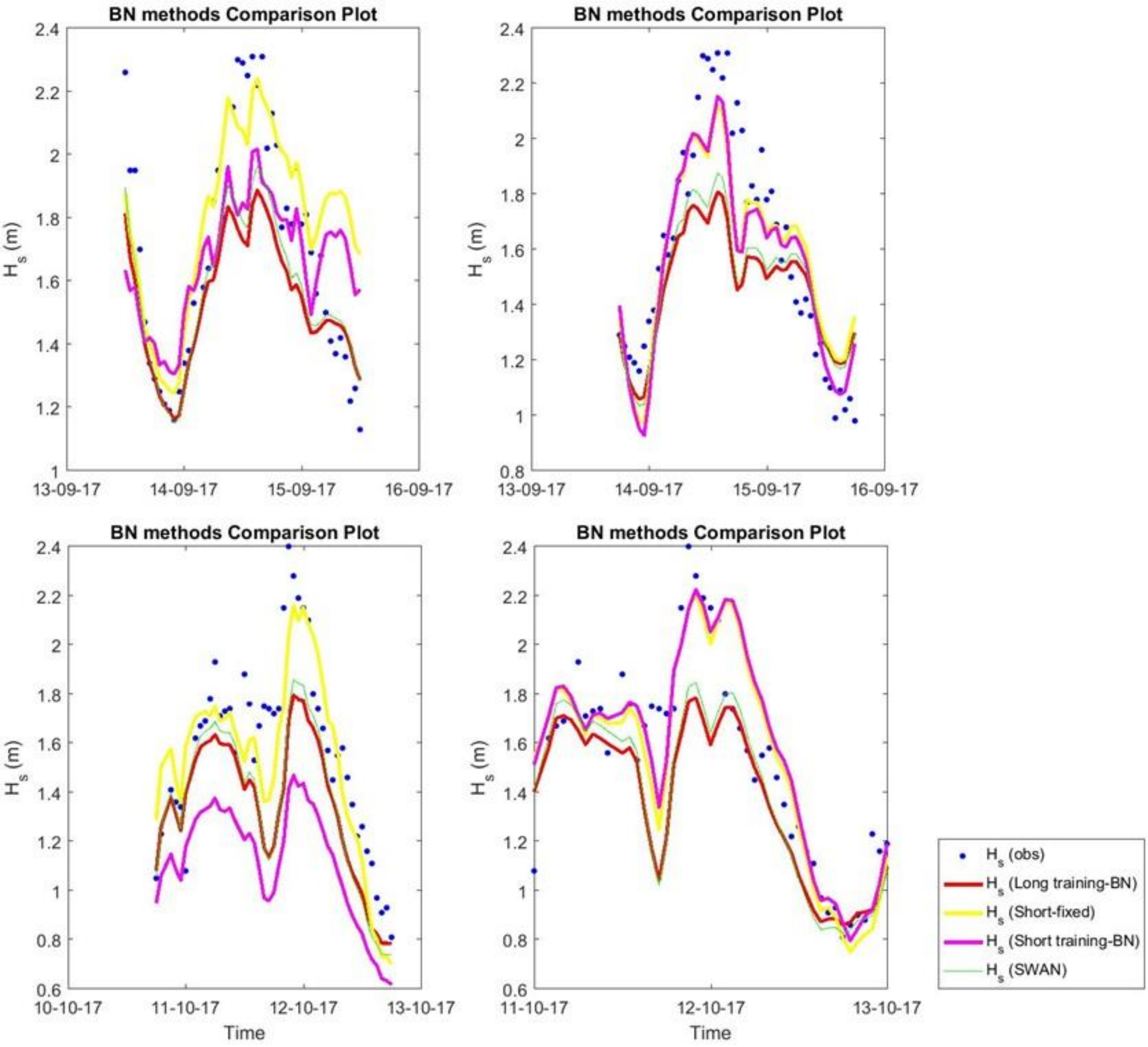


Fig. 27. Occasions in which the BN methods provided satisfying predictive accuracy (Gwynt-y-Mor).

From the cases of Figure 27, it becomes absolutely clear that all of the newly implemented BN error correction models can provide a significant accuracy enhancement in significant wave height (H_S) predictions. Also, all of the presented examples are crucial for the decision making procedures in offshore maintenance operations, supporting even further the role that the BN techniques can play in the Meteo Dashboard platform, or other similar applications.

In all the above figures, the lack of the corrections provided by the BN models would most probably condemn any offshore operations.

5.2.3. Summative (Yearly) Evaluation Metrics

To be able to establish which model serves the application better, the simulation was conducted for the whole year of 2017, as stated before. Table 6 shows the general evaluation metrics produced for the aforementioned time interval, for all the implemented techniques in the Gwynt-y-Mor wind farm case. The respective table for the case of Rhyl Flats can be found also in this section, so that the performance of the model for the whole area of the Irish Sea can also be evaluated.

Table 6. Yearly (2017) Evaluation Metrics for Gwynt-y-Mor

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	REG	ANN	Copula	SI
RMSE (m)	0.231	0.218	0.253	0.209	0.206	0.225	0.246	0.325
MAE (m)	2.410	2.360	2.607	2.728	2.907	2.397	2.566	3.999
BIAS (m)	-0.046	-0.011	-0.051	0.005	0.004	0.0365	-0.076	-0.016
URMSE (m)	0.226	0.218	0.248	0.209	0.206	0.222	0.234	0.324

Both the long-trained and custom-fixed BNs introduce an enhancement in accuracy, larger than any other method, with the exception of linear regression. Especially the custom-fixed BN’s RMSE is larger than the linear regression’s metric by 3 mm. Nevertheless, the fixed structure produce abnormal behaviour in certain occasions (such as spikes in the timeseries), which produce large errors, illustrated by the value of the MAE. The long-trained BN on the other hand shows a more consistent behaviour in terms of the maximum absolute error. In any case, both of the aforementioned techniques display satisfying performance in terms of their error distribution, which is reflected on the bias.

It has to be stressed out that evaluating the forecasting/correction performance solely based on Table 6 metrics is impossible. The metrics show the general behaviour of the models and are definitely indicative, but a more consistent evaluation has to be made, based also on application focused (case specific) indicators (see Table 7). In order to judge more robustly three different indicators were taken into account: (1) the percentage of the critically accurate predictions, i.e. the forecasts for which the measurements were higher than 1.5 m and the respective model managed to predict, (2) the false positive forecast percentage, which provides information on the amount of predictions above 1.5 m when the measurement was below, and (3) the percentage of the critically inaccurate forecasts, i.e. the amount of predictions below the 1.5 m upper boundary, when the measurement was above that limit. An important note is that the percentages were calculated over the whole time interval, i.e. in terms of the whole dataset, hence their values are small. In any case, they provide the much needed comparison in this stage.

Table 7. Application-specific Evaluation Metrics for the Gwynt-y-Mor case study.

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	REG	ANN	Copula	SI
Critically Accurate (%)	19.72	21.16	20.27	22.31	22.00	23.05	16.89	20.83
Critically Inaccurate (%)	2.55	2.82	3.79	2.10	2.34	1.90	4.72	1.96
False Positive (%)	2.26	1.93	1.50	2.10	1.97	3.01	0.82	3.01

By means of the above table (Table 7) it becomes clearer that the BN techniques provide a robust and consistent accuracy enhancement, and in combination with the metrics of Table 6, the foregoing analysis proves their suitability for the error correction scheme presented in this research. The custom-fixed BN is the one with the overall better performance in the case of the Gwynt-y-Mor wind farm. Certainly, the usefulness of Bayesian Networks is extended further if one considers the information provided by their structures, and the uncertainty estimation provided for the variable of interest. Those advantages will be clearly illustrated in the following sections.

In order to evaluate the performance of the error correction models for the whole area of interest, i.e. the Irish Sea, Tables 8 and 9 illustrate the general and application specific metrics for the case of Rhyl Flats. Once more the BN methods achieve to produce prediction of enhanced accuracy, serving the needs of the application to a satisfying degree across the spatial domain. The linear regression presents once more a better result in terms of the metrics, but the fact that it provides far less information, without any measure for the uncertainty of the variable of interest (i.e. the significant wave height), makes the BN methods more attractive and useful.

Table 8. Yearly (2017) Evaluation Metrics for Rhyl Flats.

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	REG	ANN	Copula	SI
RMSE (m)	0.203	0.178	0.200	0.201	0.163	0.212	0.187	0.275
MAE (m)	1.970	1.722	2.361	4.1731	2.361	1.991	1.769	2.991
BIAS (m)	-0.004	-0.010	-0.037	0.003	0.003	0.082	-0.054	0.002
URMSE (m)	0.203	0.178	0.196	0.201	0.163	0.196	0.179	0.275

Table 9. Application-specific Evaluation Metrics for the Rhyl Flats case study.

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	REG	ANN	Copula	SI
Critically Accurate (%)	18.02	17.01	16.04	18.82	18.02	20.64	16.64	18.49
Critically Inaccurate (%)	1.05	2.28	2.50	1.34	1.47	0.74	2.98	1.11
False Positive (%)	2.55	1.16	1.03	1.58	1.14	3.26	0.80	2.91

Although the performance achieved by the BN methods is quite satisfying, it can be seen (see Tables 8 and 9) that for the case of Rhyl Flats the network incorporating a fixed structure developed a slightly reduced performance in comparison to the respective Gwynt-y-Mor tests. It still provided a small accuracy enhancement, but it is evident from its MAE value that its behaviour in certain occasions was erratic. On the other hand, the long-trained BN method provided a satisfying correction in most occasions, nonetheless being less accurate than SWAN at certain moments.

5.2.4. Error Distribution

In Figures 28 and 29 the error distributions for the cases of Gwynt-y-Mor and Rhyl Flats are shown respectively. It is assumed that the errors follow a normal distribution[37], with the mean value being the bias, and the standard deviation the spread of the error (dispersion). To be more exact, the bias is the difference between the mean value of the distribution of errors given by the analysis and that of a standard normal distribution N (0, σ^2), i.e. one where the systematic errors are 0. In both occasions, the BN modules illustrate a small bias, i.e. a mean value for the error close to zero, and a smaller error spread in comparison to the rest of the error correction techniques and SWAN.

The erratic behaviour of the BN error correction technique incorporating a fixed structure can also be seen in Figure 30, where the different model errors' evolution in time is presented. Apart from the time interval were no data were available (January and February 2017) it can be seen that between July and September 2017 some large spikes appeared in the timeseries of the aforementioned technique. There the dependence of the H_s with the wave (D_{irp}) and wind direction (U_{dir}) becomes stronger and influences the forecast negatively, due to the large values existing in those two datasets.

The aforementioned issue will be adressed in Section 5.4, where an accuracy comparison of different BN structure configurations will be made. The respective error timeseries for the case of Gwynt-y-Mor is presented in Appendix C.

[37] According to the CLT (Central Limit Theorem) in statistics, the aggregation of a sufficiently large number of independent random variables results in a random variable which will be approximately normal. The error term can be thought of as the composite of a number of minor influences or errors. As the number of these minor influences gets larger, the distribution of the error term tends to approach the normal distribution.

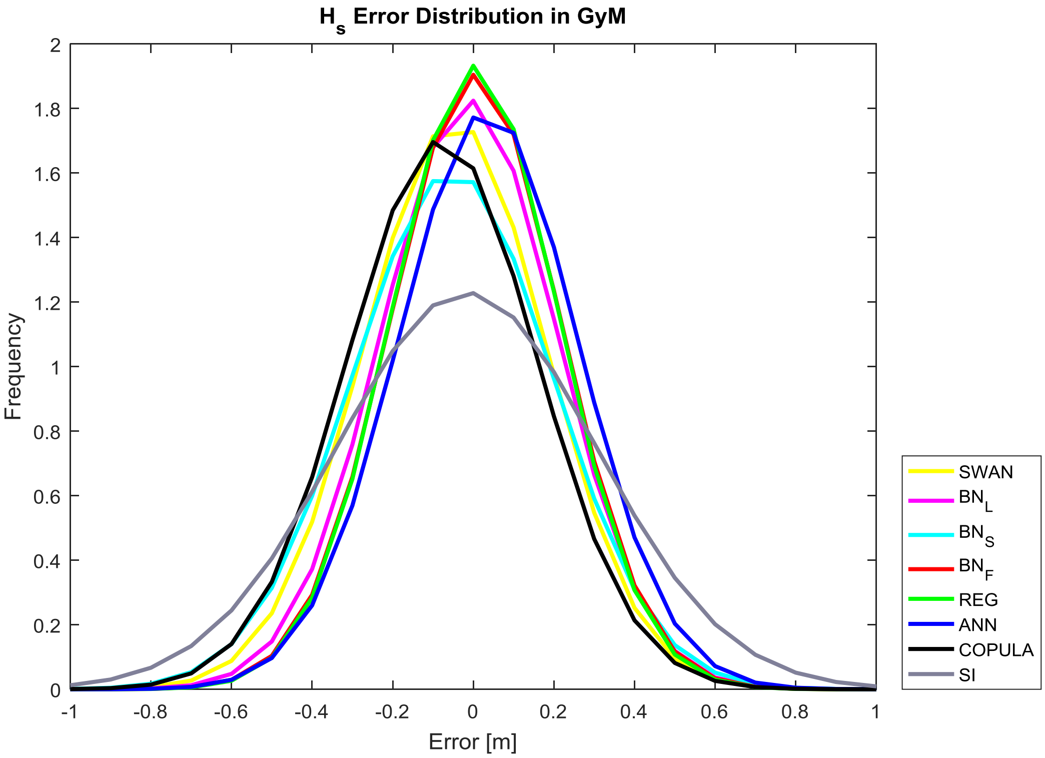


Fig. 28. Distribution of errors for the significant wave height (H_s) predictions at Gwynt-y-Mor (2017).

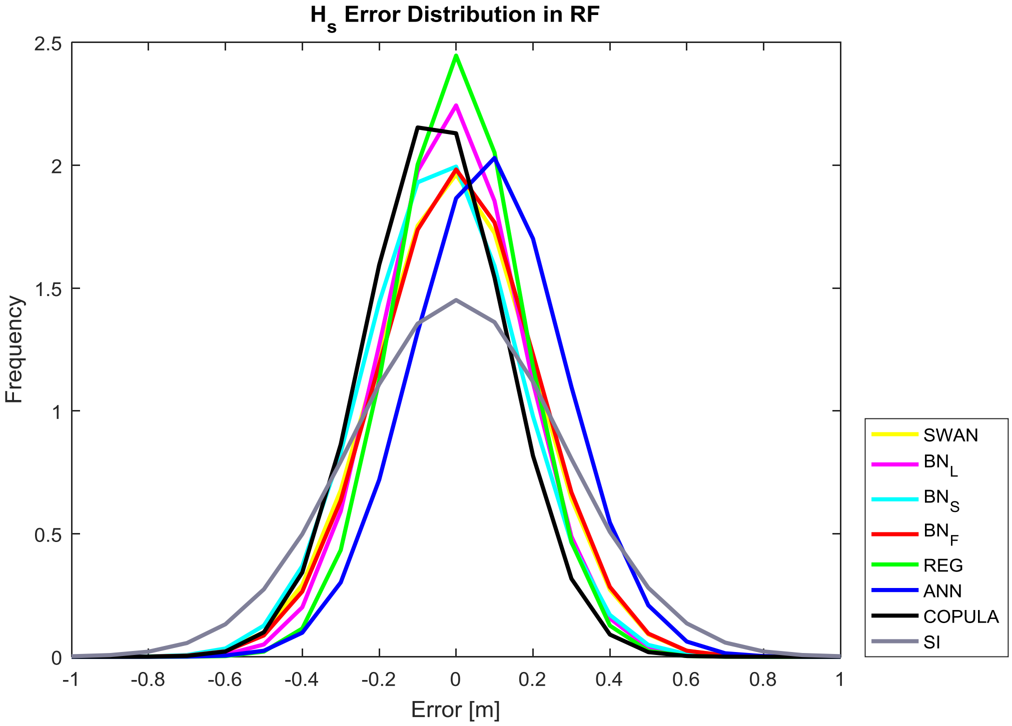


Fig. 29. Distribution of errors for the significant wave height (H_s) predictions at Rhyl Flats (2017).

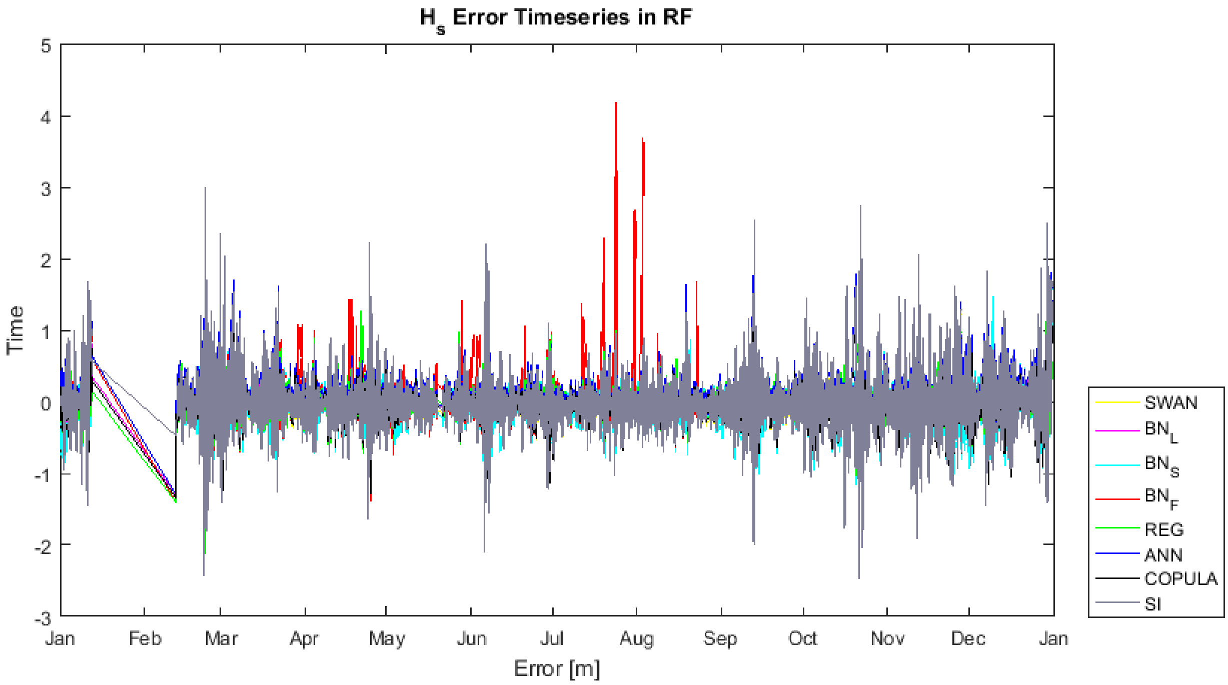


Fig. 30. Evolution of errors for the Rhyl Flats case study (2017).

5.2.5. Taylor Diagrams for the yearly predictions

Another way to evaluate the forecasting accuracy of different models is the Taylor diagram[38], which is a mathematical diagram designed to graphically indicate which of several approximate representations (models) of a phenomenon or process is the most realistic. Its main use is the quantification of the degree of correspondence between the modelled and measured behaviour in terms of three metrics: (1) the Pearson correlation coefficient, (2) the RMSE, and (3) the standard deviation. For more information the reader is referred to Taylor (2001).

Simulations that agree well with the observations will lie nearest the point marked "Observations" on the x-axis, i.e. models with relatively high correlation and low RMSE. It can be seen in Figure 31 for the case of Gwynt-y-Mor, that the linear regression and the BN model incorporating the fixed structures are the most suitable techniques according to the graph, something that was also evident from Tables 6 and 7. The respective Taylor diagram for the case of Rhyl Flats can be found in Appendix C.

[38] The Taylor Diagram was invented by Karl E. Taylor in 1994.

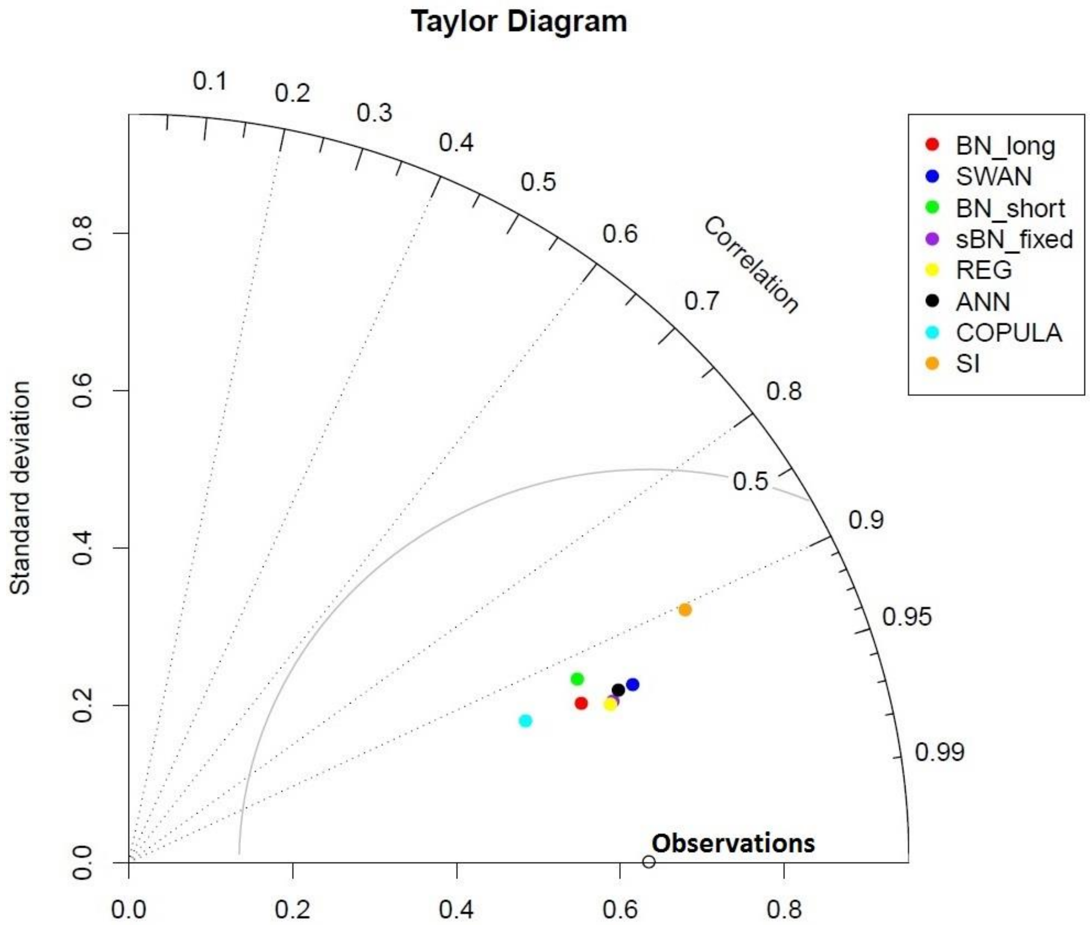


Fig. 31. Taylor diagram for significant wave height (H_s) predictions (Gwynt-y-Mor in 2017).

5.2.6. Scatterplots of EC Model Vs Observations

Any correlation between each model's predictions and the observations in hand, accompanied with a certain confidence intervals, can be distinguished by means of scatterplots (see Figures 32 and 33). The diagonal (red line) indicates a pure linear relation between the respective sets of data, and the scatter around it gives an indication of the uncertainty incorporated in the prediction.

In terms of the BN error correction models (Figure 32), the close relation between the predictions produced by the long-trained BN (top right) and the numerical model (SWAN) is evident. Regarding the uncertainty, the scatterplots of the aforementioned models are similar, while the short-trained BN model (bottom left) introduces a generally larger scatter under the diagonal. Thus it can be concluded by the graph that the short-trained BN has a tendency to under-predict the measurements.

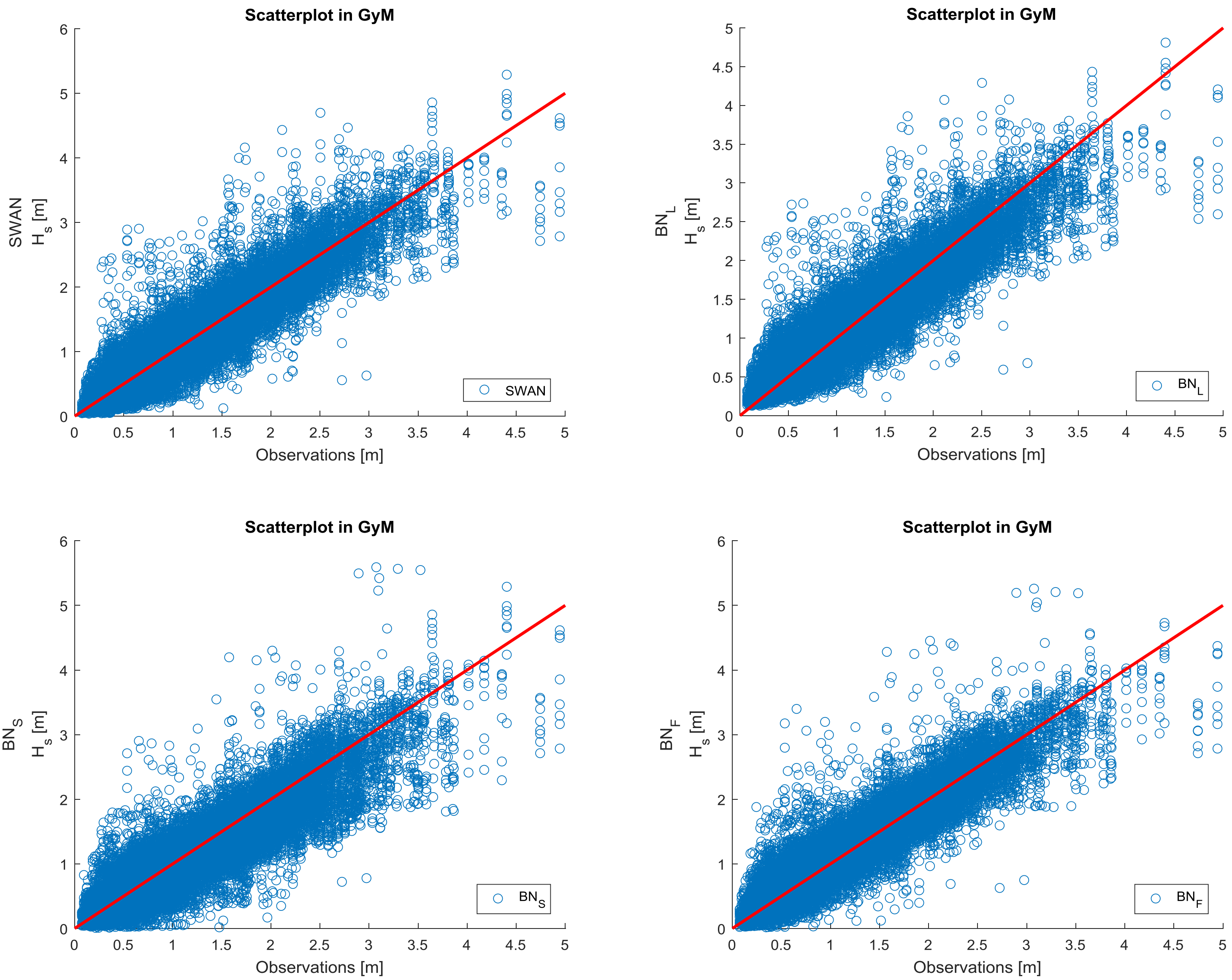


Fig. 32. Scatterplots of SWAN (upper left) and BN models in relation to the measurements (Gwynt-y-Mor).

On the other hand, the BN method incorporating the fixed structure (bottom right of Figure 32) seems to be more concentrated and equally distributed around the diagonal, introducing an accuracy enhancement, a fact also concluded in the previous sections (see Section 5.2.2). Certainly, there is some evidence of over-prediction, justified by the isolated values above the diagonal.

In Figure 33 the rest of the incorporated techniques' predictions in relation to the observations at Gwynt-y-Mor can be seen. Regarding the linear regression (top left) a similar behaviour to the fixed BN method can be distinguished. Nevertheless, some negative values were also produced by the linear regression model, leading to some unrealistic values below the diagonal, while a behaviour leading to over-prediction (outliers above the diagonal) is also presented in this case.

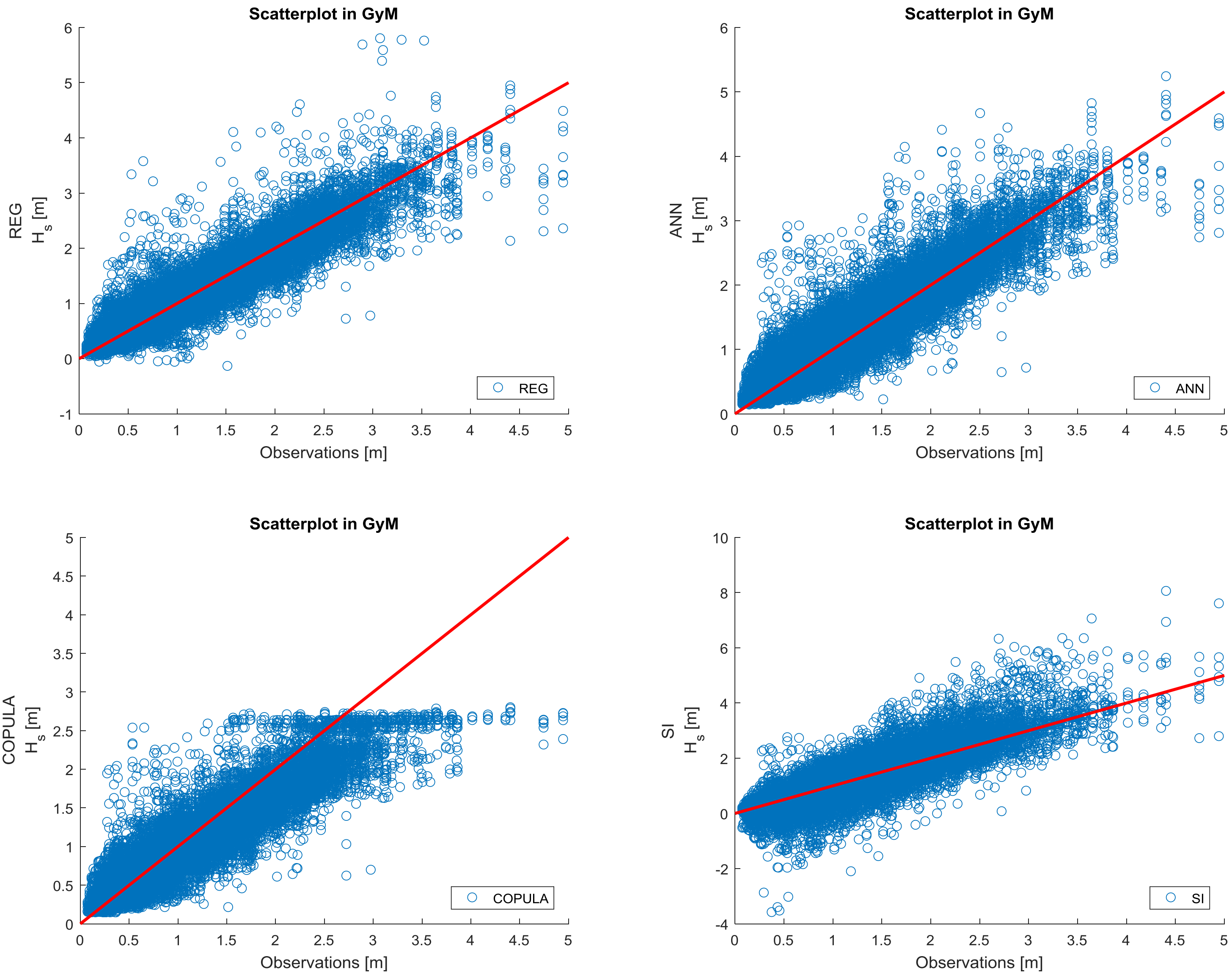


Fig. 33. Scatterplots of implemented error correction models in relation to the measurements (Gwynt-y-Mor).

The implemented Gumbel Copula seems to be unable to achieve high accuracy in the large significant wave height values, which was also reflected on the results of Table 7. Other than that, its values are concentrated around the diagonal, with a tendency of under-prediction (more values below the diagonal). The ANN model shows a tendency of predicting higher values than the ones expected, but its performance in terms of high H_s values is robust and consistent (really small number of isolated values above the diagonal).

Finally, the Bernstein Stochastic Interpolation (SI) generally over-predicts in comparison to the observed values, displaying also an unsatisfying general behaviour in lower wave height value forecasts, where unrealistic negative estimates seem to exist. Even if some of the SI model's forecasts are quite satisfying, the general behaviour of the model leads to the conclusion that this technique is not ideal for this particular application. In any case, the evidence provided in Figure 24 proves the stochastic interpolation's usefulness under certain conditions and supports its use in real-time corrections[39], since the computational load of this technique definitely allows its use, without hindering the forecasting procedure. For the respective scatterplots corresponding to the case of Rhyl Flats the reader is referred to Appendix C.

5.3. Uncertainty Evaluation

One major advantage of the newly created BN methods, in comparison to the already implemented techniques on the Meteo Dashboard platform, is their ability to provide estimates of the uncertainty governing the variable of interest; in the case of this research the significant wave height (H_s). The only one of the other techniques able to produce confidence intervals is the Gumbel Copula. Nevertheless, the assumption of a Gumbel Copula as the most suitable one for the data influences the confidence intervals' performance significantly.

Regarding the BN methods, as stated in Chapter (see Section 4.4), the assumption of multivariate normality for the conditional distribution of H_s governs the predictions. Although the above assumption could be quite restrictive in other applications, the predictions acquired by the BN models are quite satisfying, providing a correction of the numerical model (SWAN) forecast in the majority of the occasions in the time interval under consideration (Jan 2017 – Jan 2018). As a result of the aforementioned supposition (normal conditional H_s distribution), the uncertainty boundaries given by the BN models are symmetrical[40]. That of course does not condemn their performance or their usefulness, which is going to be examined thoroughly in this section. In comparison to the aforementioned confidence intervals, log-normal uncertainty bounds were also produced using a log-transformation of the data, justified by the log-normal distribution fitting the H_s dataset for both stations.

[39] Real-time corrections are connected to the individual performance on each forecast. Therefore, a generally inaccurate model might produce extremely satisfying predictions under certain conditions.

[40] The 95% uncertainty bounds are given by the 2.5th and 97.5th quantile of the conditional H_s distribution.

5.3.1. Normal and Log-Normal Uncertainty Bounds

Before moving to the comparison of the uncertainty bounds provided by the different techniques, it is helpful to present the methodology followed during the analysis. The production of the normal (symmetrical) confidence intervals is straight-forward. The use of the `cpdist` function, provided by the `bnlearn` package in R, makes things even simpler. Essentially, the conditional probability is given and by extraction of its 2.5th and 97.5th quantile, the 95% uncertainty bounds are obtained for each one of the point predictions[41].

Because the symmetrical nature of the uncertainty bounds is not quite realistic, a log-transformation also took place to obtain the log-normal boundaries corresponding to the distribution of the significant wave height (H_s). Summarily, the significant wave height (H_s) data were transformed to log-values (log-transformation) and then inserted into `bnlearn` package. The network was trained with the aforementioned transformed data and produced predictions and the conditional distribution of each point prediction. Those values were transformed back to their original form (i.e. 10^x) and that way the log-normal boundaries were obtained. Again the 2.5th and the 97.5th quantiles were used. It has to be stressed that the expected value (i.e. the point prediction) was different than the one obtained by the assumption of the multivariate normality (see Figure 34).

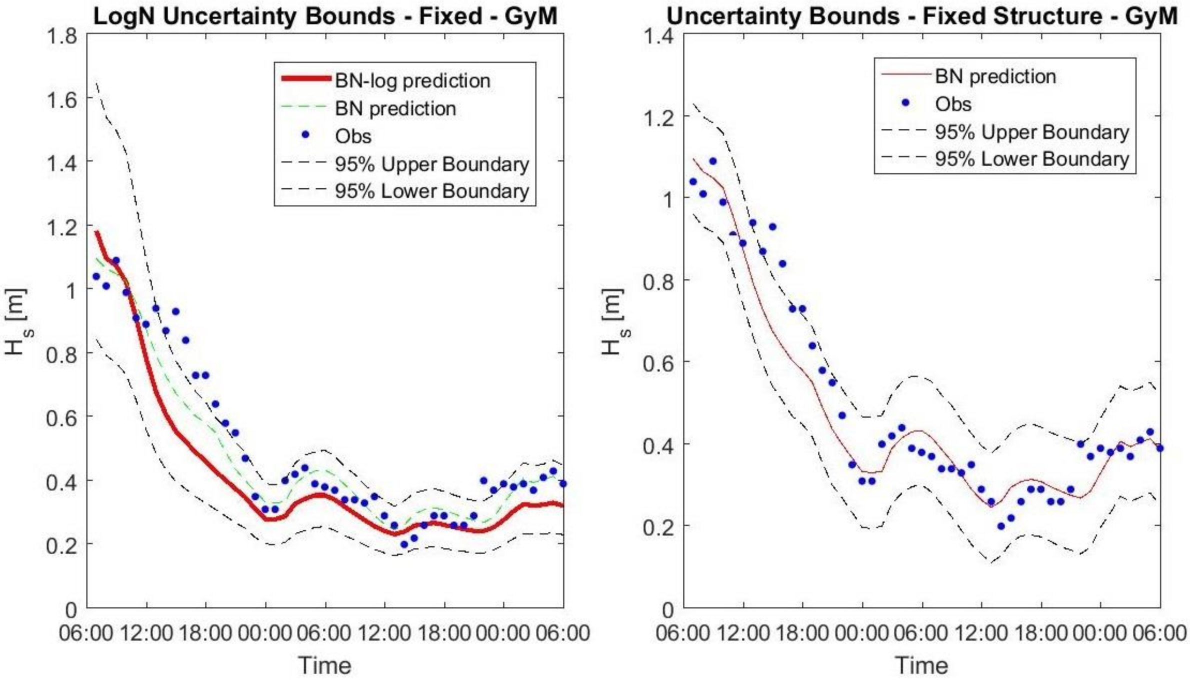


Fig. 34. Log-Normal (left) and Normal (right) 95% Uncertainty Bounds (fixed BN method for Gwynt-y-Mor on 09-03-2017).

[41] As stated in previous sections (see Section 4.4) the point predictions are the expected values of the conditional distribution.

In the case of Figure 34, the differences in terms of point predictions and the uncertainty bounds are visible. In this case the normal confidence intervals, despite their symmetrical appearance are more accurate and useful than the respective log-normal ones. That statement can be justified by taking a closer look at the coverage percentage, i.e. the amount of observations included in the interval in relation to the total number of observations, and the average length of the uncertainty interval, i.e. the average difference between the upper and lower boundaries. The normal uncertainty bounds (right graph) provide a coverage percentage equal to 90% (89.6%) of the total observations, in comparison to the 77% coverage provided by the log-normal boundaries. Regarding the average length of the uncertainty bounds, a comparison will be made incorporating the data sets for the whole year of 2017, in order to establish which uncertainty bounds serve the application satisfyingly. For the sake of completeness Figures 35 and 36 present the respective confidence intervals provided by the other two BN techniques, for both distributions.

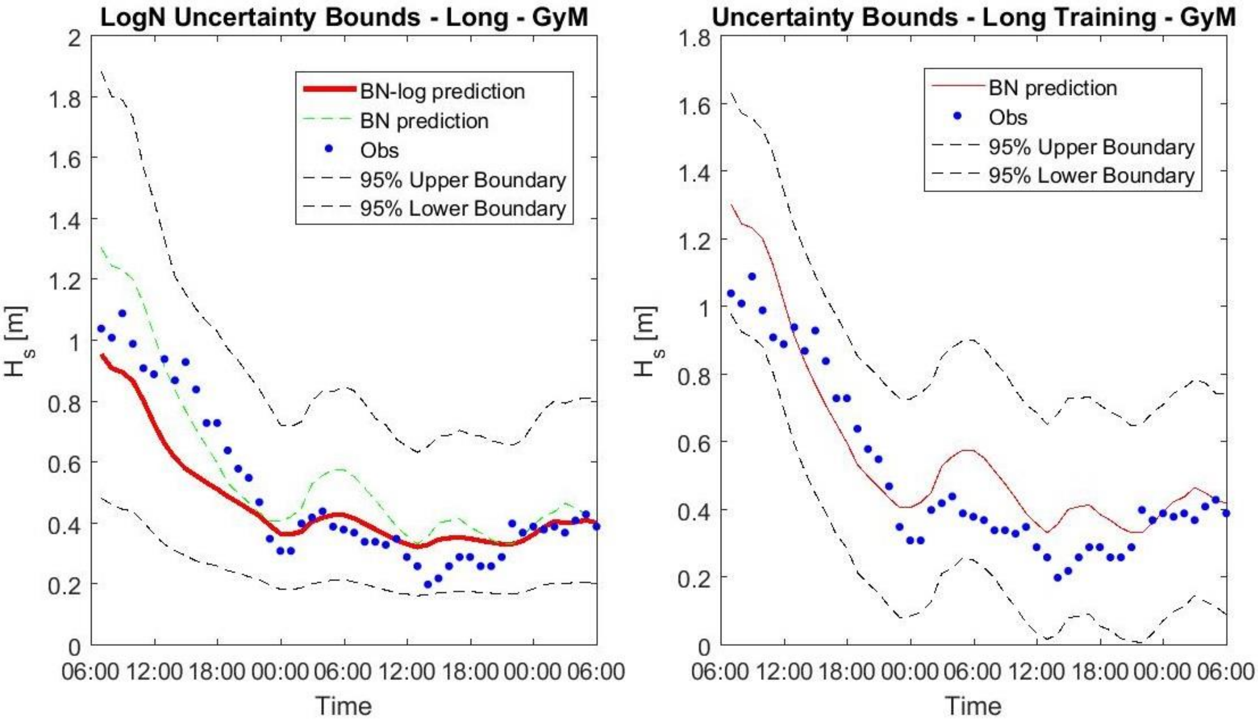


Fig. 35. Log-Normal (left) and Normal (right) 95% Uncertainty Bounds (long-trained BN for Gwynt-y-Mor on 09-03-2017).

Both intervals for the case of the long-trained BN (Figure 35) include all the observational values, providing, nevertheless, quite unrealistic uncertainty bounds with really large length. On the other hand, for the short-trained BN (Figure 36) the log-normal uncertainty bounds give a coverage percentage equal to 92%, while for the respective normal ones the coverage is 88%. Although the coverage percentage is larger for the log-normal confidence interval, its length is larger and in certain points unrealistic, which raises the question of when the uncertainty estimate is more useful.

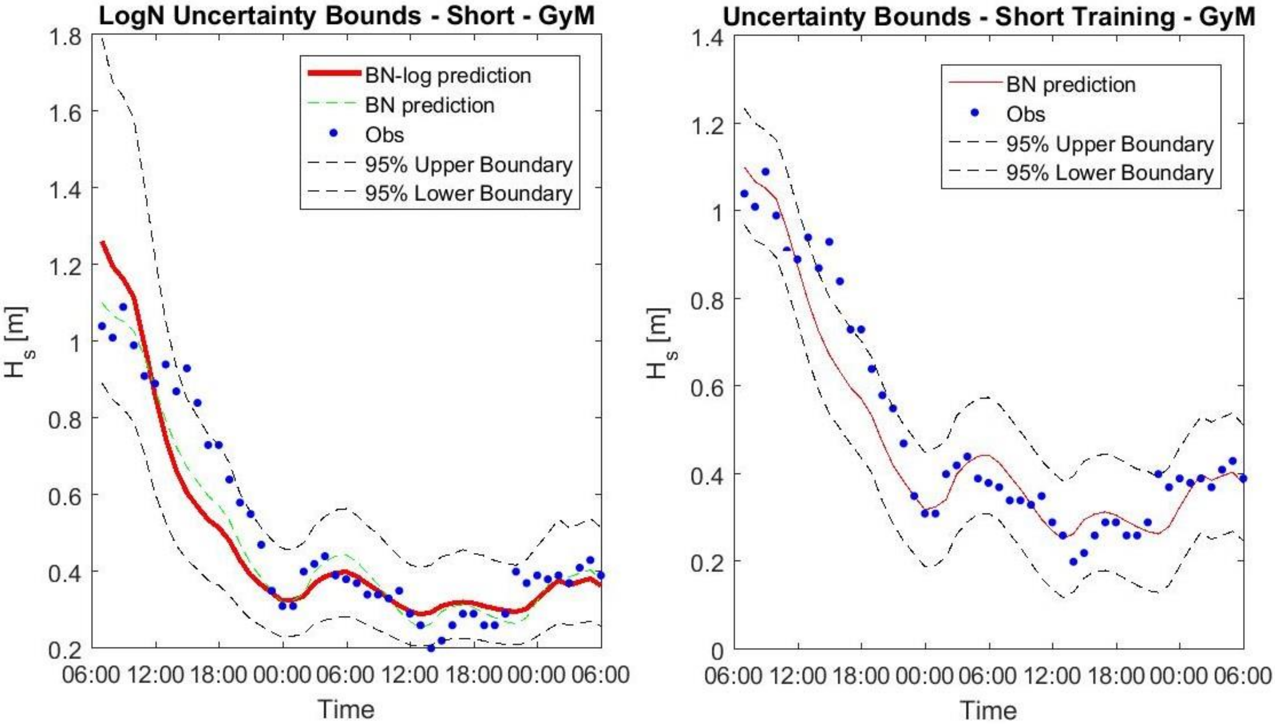


Fig. 36. Log-Normal (left) and Normal (right) 95% Uncertainty Bounds (short-trained BN for Gwynt-y-Mor on 09-03-2017).

5.3.2. Comparison of the provided Uncertainty Bounds

The following tables (Tables 10 and 11) provide the necessary information for the evaluation of the uncertainty estimates provided by each error correction technique. It can be seen that the log-normal uncertainty bound provide smaller coverage percentages with similar or larger average lengths. As a result the normal confidence intervals are more efficient and accurate. The most useful uncertainty boundaries seem to be the ones provided by the BN model incorporating the fixed structure, which have a high coverage percentage (86.1%) accompanied by a satisfying average length.

Table 10. Uncertainty Bounds comparison for the Gwynt-y-Mor case study.

Method	BN Long Training	BN Fixed Structure	BN Short Training	Copula	BN Long Training (Log-N)	BN Short Training (Log-N)	BN Fixed Structure (Log-N)
Coverage (%)	89.2	86.1	75.3	68.5	95.4	73.1	76.5
Average Length (m)	0.630	0.531	0.356	0.375	1.185	0.550	0.594

Table 11. Uncertainty Bounds comparison for the Gwynt-y-Mor case study (negative values included)t.

Method	BN Long Training	BN Fixed Structure	BN Short Training	Copula	BN Long Training (Log-N)	BN Short Training (Log-N)	BN Fixed Structure (Log-N)
Coverage (%)	89.2	86.1	75.3	68.5	95.4	73.1	76.5
Average Length (m)	0.639	0.533	0.464	0.375	1.185	0.550	0.594

Due to the nature of the data, there might be values that are close to 0. In such case the symmetrical normal boundaries will include negative values, since the expected value, i.e. the point prediction, will also be close to 0. Certainly, negative wave height values do not exist, and as a result the lower boundary was set to 0 for those occasions. Table 11 provides the values of the average length, when those negative values are included in the analysis, with the differences not being significant.

Considering the overall performance in terms of the given uncertainty, in combination with the point predictions provided in the preceding sections (see Section 5.2.3) it seems that the BN method incorporating a fixed structure, alongside with the respective normal confidence intervals, is the most suitable one for the Gwynt-y-Mor case study. The long-trained BN normal boundaries have also a steady and robust performance, which also makes them an attractive and satisfying model.

Finally, is has to be noted that the extremely large coverage given by the log-normal uncertainty bounds, for the case of the long-trained BN model, is justified by the similarly large average length of the intervals, which makes the solution less suitable. Definitely, the log-normal boundaries have a more realistic form (i.e. only positive values and a match with parametric distribution fitting the H_s well), but in case the performance is taken into account the normal confidence intervals pose many advantages. The respective tables referring to the Rhyl Flats case can be found in Appendix C.

5.4. BN Structures and Configurations

This section consists of two major parts, both of them referring to the time interval under consideration, i.e. the year of 2017. The first one includes the BN structures incorporated in the simulations so far, the relations between the implemented variables, and an analysis of the information obtained by those relations. The second part introduces different BN structures and configurations, with an altered number of incorporated variables, as well as a comparison between the predictions and uncertainty provided by them and the 6-node structure implemented in the preceding examples. Also in this part the dependence between the variables will be discussed, in order to establish a clearer view of the governing relations existing in the application.

5.4.1. BN Structures with 6 Variables

Up until now in this research, the BN structures incorporated involved 6 nodes (see Figures 37 and 38), corresponding to the following variables: (1) the observed significant wave height (H_s), (2) the numerical significant wave height ($H_{s,num}$) obtained by SWAN, (3) the wave directions (D_{irp}), (4) the zero-crossing wave period (T_z), (5) the wind velocity 10 m above the sea surface (U_{10}), and (6) the wind direction (U_{dir}). The simulations were carried out using data driven structures, i.e. structures acquired by the nature of the data and not imposed a priori. In general it was noted that trying to create a structure using general knowledge on the incorporated variables (i.e. knowledge on the underlying relations procured by the literature or by experts) only hindered the prediction/correction procedure instead of enhancing its accuracy.

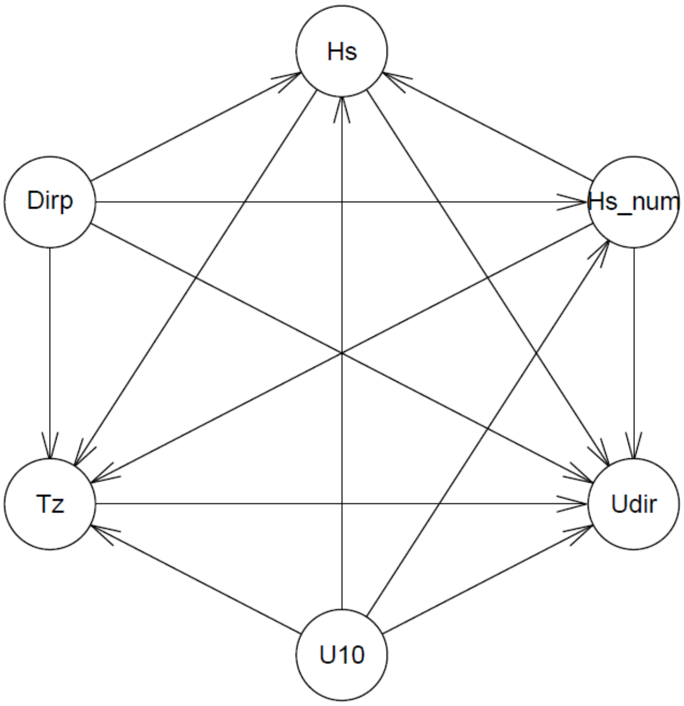


Fig. 37. Structure for the long-trained and fixed BN models, incorporating 6 variables (Gwynt-y-Mor).

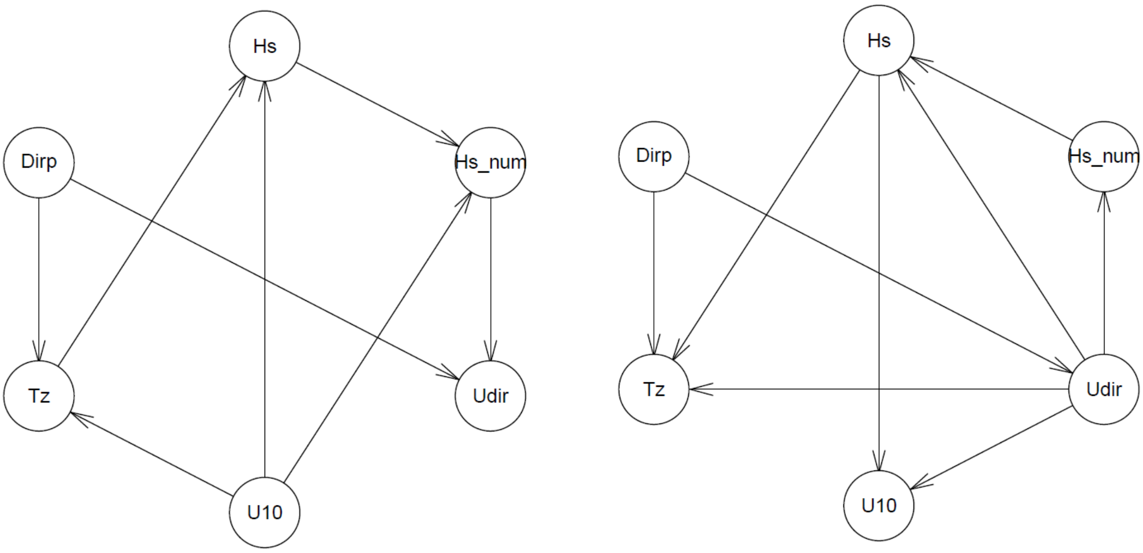


Fig. 38. Structures for the short-trained BN model incorporating 6 variables for the case of Gwynt-y-Mor.
(29-12-2017 at 18:00 and 17-11-2017 at 06:00)

The structure presented in Figure 37 is constant over time, i.e. the direction of the arcs remains the same, while the structure connected to the short-trained BN model (Figure 38) is continuously changing, since it depends solely on data retrieved 48 hours prior to the forecast. Some of the relations governing the first of the aforementioned structures (constant) are anticipated, when others oppose what would be expected by the common knowledge on the variables in hand.

The most distinctive examples here are the relations between the observed significant wave height (H_s) and the wind velocity (U_{10}), as well as the wind (U_{dir}) and wave (D_{irp}) directions. In a situation represented by the dependencies described in the literature, supported by common knowledge by experts, one would expect the wind direction to influence the wave direction, i.e. the arc connecting those two nodes to have a direction from U_{dir} to D_{irp}. In both occasions presented in the preceding figures, the opposite occurs. The data-driven analysis implies that the wind direction depends on the wave direction, something which is certainly not the case in reality. But here a quite reasonable explanation exists, justifying this kind of behaviour. The wind and wave directions are measured at the same locations, a fact that insinuates that the variables influence one another in one specific area. Still, waves are created by storms occurred many kilometres (or miles in the nautical language) away from the location of the measurement. As a result, the measured wind directions might indeed not have any influence on the wave directions. Further, the wave direction is influenced by many effects, such as diffraction due to islands or other obstacles, so their direction can be totally irrelevant to the values given by the wind direction. That of course raises the question on whether the wind direction could be omitted by the analysis, which will be addressed in the following sections.

On the other hand, the significant wave height and wave direction relation is in reality two different stories. For the case of the long training (3 years of data), presented in Figure 37, the relation is the one expected by the descriptions available in the literature, corresponding to the experts opinions. To be more exact, the wind velocity influences the significant wave height, a dependence which is highlighted by the high correlation between the variables, shown in Table 12 (a correlation coefficient equal to 0.795). In the same table other relations are also visible, as for instance the wind and wave direction relation, which justifies the structures form. Also visible is the extremely high dependency between the observed and numerically derived wave heights, which gives the character of correction instead of pure prediction to this whole research, since the quality of the numerical model (SWAN) results influence highly the long-trained model’s accuracy.

Contrarily, the short-trained BN model provides a variety of relations between the wind velocity and the observed significant wave height, due to the dynamic nature of the offshore events, which force the data to rapidly change behaviour. Figure 38 clearly illustrates that there is no clear relation between the two aforementioned variables, since the direction of the connection changes repeatedly, and in some occasion becomes even inexistent (no connection at all). That of course is again explained by the nature of wave creation by distant storms, or secondary events like diffraction or reflection, since also those two variables are measured in the same location. The correlation presented for the long training set just proves that in most occasions the wind velocity magnitude influences the waves, but does not in any case make this absolute for all day-to-day cases.

Table 12. Correlation matrix for the long-trained BN model structure (Gwynt-y-Mor).

Variable	D_{irp}	T_z	U_{10}	U_{dir}	$H_{s,num}$	H_s
D_{irp}	1.000	0.381	0.001	0.515	0.245	0.249
T_z	0.381	1.000	0.596	0.359	0.842	0.874
U_{10}	0.001	0.596	1.000	0.110	0.820	0.795
U_{dir}	0.515	0.359	0.110	1.000	0.319	0.329
$H_{s,num}$	0.245	0.842	0.820	0.319	1.000	0.964
H_s	0.249	0.874	0.795	0.329	0.964	1.000

Table 13 is an indicative example of the erratic behaviour of the short-trained BN model's structure, when it is compared to the correlation matrix given for the long training case (Table 12). It is obvious that the relation between the wind velocity and the observed significant wave height is much weaker than the one presented previously. The same holds for the relation between the wind and wave directions, which is almost inexistent (correlation coefficient equal to -0.003). Since the underlying relations (dependencies) change rapidly and very dynamically, the example correlation matrix presented in Table 13 supports the unpredictable predicting performance of the short-trained BN model, which also has an erratic character as described in previous sections.

Table 13. Correlation matrix for the short-trained BN model structure on 29-12-2017 at 18:00 (Gwynt-y-Mor).

Variable	D_{irp}	T_z	U_{10}	U_{dir}	$H_{s,num}$	H_s
D_{irp}	1.000	0.659	0.138	-0.003	0.422	0.573
T_z	0.659	1.000	-0.094	0.109	0.462	0.774
U_{10}	0.138	-0.094	1.000	0.281	0.416	0.297
U_{dir}	-0.003	0.109	0.281	1.000	0.564	0.391
$H_{s,num}$	0.422	0.462	0.416	0.564	1.000	0.687
H_s	0.573	0.774	0.297	0.391	0.687	1.000

Regarding the fixed structure BN model which retrieves the structure form (direction or even existence of certain arcs) from the large dataset of the long training (3 years), but is predicting according to the relations obtained by data acquired just 48 hours prior to the forecast, the underlying relations are essentially stable in terms of direction, since the long-trained structure remains constant. The mixing of the long and short training sets to create essentially a hybrid structure, influences the forecasting accuracy in most occasions, as shown in Section 5.2, even if under certain conditions the behaviour of this model can be varying as well. Finally, as stated previously in Section 4.5.4, it is preferable for this application to allow the structure to be completely data-driven, when a sufficiently large amount of information is available (such as the long training set), since the real relations governing the phenomenon cannot be known a priori in dynamic offshore environments. The respective tables and figures for the case of Rhyl Flats can be found in Appendix C.

5.4.2. Comparison of Alternative BN Structures

It is really interesting to examine how different configurations of the BN structures, i.e. a different number of nodes with a selection of variables or certain assigned (custom-fitted) relations, influence the predictions and the provided uncertainty, while testing them in the time interval previously presented (Jan 2017 – Jan 2018). This comparison will shed some light on whether one or more of the incorporated variables influence the models' accuracy positively, and will reveal if the erratic behaviour of the models incorporating short term past data can be casted off. A reminder here for the reader is the behaviour presented in Section 5.2.3 for the case of Rhyl Flats, where the impact of the BN models' corrections where not so clear (see also Table 8).

Figure 39 presents the case of BN structures incorporating 4 variables (nodes) for the case of Gwynt-y-Mor: (1) the observed significant wave height (H_s), (2) the numerically derived significant wave height ($H_{s,num}$), (3) the wave direction (D_{irp}), and (4) the zero-crossing wave period (T_z). Once more here the abrupt differences between long- and short-trained BN models' structures are visible, resulting from the nature of the incorporated data. It has to be reminded that on 17-11-2017 the short-trained BN model provided a quite satisfying forecast (see also Figure 25).

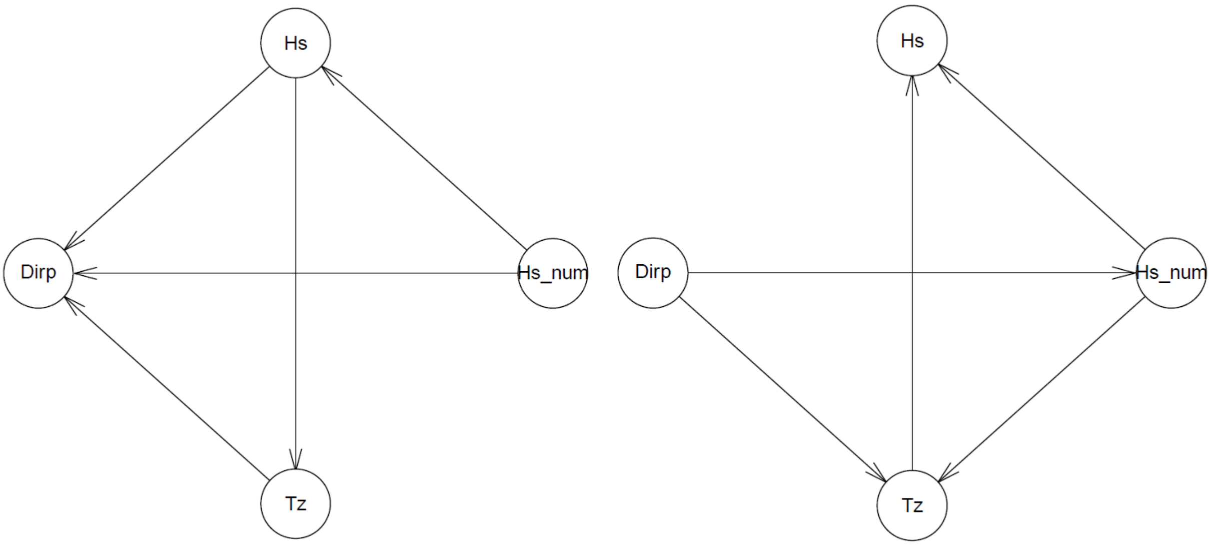


Fig. 39. Long- (left) and short-trained (right) BN models' structures for the case Gwynt-y-Mor incorporating 4 variables. (27-10-2017 at 18:00)

The resulting general and application oriented evaluation metrics for this case are presented in Tables 14 and 15. As it can be seen, the exclusion of the meteorological variables, i.e. the wind velocity and direction, only triggered a reduction of the fixed structure accuracy, to a point where it became equal to the short-trained BN models' one. A general comment could be that the behaviour of the BN models incorporating short term past data was less erratic, but still any accuracy reduction is unsatisfying.

Table 14. Evaluation metrics for the case of the 4-variable BN models (Gwynt-y-Mor).

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	BN Long Training	BN Short Training	BN Fixed Structure
		6 Nodes			4 Nodes		
RMSE (m)	0.231	0.218	0.253	0.209	0.218	0.257	0.257
MAE (m)	2.410	2.360	2.607	2.728	2.349	2.804	2.804
BIAS (m)	-0.046	-0.011	-0.051	0.005	-0.011	-0.056	-0.056
URMSE (m)	0.226	0.218	0.248	0.209	0.218	0.251	0.251

Table 15. Application specific evaluation metrics for the case of the 4-variable BN models (Gwynt-y-Mor).

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	BN Long Training	BN Short Training	BN Fixed Structure
		6 Nodes			4 Nodes		
Critically Accurate (%)	19.72	21.16	20.27	22.31	20.94	19.65	19.65
Critically Inaccurate (%)	2.55	2.82	3.79	2.10	2.88	4.06	4.06
False Positive (%)	2.26	1.93	1.50	2.10	1.93	1.35	1.35

Regarding the coverage percentage and the average length of the uncertainty bounds, again a reduction in performance is noticed (see Table 16) in the case of the fixed structure, while a small and insignificant enhancement of accuracy is observed in the short – and long-trained BN error correction models. As a result it can be concluded that for the Gwynt-y-Mor case the exclusion of the meteorological variables had a negative effect, and the 6-variable structure is suggested between the two.

Table 16. Uncertainty estimates' performance for the case of the 4-variable BN structures (Gwynt-y-Mor).

Method	BN Long Training	BN Fixed Structure	BN Short Training	Copula	BN Long Training (Log-N)	BN Short Training (Log-N)	BN Fixed Structure (Log-N)
Coverage (%)	89.3	77.6	77.6	68.5	95.4	73.1	78.8
Average Length (m)	0.630	0.553	0.553	0.375	1.185	0.637	0.550

For the Rhyl Flats case, whose metrics are shown in Tables 17 and 18 in comparison to the 6-variable BN models presented in preceding sections, it is evident that an enhancement of the models incorporating short-term past data has been achieved. This improvement is noticed in terms of the general metrics (see RMSE and URMSE in Table 17), as well as for some of the application specific metrics, such as the false positive percentage, and results from the more predictable and stable performance of the models.

Table 17. Comparison of evaluation metrics for the 4- and 6-variable BN models (Rhyl Flats).

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	BN Long Training	BN Short Training	BN Fixed Structure
		6 Nodes			4 Nodes		
RMSE (m)	0.203	0.178	0.200	0.201	0.178	0.192	0.192
MAE (m)	1.970	1.722	2.361	4.1731	1.733	2.377	2.377
BIAS (m)	-0.004	-0.010	-0.037	0.003	-0.013	-0.030	-0.030
URMSE (m)	0.203	0.178	0.196	0.201	0.177	0.189	0.189

Table 18. Comparison of application specific evaluation metrics for the 4- and 6-variable BN models (Rhyl Flats).

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	BN Long Training	BN Short Training	BN Fixed Structure
		6 Nodes			4 Nodes		
Critically Accurate (%)	18.02	17.01	16.04	18.82	16.87	16.40	16.40
Critically Inaccurate (%)	1.05	2.28	2.50	1.34	2.30	2.48	2.48
False Positive (%)	2.55	1.16	1.03	1.58	1.14	0.84	0.84

Nevertheless, the performance of the models has not been improved dramatically, while in some occasions the results are still better for the case of the 6-variable BN structures (see Table 18). Therefore it can be concluded that the 4-variable structure does not perform any better than the 6-variable one, which has to be adopted after this comparison. The uncertainty estimates for the case of Rhyl Flats, provided by the 4-variable structure are available in Appendix C.

Further testing was conducted with a 5-variable BN structure, incorporating supplementary the wind velocity (U_{10}). Examples of the arc directions for the case of Gwynt-y-Mor are shown in Figure 40, where the relations discussed previously between the meteorological and the hydrodynamic variables are again varying depending on the training of the BN model (long or short training). The explanation here is quite the same, since for the largest part of the year the wind velocity can in general influence the significant wave height, while in certain occasions this might not happen due to the origins of the waves.

Interestingly and importantly, the performance of the models is enhanced slightly, while being more consistent for the BNs incorporating short-term past data. Tables 19 and 20 illustrate the accuracy improvement in terms of the general and application specific metrics respectively. It is shown that the RMSE values are smaller for all BN models, with the new value provided by the one including the fixed structure being the smaller in comparison to the rest of the error correction techniques. The accuracy in predictions close to the critical boundary has also increased, particularly in terms of the critically accurate and false positive percentages.

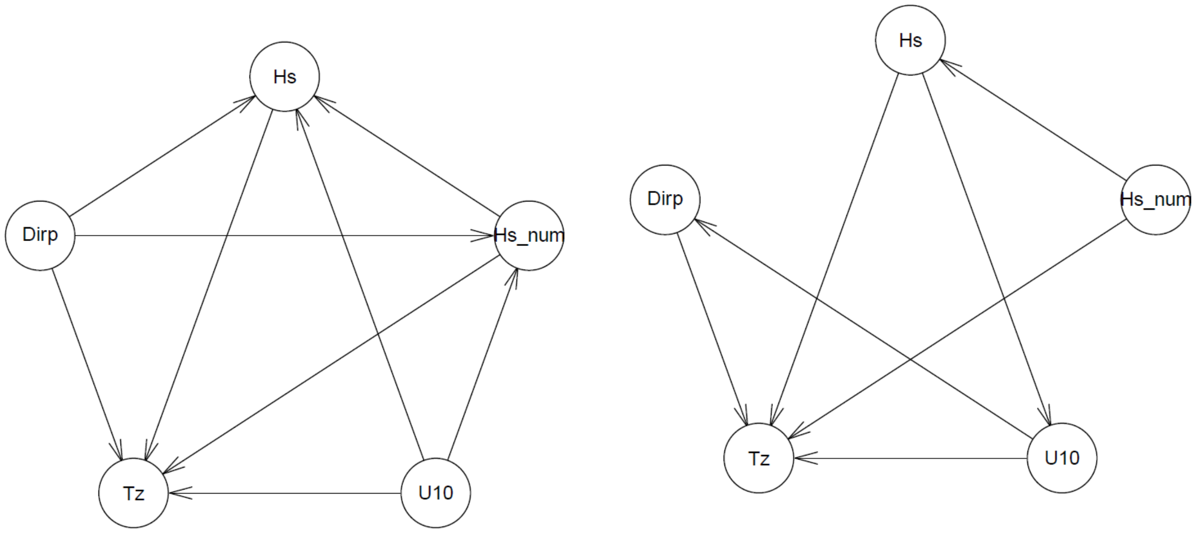


Fig. 40. Long- (left) and short-trained (right) BN models’ structures for the case Gwynt-y-Mor incorporating 5 variables. (27-10-2017 at 18:00)

Table 19. Evaluation metrics for the case of the 5-variable BN models (Gwynt-y-Mor).

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	BN Long Training	BN Short Training	BN Fixed Structure
		6 Nodes			5 Nodes		
RMSE (m)	0.231	0.218	0.253	0.209	0.219	0.248	0.208
MAE (m)	2.410	2.360	2.607	2.728	2.360	2.801	2.809
BIAS (m)	-0.046	-0.011	-0.051	0.005	-0.012	-0.055	0.002
URMSE (m)	0.226	0.218	0.248	0.209	0.219	0.248	0.208

Table 20. Application specific evaluation metrics for the case of the 5-variable BN models (Gwynt-y-Mor).

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	BN Long Training	BN Short Training	BN Fixed Structure
		6 Nodes			5 Nodes		
Critically Accurate (%)	19.72	21.16	20.27	22.31	21.08	19.82	22.19
Critically Inaccurate (%)	2.55	2.82	3.79	2.10	2.87	3.87	2.34
False Positive (%)	2.26	1.93	1.50	2.10	1.96	1.39	1.95

Regarding the uncertainty estimates, the coverage percentages and the lengths are similar to the 6-variable BN models' figures, without any improvement to the average length of the long-trained log-normal confidence intervals. Yet, the extremely high coverage percentage reaching nearly 96% of the total observations, as well as the relatively realistic behaviour of the boundaries are factors that cannot be overlooked. It is truly difficult to determine which boundary is the most suitable and it always depends on the applications needs. Nevertheless, both kinds of confidence intervals display an improvement when compared to the already existent uncertainty estimates given by the Gumbel Copula.

Table 21. Uncertainty estimates' performance for the case of the 5-variable BN structures (Gwynt-y-Mor).

Method	BN Long Training	BN Fixed Structure	BN Short Training	Copula	BN Long Training (Log-N)	BN Short Training (Log-N)	BN Fixed Structure (Log-N)
Coverage (%)	89.2	87.6	76.0	68.5	95.4	77.0	73.1
Average Length (m)	0.631	0.551	0.530	0.375	1.185	0.610	0.550

Even more interesting are the results produced for the case of Rhyl Flats. As shown in Table 22, there is a significant improvement in terms of all the metrics, to a degree that the BN model incorporating the fixed structure becomes the error correction technique serving the application better. Table 23 illustrates that also in terms of critical performance, around the 1.5 m upper boundary, the fixed structure BN model's performance is enhanced. Moreover, the behaviour of the 5-variable structures regarding models including short-term past data (i.e. 48 hours prior to the forecast), is quite consistent and robust in comparison to the structure incorporating 6 variables. Here, the point that the wind direction causes unsteadiness to the predictions, inducing a completely erratic and unpredictable performance in certain occasions, is proved. Because the uncertainty estimates display large improvement as well, it seemed fit to present them here in comparison to the results given by the 6-variable BN structure (see Table 24). The normal confidence intervals of the fixed-structured BN reach a coverage percentage of nearly 91% of the total observations, with an average length of just 49 cm. Certainly, the form of the boundaries is not ideal, since they are symmetrical, but still their performance provides a significant enhancement in accuracy, making the BN models a valuable correction tool for this application. The long-trained BN model is equally good in terms of accuracy whichever the configuration may be, making it also a robust and reliable tool, which with the inclusion of its uncertainty bounds introduces a significant improvement of the significant wave height (H_s) predictions. It can be concluded that the 5-variable BN models should be used for the case of Rhyl Flats, due to its robust behaviour, in comparison to similar techniques incorporating 4 or 6 variables.

Table 22. Evaluation metrics for the case of the 5-variable BN models (Rhyl Flats).

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	BN Long Training	BN Short Training	BN Fixed Structure
		6 Nodes			5 Nodes		
RMSE (m)	0.203	0.178	0.200	0.201	0.178	0.195	0.163
MAE (m)	1.970	1.722	2.361	4.1731	1.733	2.377	2.361
BIAS (m)	-0.004	-0.010	-0.037	0.003	-0.013	-0.038	0.003
URMSE (m)	0.203	0.178	0.196	0.201	0.177	0.191	0.163

Table 23. Application specific evaluation metrics for the case of the 5-variable BN models (Rhyl Flats).

Method	SWAN	BN Long Training	BN Short Training	BN Fixed Structure	BN Long Training	BN Short Training	BN Fixed Structure
		6 Nodes			5 Nodes		
Critically Accurate (%)	18.02	17.01	16.04	18.82	16.87	16.08	18.03
Critically Inaccurate (%)	1.05	2.28	2.50	1.34	2.30	2.45	1.47
False Positive (%)	2.55	1.16	1.03	1.58	1.14	0.84	1.14

Table 24. Uncertainty estimates' performance for the case of a 5-variable BN structure (Rhyl Flats).

Method	BN Long Training	BN Fixed Structure	BN Short Training	Copula	BN Long Training (Log-N)	BN Short Training (Log-N)	BN Fixed Structure (Log-N)
5 Variables							
Coverage (%)	89.6	90.8	77.2	70.9	95.0	77.1	73.2
Average Length (m)	0.527	0.489	0.430	0.327	1.024	0.505	0.460
6 Variables							
Coverage (%)	89.7	64.7	69.8	70.9	94.7	68.9	61.0
Average Length (m)	0.527	0.491	0.427	0.327	0.948	0.466	0.425

5.5. Remarks – Comments

Summarily, in this chapter the performance of the newly implemented BN methods in tests including data for the year of 2017 became evident. The BN method incorporating a fixed structure, as well as short-term past data, seems to be the best overall, out-performing any other error correction technique. The configuration that the BN structure should have depends highly on the available data. In Gwynt-y-Mor the BN models incorporating 6 variables, namely the observed significant wave height (H_s), the numerically produced significant wave height ($H_{s,num}$), the wave direction (D_{irp}), the zero-crossing wave period (T_z), the wind velocity (U_{10}), and the wind direction (U_{dir}), seems to serve the application equally good to the 5-variable structure, where the wind direction is excluded. A general comment is that the 6-variable BN structures behave erratically in certain occasions, when short-term past data (i.e. data retrieved 48 hours prior to the forecast) are incorporated.

On the other hand, for the Rhyl Flats dataset, the exclusion of the wind direction is imperative in order for all the BN models to be able to produce results of enhanced accuracy, due to the condition of the aforementioned variable's dataset. Certainty, the long-trained BN provides robust and consistent results for both stations, and with the inclusion of the uncertainty estimates provided, it becomes also a very attractive and equally suitable error correction technique. The BN structures were all produced by data-driven procedures, based on score-based tests provided by `bnlearn`. It is recommended that the data should determine the relations in the BN structure, and no pre-determined arc direction should exist. The relations between the meteorological and hydrodynamic data seem odd in specific occasions, but the nature of the locations under consideration justifies the produced results. Since the origins of waves and wind are unknown in most occasions the data-driven procedure seemed and was fitter and more accurate. For examples of results given by structures in which the relation between wind velocity and significant wave height was imposed a priori the reader is referred to Appendix C, where it becomes evident why the data-driven structures serve the application better.

All in all, it can be concluded that the BN methods provide the most suitable solution in terms of error correction for the examples in the Irish Sea. A major advantage is the information available by its structures and uncertainty estimates, which can be either provided in normal or log-normal format. The normal confidence intervals seem to be the most suitable for this application, since they provide more information, especially in terms of the higher boundary. Moreover, they introduce a really acceptable average length in comparison to their log-normal counterparts. Still, the log-normal uncertainty bounds grant behaviours close to reality, since they only provide positive estimates, but their average length, especially the one given by the long-trained technique, is totally unrealistic. Generally though, it can be said that all BN methods enhance the uncertainty estimates' performance in comparison to the already existent Gumbel Copula.

6. Conclusions

In this thesis, the behaviour and performance of several statistical and stochastic techniques while providing significant wave height (H_s) predictions/corrections was evaluated, and compared to the forecasts given by a numerical model (SWAN), with special attention given to the Bayesian Network models and the information granted by their structures. Uncertainty estimates for the aforementioned variable of interest were also given by the underlying conditional distribution of the H_s, leading to a quite satisfying mapping of the variable's behaviour. The provided confidence intervals were compared to estimates given by an assumed Gumbel Copula, showing a significant increase in forecasting performance. Because of the underlying multivariate normality assumption in the used `bnlearn` package of R, the given uncertainty boundaries were symmetrical above and below the point predictions. To assess that, and possibly reach a more realistic behaviour for the significant wave height estimates, a logarithmic transformation took place, to obtain log-normal uncertainty boundaries, derived by a parametric distribution fitting the H_s data better. Nevertheless, the estimates given the normal conditional distribution outperformed the ones provided by their log-normal counterparts, with equal or smaller average lengths, i.e. differences between the upper and lower estimates averaged over a whole year. It has to be stressed out that all the tests and validation were conducted bearing in mind the operational nature of the application, i.e. that the models had to perform well in conditions where the measurements and numerical model (SWAN) results were obtained in real-time. As a result, the ability of the models to provide operation (real-time) predictions of enhanced accuracy was also assessed.

6.1. Overall Model Performance

For the evaluation of the error correction techniques, metrics describing the general performance of the models and the quality of the predictions, as well as application-specific metrics focusing on constraints and boundaries provided by the application in hand (maintenance and installation operations in offshore wind farms), were used. For both stations under consideration in the Irish Sea (Gwynt-y-Mor and Rhyl Flats), the BN model incorporating a long-trained structure (the so-called fixed structure throughout this research) in combination with short-term past data (48-hrs prior to the forecast) performs better than any other model, including SWAN. This conclusion is reflected in all of the evaluation metrics. Also, the long-trained BN model produces results of really good accuracy, with a stable and robust performance, regardless the number of the variables. Given also that both of these models provide uncertainty estimates, which cover nearly 90% of the total number of measurements, it can be concluded that the BN models provide the most suitable solution in terms of error correction, enhancing the SWAN forecasts significantly and ensuring nautical and operational safety in a large number of occasions.

6.2. Analysis of the BN Structures

The results provided by statistical methods are largely dependent on the data quality or suitability. Due to the topology (Irish Sea) which induces secondary events in terms of hydrodynamics (reflection, diffraction, etc.), some direct variable relations that would seem obvious are not so trivial after all. For instance, as discussed extensively in the previous chapters, the dependencies between meteorological and hydrodynamic variables, as the wind (U_{dir}) and wave directions (D_{irp}), do not seem to exist when the analysis is data-driven. Tests conducted with pre-determined relations in a structure, i.e. the meteorological variables (U_{dir} and U_{10}) influencing the hydrodynamic ones (H_s, T_z, and D_{irp}) and not the other way around, did not produce satisfying results. Thus, data-driven approaches were used and are recommended when the morphology of the area, or the way the measurements are collected (e.g. with wave-rider buoys and met-masts), include many uncertainties of their own.

The previously described sensitivity on the data suitability can influence the results of the BN models significantly, and make their behaviour erratic. For the case of the long-trained BN model, the 6-variable structure seems to perform equally well for both case studies, but the BN incorporating the fixed structure (which actually produced the best results overall) seems to be greatly influenced, especially in the case of Rhyl Flats. In offline mode it is easy to establish and recognise which variable/s reduces the respective models' accuracy, but when the models run operationally it is impossible to interfere. In that regard the 5-variable BN structures, excluding the wind direction (U_{dir}) are the most suitable to be used in operational (real-time) conditions.

6.3. Evaluation of the Uncertainty Estimates

Regarding the uncertainty estimates, which are provided in the form of 95% confidence intervals extracted from the conditional distribution of the significant wave height (H_s), the BN models in all cases outperform the assumed Gumbel Copula. The coverage percentage throughout the year of 2017 in both wind farms reaches approximately 90% of the total number of measurements, providing reasonable average lengths of 50-60 cm, for the case of the normal conditional distribution, i.e. 25-30 cm upper and lower uncertainty. The log-normal uncertainty boundaries, despite their more realistic appearance and behaviour, provide larger average lengths with smaller coverage percentages, except for the case of the long-trained BN model, which on one hand provides coverage of nearly 95% of the total number of measurements, but in the same time displays an average length of 1.18 m for its confidence intervals. As a result, it can be concluded that the normal uncertainty boundaries of the BN incorporating short- and long-term past data (fixed structure) are the ones providing the best results.

Again in this occasion, the 5-variable structure (excluding the wind direction) has to be incorporated in order to achieve enhanced accuracy with the fixed-structured BN model. Therefore, in real-time this structure should be chosen above the 6-variable one, due to its far more stable performance.

6.4. Operational Functionality

As described in all the previous paragraphs and chapters, the final goal is to manage to emulate the real-time nature of the application and draw conclusions for the applicability of the methods under consideration in operational environments. In that regard, the fixed-structured 5-variable BN model outperforms any other statistical or stochastic technique, not only in terms of point predictions/corrections, but also in terms of the uncertainty estimates which encapsulate nearly 90% of the measurements in 2017, in both stations (Gwynt-y-Mor and Rhyl Flats). Certainly this kind of model has one major disadvantage; the fact that it needs short-term past data (48-hrs prior to the forecast) makes it unable to produce predictions/corrections in the absence of recent observations. This effect is not an issue with the long-trained BN model, which has equally good general metrics, with RMSE values close to fixed-structure BN (see Chapter 5), but is underperforming in terms of the critical situations (close to the upper 1.5 m boundary), i.e. displays a more conservative behaviour which makes it unable in certain occasions to predict significant wave height peaks (see application-specific metrics of Chapter 5).

Consequently, it is really a matter of subject which model is better in terms of the operational performance. Surely, the ability of the long-trained BN model to produce forecasts of enhanced accuracy constantly, even in the absence of recent observations makes it really attractive for real-time use. Yet, one cannot in any case overlook the really good performance of the fixed-structured BN, especially in producing critical predictions (close the application's upper boundary), which constitutes probably the most important benefit of using it. As a result, it is clearly up to the user and the nature of the application, i.e. whether the variable of interest displays a really dynamic behaviour, such as in the case of the H_s, or not, to determine which of the aforementioned models is the most suitable one. In a nutshell, the

BN models are really suitable for operational use, but which one has to be used is largely dependent on the needs of the application.

6.5. Research Directions

To be able to draw more concrete conclusions on the operational performance of the methods, real-time testing has to be performed for an extensive period of time. All models were tested successfully in terms of their functionality in operational environments, but surely a concise and consistent validation of the produced results has to take place.

Regarding the wind direction, and its suitability in terms of use, possibly it could be discretized rather than used as an additional continuous variable, hence transforming the BN network into a hybrid one. Also the accuracy of the models could be evaluated per season, or type of event, e.g. for wind coming from NW in comparison to SE, depending on the main wind directions generally occurring in the Irish Sea, or for varying training sets of 1,2, or 4 years of data.

Finally, concerning the impact that the corrections have, which is not always easy to find out just by studying metrics and percentages, an application-based stochastic impact assessment can take place, revealing the importance of the models in monetary and risk terms. For more information, the reader is referred to Bastola et al. (2011).

Appendix A

A.1. Preliminary Analysis BN Structures

Here, the Bayesian Networks' (BN) structures corresponding to the preliminary analysis results are shown. The structures of the short- and long – trained BNs are sometimes similar in terms of their arc direction. Nevertheless, the correlation between the variables differs and, as a result, the predictions provided by each network display disparate accuracy.

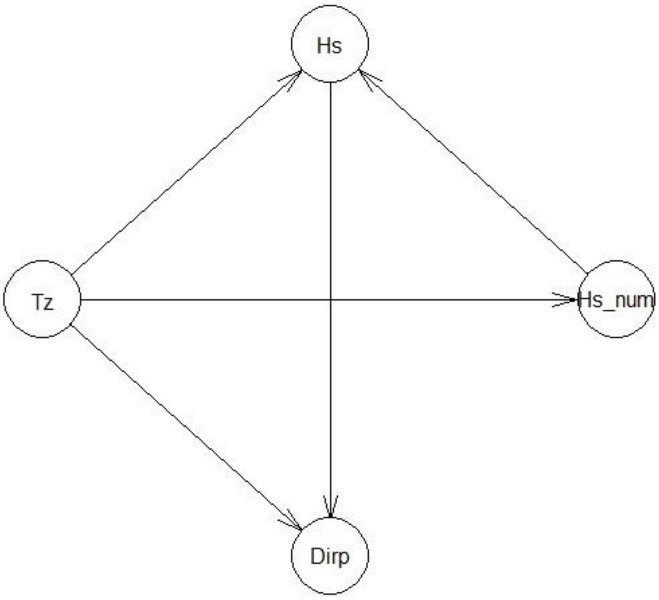


Fig. A. 1. BN structure for short and long training on 28-30 May 2015 (Rhyl Flats - Irish Sea).

Table A. 1. Long-trained BN correlation matrix on 28-30 May 2015 (Rhyl Flats - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.4281	0.6865	0.6814
D_{irp}	0.4281	1	0.2359	0.2012
$H_{s,num}$	0.6865	0.2359	1	0.9159
H_s	0.6814	0.2012	0.9159	1

Table A. 2. Short-trained BN correlation matrix on 28-30 May 2015 (Rhyl Flats - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.2866	0.5827	0.5198
D_{irp}	0.2866	1	-0.1689	-0.3572
$H_{s,num}$	0.5827	-0.1689	1	0.7676
H_s	0.51988	-0.3572	0.7676	1

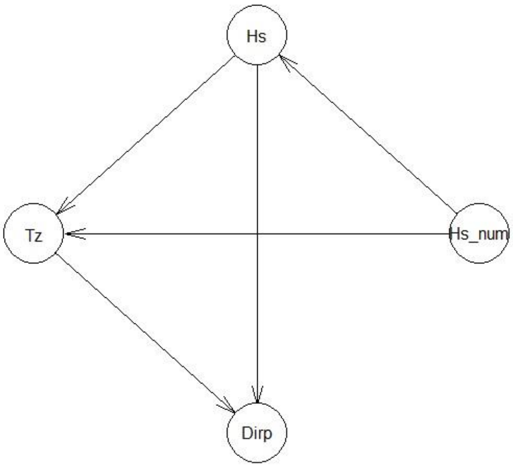


Fig. A. 2. BN structure for long training on 28-30 May 2015 (Gwynt-y-Mor - Irish Sea).

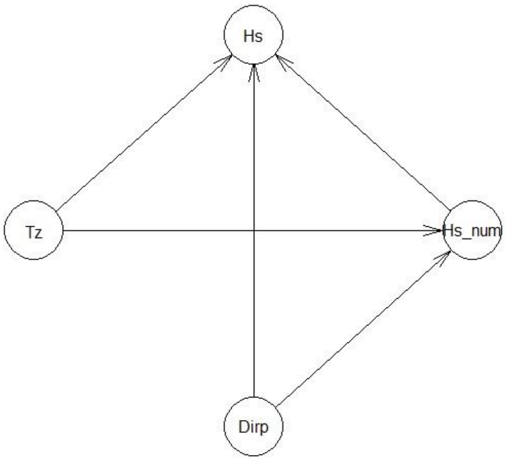


Fig. A. 3. BN structure for short training on 28-30 May 2015 (Gwynt-y-Mor - Irish Sea).

Table A. 3. Long-trained BN correlation matrix on 28-30 May 2015 (Gwynt-y-Mor - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	-0.0692	0.2677	0.8191
D_{irp}	-0.0692	1	-0.8905	-0.3947
$H_{s,num}$	0.2677	-0.8905	1	0.6401
H_s	0.8191	-0.3947	0.6401	1

Table A. 4. Short-trained BN correlation matrix on 28-30 May 2015 (Gwynt-y-Mor - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.4127	0.5481	0.7525
D_{irp}	0.4127	1	0.1129	0.2028
$H_{s,num}$	0.5480	0.1129	1	0.7972
H_s	0.7525	0.2028	0.7972	1

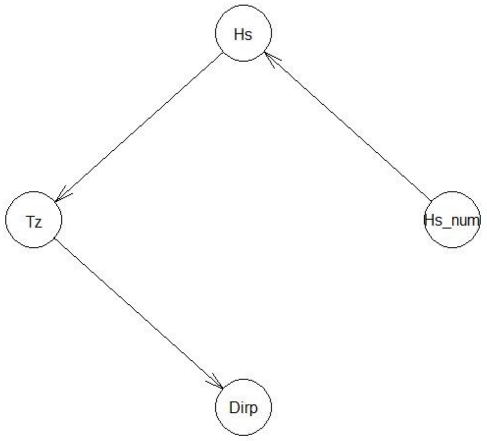


Fig. A. 4. BN structure for long training on 29-31 July 2015 (Rhyl Flats - Irish Sea).

Table A. 5. Long-trained BN correlation matrix on 29-31 July 2015 (Rhyl Flats - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.5284	0.5953	0.6939
D_{irp}	0.5284	1	0.2789	0.3603
$H_{s,num}$	0.5953	0.2789	1	0.8528
H_s	0.6939	0.3603	0.8528	1

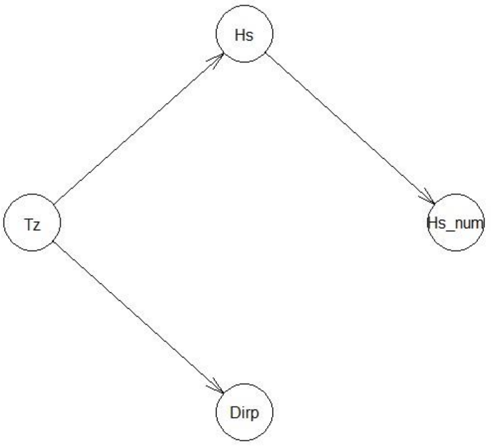


Fig. A. 5. BN structure for short on 29-31 July 2015 (Rhyl Flats - Irish Sea).

Table A. 6. Short-trained BN correlation matrix on 29-31 July 2015 (Rhyl Flats - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.5547	0.4499	0.8952
D_{irp}	0.5547	1	0.1429	0.5055
$H_{s,num}$	0.4499	0.1429	1	0.5907
H_s	0.8952	0.5055	0.5907	1

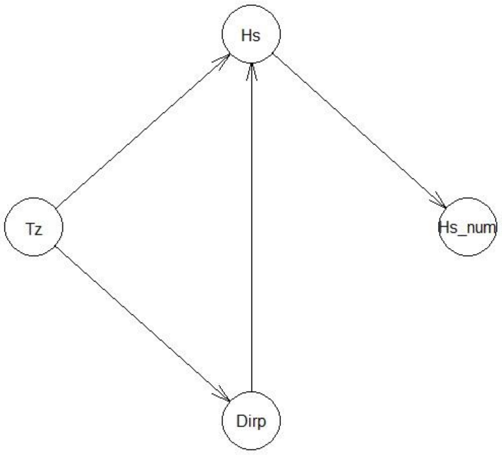


Fig. A. 6. BN structure for long training on 29-31 July 2015 (Gwynt-y-Mor - Irish Sea).

Table A. 7. Long-trained BN correlation matrix on 29-31 July 2015 (Gwynt-y-Mor - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.4409	0.4958	0.7472
D_{irp}	0.4409	1	0.1003	0.2369
$H_{s,num}$	0.4958	0.1003	1	0.6941
H_s	0.7472	0.2369	0.6941	1

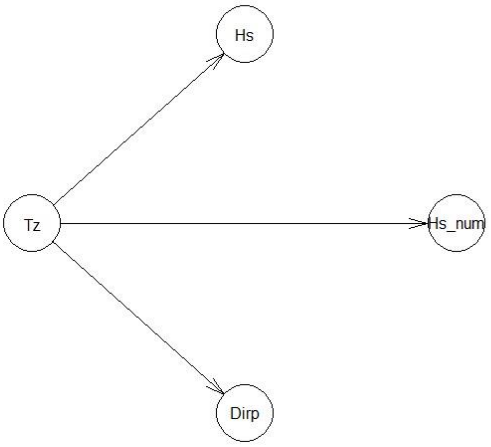


Fig. A. 7. BN structure for short on 29-31 July 2015 (Gwynt-y-Mor - Irish Sea).

Table A. 8. Short-trained BN correlation matrix on 29-31 July 2015 (Gwynt-y-Mor - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.3723	0.6074	0.8674
D_{irp}	0.3723	1	0.3631	0.2556
$H_{s,num}$	0.6074	0.3631	1	0.5176
H_s	0.8674	0.2556	0.5176	1

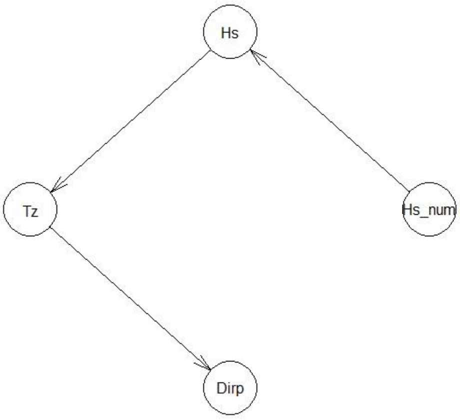


Fig. A. 8. BN structure for long training on 28-30 August 2015 (Rhyl Flats - Irish Sea).

Table A. 9. Long-trained BN correlation matrix on 28-30 August 2015 (Rhyl Flats - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.3512	0.6222	0.6614
D_{irp}	0.3512	1	0.1969	0.2160
$H_{s,num}$	0.6222	0.1969	1	0.9351
H_s	0.6614	0.2160	0.9351	1

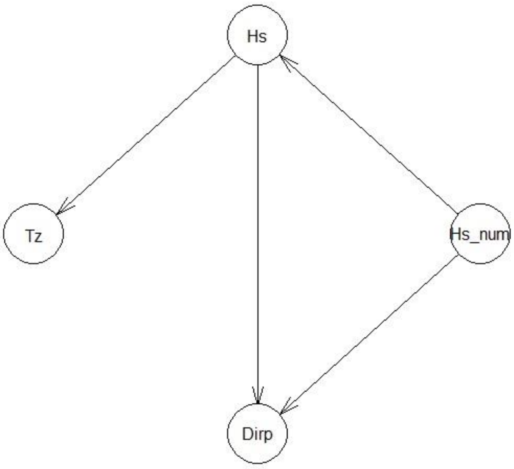


Fig. A. 9. BN structure for short on 28-30 August 2015 (Rhyl Flats - Irish Sea).

Table A. 10. Short-trained BN correlation matrix on 28-30 August 2015 (Rhyl Flats - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.4684	0.5401	0.6317
D_{irp}	0.4684	1	0.5585	0.5689
$H_{s,num}$	0.5401	0.5585	1	0.6485
H_s	0.6317	0.5689	0.6485	1

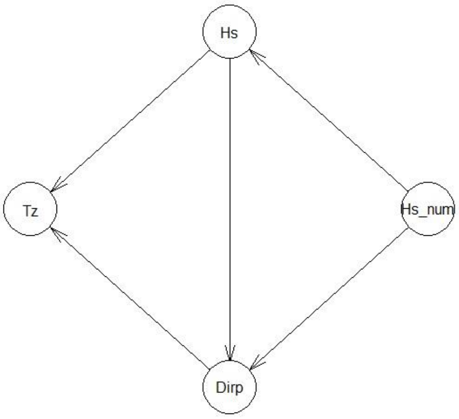


Fig. A. 10. BN structure for long training on 28-30 August 2015 (Gwynt-y-Mor - Irish Sea).

Table A. 11. Long-trained BN correlation matrix on 28-30 August 2015 (Gwynt-y-Mor - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.4032	0.3396	0.5076
D_{irp}	0.4032	1	0.0960	0.3338
$H_{s,num}$	0.3396	0.0960	1	0.7529
H_s	0.5076	0.3338	0.7529	1

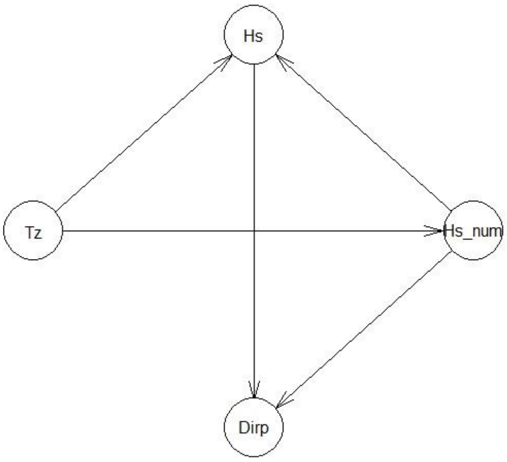


Fig. A. 11. BN structure for short on 28-30 August 2015 (Gwynt-y-Mor - Irish Sea).

Table A. 12. Short-trained BN correlation matrix on 28-30 August 2015 (Gwynt-y-Mor - Irish Sea).

Variable	T_z	D_{irp}	$H_{s,num}$	H_s
T_z	1	0.4719	-0.3082	0.6022
D_{irp}	0.4719	1	-0.3110	0.5501
$H_{s,num}$	-0.3082	-0.3110	1	0.0701
H_s	0.6022	0.5501	0.0701	1

A.2. Periodograms

As described in the main text (Section 3.2), in order to retrieve the seasonality of the original dataset, to create a timeseries data frame in R, a Fourier Transform was carried out to distinguish the main frequencies. The results of such a procedure can be displayed by means of periodograms[42] (see below), which clearly indicate the required frequencies. A periodogram is an estimate of the spectral density of a signal. By definition[43], the power spectral density of a continuous function, $x(t)$, is the Fourier Transform of its auto-correlation function (see e.g. Box and Jenkins, 1976; Fulop and Fitz, 2006; Auger and Flandrin, 1995; Schuster, 1898):

$$\mathcal{F}\{x(t) * x(-t)\} = X(f) \cdot X^{*}(f) = |X(f)|^{2} \tag{1.A}$$

In many applications, periodogram-based techniques introduce small biases, which are unacceptable. Another deficiency of the periodogram is that the variance at a given frequency does not decrease as the number of samples used in the computation increases. The averaging needed to analyse noise-like signals, or even sinusoids at low signal-to-noise ratios, is not provided, resulting in the need for more sophisticated methods of spectral estimation. To expand further on the subject of spectral estimation, is far from the scope of this research. Consequently, the results derived from the analysis that took place, to obtain finally the needed ARIMA model, are presented in the following figures.

[42] The term "*periodgram*" was invented by Arthur Schuster in 1898.

[43] See also the Cross-correlation theorem (*http://mathworld.wolfram.com/Cross-CorrelationTheorem.html*).

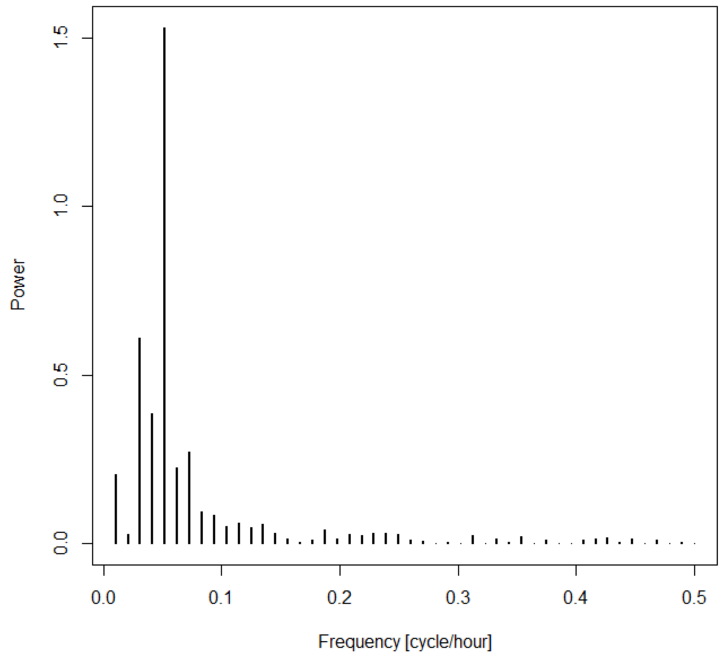


Fig. A. 12. Periodogram corresponding to the 28-30 May 2015 data (Rhyl Flats - Irish Sea).

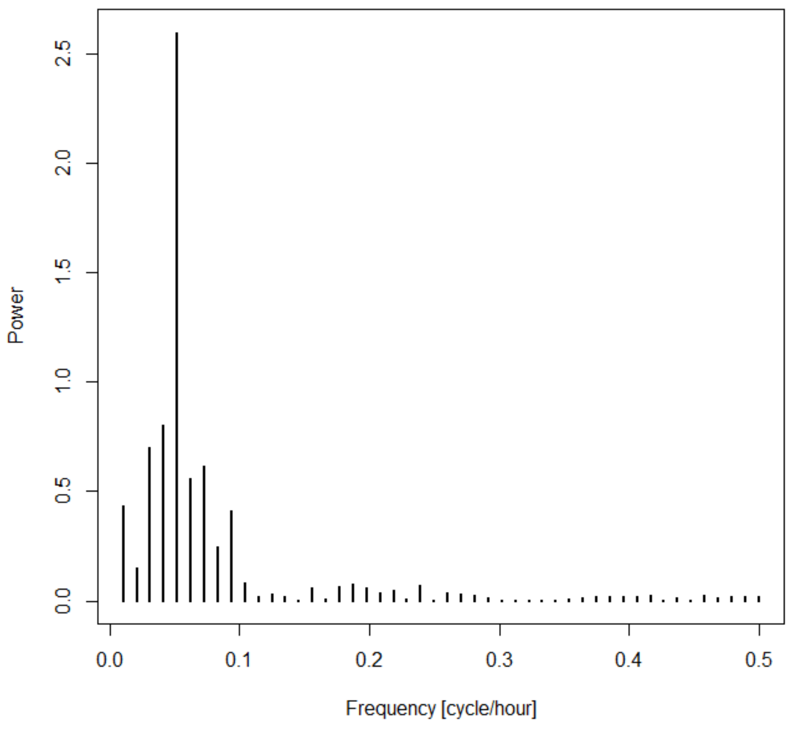


Fig. A. 13. Periodogram corresponding to the 28-30 May 2015 data (Gwynt-y-Mor - Irish Sea).

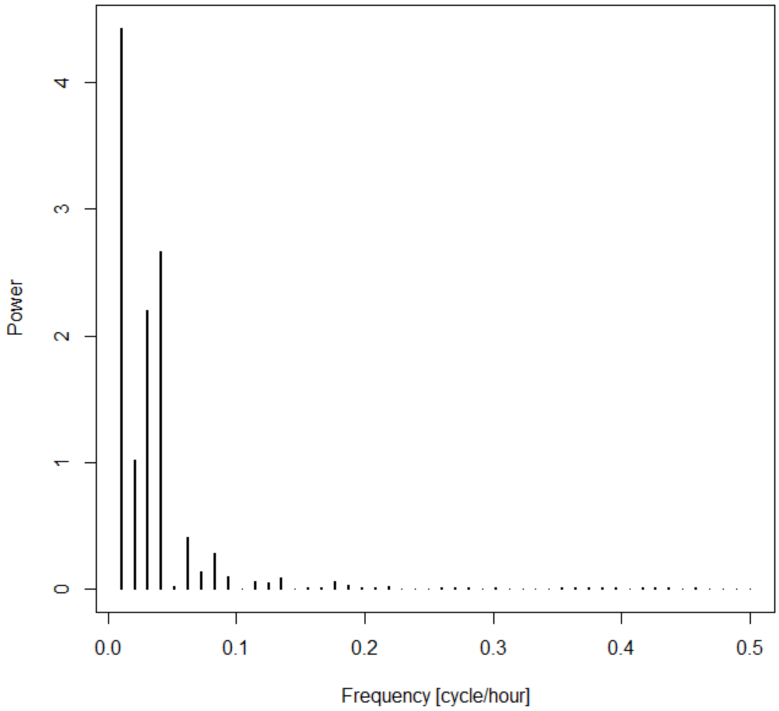


Fig. A. 14. Periodogram corresponding to the 29-31 July 2015 data (Rhyl Flats - Irish Sea).

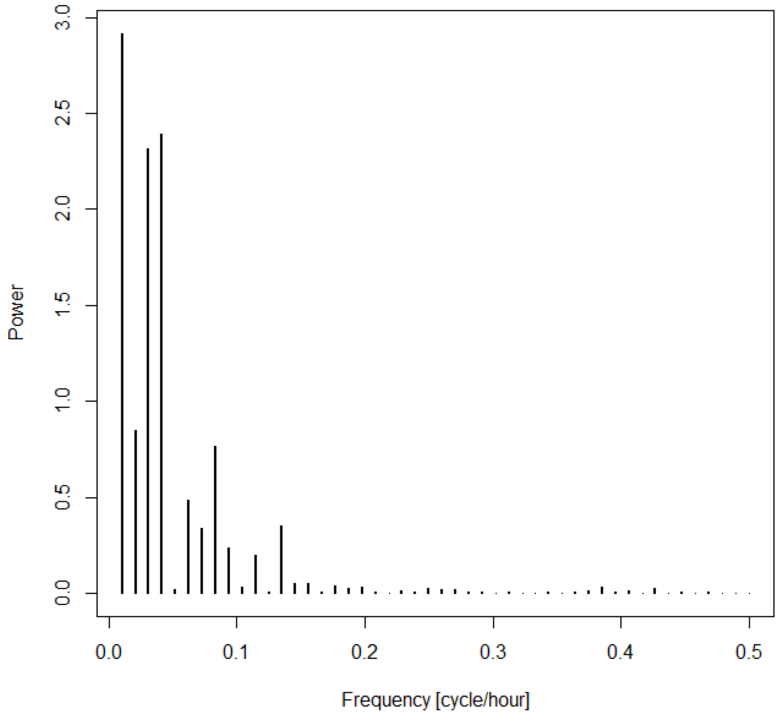


Fig. A. 15. Periodogram corresponding to the 29-31 July 2015 data (Gwynt-y-Mor - Irish Sea).

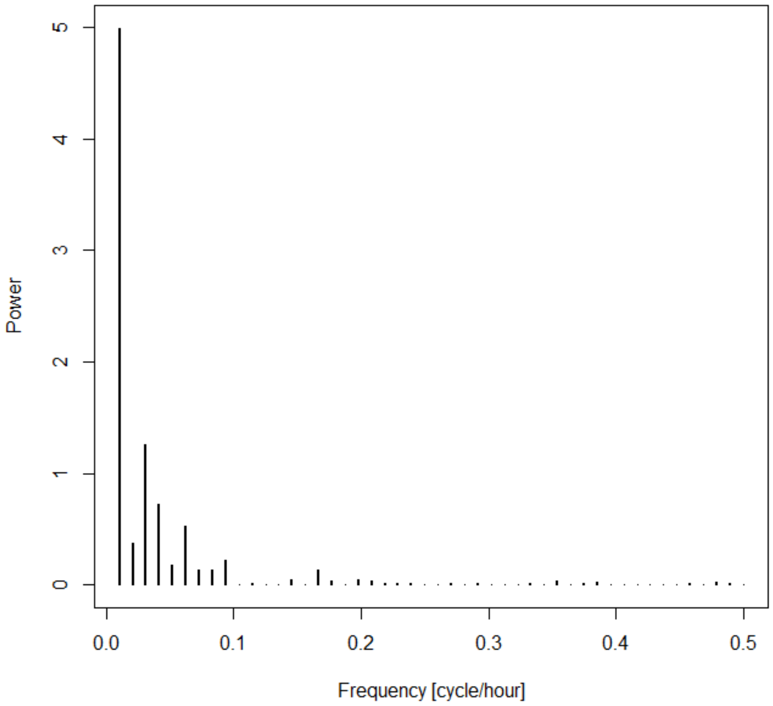


Fig. A. 16. Periodogram corresponding to the 28-30 August 2015 data (Rhyl Flats - Irish Sea).

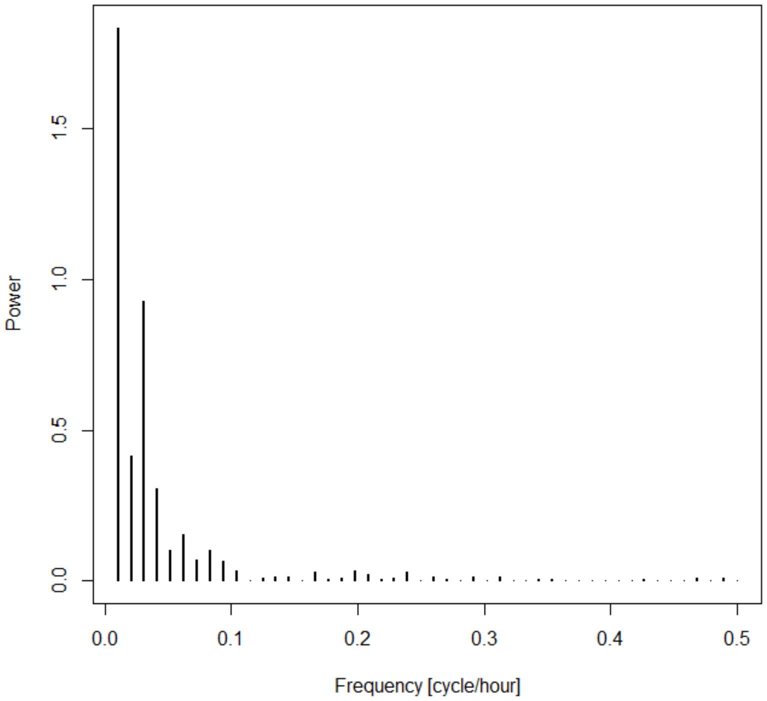


Fig. A. 17. Periodogram corresponding to the 28-30 August 2015 data (Gwynt-y-Mor - Irish Sea).

From the previously displayed periodograms, the seasonality of the dataset was computed, in order for the required timeseries data frame to be created. In case the seasonality was larger than the time interval indicated by the dataset (i.e. 48 hours), the major frequency was rejected and the secondary major one was chosen.

A.3. Functionality Test Results

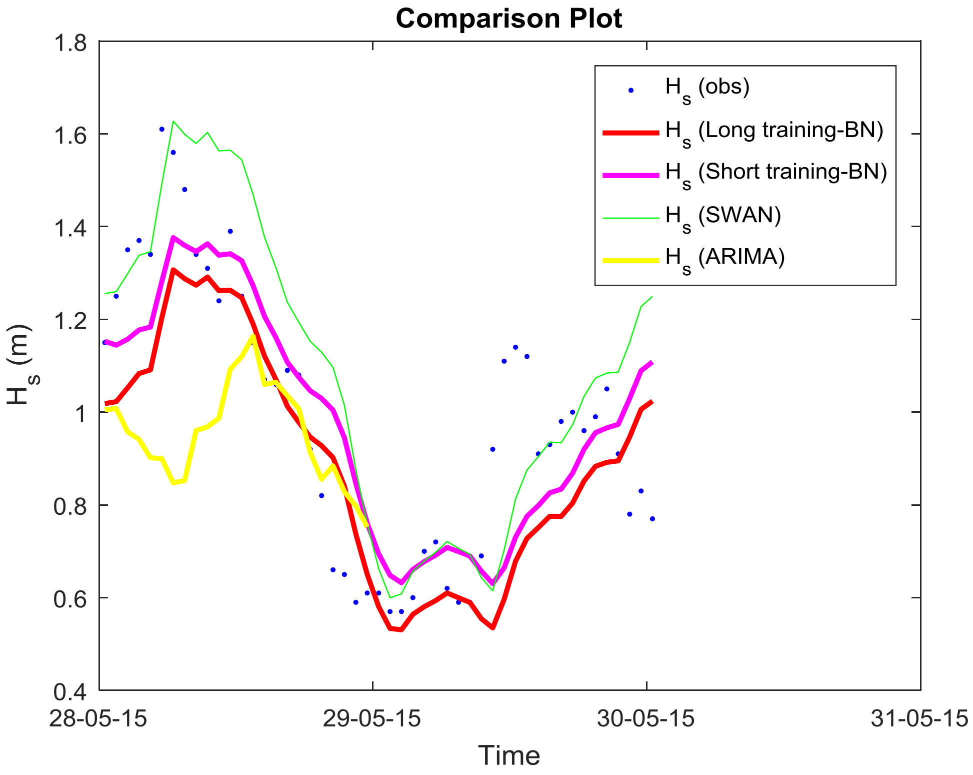


Fig. A. 18. Comparison between the BN and ARIMA techniques for 28-30 May 2015 (Rhyl Flats - Irish Sea).

Table A. 13. Preliminary Evaluation Metrics (RMSE - MAE table) for 28-30 May 2015, at Rhyl Flats (Irish Sea).

Method	SWAN	Long-trained BN	Short-trained BN	ARIMA[44]
RMSE (48 hours)	0.2179 m	0.1821 m	0.1955 m	-
RMSE (24 hours)	0.2312 m	0.1756 m	0.1688 m	0.3292 m
Maximum Absolute Error	0.4797 m	0.5135 m	0.4459 m	0.7125 m

[44] ARIMA: (1,0,0)(1,1,1) with a seasonality equal to 19 hours.

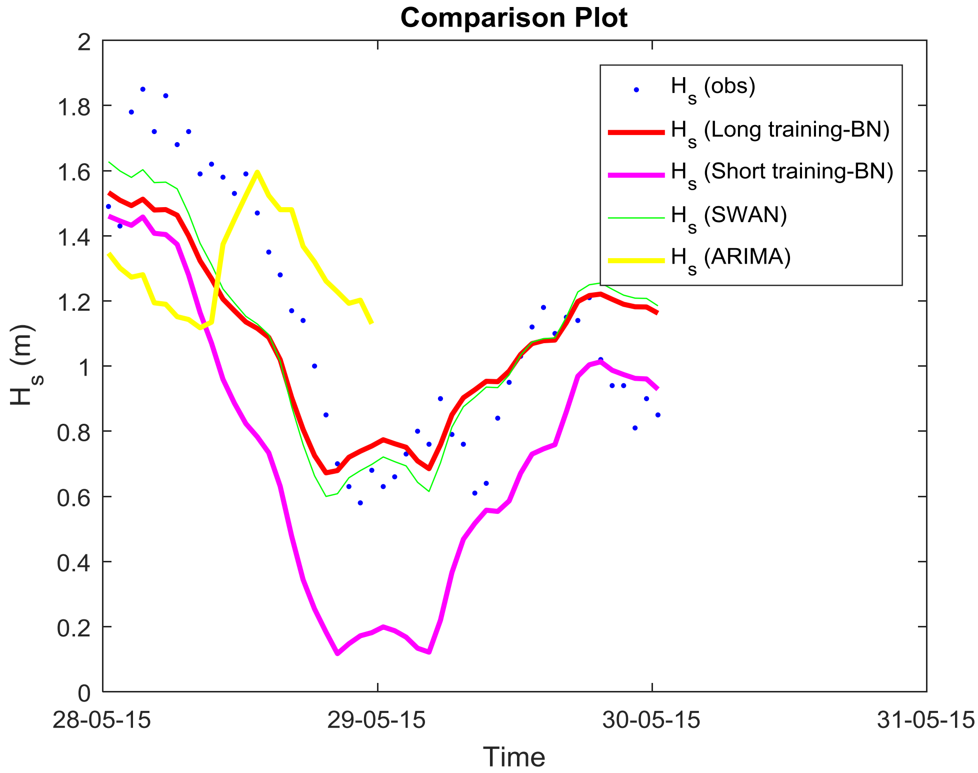


Fig. A. 19. Comparison between the BN and ARIMA techniques for 28-30 May 2015 (Gwynt-y-Mor - Irish Sea).

Table A. 14. Preliminary Evaluation Metrics (RMSE - MAE table) for 28-30 May 2015, at Gwynt-y-Mor (Irish Sea).

Method	SWAN	Long-trained BN	Short-trained BN	ARIMA[45]
RMSE (48 hours)	0.2179 m	0.1821 m	0.1955 m	-
RMSE (24 hours)	0.2312 m	0.1756 m	0.1688 m	0.3292 m
Maximum Absolute Error	0.4372 m	0.4546 m	0.7957 m	0.6402 m

[45] ARIMA: (1,0,0)(1,1,1) with a seasonality equal to 19 hours.

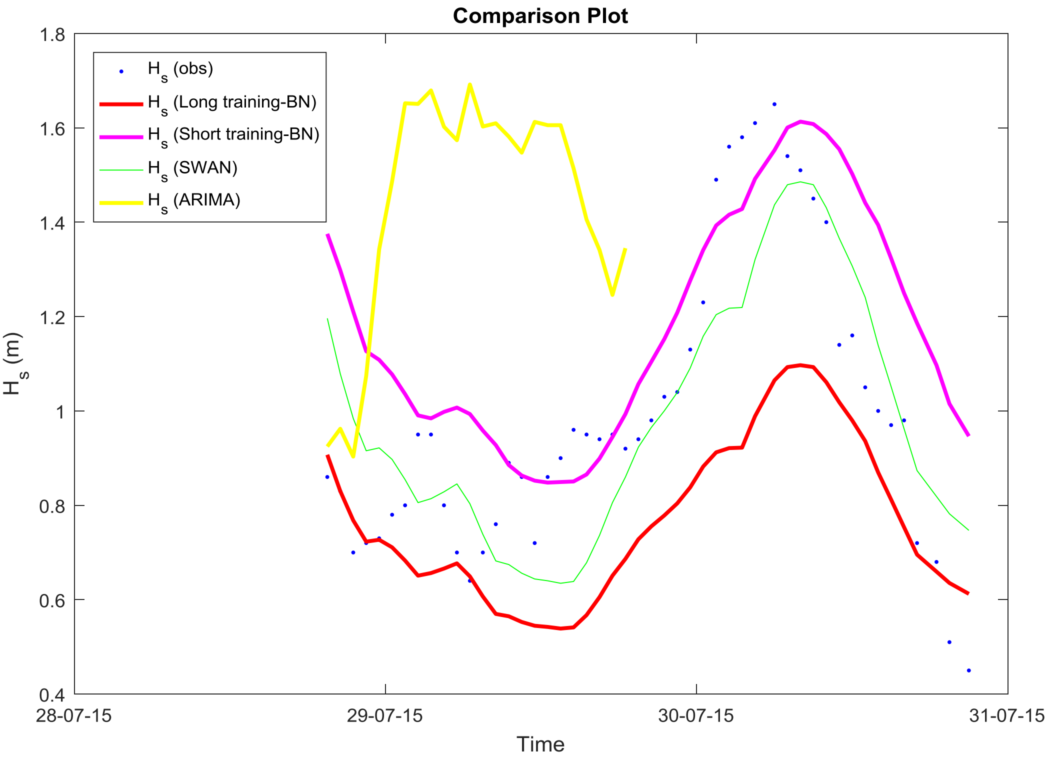


Fig. A. 20. Comparison between the BN and ARIMA techniques for 29-31 July 2015 (Rhyl Flats - Irish Sea).

Table A. 15. Preliminary Evaluation Metrics (RMSE - MAE table) for 29-31 July 2015, at Rhyl Flats (Irish Sea).

Method	SWAN	Long-trained BN	Short-trained BN	ARIMA[46]
RMSE (48 hours)	0.1868 m	0.3009 m	0.2690 m	-
RMSE (24 hours)	0.1935 m	0.2331 m	0.2582 m	0.6647 m
Maximum Absolute Error	0.3608 m	0.6578 m	0.5153 m	0.7125 m

[46] ARIMA: (1,1,0)(1,1,0) with a seasonality equal to 24 hours.

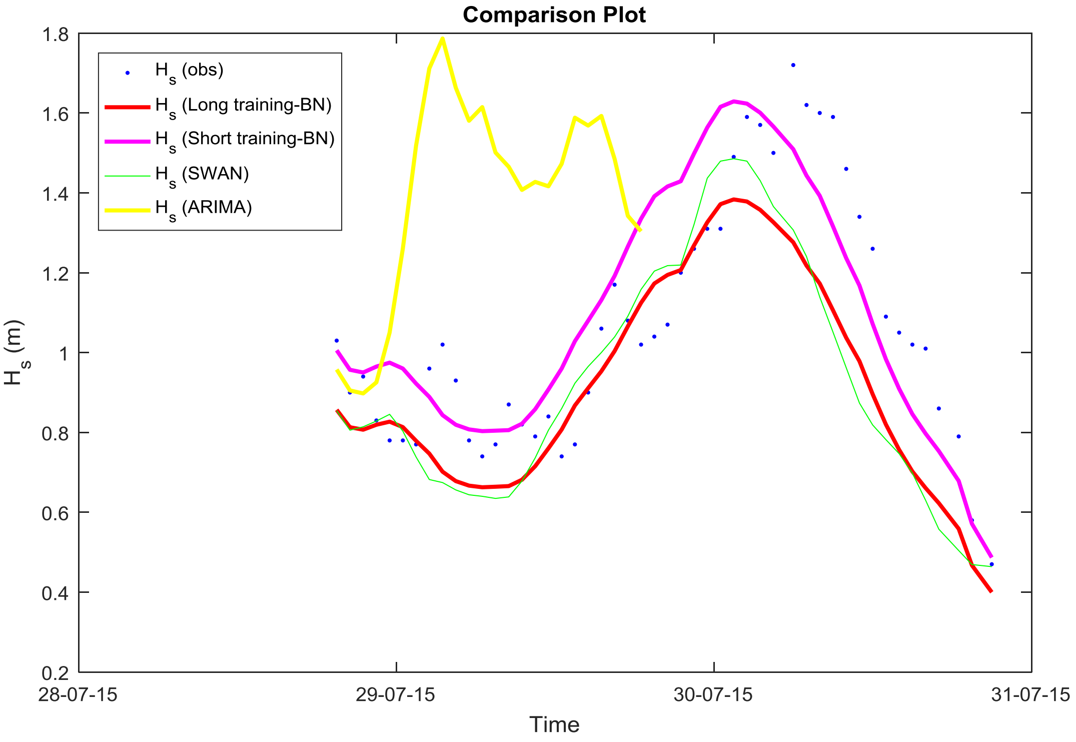


Fig. A. 21. Comparison between the BN and ARIMA techniques for 29-31 July 2015 (Gwynt-y-Mor - Irish Sea).

Table A. 16. Preliminary Evaluation Metrics (RMSE - MAE table) for 29-31 July 2015, at Gwynt-y-Mor (Irish Sea).

Method	SWAN	Long-trained BN	Short-trained BN	ARIMA[47]
RMSE (48 hours)	0.2383 m	0.2181 m	0.1718 m	-
RMSE (24 hours)	0.1499 m	0.1348 m	0.1400 m	0.5821 m
Maximum Absolute Error	0.5386 m	0.4839 m	0.3513 m	0.8747 m

[47] ARIMA: (0,1,2)(0,1,1) with a seasonality equal to 24 hours.

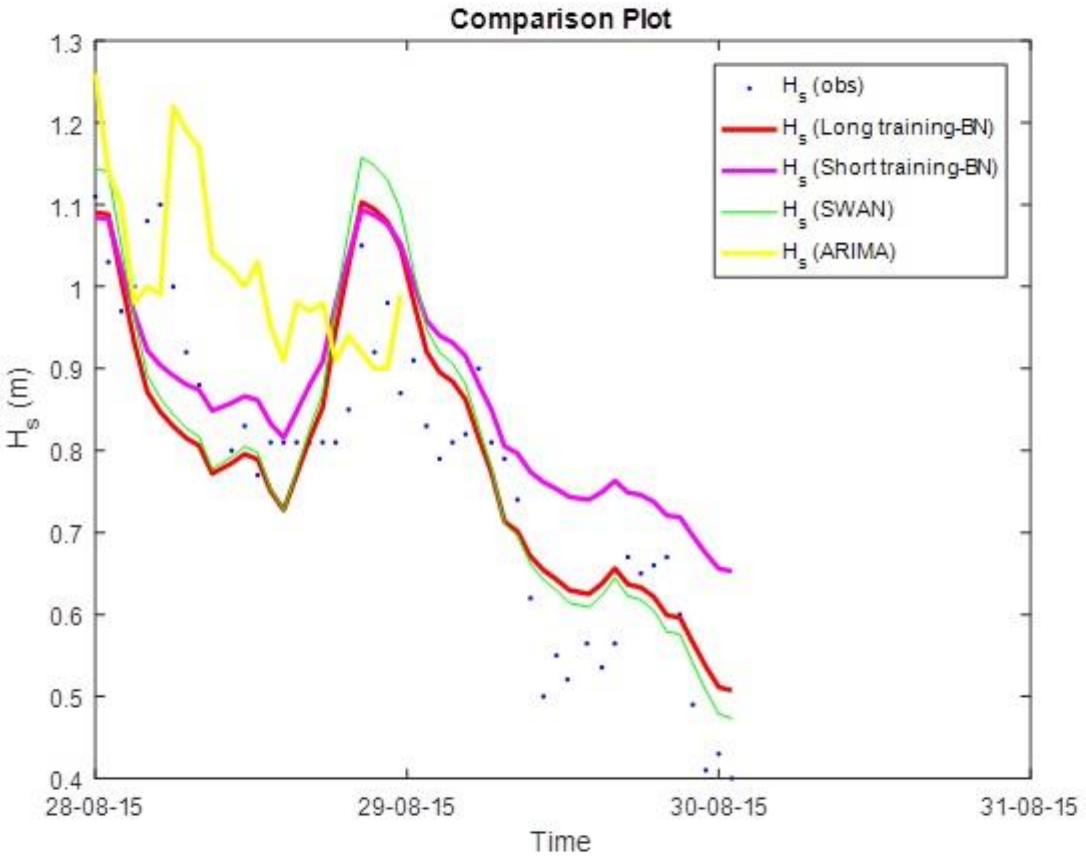


Fig. A. 22. Comparison between the BN and ARIMA techniques for 28-30 August 2015 (Rhyl Flats - Irish Sea).

Table A. 17. Preliminary Evaluation Metrics (RMSE - MAE table) for 28-30 August 2015, at Rhyl Flats (Irish Sea).

Method	SWAN	Long-trained BN	Short-trained BN	ARIMA[48]
RMSE (48 hours)	0.1048 m	0.0979 m	0.1343 m	-
RMSE (24 hours)	0.1264 m	0.1134 m	0.1018 m	0.1615 m
Maximum Absolute Error	0.2363 m	0.2530 m	0.2646 m	0.2900 m

[48] ARIMA: (0,1,0)(0,1,0) with a seasonality equal to 32 hours.

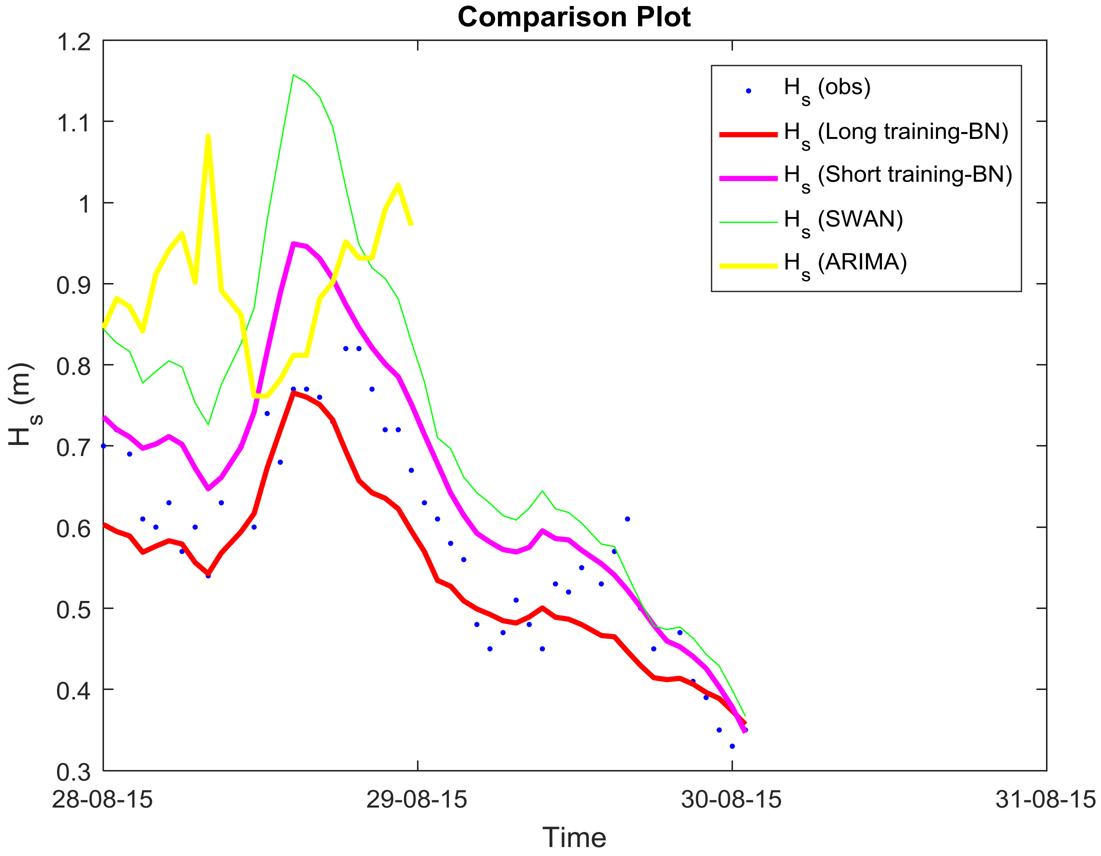


Fig. A. 23. Comparison between the BN and ARIMA techniques for 28-30 August 2015 (Gwynt-y-Mor - Irish Sea).

Table A. 18. Preliminary Evaluation Metrics (RMSE - MAE table) for 28-30 August 2015, at Gwynt-y-Mor (Irish Sea).

Method	SWAN	Long-trained BN	Short-trained BN	ARIMA[49]
RMSE (48 hours)	0.1777 m	0.0684 m	0.0899 m	-
RMSE (24 hours)	0.2325 m	0.0775 m	0.1076 m	0.2377 m
Maximum Absolute Error	0.3889 m	0.1633 m	0.2084 m	0.5416 m

[49] ARIMA: (0,0,1)(0,1,0) with a seasonality equal to 32 hours.

Appendix B

B.1. Numerical Model (SWAN) Timeseries

For the reader to be more acquainted with the numerical model (SWAN) data used during the yearly simulations of the error correction model, the timeseries of the variables of interest for both stations in the Irish Sea, i.e. the ones corresponding to the Gwynt-y-Mor and Rhyl Flats wind farms, are displayed below (Figures B.1 - B.2 and B.5 - B.6). Alongside with the variables used in the analysis, the timeseries of the swell parameters (Figures B.4 and B.7) are also collated. It is clear by looking at the graphs, that the differences the swell creates at the hydrodynamic data can be significant, and certainly not negligible.

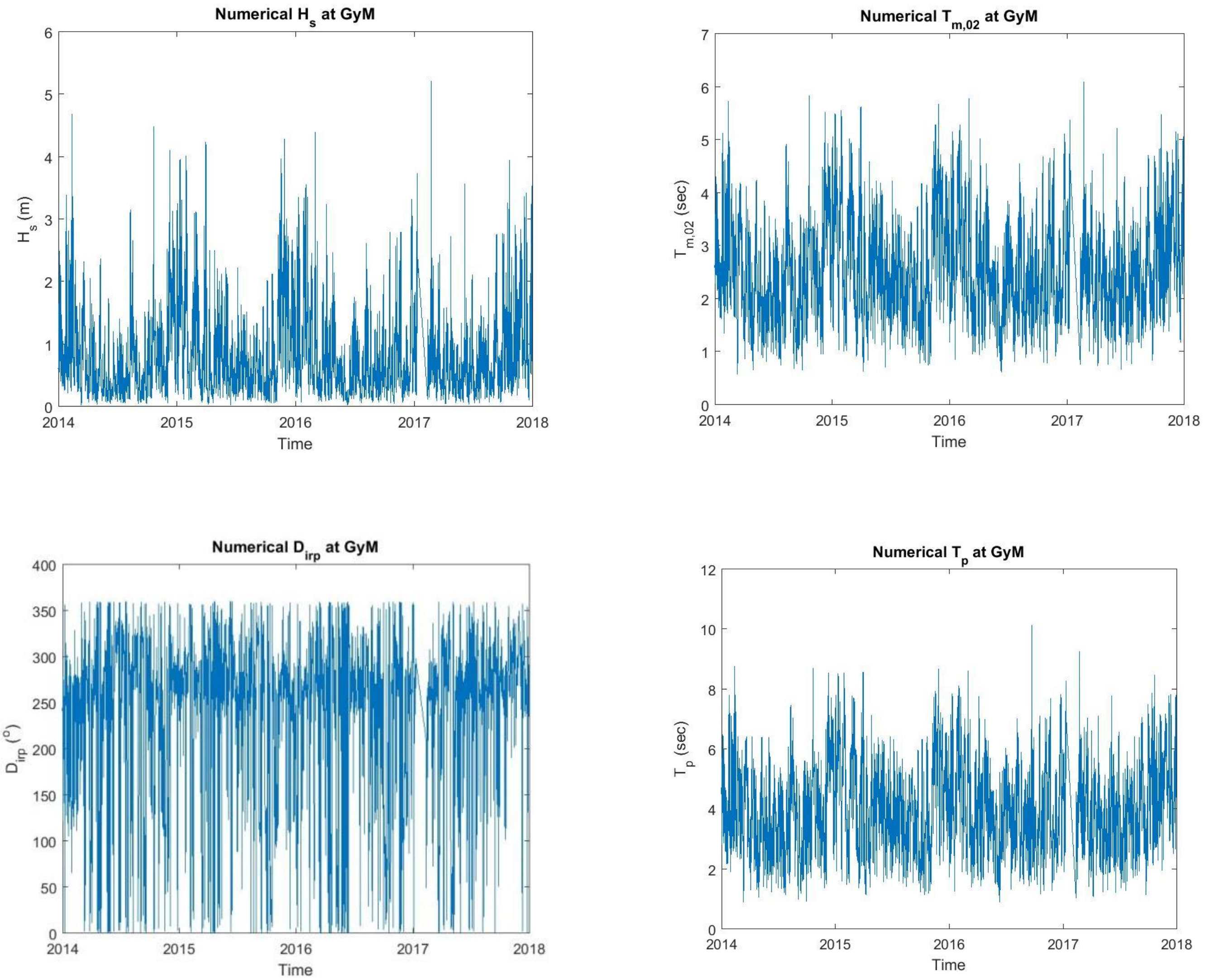


Fig. B. 1. Significant wave height (top left), zero-crossing wave period (top right), wave direction (bottom left), and peak wave period (bottom right) produced by SWAN (Gwynt-y-Mor).

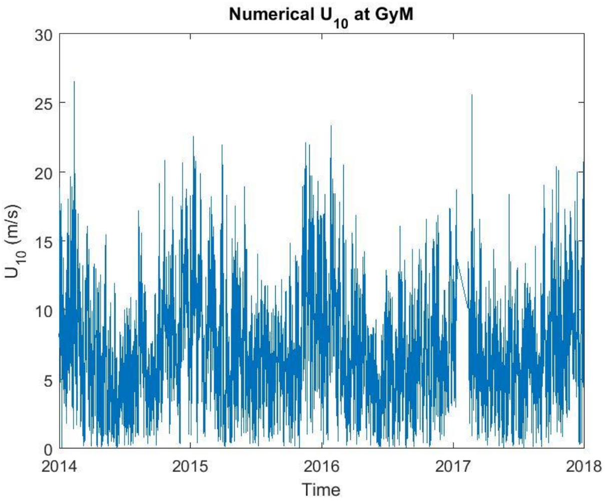


Fig. B. 2. Numerical model (HIRLAM) wind velocity data, which served as input data to SWAN (Gwynt-y-Mor).

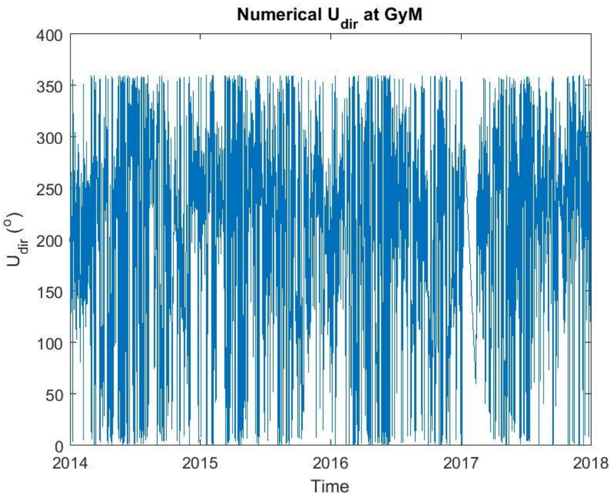


Fig. B. 3. Numerical model (HIRLAM) wind direction data, which
served as input data to SWAN (Gwynt-y-Mor).

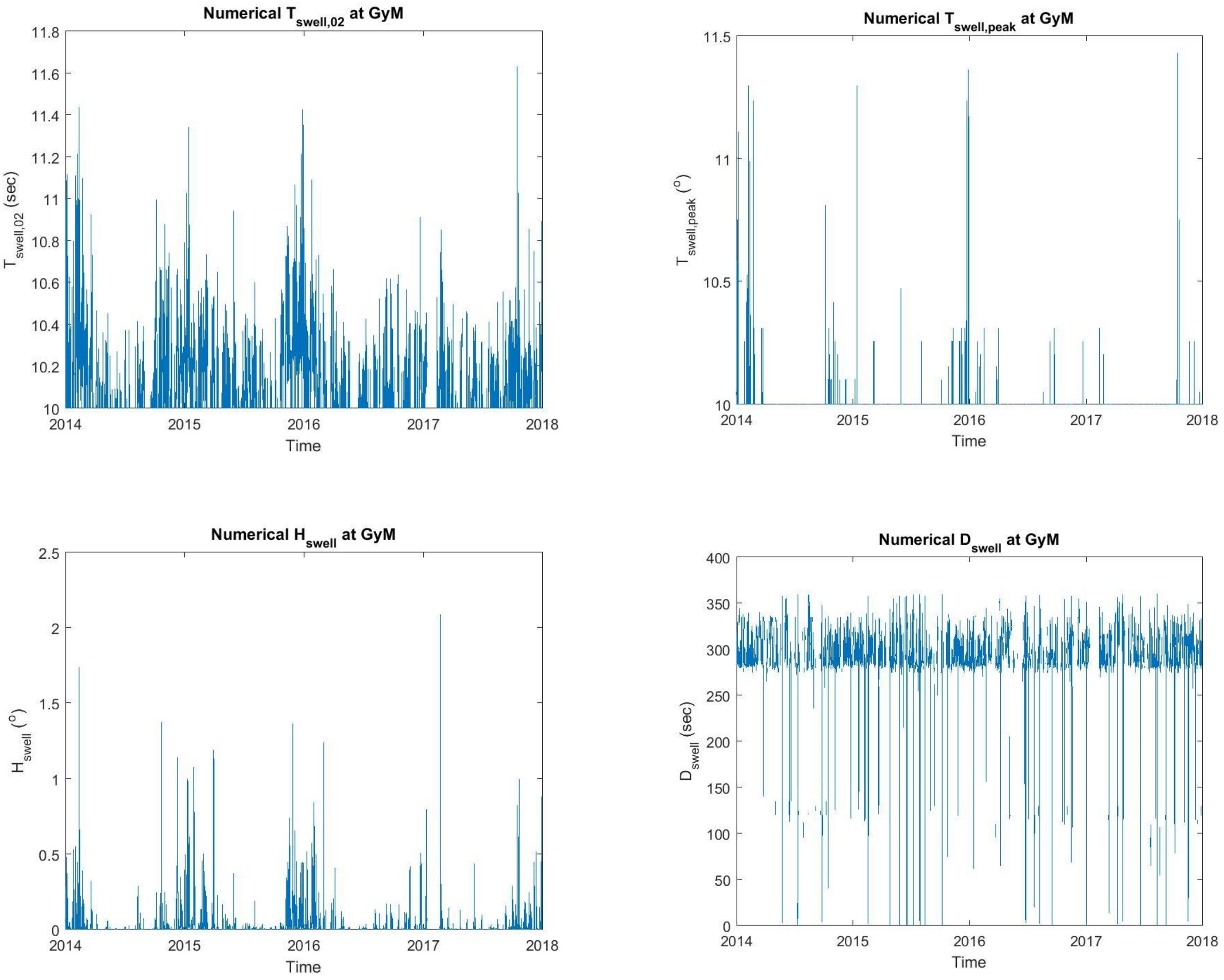


Fig. B. 4. Timeseries of the swell components, as produced by SWAN (Gwynt-y-Mor).

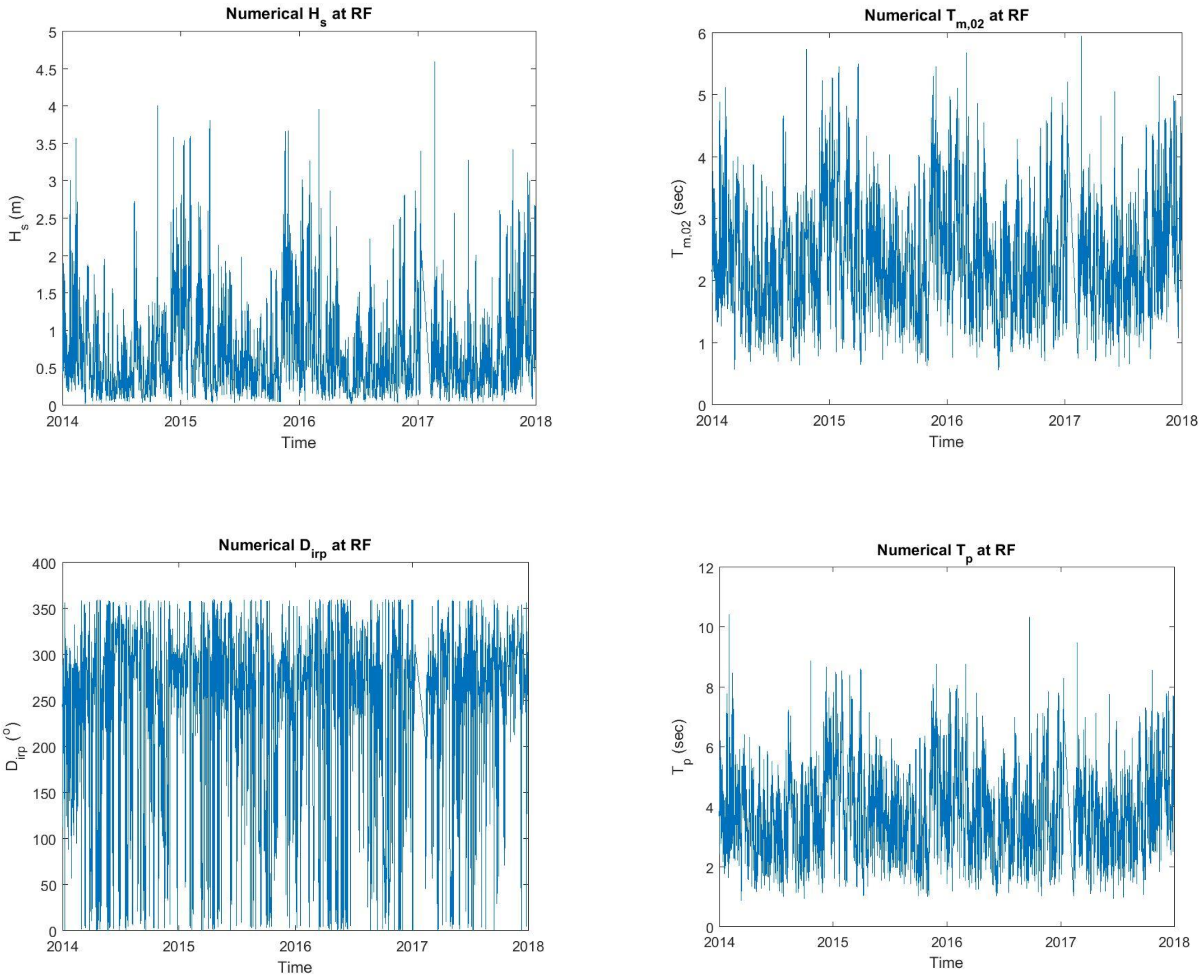


Fig. B. 5. Hydrodynamic data timeseries as produced by SWAN for the case of Rhyl Flats.

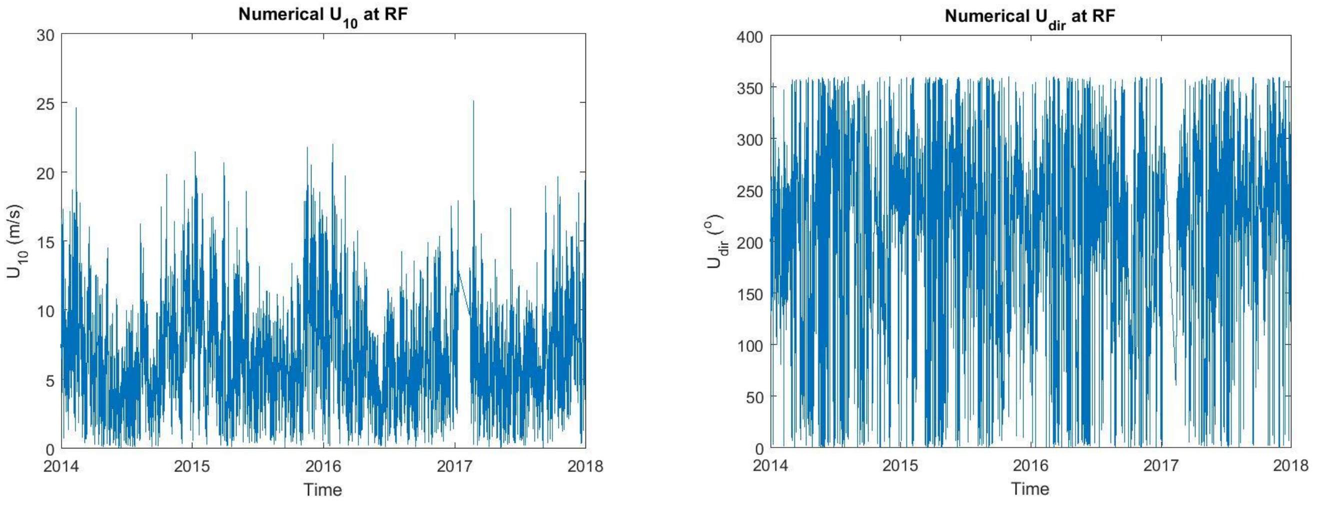


Fig. B. 6. Meteorological data timeseries as used in SWAN for the case of Rhyl Flats (produced by HIRLAM).

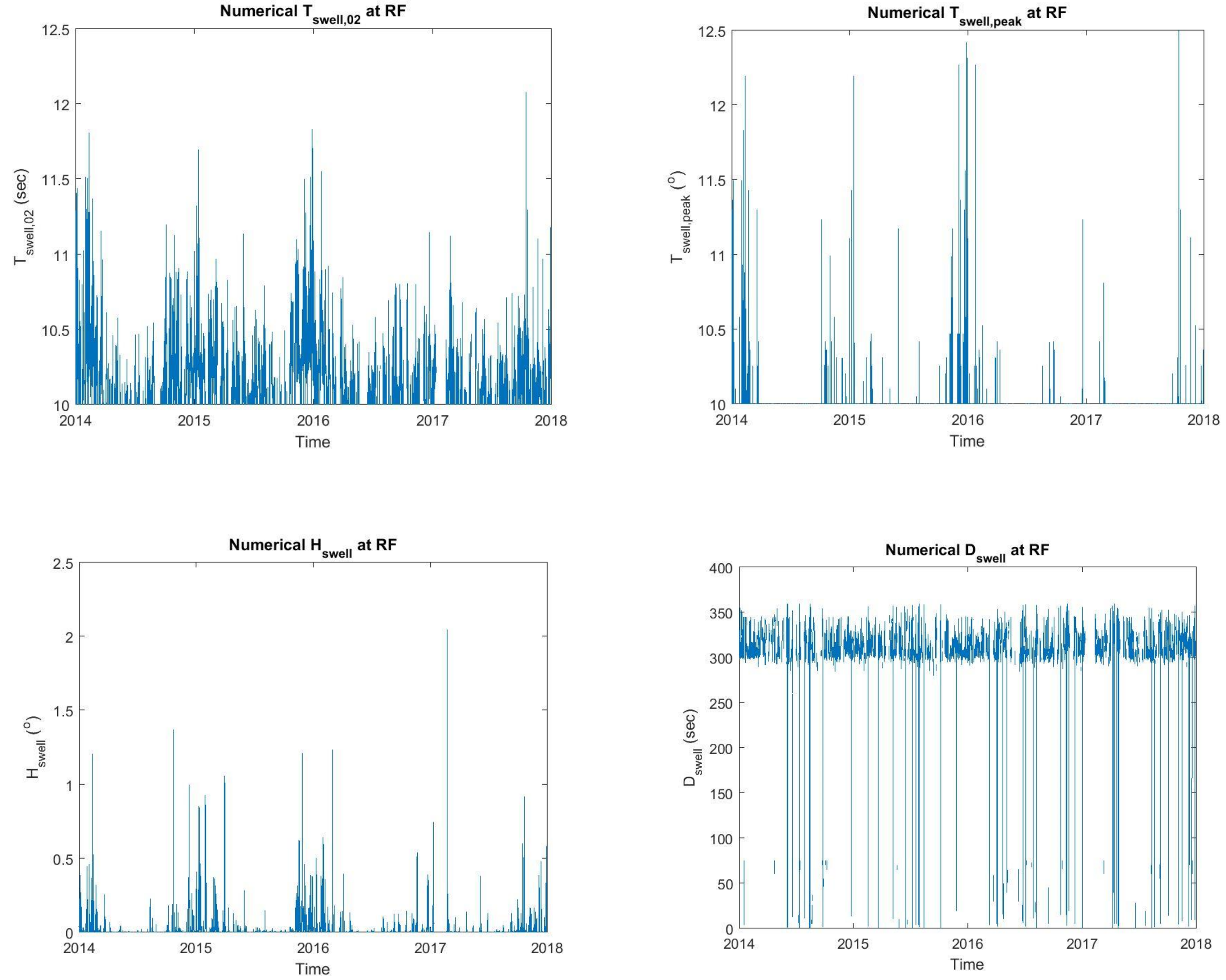


Fig. B. 7. Timeseries of the swell components, as produced by SWAN (Rhyl Flats).

B.2. Raw and Clean Observational Data Timeseries

In order for the reader to have a complete picture of the application, the raw and clean measurement's timeseries are presented in the following figures (Figures B.8 - B.11) for the Rhyl Flats offshore wind farm. The issues displayed during the peak wave period cleaning are also visible (Figure B.11).

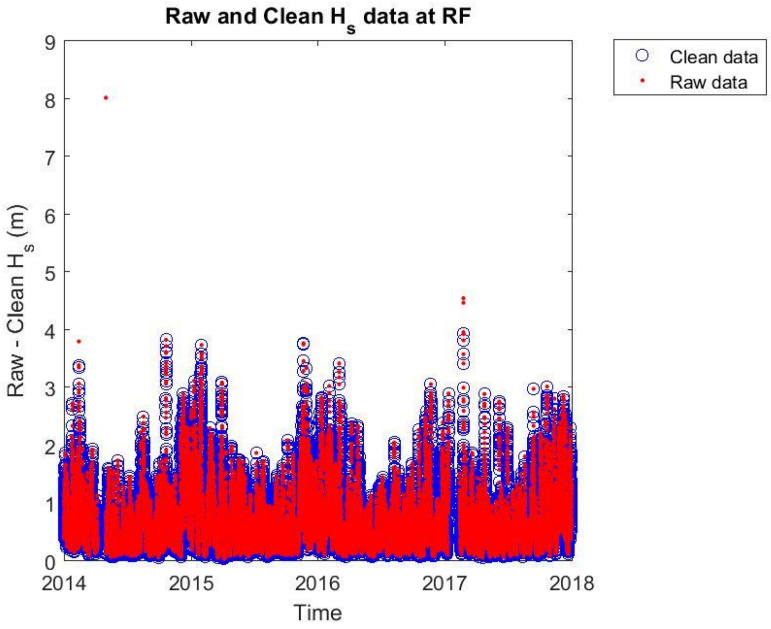


Fig. B. 8. Clean (blue circles) and raw (red dots) significant wave height (H_S) timeseries for the Rhyl Flats case.

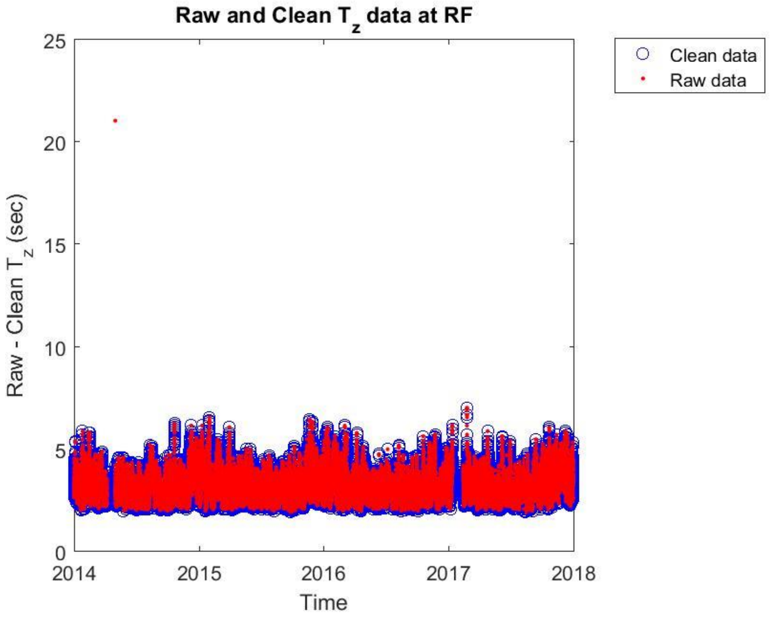


Fig. B. 9. Clean (blue circles) and raw (red dots) zero-crossing wave period (T_Z) timeseries for the Rhyl Flats case.

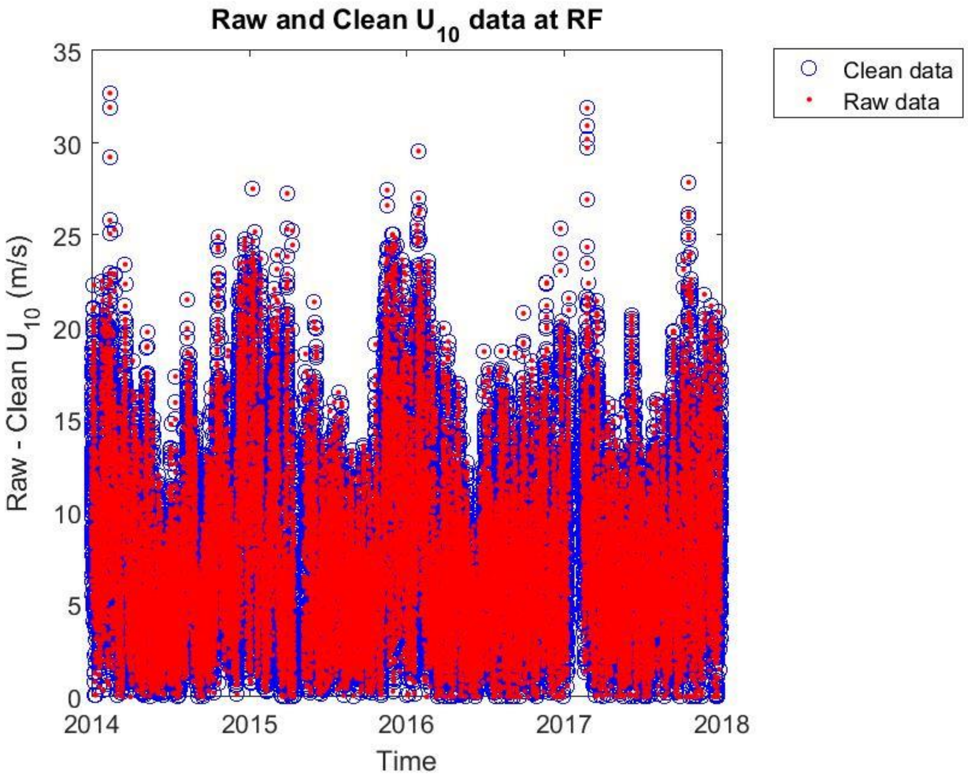


Fig. B. 10. Clean (blue circles) and raw (red dots) wind velocity (U_{10}) timeseries for the Rhyl Flats wind farm.

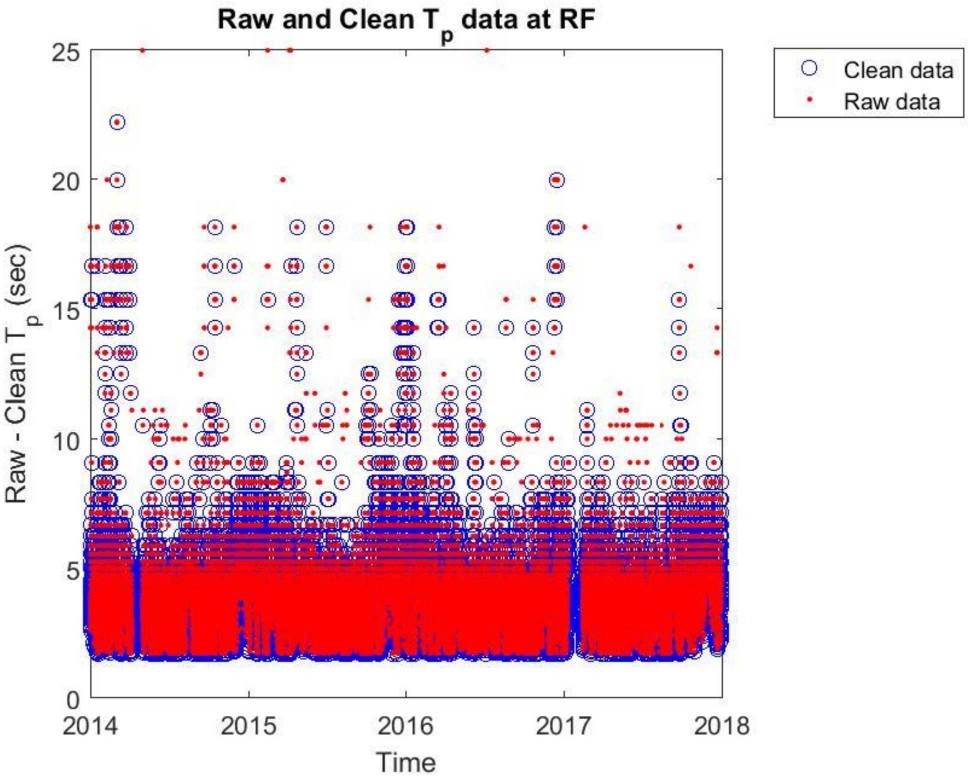


Fig. B. 11. Clean (blue circles) and raw (red dots) peak wave period (T_p) timeseries for the Rhyl Flats wind farm.

As mentioned in Chapter 3, the nature of the wave and wind direction timeseries makes their presentation almost superfluous. Nevertheless, for the sake of consistency in the presentation of the variables' timeseries and their cleaning procedure, the raw and clean data of the aforementioned parameters can be seen below (for both stations). Because the direction is measured in degrees (^{o}), the values essentially create a circle. As a result it is virtually impossible to extract something useful from the timeseries, except from the behaviour and the availability of the data as a whole.

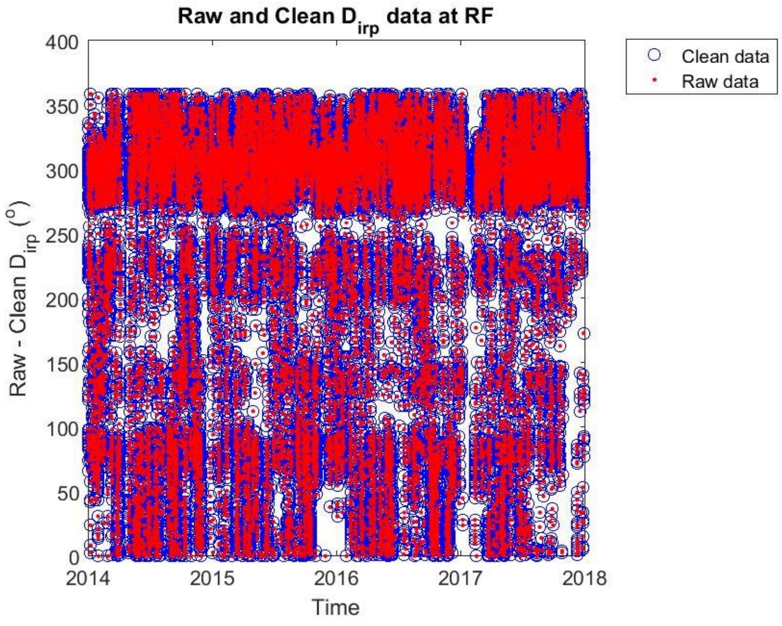


Fig. B. 12. Clean (blue circles) and raw (red dots) wave direction (D$_{irp}$) timeseries for the Rhyl Flats wind farm.

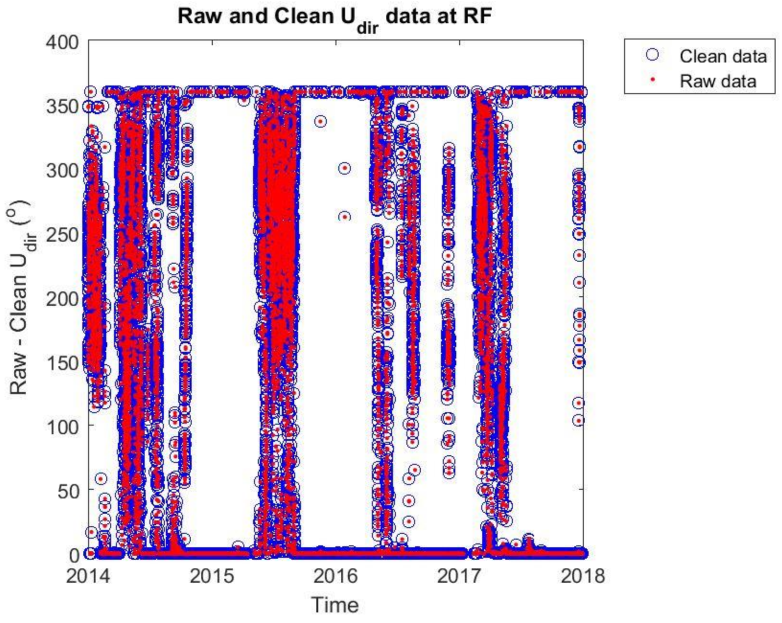


Fig. B. 13. Clean (blue circles) and raw (red dots) wind direction (U$_{dir}$) timeseries for the Rhyl Flats wind farm.

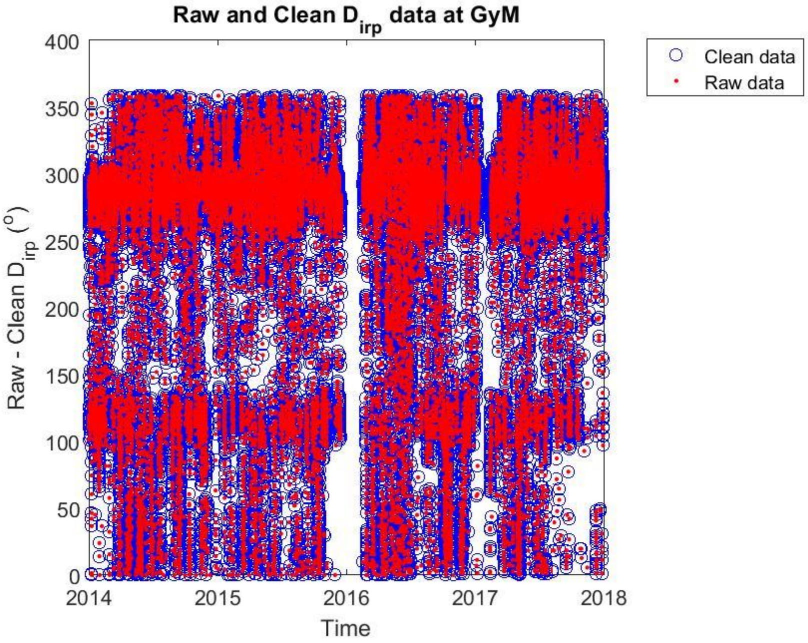


Fig. B. 14. Clean (blue circles) and raw (red dots) wave direction (D_{irp}) timeseries for the Gwynt-y-Mor wind farm.

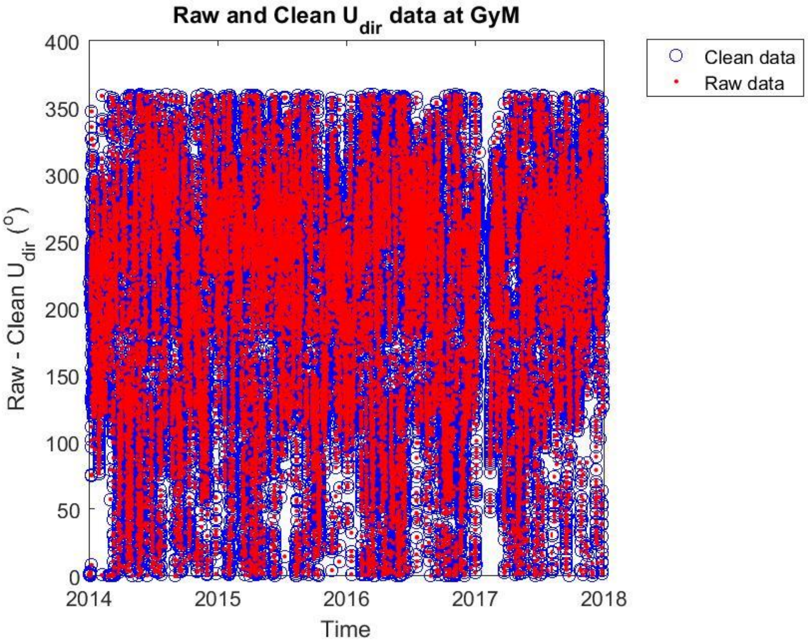


Fig. B. 15. Clean (blue circles) and raw (red dots) wind direction (U_{dir}) timeseries for the Gwynt-y-Mor wind farm.

As a matter of fact the general behaviour of the data, present an interesting element. As it can be seen in Figure B.13, the wind direction data are concentrated in one direction. Thus, there is a clear, much more dominant than any other, wind direction, while the wave direction is more scattered. The wind roses produced by means of the ORCA tool, also display this kind of behaviour (see Figures B.16 and B.17). The aforementioned behaviour plays an extremely important role in the way the short-trained BN methods conduct a prediction (see also Chapter 5). To be more specific, the dominant nature of one wind direction makes the forecast really inaccurate in many occasions. This is one of the reasons why different configurations of BNs were tested, with the results presented in Chapter 5 and Appendix C.

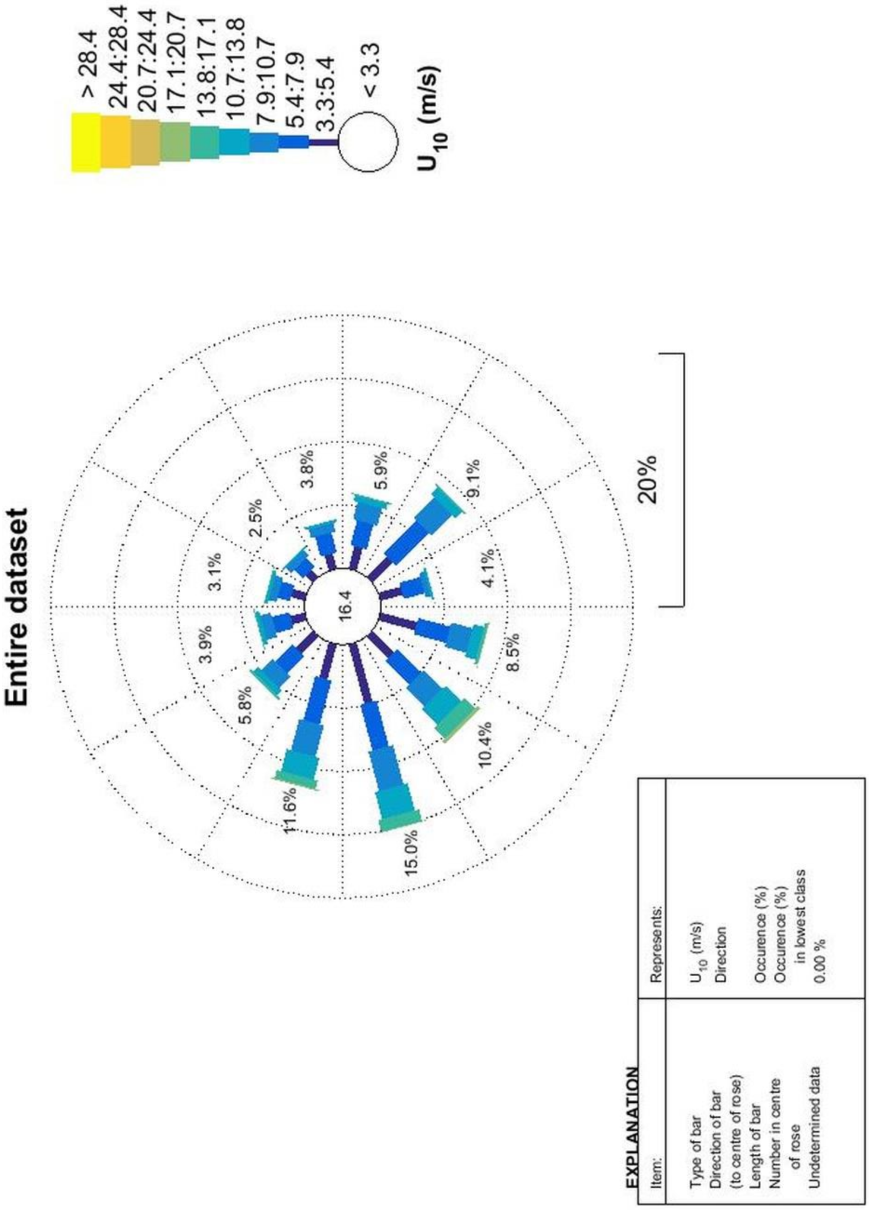


Fig. B. 16. Wind rose for the case of Gwynt-y-Mor.

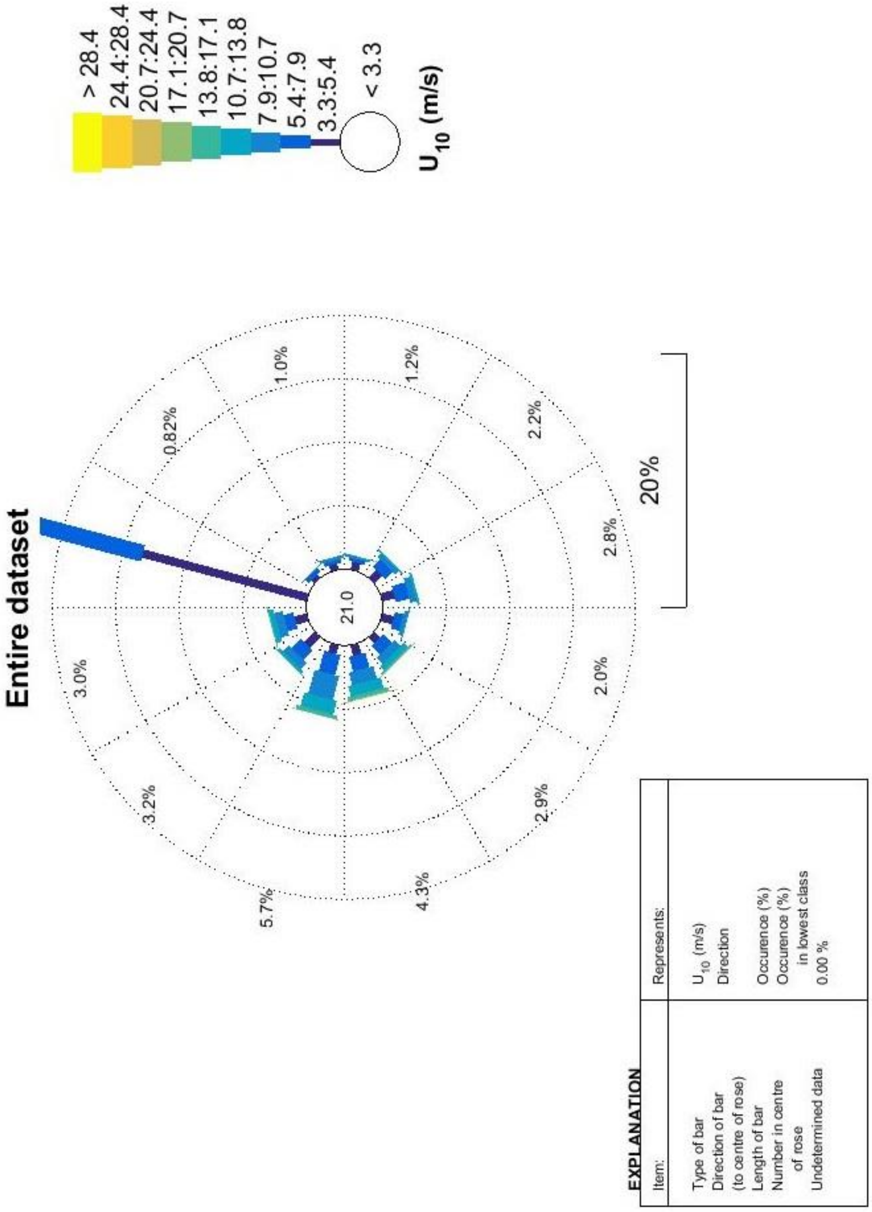
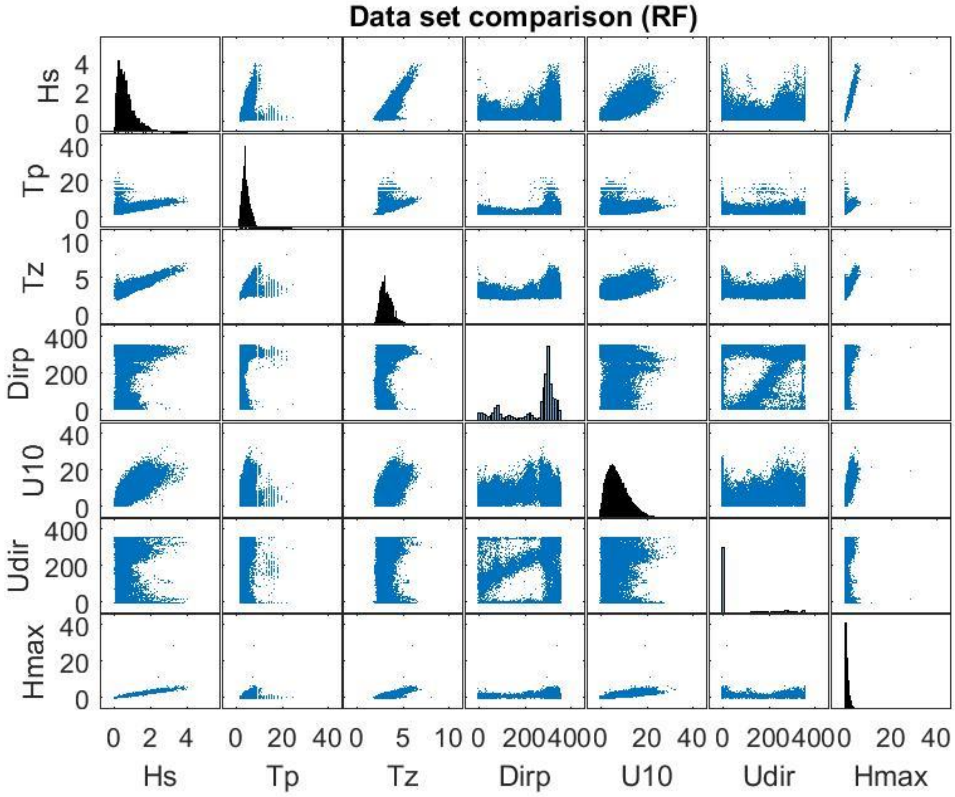


Fig. B. 17. Wind rose for the case of Rhyl Flats

Regarding the individual variables' distributions and the joint occurrence of wind and waves in the location of Rhyl Flats, similar graphs were produced as the ones presented for Gwynt-y-Mor. Again in this occasion, the individual distributions of the significant wave height (H_s) and the wind velocity (U_{10}) resemble a Rayleigh distribution, in accordance with the literature (see Figure B.18). Supplementary, the variables which have a clear, almost linear relation can be distinguished, while from Figure B.19 it is evident that the wind velocity and the significant wave height are highly correlated (it can be seen that periods with high winds correspond to time periods of high waves).



.

Fig. B. 18. Joint and individual distributions of the hydrodynamic and meteorological variables at Rhyl Flats.

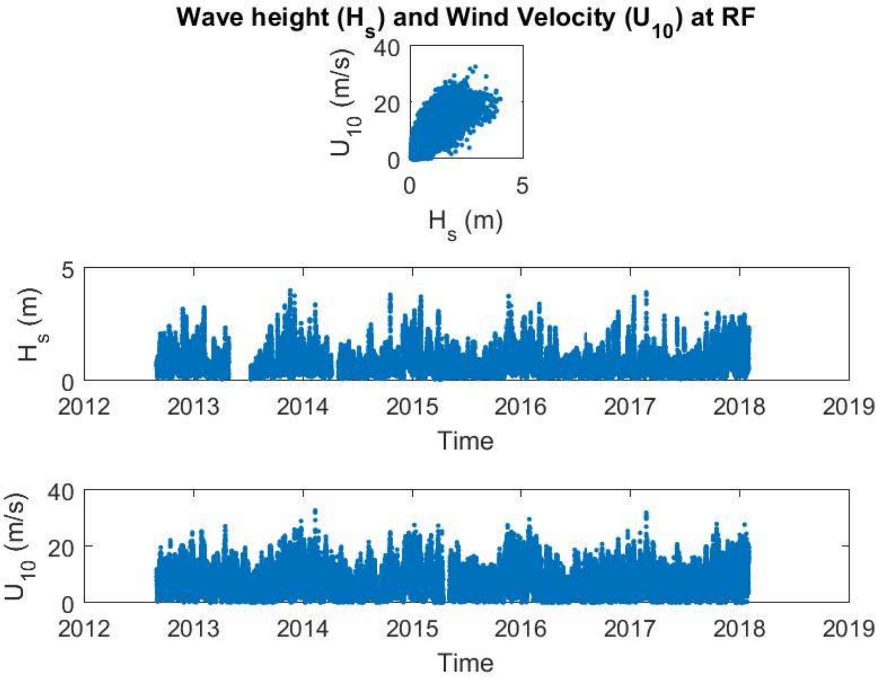


Fig. B. 19. Joint occurrence of wind and waves at Rhyl Flats.

B.3. Numerical Model Data (SWAN) Vs Measurements

The relation between the measurement data and the datasets produced by SWAN or HIRLAM can be visualised in scatterplots. Below, the scatterplots between numerical and observational data for the Rhyl Flats' measurement station are displayed (see Figure B.20). Again in this case, a distinctive relation, almost linear, for the H_s, the T_z, and the U_{10} is observed, justifying a large correlation between the model and observational variables. Regarding the cleaning procedure, the relations displayed below solidify that a satisfying cleaning procedure took place for that station as well.

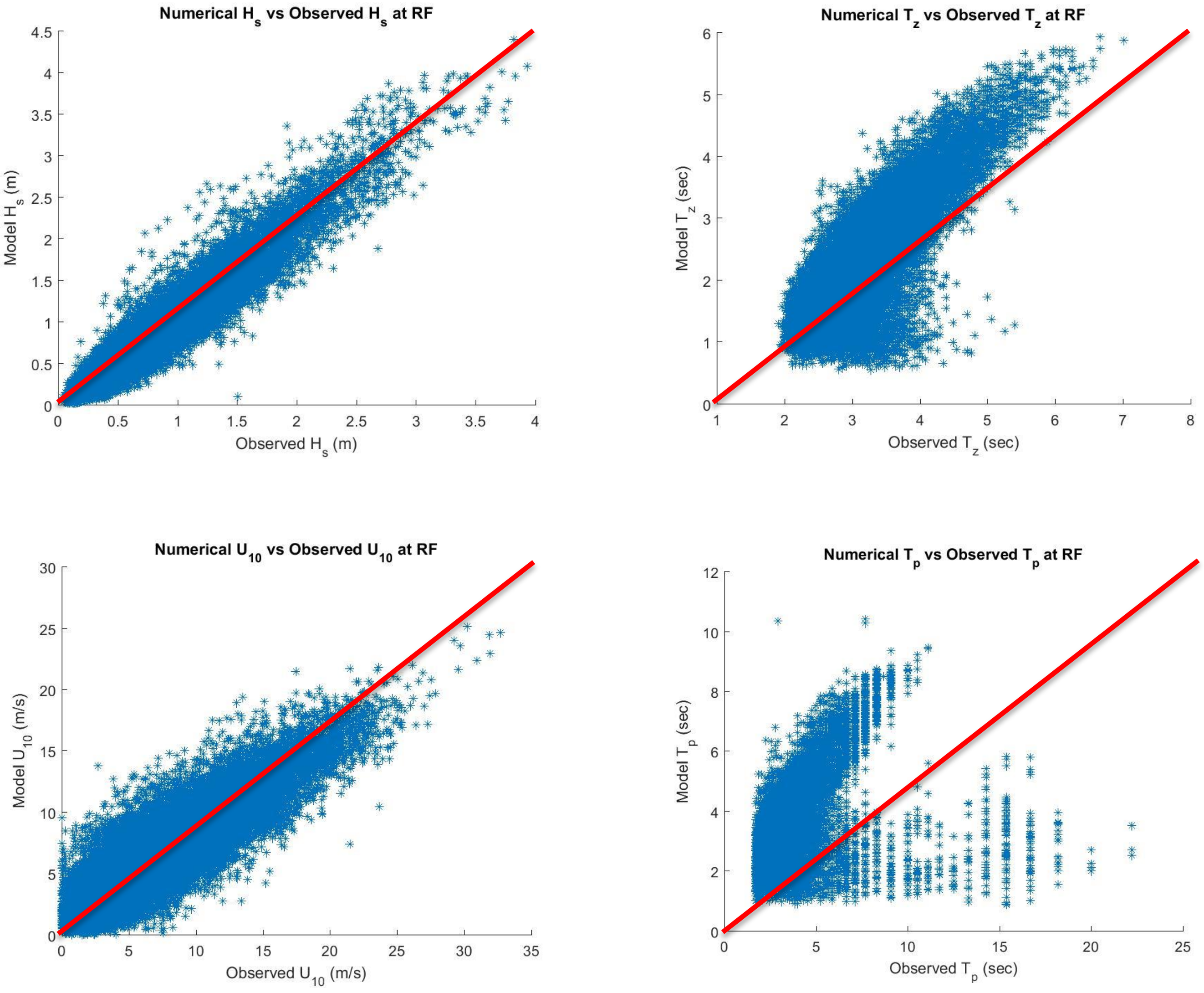


Fig. B. 20. Numerical data vs Measurements for the Rhyl Flats' meteorological and hydrodynamic datasets.

For the case of the peak period (T_p), seen in the bottom right of Figure B.20, a behaviour which definitely makes the use of that variable unsafe is shown. Again, the observational data involve swell parameters, which are separated from the pure variable's timeseries in the case of numerical modelling. As a result, significant differences are displayed, alongside with the known behaviour due to discretization. Consequently, the Rhyl Flats datasets verify and support the decision of not including the peak wave period in the simulations of the BN model. Supplementary, the scatterplots concerning the wave (D_{irp}) and wind (U_{dir}) directions are presented in Figure B.21, for the Rhyl Flats (top row) and Gwynt-y-Mor stations (bottom row). The strange nature of the scatterplots is due to the fact that the directions are measured in degrees (^{o}), i.e. their values are points on a circle. For example, if the model direction value is 1^{o} and the measurement is 359^{o} the difference in reality is insignificant, while seemingly on the scatterplot is large. As a result, the concentrated values around the diagonal, in combination with the values on the upper left and bottom right corners are quite satisfactory. Certainly, due to the large uncertainty incorporated in the value, many values pose differences, which don't create disturbances in the prediction accuracy of the BN model.

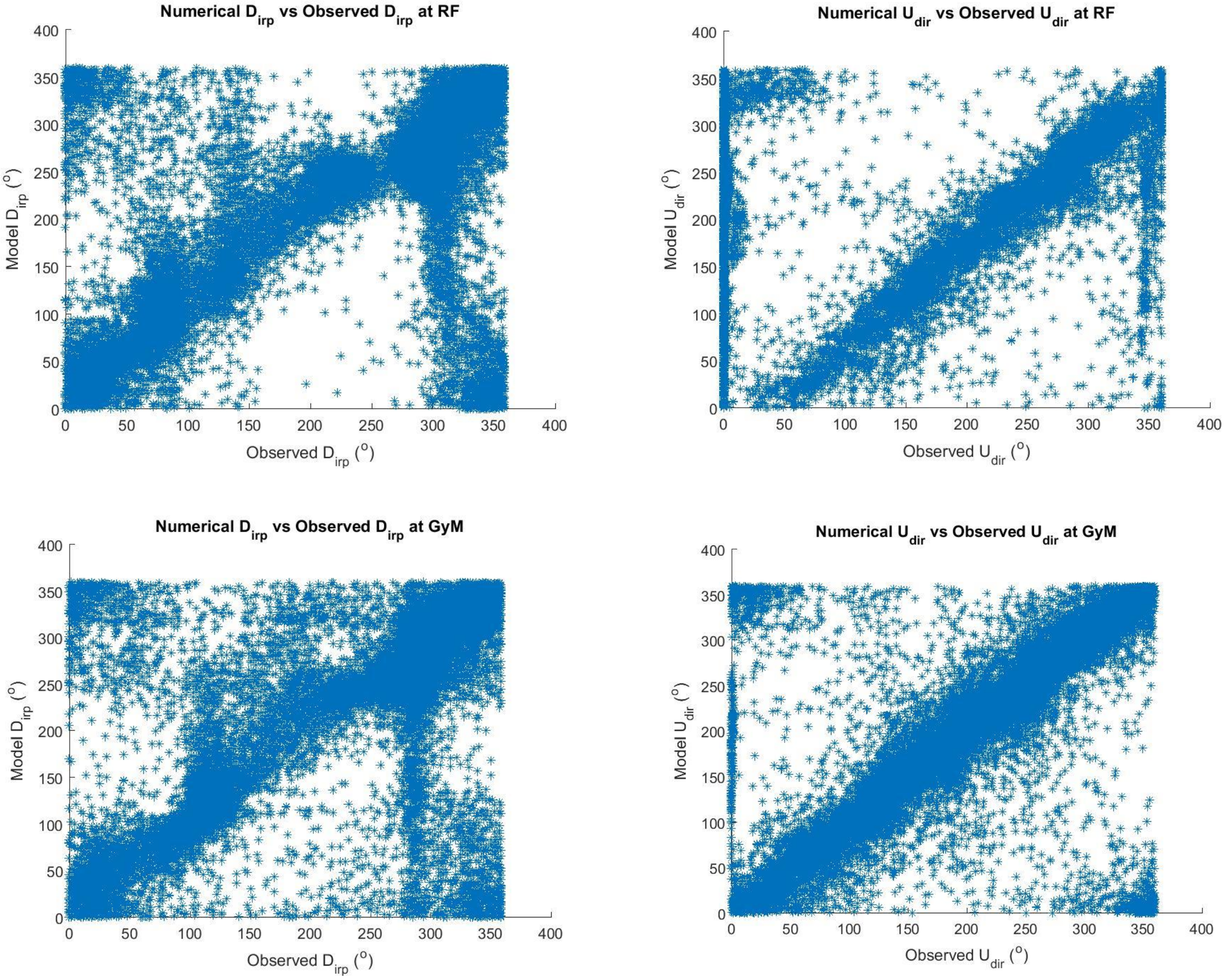


Fig. B. 21. Numerical data vs Measurements of wind (U_{dir}) and wave (D_{irp}) direction at Rhyl Flats and Gwynt-y-Mor.

B.4. Density scatterplot for GyM

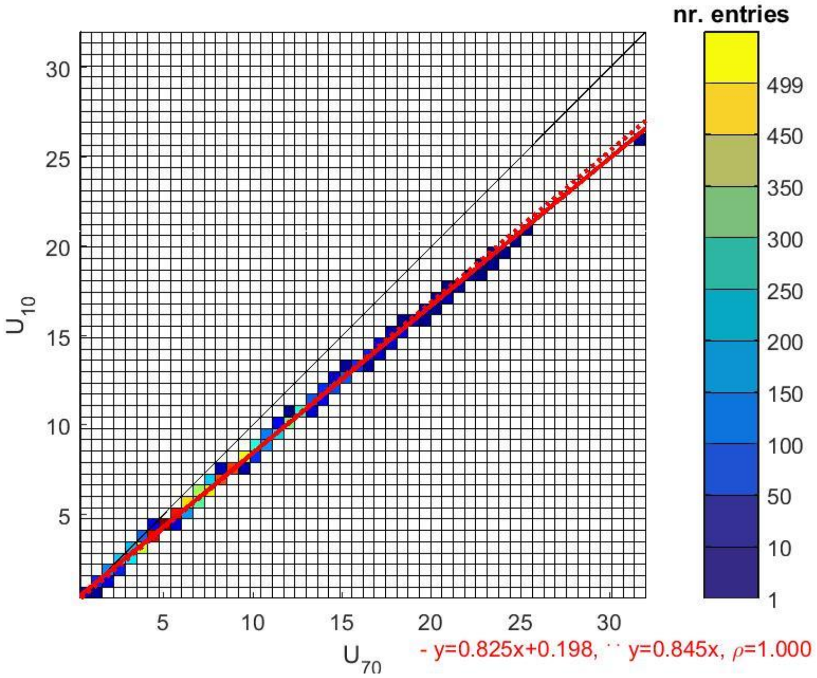


Fig. B. 22. Density scatterplot illustrating the relation between wind velocities at different levels (Gwynt-y-Mor).

B.5. Data Cleansing Filters

In statistics, a moving average is a calculation to analyse data points, by creating series of averages of different subsets of the complete dataset. It is the most commonly used filter in signal processing, mainly because it is the easiest digital filter to understand and use. The equation of the moving average filter is given as follows (see also Smith, 1999):

$$y[i] = \frac{1}{M} \cdot \sum_{j=0}^{M-1} x[i+j] \qquad \text{(B.1)}$$

where, $x[\,]$ is the input signal, $y[\,]$ is the output signal, and *M* is the number of points in the average. Due to the intense variability characterising the hydrodynamic and meteorological data, a moving average of 1 week was used, with a boundary for removing values set at 4 times the average value of the corresponding time interval.

The second, stricter filter is based on a sliding time window, over which the standard deviation is calculated[50]. For a random variable vector *A* made up of N scalar observations, the standard deviation is defined as:

$$S = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{N} |A_i - \mu|^2} \qquad \text{(B.2)}$$

[50] See also: https://nl.mathworks.com/help/matlab/ref/movstd.html#d119e754971.

where, μ is the mean of *A*. The stricter nature of this filter stems from the definition of the value removing boundary, set at 2 times the standard deviation calculated in the respective time window. Here, again, the moving window was set equal to 1 week. In both filters, the set time frame could be even smaller, e.g. 2 days, which would be reasonable thinking of the variability and dynamic behaviour that the variables have. Nevertheless, cleaning would not be as consistent, since a possible malfunction of the measurement station, or continuous false observations that could take place for various reasons, would not be able to be detected and removed. As a result, the 1 week moving time window was considered the most suitable for this case.

B.6. Peak Wave Period (T_p) Cleaning Procedure

A procedure of particular interest, due to the nature of the variable, is the cleaning of the peak wave period (T_p) data. The peak wave period is the wave period with the highest energy, which is extracted from the wave spectra[51]. The peak wave period data include swell components[52], which create extremely large values for the period (in the order of 20-30 seconds). As a result, the cleaning process of the data cannot be consistent, especially if we consider the peak wave period data retrieved from the numerical model. As becomes evident in the preceding sections of the main text (Section 3.3), the differences between the clean observational and numerical model peak wave period data can be significant, and as a result unsafe to be used in the error correction simulations.

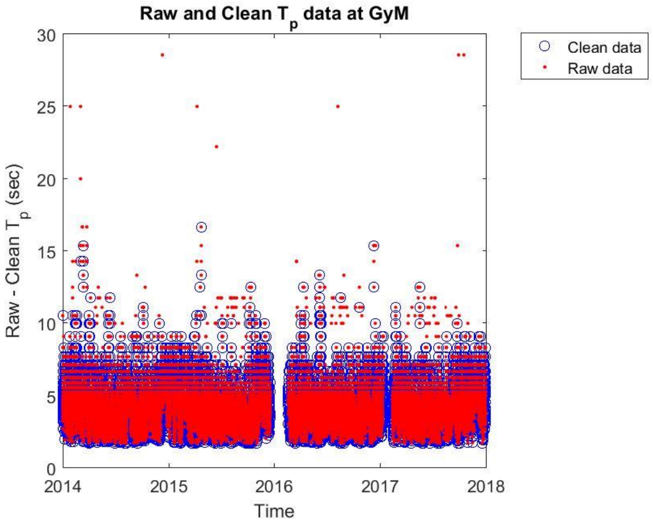


Fig. B. 23. Clean and raw peak wave period (Tp) datasets, as resulted from the ORCA cleaning procedure (Gwynt-y-Mor).

[51] The analysis of the distribution of the wave energy, as a function of wave frequency (period^{-1}) for a time-series of individual waves, is referred to as a spectral analysis. Wind wave periods (frequencies) often follow the so-called JONSWAP and Pierson-Moskowitz spectra.

[52] *Swell waves* is a series of surface gravity waves, which are not generated by the immediate local wind, but instead provoked by distant weather systems, where wind blows for a duration of time over a fetch of water.

Figure B.23 clearly illustrates that during the cleaning process, a large amount of data is removed, without proper physical justification. Also, a strange behaviour of continuous equal values is displayed, most probably due to the discretization done while deriving the raw data. As a result, either the data had to remain intact in order to be used, or their usage had to be abandoned. If the first one is the case, the differences with the numerical model wave period would still be significant.

B.7. Joint Occurrence of Wind and Waves

In order to be able to evaluate the dependence and the relations of the data-driven techniques, which incorporate the wind velocity and direction, alongside with the significant wave height, it is important to check their joint occurrence, i.e. whether high winds provoke high waves in the area under consideration. This can be illustrated by the scatter between the two observed variables, as well as a direct comparison of their “clean” timeseries (Figure B.24).

From the scatterplot (upper part of Figure B.24) it is evident that there is a direct relation between wind velocity and significant wave height, since the scatter is concentrated around the diagonal. From the timeseries it is also visible that, in many occasions, high wind velocities provoke large significant wave heights (e.g. in the start of 2017). This behaviour maybe seems obvious but in reality is far from it. High wind waves might be produced by storms, far away from the area of interest, but still get measured by the devices. It has to be stressed here that the wind velocity and the wave height are measured by two completely different devices, which have some distance between them.

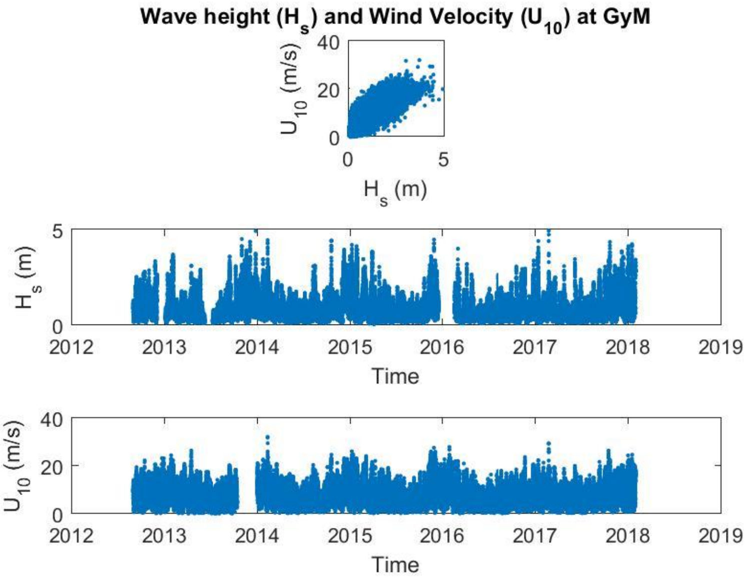


Fig. B. 24. Joint-occurrence of wind and waves at Gwynt-y-Mor.

Appendix C

C.1. Error Timeseries for the Gwynt-y-Mor wind farm

The respective error timeseries for the case of Gwynt-y-Mor is presented in the following figure (Figure C.1). It is shown that the behaviour of the models is far more stable in comparison to the one observed for the Rhyl Flats case. Probably the condition of the acquired data (observations), which was better for the case of Gwynt-y-Mor for the variable of wind direction (see also Appendix B), led to a more erratic type of behaviour for the Rhyl Flats dataset, which was counteracted by removing that node from the structure (see also Section 5.4.2).

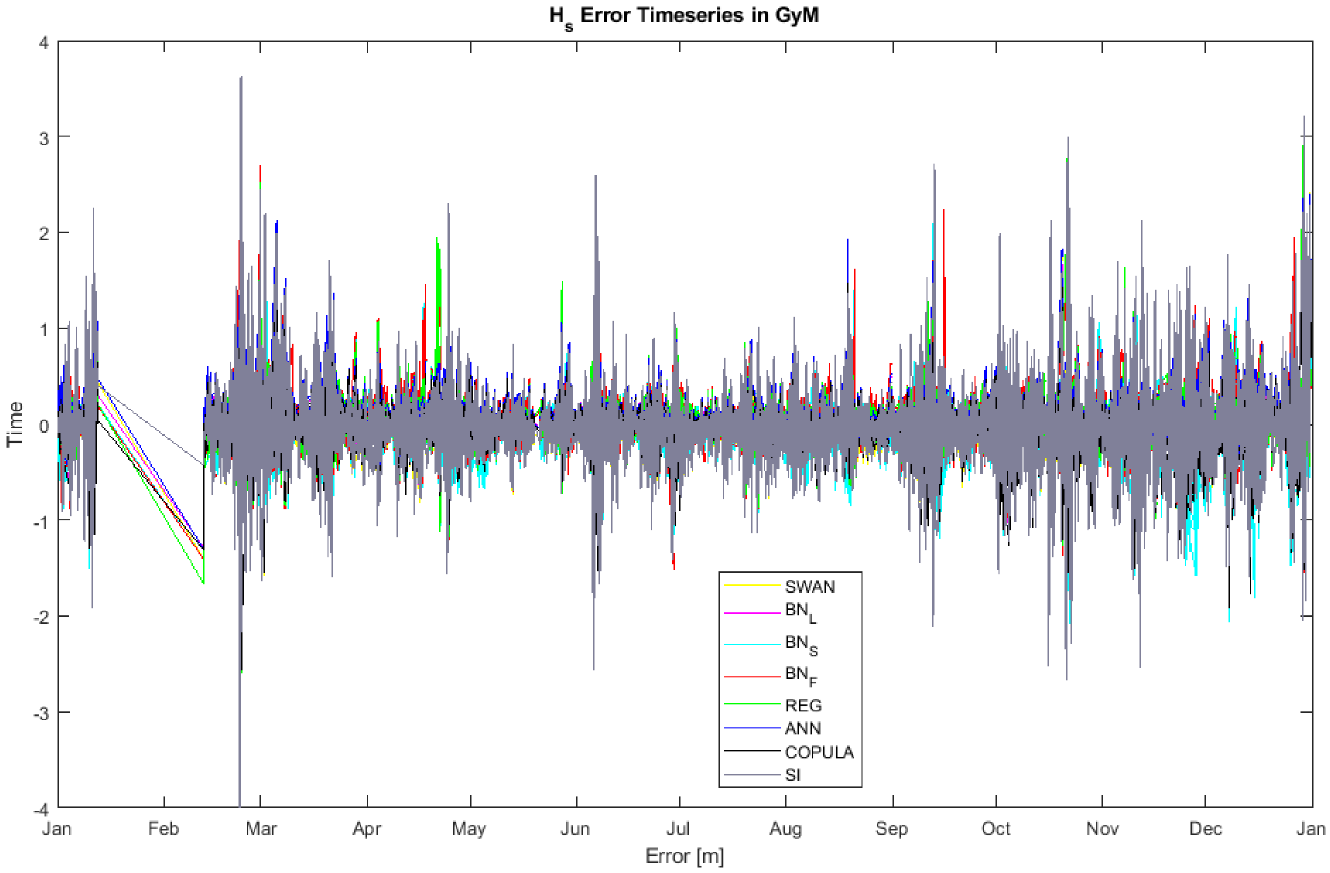


Fig.C. 1. Evolution of errors in relation with time for the Gwynt-y-Mor case study.

C.2. Scatterplots for the Rhyl Flats wind farm

The scatterplots presented below (Figures C.2 and C.3) illustrate the relation between the significant wave height (H_s) measurements acquired for the Rhyl Flats case, and the error correction and numerical models' results for the year of 2017. The erratic behaviour introduced for the BN model incorporating a fixed structure is evident by the outlying data above the diagonal in the lower right graph of Figure C.2. Most probably, responsible for that behaviour are the wind direction data (U_{dir}), since following tests without that variable revealed a more stable and accurate performance for this specific model.

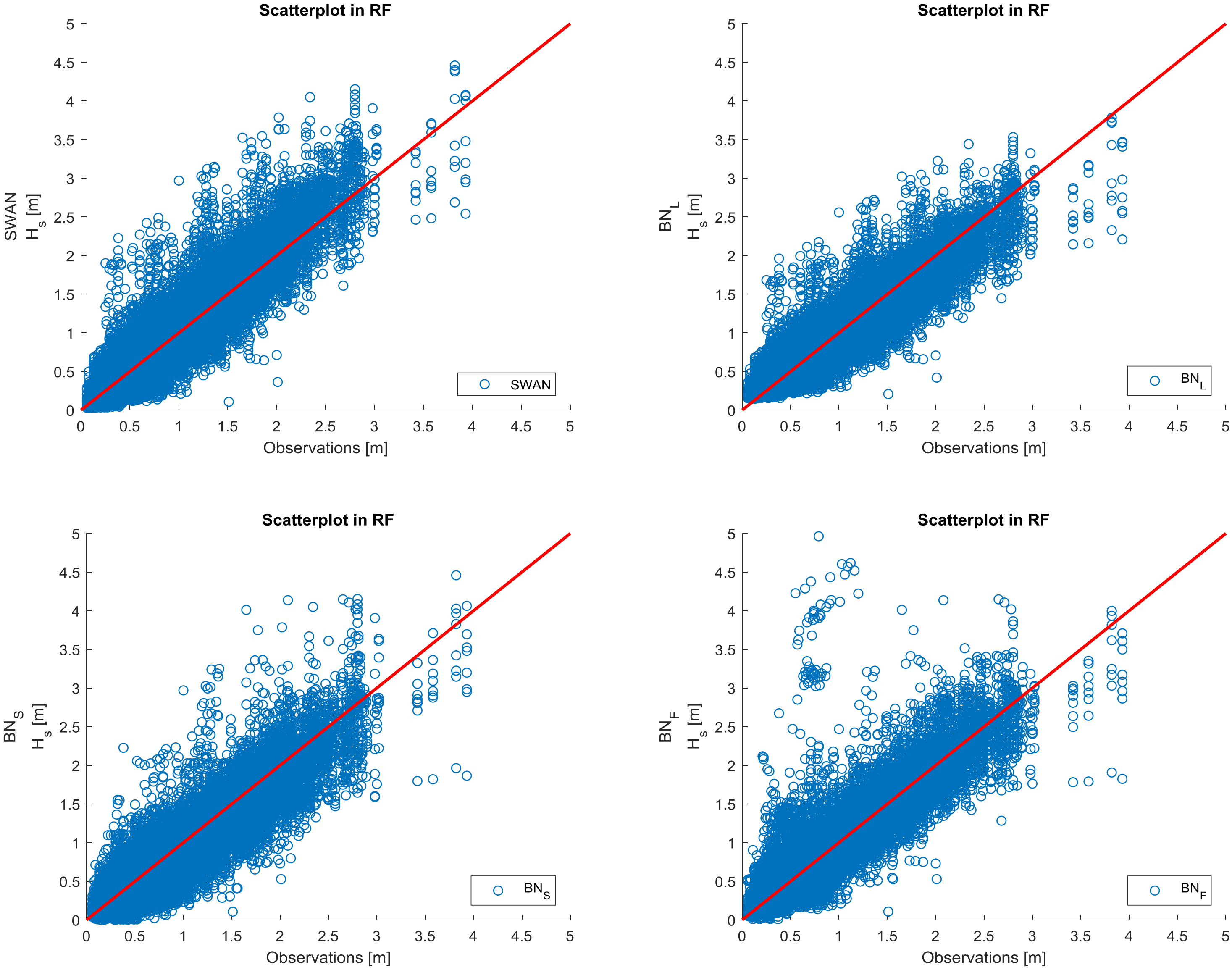


Fig.C. 2. Scatterplots of implemented BN models compared with SWAN results in relation to the observations in Rhyl Flats. (Jan 2017 – Jan 2018)

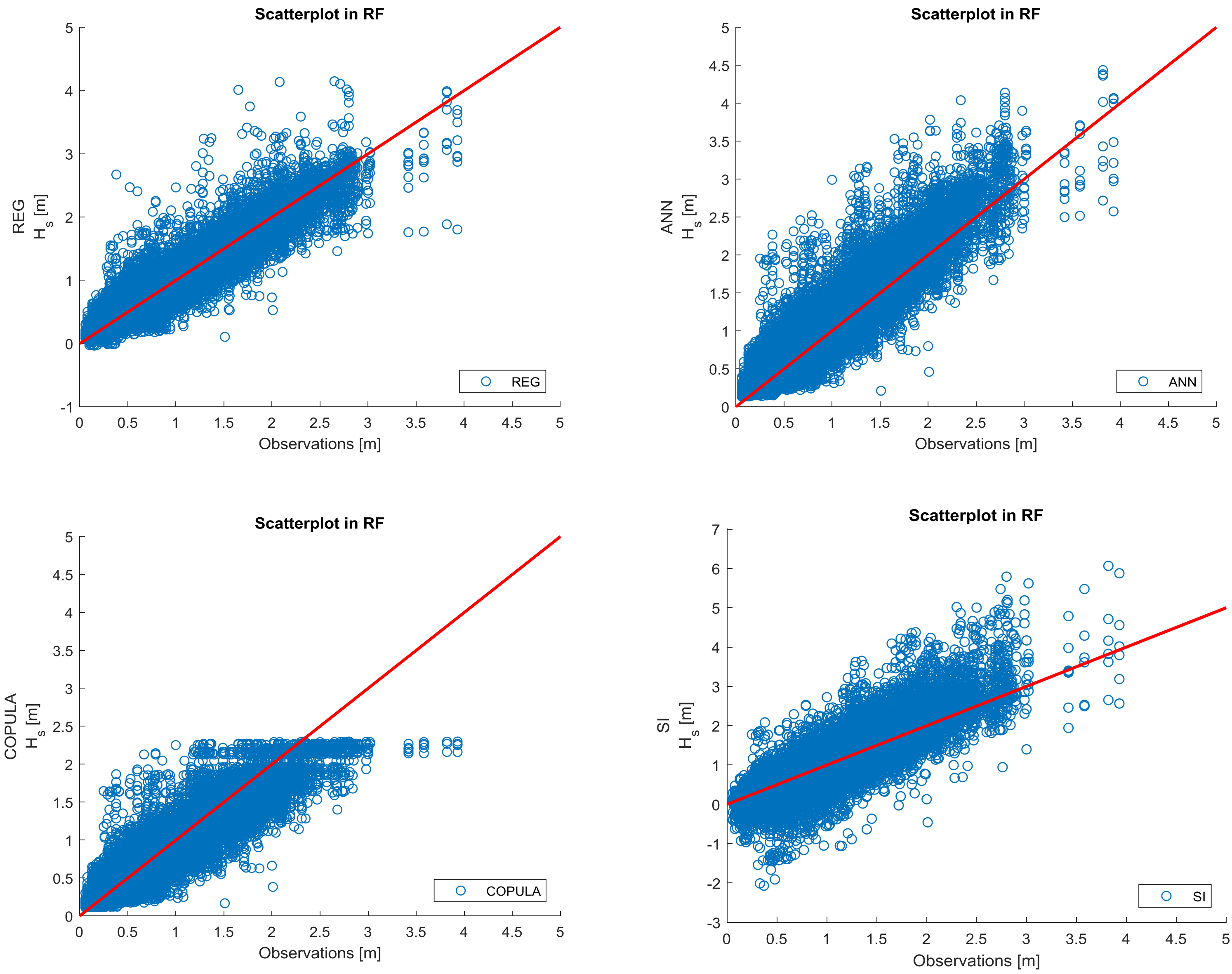


Fig.C. 3. Scatterplots of already operation error correction techniques in relation to Rhyl Flats' observations. (Jan 2017 – Jan 2018)

C.3. Taylor diagram for the Rhyl Flats wind farm

The Taylor diagrams of the 5-variable and 6-variable structures for the case of Rhyl Flats (Figures C.4 and C.5) display clearly the suitability of the first one for the application.

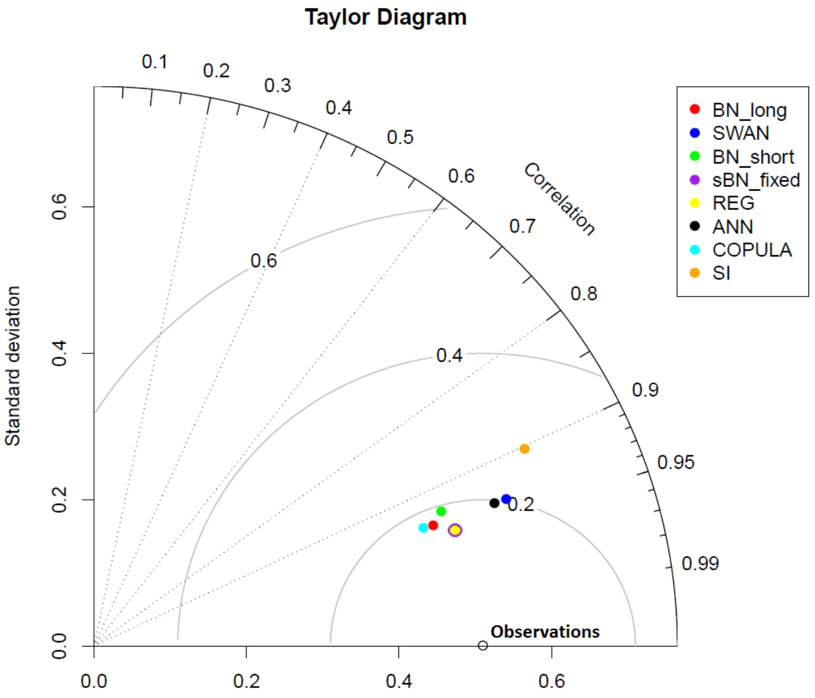


Fig.C. 4. Taylor diagram for the case where 5-variable BN models are incorporated (Rhyl Flats).

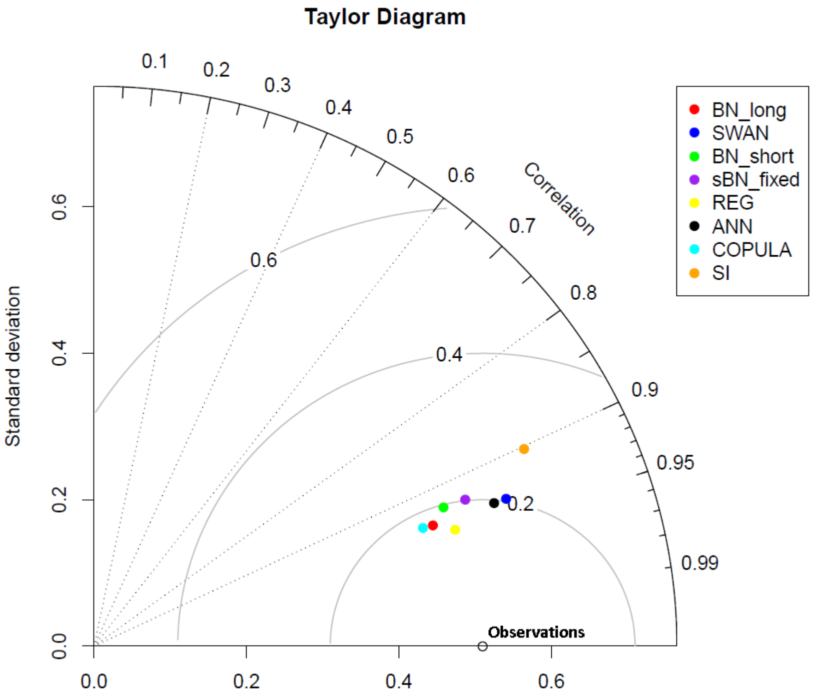


Fig.C. 5. Taylor diagram for the case where 6-variable BN models are incorporated (Rhyl Flats).

C.4. Uncertainty estimates for the Rhyl Flats wind farm

The uncertainty boundaries provided for the case of Rhyl Flats, for different BN structure configurations, are presented in the following tables (Tables C.1 and C.2). A general comment is that the log-normal uncertainty bounds of the long-trained BN model demonstrate the highest coverage percentage, with the cost of an also large average length. Its normal counterpart displays also a large coverage performance, with a much smaller (half the size) average length. As stated in the main text (Section 5.4.2) it is extremely difficult to distinguish which kind of confidence estimate is better, since both exhibit an equal amount of advantages. The most realistic ones are the log-normal bounds, since they solely produce positive values, but the most informative ones seem to be the normal intervals. Still, despite their ability to avoid negative values, the log-normal boundaries provide in certain occasions extremely large estimates which can be deceiving in this application, i.e. provide an upper boundary way above 1.5 meters, while the observations are below, hence endangering the normal conduct of the maintenance operations. As a result a general conclusion for this case, as previously stated for Gwynt-y-Mor, is that the normal uncertainty bounds fit and serve the application better, since the upper confidence boundary is essentially the one with the highest importance.

Table C. 1. Uncertainty estimates' performance for the case of the 4-variable BN structures (Rhyl Flats).

Method	BN Long Training	BN Fixed Structure	BN Short Training	Copula	BN Long Training (Log-N)	BN Short Training (Log-N)	BN Fixed Structure (Log-N)
Coverage (%)	89.6	80.0	80.0	70.9	95.0	80.5	73.2
Average Length (m)	0.527	0.450	0.450	0.327	1.024	0.460	0.536

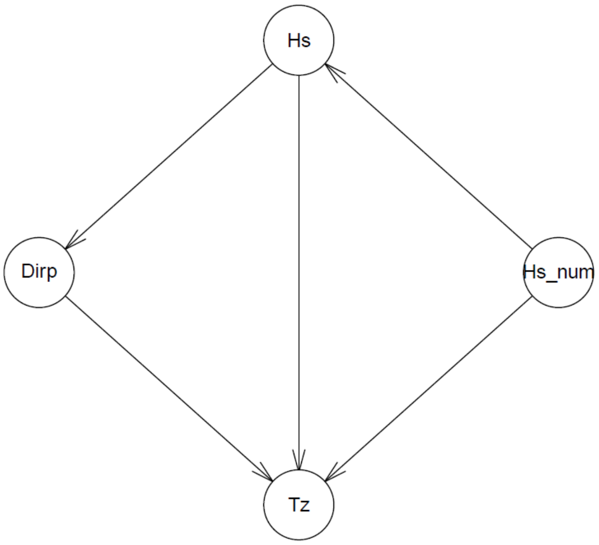
Table C. 2. Uncertainty estimates' performance for the case of the 6-variable BN structures (Rhyl Flats).

Method	BN Long Training	BN Fixed Structure	BN Short Training	Copula	BN Long Training (Log-N)	BN Short Training (Log-N)	BN Fixed Structure (Log-N)
Coverage (%)	89.7	64.7	69.8	70.9	94.7	68.9	61.0
Average Length (m)	0.527	0.491	0.427	0.327	0.948	0.466	0.425

What can also be concluded is that the exclusion of the wind direction in Rhyl Flats is extremely beneficial for the uncertainty estimates. Due to the erratic behaviour of the predictions the confidence intervals are also inaccurate for those occasions, and as a result the coverage percentage drops significantly. The long-trained BN model continues to be the most consistent and robust in both cases.

C.5. BN structures for the Rhyl Flats wind farm

For the sake of completeness some examples of BN structures, for the case of Rhyl Flats, are presented below (see Figures C.6 to C.8). Also in this case the differences between short- and long-trained BN models are evident. For the cases of BN models incorporating one or more meteorological variables, i.e. the wind velocity (U_{10}) or/and the wind direction (U_{dir}), the relations with hydrodynamic variables, such as the significant wave height (H_s) or the wave direction (D_{irp}), can be examined by mean of Tables C.3 to C.5.



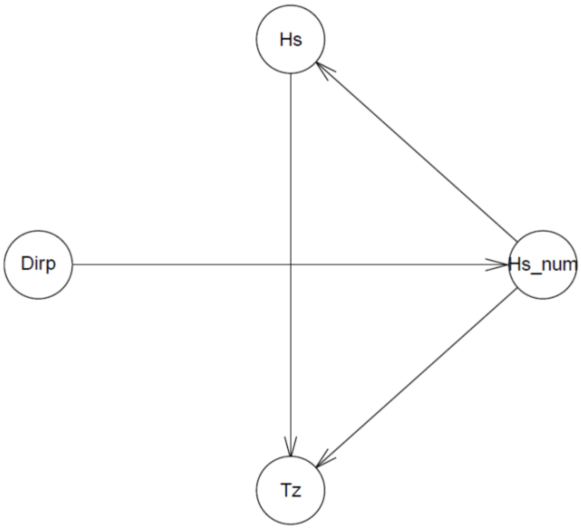


Fig.C. 6. Long-trained (left) and short-trained (right) BN model 4-variable structures for the Rhyl Flats case. (21-02-2017 at 06:00)

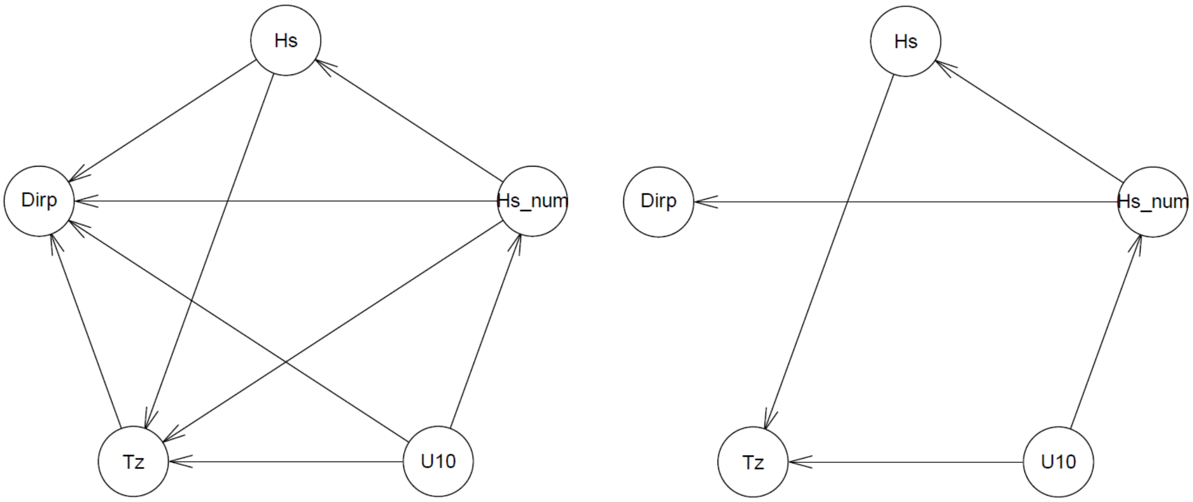


Fig.C. 7. Long-trained (left) and short-trained (right) BN model 5-variable structures for the Rhyl Flats case. (21-02-2017 at 06:00)

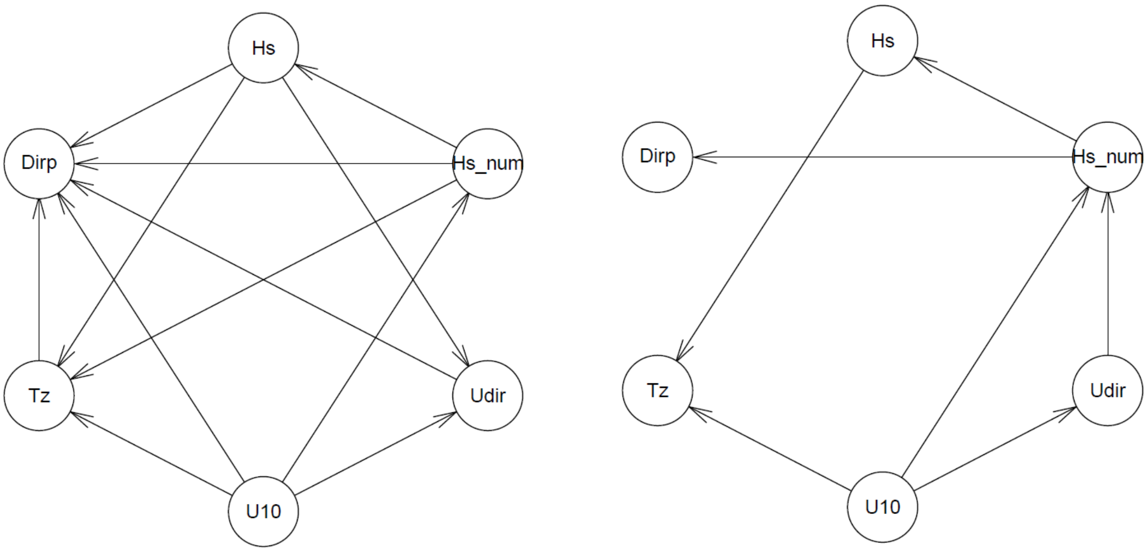


Fig.C. 8. Long-trained (left) and short-trained (right) BN model 6-variable structures for the Rhyl Flats case. (21-02-2017 at 06:00)

A very interesting point for discussion is the inexistent connection of the wind velocity (U_{10}) and the observed significant wave height (H_s). In none of the two long-trained (constant over time) structures incorporating the meteorological parameters the two aforementioned variables are connected. The same holds for the short-trained BN models. It has to be stressed out that all of the preceding structures for the short-training correspond to dates and times in which the correction was quite satisfying. Certainty, the significant wave height is indirectly dependent (or conditionally independent) on the wind velocity through the numerical significant wave height ($H_{s,num}$), but still the data show a completely different behaviour in comparison to the Gwynt-y-Mor case, even if the locations are both in Liverpool Bay (Irish Sea).

Table C. 3. Correlation matrix for the short-trained BN model 4-variable structure (Rhyl Flats).

Variable	D_{irp}	T_z	$H_{s,num}$	H_s
D_{irp}	1.000	-0.444	-0.911	-0.843
T_z	-0.444	1.000	0.386	0.604
$H_{s,num}$	-0.813	0.338	0.935	0.858
H_s	0.150	0.072	-0.204	-0.268

Examining Tables C.4 and C.5 it is shown that the correlation coefficient between the wind velocity (U_{10}) and the significant wave height (H_s) is quite large. Nevertheless, no connection is visible on the corresponding long- or short-trained structures, except their indirect dependence (indirect causal effects) through the significant wave height produced by SWAN ($H_{s,num}$). Since the $H_{s,num}$ is essentially given, the wind velocity and the significant wave height are conditionally independent throughout the tests, and that is most probably the reason that a decrease in accuracy and performance is noticed when custom-fitted relations (i.e. an imposed dependency between the wind velocity and the significant wave height) come into play.

Table C. 4. Correlation matrix for the short-trained BN model 5-variable structure (Rhyl Flats).

Variable	D_{irp}	T_z	U_{10}	$H_{s,num}$	H_s
D_{irp}	1.000	0.444	-0.813	-0.911	-0.843
T_z	-0.444	1.000	0.338	0.386	0.604
U_{10}	-0.813	0.338	1.000	0.935	0.858
$H_{s,num}$	0.150	0.072	-0.359	-0.204	-0.268

H_s	-0.911	0.386	0.935	1.000	0.875

Table C. 5. Correlation matrix for the long-trained BN models (Rhyl Flats).

Variable	D_{irp}	T_z	U_{10}	U_{dir}	$H_{s,num}$	H_s
D_{irp}	1.000	0.429	0.106	0.072	0.278	0.285
T_z	0.429	1.000	0.494	-0.013	0.797	0.834
U_{10}	0.106	0.494	1.000	-0.055	0.782	0.751
U_{dir}	0.072	-0.013	-0.055	1.000	-0.033	-0.025
$H_{s,num}$	0.278	0.797	0.782	-0.033	1.000	0.961
H_s	0.285	0.834	0.751	-0.025	0.961	1.000

C.6. Comparison of data-driven and imposed structures

In case a relation was imposed between the hydrodynamic and meteorological data, e.g. the observed significant wave height (H_s) being dependent on the wind velocity (U_{10}) the results would not be as satisfying as the ones produced by the data driven procedure. Figure C.9 illustrates an example where the data-driven 5-variable structure for Gwynt-y-Mor outperforms its imposed counterpart. The example presented below serves just presentation purposes and displays the general tendency of the models on a randomly selected day.

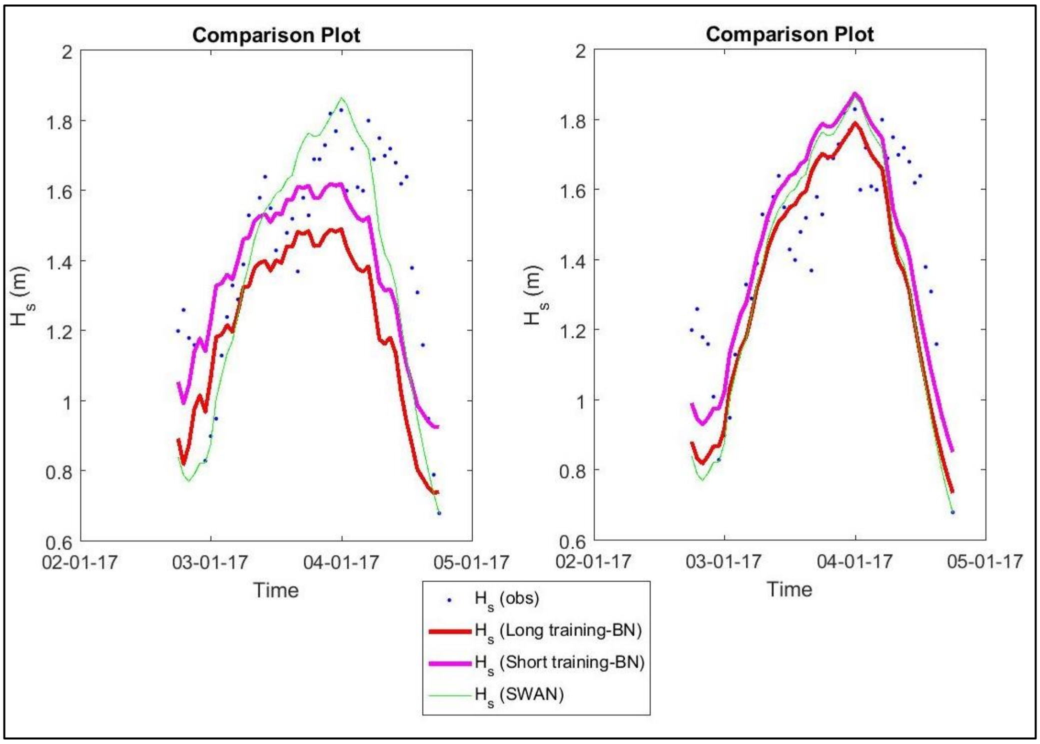


Fig.C. 9. Example of differences between a data-driven and an imposed structure results for Gwynt-y-Mor.

Appendix D

D.1. Copula Families

The Gaussian copula[53] is formulated as follows:

$$C(u, v) = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right) \tag{D.1}$$

where $u, v \in [0,1]$, Φ denotes the standard normal distribution function and Φ_ρ the standard bivariate normal distribution function with linear correlation coefficient ρ. The Gumbel and Clayton copulas are one parameter Archimedean copulas, which are defined as:

$$C(u, v) = \varphi^{-1}\left(\varphi(u) + \varphi(v)\right) \tag{D.2}$$

where φ is the generator function of the respective copula. The generator functions of Gumbel and Clayton copulas are given respectively by (see also Nelsen, 2003):

$$\varphi(u) = (-\ln(u))^\theta,\ \theta\ \in [1, \infty) \tag{D.3}$$

[53] The Gaussian copula alongside with the student-t, are elliptical copula families (see Fang et al., 2002; 2005).

$$\varphi(u) = \frac{(u^{-\beta} - 1)}{\beta}, \beta \in [-1, \infty) \tag{D.4}$$

As a result the Gumbel copula is formed as:

$$C(u, v; \theta) = exp\left\{-\left[(-\ln(u))^{\theta} + (-\ln(v))^{\theta}\right]^{\frac{1}{\theta}}\right\} \tag{D.5}$$

While the Clayton copula can be written as:

$$C(u, v; \beta) = \left(u^{-\beta} + v^{-\beta} - 1\right)^{-\frac{1}{\beta}} \tag{D.6}$$

D.2. Cramer-von Mises

The Cramer-von Mises statistic is formulated as follows (see also Remillard, 2010):

$$S_n = \int_{[0,1]^d} A_n^2(u) \cdot d \cdot C_n(u) = \sum\left\{C_n(u) - C_{\theta_n}(u)\right\}^2 \tag{D.7}$$

$$A_n = \sqrt{n} \cdot \left(C_n - C_{\theta_n}\right) \tag{D.8}$$

where C_n is the empirical copula and $C_{\theta n}$ the estimated theoretical copula.

The empirical copula C_n, according to Deheuvels (1979), is defined as follows:

$$C_n(u) = \frac{1}{n} \cdot \sum_{i=1}^{n} 1 \cdot (U_{i1} \le u_1, \dots, U_{id} \le u_d),\ with\ \ u = (u_1, \dots, u_d) \in [0,1]^d \tag{D.9}$$

D.3. Artificial Neural Networks Basics

The data are inserted to the input layer nodes, which transmit them to the hidden layer. The hidden layer nodes sum up the received values, add a bias to this sum, and then pass them through a nonlinear transfer function, like the log sigmoid or the hyperbolic tangent sigmoid. Those activation functions are used in order to make the ANN capable of representing the non-linear dependencies. The aforementioned procedure's result is transferred to the output nodes, which operate identically to the hidden nodes.

The feed forward network can be expressed mathematically in the following form:

$$y_k(x) = \sum_{j=1}^{M} w_{kj} \times T_r(z) + b_{ko} \tag{D.10}$$

$$z = \sum_{i=1}^{D} w_{ji} \times x_i + b_{ji} \tag{D.11}$$

where x is the original parameter space of dimension D, w_{kj} and w_{ji} are the weighting parameters, b_{ko} and b_{ji} are bias parameters, M is the number of the hidden nodes, and T_r (z) is the transfer (activation) function. The four more commonly used transfer functions are the unit step, the sigmoid, the piecewise linear and the Gaussian functions.

To achieve the required accuracy, an iterative procedure for minimizing the global errors between the observed and the network predicted values is used. To do so, sufficient training[54] of the network is required, in order for the weights and biases to be calculated. Analytically, the above statement can be expressed as follows:

$$E_p = \frac{1}{2} \cdot \sum_{k=0}^{N} (O_k - t_k)^2 \tag{D.12}$$

Where N is the total number of nodes in the output layer, O_k is the output of the kth node, and t_k is the target output at the kth node.

The back-propagation algorithm has been applied extensively in various engineering problems (see e.g. Goh, 1995; Yagawa and Okuda, 1996; Tsai and Lee, 1999; Kerh and Yee, 2000) throughout the years as a training technique, in which the steepest descent is used and the weights and biases are adjusted by moving a small step in the direction of negative gradient of the error function in each iteration, until convergence is reached. Other training methods are the conjugate gradient algorithm (see Fitch et al., 1991; Fletcher and Reeves, 1964), where the gradient descent is made along a direction which is conjugate or orthogonal to the previous step, and the cascade correlation algorithm (see Fahlman and Lebiere, 1990), through which training efficiency is achieved by optimization of the weights using the gradient ascent method[55]. For more information on theoretical concepts of neural networks the reader is referred to Kosko (1992), Wu (1994), Bose and Liang (1998), Wasserman (1993), Maier and Dandy (2000), Dawson and Wilby (2001) and the ASCE Task Committee (2000).

D.4. Applications of ARMA

The ARMA is a useful tool for analysing, modelling, and forecasting met-ocean data, such as the wave height. Hence, considerable attention has been given to this method in many ocean engineering applications. The Box – Jenkins autoregressive model has been used extensively to simulate time series of the significant wave height (see e.g. Guedes Soares and Ferreira, 1996). Sobey (1996) proposed that the sequences of individual waves can be described decently as a first order ARMA process. Non-stationary time series of wave spectral parameters with missing values were analysed, simulated and completed using

[54] This method is called *learning*. The various training sets incorporated to train the ANN are called *epochs*.

[55] In the gradient ascent method the correlation between output of a hidden node and the residual error of the network is maximized.

ARMA models (see Stefanakos and Athanassoulis, 2001; Ho and Yim, 2005), while the wave processes and the fluctuations of the wave height of certain time histories were simulated by Li and Kareem (1993). Pena-Sanchez and Ringwood (2017) made a critical comparison of the AR and ARMA models in short-term wave forecasting and concluded that the AR model can give equally satisfying results with the ARMA model, despite ARMA's complexity (see also Fusco and Ringwood, 2010). Nevertheless, in the work of Ming Ge and Kerrigan (2016), it was suggested that the AR model may not be the best option in case the data set exceeds a certain length, and that the ARMA model achieves better results. The modelling advantages of the ARMA model over the AR and MA models are described also by Makridakis et al. (1998). Three different algorithms for the AR, MA and ARMA models were presented by Spanos (1983), mainly focusing on the simulation of time series compatible with a given power spectrum of ocean waves. In the same work, the applicability of those algorithms in offshore engineering problems was also described. Finally, Martzikos and Soukissian (2017) tried various ARMA models, in order to model the sea surface elevation using data from the Greek Seas. After the best-fit model was found, a forecast of the free surface elevation was carried out, confirming the fair forecasting capabilities of the model in the estimation of forecast errors. For more details on the capabilities of an ARMA model in forecasting future values of an observed time series, the reader is referred to Chatfield (2000) and Montgomery et al. (2015).

Bibliography

Manouchehr Vaziri (1997) *Predicting Caspian Sea Surface Water Level by ANN and ARIMA Models*, Journal of Waterway, Port, Coastal, and Ocean Engineering, Vol. 123, No.4, July/August, 1997, ©ASCE, ISSN: 0733-950X/97/0004-01580162.

G.P. Zhang (2003), *Time series forecasting using hybrid ARIMA and neural network model*, Neurocomputing, Volume 50, (2003), pp. 159 – 175, PII: S0925-2312(01)00702-0, DOI: https://doi.org/10.1016/S0925-2312(01)00702-0.

M. Khashei, M. Bijari (2010) *An artificial neural network (p,d,q) model for timeseries forecasting*, Expert Systems with Applications, 37, (2010), 479–489, DOI: 10.1016/j.eswa.2009.05.044.

P. Delicado, A. Justel (1999) *Forecasting with Missing Data: Application to Coastal Wave Heights*, Journal of Forecasting, 18, (1999), 285-298, CCC: 0277-6693/99/040285-14.

Yousun Li, A. Kareem (1993) *Parametric modelling of stochastic wave effects on offshore platforms*, Applied Ocean Research, 15, (1993), 63-83, 0141-1187/93/$06 00 © 1993 Elsevier Science Publishers Ltd.

R.J. Rodney (1996) *Correlation between individual waves in real sea state,* Coastal Engineering, 27, (1996), 223-242, 0378-3839/96/$15.00, Published by Elsevier Science B.V., PII: S0378-3839(96)00010-10-5, DOI: https://doi.org/10.1016/0378-3839(96)00010-5.

N.T. Martzikos, T.H. Soukissian (2017) *Modelling of the sea surface elevation based on a data analysis in the Greek seas*, Applied Ocean Research, 69, (2017), 76–86, DOI: https://doi.org/10.1016/j.apor.2017.10.008.

Y. Pena Sanchez, J.V. Ringwood (2017) *A Critical Comparison of AR and ARMA Models for Short-term Wave Forecasting*, Proceedings of the 12th European Wave and Tidal Energy Conference 27th Aug -1st Sept 2017, Cork, Ireland, ISSN: 2309-1983.

Iman Malekmohamadi, Mohammad Reza Bazargan-Lari, Reza Kerachian, Mohammad Reza Nikoo, Mahsa Fallahnia (2011) *Evaluating the efficacy of SVMs, BNs, ANNs and ANFIS in wave height prediction*, Ocean Engineering, 38, (2011), 487–497, DOI: 10.1016/j.oceaneng.2010.11.020.

A. Altunkaynak, M. Ozger (2004) *Temporal significant wave height estimation from wind speed by perceptron Kalman filtering*, Ocean Engineering, 31, (2004), 1245–1255, DOI: 10.1016/j.oceaneng.2003.12.008.

S. Asma, A. Sezer, O. Ozdemir (2012) *MLR and ANN models of significant wave height on the west coast of India*, Computers & Geosciences, 49, (2012), 231–237, DOI: http://dx.doi.org/10.1016/j.cageo.2012.05.032.

G. Leontaris, O. Morales-Napoles, A.R.M. (Rogier) Wolfert (2016) *Probabilistic scheduling of offshore operations using copula based environmental time series – An application for cable installation management for offshore wind farms*, Ocean Engineering, 125, (2016), 328–341, DOI: http://dx.doi.org/10.1016/j.oceaneng.2016.08.029.

R. Jane, L. Dalla Valle, D. Simmonds, A. Raby (2016) *A copula-based approach for the estimation of wave height records through spatial correlation*, Coastal Engineering, 117, (2016), 1–18, DOI: http://dx.doi.org/10.1016/j.coastaleng.2016.06.008.

M.C. Deo, A. Jha, A.S. Chaphekar, K. Ravikant (2001) *Neural networks for wave forecasting, Ocean Engineering*, 28, (2001), 889–898, PII: S00 29 -8018(00)00027-5, DOI: https://doi.org/10.1016/S0029-8018(00)00027-5.

Ching-Piao Tsai, Chang Lin, Jia-N Shen (2002) *Neural network for wave forecasting among multi-stations*, Ocean Engineering, 29, (2002), 1683–1695, PII: S00 29 -8018(01)00112-3, DOI: https://doi.org/10.1016/S0029-8018(01)00112-3.

O. Makarynskyy (2004) *Improving wave predictions with artificial neural networks*, Ocean Engineering, 31, (2004), 709–724, DOI:10.1016/j.oceaneng.2003.05.003.

A. Malekmohamadi, R. Ghiassi, M.J. Yazdanpanah (2008) *Wave hindcasting by coupling numerical model and artificial neural networks*, Ocean Engineering, 35, (2008), 417–425, DOI: 10.1016/j.oceaneng.2007.09.003.

N. Krishna Kumar, R. Savitha, Abdullah Al Mamun (2017) *Regional ocean wave height prediction using sequential learning neural networks*, Ocean Engineering, 129, (2017), 605–612, DOI: http://dx.doi.org/10.1016/j.oceaneng.2016.10.033.

M.C. Deo, C. Sridhar Naidu (1999) *Real time wave forecasting using neural networks*, Ocean Engineering, 26, (1999), 191–203, PII: S0029-8018(97)10025-7, DOI: https://doi.org/10.1016/S0029-8018(97)10025-7.

O. Makarynskyy, A.A. Pires-Silva, D. Makarynska, C. Ventura-Soares (2005) *Artificial neural networks in wave predictions at the west coast of Portugal*, Computers & Geosciences, 31, (2005), 415–424, DOI: 10.1016/j.cageo.2004.10.005.

S.N. Londhe, Shalaka Shah, P.R. Dixit, T.M. Balakrishnan Nair, P. Sirisha, Rohit Jain (2016) A Coupled Numerical and Artificial Neural Network Model for Improving Location Specific Wave Forecast, Applied Ocean Research, 59, (2016), 483–491, DOI: http://dx.doi.org/10.1016/j.apor.2016.07.004.

J.D. Agrawal, M.C. Deo (2002) *On-line wave prediction*, Marine Structures, 15, (2002), 57–74, PII: S0951-8339(01)00014–4, DOI: https://doi.org/10.1016/S0951-8339(01)00014-4.

S. Mandal, Subba Rao, D.H. Raju (2005) *Ocean wave parameters estimation using backpropagation neural networks*, Marine Structures, 18, (2005), 301–318, DOI: 10.1016/j.marstruc.2005.09.002.

S.N. Lodhe, V. Panchang (2005) *One-day wave forecasts using buoy data and artificial neural networks*, Conference Paper, February 2005, DOI: 10.1109/OCEANS.2005.1640074, Source: IEEE Xplore.

Zhixu Zhang, Chi-Wai Li, Yiquan Qi and Yok-Sheung Li (2006) *Incorporation of artificial neural networks and data assimilation techniques into a third-generation wind–wave model for wave forecasting*, Journal of Hydroinformatics, 08.1, (2006), DOI: 10.2166/jh.2006.005.

Aditya N. Deshmukh, M. C. Deo, Prasad K. Bhaskaran, T. M. Balakrishnan Nair, and K. G. Sandhya (2016) *Neural-Network-Based Data Assimilation to Improve Numerical Ocean Wave Forecast*, IEEE Journal of Oceanic Engineering, Vol. 41, No. 4, (2016), DOI: 10.1109/JOE.2016.2521222.

O. Makarynskyy (2007) *Artificial neural networks in merging wind wave forecasts with field observations*, Indian Journal of Marine Sciences, Vol. 36(1), (2007), pp. 7-17, IPC Code: Int. CI. (2006) G06F 7/20; G06Q 99/00.

S. N. Londhe, Vihay Panchang (2006) *One-Day Wave Forecasts Based on Artificial Neural Networks*, Journal of Atmospheric and Oceanic Technology, Vol. 23, 1593-1603, © 2006 American Meteorological Society, DOI: https://doi.org/10.1175/JTECH1932.1.

O. Makarynskyy (2005) *Neural pattern recognition and prediction for wind wave data assimilation*, Pacific Oceanography, Vol. 3, No. 2, (2006), https://www.researchgate.net/publication/288867692.

Elzbieta M. Bitner-Gregersen, Oistein Hagen (1990) *Uncertainties in Data for the Offshore Environment*, Structural Safety, 7, (1990), 11-34, 0167-4730/90/$03.50, DOI: https://doi.org/10.1016/0167-4730(90)90010-M.

S. Haver, T. Moan (1983) *On some uncertainties related to the short term stochastic modelling of ocean waves*, Applied Ocean Research, 1983, Vol. 5, No. 2, 93-108, 0309-1708/83/020093-16 $2.00, DOI: https://doi.org/10.1016/0141-1187(83)90021-4.

E.M. Bitner-Gregersen, S.K. Bhattacharya, I.K. Chatjigeorgiou, I. Eames, K. Ellermann, K. Ewans, G. Hermanski, M.C. Johnson, N. Ma, C. Maisondieu, A. Nilva, I. Rychlik, T. Waseda (2014), *Recent developments of ocean environmental description with focus on uncertainties*, Ocean Eng., 86, (2014), pp. 26-46, DOI: http://dx.doi.org/10.1016/j.oceaneng.2014.03.002.

E.M. Bitner-Gregersen, K. Ewans, M.C. Johnson (2014) *Some uncertainties associated with wind and wave description and their importance for engineering applications*, Ocean Engineering, 86, (2014), 11–25, DOI: http://dx.doi.org/10.1016/j.oceaneng.2014.05.002.

F. Fusco and J. Ringwood (2010) *Short-Term Wave Forecasting for Real-Time Control of Wave Energy Converters*, IEEE Transactions on Sustainable Energy, vol. 1, no. 2, pp. 99–106, 2010, DOI: 10.1109/ISIE.2010.5637714.

Ming Ge, E. C. Kerrigan (2016) *Short-term Ocean Wave Forecasting Using an Autoregressive Moving Average Model*, 2016 UKACC 11th International Conference on Control (CONTROL), Electronic ISBN: 978-1-4673-9891-6, DOI: 10.1109/CONTROL.2016.7737594.

R.J. Sobey (1996) *Correlation between individual waves in a real sea state*, Coast. Eng., 27, (3-4), (1996) 223–242, http://dx.doi.org/10.1016/0378-3839(96)00010-5.

C. Guedes Soares, A.M. Ferreira (1996) *Representation of non-stationary time series of significant wave height with autoregressive models*, Probabilistic Eng.Mech., 11(3), (1996), 139–148, http://dx.doi.org/10.1016/0266-8920(96)00004-5.

C.N. Stefanakos, G.A. Athanassoulis (2001) *A unified methodology for the analysis,completion and simulation of nonstationary time series with missing values with application to wave data*, Appl. Ocean Res., 23 (4), (2001), 207–220, http://dx.doi.org/10.1016/S0141-1187(01)00017-7.

Y. Li, A. Kareem (1993) *Parametric modelling of stochastic wave effects on offshore platforms*, Appl. Ocean Res., 15 (2), (1993), 63–83, http://dx.doi.org/10.1016/0141-1187(93)90022-P.

P.D. Spanos (1983) *ARMA algorithms for ocean wave modeling*, J. Energy Resour.Technol., 105, (1983) 300–309, http://dx.doi.org/10.1115/1.3230919.

P.C. Ho, J.Z. Yim (2005) *A study of the data transferability between twowave-measuring stations*, Coast. Eng., 52 (4), (2005), 313–329, http://dx.doi.org/10.1016/j.coastaleng.2004.12.003.

G.E.P. Box, G.M. Jenkins, G.C. Reinsel (1994) *Time Series Analysis, Forecasting and Control*, third ed., Prentice-Hall, Englewood Cliffs, N.J, 1994, ISBN: 978-1-118-67502-1.

C. Chatfield (2000) *Time-Series Forecasting*, Chapman & Hall, Boca Raton, 2000, ISBN-13: 978-1584880639, ISBN-10: 1584880635.

D.C. Montgomery, C.L. Jennings, M. Kulahci (2015) *Introduction to Time Series Analysis and Forecasting*, Wiley, Hoboken, New Jersey, 2015, ISBN: 978-0-471-65397-4.

S. Makridakis, S.C. Wheelwright, R.J. Hyndman (1998) *Forecasting: Methods and Applications*, John Wiley & Sons, New York, 1998, ISBN: 978-0-471-53233-0.

Box, G. E., Jenkins, G. M. (1976). *Time series analysis: forecasting and control*. Holden-Day, Oakland, Calif., ISBN: 0816211043.

C. Chatfield (1996) *Model uncertainty and forecast accuracy*, Journal of Forecasting, Volume 15, pp. 495–508, DOI: https://doi.org/10.1002/(SICI)1099-131X(199612)15:7<495::AID-FOR640>3.0.CO;2-O.

G.M. Jenkins (1982) *Some practical aspects of forecasting in organisations*, J. Forecasting, 1, (1982), 3–21, DOI: https://doi.org/10.1002/for.3980010103.

S. Makridakis, A. Anderson, R. Carbone, R. Fildes, M. Hibdon, R. Lewandowski, J. Newton, E. Parzen, R. Winkler (1982) *The accuracy of extrapolation (time series) methods: results of a forecasting competition*, J. Forecasting, 1, (1982), 111–153, DOI: DOI: 10.1002/for.3980010202.

J.K Wu(1994) *Neural Network and Simulation Methods*, Marcel Dekker, New York, ISBN: 0824791819.

B. Kosko (1992) *Neural Networks and Fuzzy Systems*, Prentice Hall, Englewood Cliffs, NJ, ISBN: 0-13-611435-0.

W.S McCulloch, W. Pitts (1943) *A logical calculus of the ideas immanent in nervous activity*, Bulletin of Mathematical Biophysics, 5, 115–133, DOI: https://doi.org/10.1007/BF02478259.

D.E. Rumelhart, G.E. Hinton, R.J. Williams (1986) *Learning representations by back-propagating errors*, Nature, 323, 533–536, DOI: https://doi.org/10.1038/323533a0.

C. Peterson, J.R. Anderson (1987) *A mean field learning algorithm for neural networks*, Complex Systems, 1, 995–1019, DOI: https://doi.org/10.1007/978-94-011-5014-9_20.

D. Psaltis, A. Sideris, A.A. Yamamura (1988) *A multilayered neural network controller*, IEEE Control System Magazine, 8, 17–21, DOI: 10.1109/37.1868.

K. Thirumalaiah, M.C. Deo (1998) *River stage forecasting using artificial neural networks,* ASCE Journal of Hydrologic Engineering, Vol. 3 (1), pp. 26–32, DOI: https://doi.org/10.1061/(ASCE)1084-0699(1998)3:1(26).

K. Thirumalaiah, M.C. Deo (2000) *Some studies on hydrological forecasting using neural networks*, ASCE Journal of Hydrologic Engineering, 5 (2), 180–189, DOI: 10.1061/(ASCE)1084-0699(2000)5:2(180).

J. Kasperkiewiecz, J. Racz, A. Dubrawski (1995) *HPC strength prediction using artificial neural networks*, ASCE Journal of Computing in Civil Engineering, 9 (4), 279–284, DOI: https://doi.org/10.1061/(ASCE)0887-3801(1995)9:4(279).

J.P. Grubert (1995) *Prediction of estuarine instabilities with artificial neural networks*, ASCE Journal of Computing in Civil Engineering, 9 (4), 266–274, DOI: https://doi.org/10.1061/(ASCE)0887-3801(1995)9:4(266).

M.N. French, W.F. Krajewski, R.R. Cuykendall (1992) *Rainfall forecasting in space and time using neural networks*, Journal of Hydrology 137, 1–31, DOI: https://doi.org/10.1016/0022-1694(92)90046-X.

M.C. Deo, N. Kiran Kumar (1999) *Interpolation of wave heights*, Ocean Engineering, 27 (9), 907–919, DOI: https://doi.org/10.1016/S0029-8018(99)00023-2.

Y.C. Yeh, Y.H. Kuo, D.S. Hsu (1993) *Building KBES for diagnosing PC piles with artificial neural networks*, ASCE Journal of Computing in Civil Engineering, 7 (1), 71–93, DOI: 10.1061/(ASCE)0887-3801(1993)7:1(71).

J.P. Fitch, S.K. Lehman, F.U. Dowla, S.K. Lu, E.M. Johansson, D.M. Goodman (1991) *Ship wake detection procedure using conjugate gradient trained artificial neural network*, IEEE Transactions on Geosciences and Remote Sensing, 9 (5), 718–725, DOI: 10.1109/36.83986.

R. Fletcher, C.M. Reeves (1964) *Function minimization by conjugate gradients*, Computer Journal, 149–153, DOI: https://doi.org/10.1093/comjnl/7.2.149.

S.E. Fahlman, C. Lebiere (1990) *The cascade corrlation training architecture*, In Advances in Neural Engineering Processing Systems, vol. 2. Morgan Kaufmann, San Mateo, CA, URL: https://papers.nips.cc/paper/207-the-cascade-correlation-learning-architecture.pdf.

A.T.C. Goh (1995) *Evaluation of seismic liquefaction using neural networks*, In: Topping, B.H.V. (Ed.), Developments in Neural Networks and Evolutionary Computing for Civil and Structural Engineering, Civil-Comp Press, Edinburgh, pp. 121–126, DOI: https://doi.org/10.1007/s11803-007-0766-7.

C.P. Tsai, T.L. Lee (1999) *Back-propagation neural network in tidal-level forecasting*, Journal of Waterway, Port, Coastal and Ocean Engineering, ASCE, 125, 195–202, DOI: https://doi.org/10.1061/(ASCE)0733-950X(1999)125:4(195).

G. Yagawa, H. Okuda (1996) *Finite element solutions with feedback network mechanism through direct minimization of energy functionals*, International Journal for Numerical Methods in Engineering, 39, 867–883, DOI: https://doi.org/10.1002/(SICI)1097-0207(19960315)39:5<867::AID-NME886>3.0.CO;2-Q.

T. Kerh, Y.C. Yee (2000) *Analysis of a deformed three-dimensional culvert structure using neural networks*, Advances in Engineering Software, 31, 367–375, DOI: https://doi.org/10.1016/S0965-9978(99)00058-7.

N.K. Bose, P. Liang (1998) *Neural Network Fundamentals with Graphs, Algorithms and Applications*, Tata McGraw-Hill Publication, ISBN-10: 0070066183.

C.W. Dawson, R.L. Wilby (2001) *Hydrological modeling using artificial neural networks*, Prog. Phys. Geogr., 25 (1) (2001), 80–108, DOI: https://doi.org/10.1177/030913330102500104.

H.R. Maier, G.C. Dandy (2000) *Neural networks for prediction and forecasting of water resources variables: a review of modelling issues and applications*, Elsevier Environ. Modell. Software, 15 (2000), 101–124, DOI: https://doi.org/10.1016/S1364-8152(99)00007-9.

P.D. Wasserman (1993) *Advanced Methods in Neural Computing*, Van NostrandReinhold, New York, 1993, pp. 255 (www.mikepoweredbydhi.com/products/mike-21) www.incois.gov.in.

The ASCE Task Committee (2000) *Artificial neural networks in hydrology I: preliminary concepts*, J. Hydrol. Eng., 5 (2) (2000), 115–123, DOI: https://doi.org/10.1061/(ASCE)1084-0699(2000)5:2(115).

Joe, H. (1997) *Multivariate models and dependence concepts*, Chapman and Hall, London, ISBN: 978-0-412-07331-1.

Nelsen, R. B. (1999) *An introduction to copulas*, Springer, New York, ISBN: 978-0-387-28678-5.

Nelsen, R.B. (2003) *Properties and applications of copulas: a brief survey*, In: Proceedings of the First Brazilian Conference on Statistical Modelling in Insurance and finance, University of São Paulo, URL: http://w4.stern.nyu.edu/ioms/docs/sg/seminars/nelsen.pdf.

Joe, H. (2014) *Dependence Modeling with Copulas*, CRCPress, Taylor & Francis Group, Vancouver, Canada, ISBN: 9781466583221.

P. Embrechts, F. Lindskog, A. McNeil (2003) *Modelling dependence with copulas and applications to risk management*, In: Rachev, S.T.(Ed.), Handbook of Heavy Tailed distributions in Finance, Elsevier/North Holland), Amsterdam, pp. 329–384, DOI: https://doi.org/10.1016/B978-044450896-6.50010-8.

G. Salvadori, G.R. Tomasicchio, F. D′Alessandro (2013) *Multivariate approach to design coastal and off-shore structures*, In: Conley, D.C., Masselink, G., Russell, P. E. and O′Hare, T.J. (eds.). Proceedings 12th International Coastal Symposium (Plymouth, England), Journal of Coastal Research, Special Issue No. 65, pp. 386 – 391, DOI: http://www.jcronline.org/doi/pdf/10.2112/SI65-066.1?code=cerf-site.

X. Yang, Q. Zhang (2013) *Joint probability distribution of winds and waves from wave simulation of 20 years (1989–2008) in Bohai Bay*, Water Sci. Eng., 6(3), 296–307, DOI: https://doi.org/10.3882/j.issn.1674-2370.2013.03.006.

H.-B. Fang, K.-T. Fang, S. Kotz (2002) *The meta-elliptical distributions with given marginals*, Journal of Multivariate Analysis, 82, 1–16, DOI: https://doi.org/10.1006/jmva.2001.2017.

H.-B. Fang, K.-T. Fang, S. Kotz (2005) *Corrigendum to: “The meta-elliptical distributions with given marginals” [J. Multivariate Anal. 82: 1–16 (2002)]*, Journal of Multivariate Analysis, 94, 222–223, DOI: 10.1016/j.jmva.2004.10.001.

H. Kazianka (2012) *spatial Copula: a Matlab toolbox for copula-based spatial analysis*, Stoch. Environ. Res. Risk Assess, DOI: 10.1007/ s00477-012-0571-3 (in press).

L. Poelhekke, W. S. Jäger, A. van Dongere, T. A. Plomaritis, R. McCall, Ó. Ferreira (2016) *Predicting coastal hazards for sandy coasts with a Bayesian Network*, Coastal Engineering, 118, (2016), 21–34, DOI: http://dx.doi.org/10.1016/j.coastaleng.2016.08.011.

O. Morales-Nápoles, R.D.J.M. Steenbergen (2014) *Analysis of axle and vehicle load properties through Bayesian Networks based on Weigh-in-Motion data*, Reliability Engineering and System Safety, 125, (2014), 153–164, DOI: http://dx.doi.org/10.1016/j.ress.2014.01.018.

P. Weber, G. Medina-Oliva, C. Simon, B. Iung (2012) *Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas*, Engineering

Applications of Artificial Intelligence, 25, (2012), 671–682, DOI: 10.1016/j.engappai.2010.06.002.

W.S. Jager, E.K. Christie, A.M. Hanea, C. den Heijer, T. Spencer (2017) *A Bayesian network approach for coastal risk analysis and decision making, Coastal Engineering*, 2017, 1–14, DOI: http://dx.doi.org/10.1016/j.coastaleng.2017.05.004 (in press).

A. Hanea, O. Morales-Napoles, D. Ababei (2015) *Non-parametric Bayesian networks: Improving theory and reviewing applications*, Reliability Engineering and System Safety, 144, (2015), 265–284, DOI: http://dx.doi.org/10.1016/j.ress.2015.07.027.

G. Medina Oliva, P. Weber, C. Simon, B. Iung (2009) *Bayesian networks Applications on Dependability, Risk Analysis and Maintenance*, 2nd IFAC Workshop on Dependable Control of Discrete Systems DCDS'09, Bari, Italy, June 10-12, 2009, DOI: 10.3182/20090610-3-IT-4004.0040.

A. Kroon, M. de Schipper, K. den Heijer, S. Aarninkhof, P. van Gelder (2017) *Uncertainty assessment in coastal morphology prediction with a bayesian network*, In T. Aagaard, R. Deigaard, & D. Fuhrman (Eds.), Proceedings of Coastal Dynamics 2017: Helsingør, Denmark, pp. 1909-1920, Paper No. 254, URL: https://repository.tudelft.nl/islandora/object/uuid%3A424c4f19-4bd2-48cd-9221-b6e3e8426636.

K. P. Murphy (2001) *The Bayes Net Toolbox for Matlab*, Department of Computer Science, University of California, Berkeley, Berkeley, California, 94720-1776, URL: https://www.cs.ubc.ca/~murphyk/Papers/bnt.pdf.

C. K. Wikle, L. M. Berliner (2007) *A Bayesian tutorial for data assimilation*, Physica D, 230, (2007), 1–16, DOI: 10.1016/j.physd.2006.09.017.

M. A. Parrish, H. Moradkhani, C. M. DeChant (2012) *Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation*, Water Resources Research, Vol. 48, W03519, DOI: 10.1029/2011WR011116.

N. G. Plant, K. T. Holland (2011) *Prediction and assimilation of surf-zone processes using a Bayesian network Part I: Forward models, Coastal Engineering*, 58, (2011), 119–130, DOI: 10.1016/j.coastaleng.2010.09.003.

N. G. Plant, K. T. Holland (2011) *Prediction and assimilation of surf-zone processes using a Bayesian network Part II: Inverse models*, Coastal Engineering, 58, (2011), 256–266, DOI: 10.1016/j.coastaleng.2010.11.002.

K. E.Wilson, P. N. Adams, C. J. Hapke, E. E. Lentz, O. Brenner (2015) *Application of Bayesian Networks to hindcast barrier island morphodynamics*, Coastal Engineering, 102, (2015), 30–43, DOI: http://dx.doi.org/10.1016/j.coastaleng.2015.04.006.

P.A. Aguilera, A. Fernández, R. Fernández, R. Rumí, A. Salmerón (2011) *Bayesian networks in environmental modelling*, Environmental Modelling & Software, 26, (2011), 1376-1388, DOI: 10.1016/j.envsoft.2011.06.004.

S. H. Chen, C. A. Pollino (2012) *Good practice in Bayesian network modelling*, Environmental Modelling & Software, 37, (2012), 134-145, DOI: 10.1016/j.envsoft.2012.03.012.

G. Pineiro, S. Perelman, J. P. Guerschman, J. M. Paruelo (2008) *How to evaluate models: Observed vs. predicted or predicted vs. observed?*, Ecological modelling, 216, (2008), 316–322, DOI: 10.1016/j.ecolmodel.2008.05.006.

T. Soukissian, C. Kechris (2007) *About applying linear structural method on ocean data: Adjustment of satellite wave data*, Ocean Engineering, 34, (2007), 371–389, DOI: 10.1016/j.oceaneng.2006.04.002.

J.O. Rawlings, S.G. Pantula, D.D. Dickey (1998) *Applied Regression Analysis, A Research Tool*, 2nd edition, Springer, Amsterdam, ISBN: 0-387-98454-2.

R. Webster (1997) *Regression and functional relations*, European Journal of Soil Science, 48, 557–566, DOI: https://doi.org/10.1111/j.1365-2389.1997.tb00222.x.

L. Rusu, C. Guedes Soares (2016) *Comparison of various data assimilation methods to improve the wave predictions in the Portuguese coastal environment*, Maritime Technology and Engineering 3 – Guedes Soares & Santos (Eds), © 2016 Taylor & Francis Group, London, ISBN: 978-1-138-03000-8.

A. Raileanu, L. Rusu, E. Rusu (2015) *Wave modelling with data assimilation in the Romanian nearshore*, Towards Green Marine Technology and Transport – Guedes Soares, Dejhalla & Pavleti (Eds), © 2015 Taylor & Francis Group, London, ISBN: 978-1-138-02887-6.

S. Londhe, V. Panchang (2005) *One-day wave forecasts using buoy data and artificial neural networks*, OCEANS, 2005, Proceedings of MTS/IEEE, Washington, DC, USA, ISBN: 0-933957-34-3, DOI: 10.1109/OCEANS.2005.1640074.

V. M. Krasnopolsky, D. V. Chalikov, H. L. Tolman (2002) *A neural network technique to improve computational efficiency of numerical oceanic models*, Ocean Modelling, 4, (2002), 363–383, PII: S1463-5003(02)00010-0, DOI: https://doi.org/10.1016/S1463-5003(02)00010-0.

P. Jain, M.C. Deo (2007) *Real-time wave forecasts off the western Indian coast*, Applied Ocean Research, 29, (2007), 72–79, DOI: 10.1016/j.apor.2007.05.003.

S.N. Londhe (2008) *Soft computing approach for real-time estimation of missing wave heights*, Ocean Engineering, 35, (2008), 1080– 1089, DOI: 10.1016/j.oceaneng.2008.05.003.

P. Jain, M.C. Deo, G. Latha, V. Rajendran (2011) *Real time wave forecasting using wind time history and numerical model*, Ocean Modelling, 36, (2011), 26–39, DOI: 10.1016/j.ocemod.2010.07.006.

C. Genest, A.C. Favre (2007) *Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask*, Journal of Hydrologic Engineering, Vol. 12, Issue 4, 347-368, DOI: https://doi.org/10.1061/(ASCE)1084-0699(2007)12:4(347).

P. Embrechts, F. Lindskog, A. McNeil (2001) *Modelling Dependence with Copulas and Applications to Risk Management, Department of Mathematics*, ETHZ, CH-8092 Zurich, Switzerland, https://people.math.ethz.ch/~embrecht/ftp/copchapter.pdf.

R. B. Nelsen (2006) *An Introduction to Copulas*, 2nd Edition, Springer Series in Statistics, Springer-Verlag New York, 2006, ISBN: 978-0-387-28678-5, DOI: 10.1007/0-387-28678-0.

M.G. Scotto, C. Guedes Soares (2007) *Bayesian inference for long-term prediction of significant wave height*, Coastal Engineering, 54, 393-400, DOI: https://doi.org/10.1016/j.coastaleng.2006.11.003.

A.G. Sebastian, E.J.C. Dupuits, O. Morales-Napoles (2017) *Applying a Bayesian network based on Gaussian copulas to model the hydraulic boundary conditions for hurricane flood risk analysis in a coastal watershed*, Coastal Engineering, volume 125, 42-50, DOI: https://doi.org/10.1016/j.coastaleng.2017.03.008.

R. H. Shumway, D. S. Stoffer (2011) *Time series analysis and its applications: With R examples*, 4th edition, Springer Texts in Statistics, Springer International Publishing, New York: Springer, ISBN: 978-3-319-52452-8, DOI: 10.1007/978-3-319-52452-8.

J.L. Torres, A. Garcia, M. De Blas, A. De Francisco (2005) *Forecast of hourly average wind speed with ARMA models in Navarre (Spain)*, Solar Energy, 79, (2005), 65–77, DOI: 10.1016/j.solener.2004.09.013.

L. Kamal, Y. Zahra Jafri (1997) Time *series models to simulate and forecast hourly averaged wind speed in Quetta, Pakistan*, Solar Energy, Vol. 61, 1, 23-32, DOI: https://doi.org/10.1016/S0038-092X(97)00037-6.

E. Erdem, J. Shi (2011) *ARMA Based Approaches for Forecasting the Tuple of Wind Speed and Direction*, Applied Energy, 88, 1405-1414, DOI: http://dx.doi.org/10.1016/j.apenergy.2010.10.031.

P. Chen, T. Pedersen, B. Bak-Jensen, Z. Chen (2010) *ARIMA-Based Time Series Model of Stochastic Wind Power Generation*, IEEE Transactions on Power Systems, Vol. 25, Issue 2, May 2010, DOI: 10.1109/TPWRS.2009.2033277.

R.G. Kavasseri, K. Seetharaman (2009) *Day-ahead wind speed forecasting using f-ARIMA models*, Renewable Energy, 34, 5, pp. 1388-1393, DOI: 10.1016/j.renene.2008.09.006.

E Cadenas, W Rivera (2010) *Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA–ANN model*, Renewable Energy, 35, 12, 2732-2738, DOI: https://doi.org/10.1016/j.renene.2010.04.022.

M. Lei, L. Shiyan, J. Chuanwen, L. Hongling, Z. Yan (2009) *A review on the forecasting of wind speed and generated power*, Renewable and Sustainable Energy Reviews, vol. 13, 4, pp. 915-920, May 2009, DOI: https://doi.org/10.1016/j.rser.2008.02.002.

H. Langseth, T. Nielsen, R. Rumí, A. Salmerón (2009) *Inference in hybrid Bayesian networks*, Reliability Engineering & System Safety, 94, 4, 1499-1509, DOI: https://doi.org/10.1016/j.ress.2009.02.027.

H. Langseth, T. Nielsen, R. Rumi, A. Salmeron (2012) *Mixtures of truncated basis functions*, International Journal of Approximate Reasoning, 53, 2, 212–227, DOI: https://doi.org/10.1016/j.ijar.2011.10.004.

S.L. Lauritzen (1992) *Propagation of probabilities, means and variances in mixed graphical association models*, Journal of the American Statistical Association, 87, 420, 1098–1108, DOI: 10.1080/01621459.1992.10476265.

J. Pearl (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann Series in Representation and Reasoning, Morgan Kaufmann Publishers, 1st edition, ISBN: 978-0-08-051489-5.

A. Hanea, D. Kurowicka, R. Cooke (2006) H*ybrid method for quantifying and analyzing Bayesian belief nets*, Quality and Reliability Engineering International, 22, 6, 613–729, DOI: 10.1002/qre.808.

A. Hanea, D. Kurowicka, R. Cooke, D. Ababei (2010) *Mining and visualizing ordinal data with non-parametric continuous BBNs*, Computational Statistics & Data Analysis, 54, 3, 668–87, DOI: https://doi.org/10.1016/j.csda.2008.09.032.

D. Kurowicka, R. Cooke (2004) *Distribution-free continuous Bayesian belief nets*, In: Proceedings mathematical methods in reliability conference, DOI: 10.1142/9789812703378_0022.

K. Murphy (2002) *An introduction to graphical models*, Technical Report, Intel Research, 2002.

A. Kozlov, D. Koller (1997) *Non uniform dynamic discretization in hybrid networks*, In: Proceedings of the13th conference on uncertainty in artificial intelligence, 1997, UAI-P-1997-PG-314-325, arXiv: 1302.1555 [cs.AI].

U. Lerner (2002) *Hybrid Bayesian networks for reasoning about complex systems*, Ph.D. thesis, Computer Science Department, Stanford University, 2002, 279 pages.

R. Shachter, C. Kenley (1989) *Gaussian influence diagrams*, Management Science, 35, 5, 527–550, DOI: https://doi.org/10.1287/mnsc.35.5.527.

Philips R. (1998) *Multi scale modelling in the mechanics of materials*, Current Opinion in Solid State and Materials Science, 3, 6, 526–532, DOI: https://doi.org/10.1016/S1359-0286(98)80020-X.

N. Zhang, D. Poole (1994) *A simple approach to Bayesian network computations,* In: Proceedings of the tenth Canadian conference on artificial intelligence, 1994, p. 171-178, http://hdl.handle.net/1783.1/757.

D. Worm, O. Morales-Nápoles, Wvd. Haak, T. Bakri (2011) *Continuous dynamic non-parametric Bayesian networks: application to traffic prediction*, TNO-report, project number: 043.01000;2011.

D Straub, A. der Kiureghian (2010) *Bayesian Network Enhanced with Structural Reliability Methods: Methodology*, Journal of Engineering Mechanics, 136, 10, 1248–58, DOI: https://doi.org/10.1061/(ASCE)EM.1943-7889.0000173.

P.P. Shenoy, J.C. West (2011) *Inference in hybrid Bayesian networks using mixtures of polynomials*, International Journal of Approximate Reasoning, 52, 5, 641-657, DOI: https://doi.org/10.1016/j.ijar.2010.09.003.

Cooke R. (1991) *Experts in uncertainty: opinion and subjective probability in science*, Environmental ethics and science policy series, Oxford University Press, 1991, 336 pages, ISBN: 0195362373.

O. Morales-Napoles, D. Worm, P. van den Haak, A. Hanea, W. Courage, S. Miraglia (2013) *Reader for course: Introduction to Bayesian Networks*, TNO report, reference: TNO-060-DTM-2013-01115.

J. Whittaker (2009) *Graphical models in applied multivariate statistics*, Wiley Publishing, 462 pages, ISBN: 0470743662.

F.V. Jensen, T.D. Nielsen (2007) *Bayesian Networks and Decision Graphs*, Information Science and Statistics , Springer Verlag, New York, 448 pages, ISBN: 978-0-387-68282-2, DOI: 10.1007/978-0-387-68282-2.

O. Morales, D. Kurowicka, A. Roelen (2008) *Eliciting conditional and unconditional rank correlations from conditional probabilities*, Reliability Engineering & System Safety, 93, 5, 699–710 DOI: https://doi.org/10.1016/j.ress.2007.03.020.

F.V. Jensen (1996) *Introduction to Bayesian Networks*, 1st Edition, Springer-Verlag, New York, 186 pages, ISBN: 0387915028.

B. Stewart-Koster, S.E. Bunn, S.J. Mackay, N.L. Poff, R.J. Naiman, P.S. Lake (2010) *The use of Bayesian networks to guide investments in flow and catchment restoration for impaired river ecosystems*, Freshwater Biology, 55, 1, 243-260, DOI: https://doi.org/10.1111/j.1365-2427.2009.02219.x.

M.L. Palmsten, K.T. Holland, N.G. Plant (2013) *Velocity estimation using a Bayesian network in a critical‐habitat reach of the Kootenai River, Idaho*, Water Resources Research, 49, 9, 5865-5879, DOI: https://doi.org/10.1002/wrcr.20361.

M.L. Palmsten , K.D. Splinter , N.G. Plant, H.F. Stockdon (2014) *Probabilistic estimation of dune retreat on the Gold Coast, Australia*, Shore & Beach, 82, 4, 35-43, URL: https://pubs.er.usgs.gov/publication/70159345.

J. Kolibal, D. Howard (2006) *MALDI-TOF Baseline Drift Removal Using Stochastic Bernstein Approximation*, EURASIP Journal on Applied Signal Processing, Vol. 2006, Article ID 63582, Pages 1–9, DOI: 10.1155/ASP/2006/63582.

J. Kolibal, D. Howard (2006) *The Novel Stochastic Bernstein Method of Functional Approximation*, First NASA/ESA Conference on Adaptive Hardware and Systems (AHS'06), 15-18 June 2006, Istanbul, Turkey, ISBN: 0-7695-2614-4, DOI: 10.1109/AHS.2006.73.

J. Kolibal, D. Howard (2008) *Alternative Parametric Boundary Reconstruction Method for Biomedical Imaging*, Journal of Biomedicine and Biotechnology, Vol. 2008, Article ID 623475, 7 pages, DOI:10.1155/2008/623475.

R. Seyfarth, J. Kolibal, D. Howard (2006) *New Mathematical Method for Computer Graphics*, In: 2006 International Conference on Hybrid Information Technology, 9-11 Nov. 2006, Cheju Island, South Korea, ISBN: 0-7695-2674-8, DOI: 10.1109/ICHIT.2006.253457.

M. Kim, H. Ko (2011) *Resolution enhancement of ROI from surveillance video using Bernstein interpolation*, 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 30 Aug.-2 Sept. 2011, Klagenfurt, Austria, ISBN: 978-1-4577-0845-9, DOI: 10.1109/AVSS.2011.6027346.

C. E. Rasmussen, C. K. I. Williams (2006) *Gaussian Processes for Machine Learning*, MIT Press, 2006, ISBN: 026218253X, URL: http://www.gaussianprocess.org/gpml/.

A. Sklar (1959) *Fonctions de Répartition à n Dimensions et Leurs Marges*, Publications del′ Institut statistique del′ Université de Paris, 8, pp. 229–231.

H. Joe (2014) *Dependence Modeling with Copulas*, CRC Press, Taylor & Francis Group, Vancouver, Canada, 480 pages, ISBN: 9781466583221.

C. Genest, B. Remillard, D. Beaudoin (2009) *Goodness-of-fit tests for copulas: a review and a power study*, Insurance: Mathematics and Economics, 44, 2, 199–213, DOI: https://doi.org/10.1016/j.insmatheco.2007.10.005.

H. Joe (1997) *Multivariate models and dependence concepts*, Monographs on Statistics & Applied Probability, Chapman & Hall, London, 1997, ISBN: 9780412073311.

K. Aas, C. Czado, A. Frigessi, H. Bakken (2009) *Pair-copula constructions of multiple dependence*, Insurance: Mathematics and Economics, 44, 2, 182–198, DOI: https://doi.org/10.1016/j.insmatheco.2007.02.001.

B. Remillard (2010) *Goodness-of-fit test for copulas of multivariate timeseries*, SSRN Electronic Journals, 32 pages, DOI: http://dx.doi.org/10.2139/ssrn.1729982.

E.C. Brechmann, C. Czado (2015) *COPAR-multivariate time series modelling using the copula autoregressive model*, Applied Stochastic Models in Business and Industry, 31, 4, 495–514, DOI: https://doi.org/10.1002/asmb.2043.

M.S. Smith (2015) *Copula modelling of dependence in multivariate time series*, International Journal of Forecasting, 31, 3, 815–833, DOI: https://doi.org/10.1016/j.ijforecast.2014.04.003.

D. Kurowicka, R. Cooke (2006) *Uncertainty Analysis with High Dimensional Dependence Modelling*, John Wiley & Sons, Ltd, Chichester, UK, 302 pages, ISBN: 978-0-470-86306-0.

G. Salvadori, G.R. Tomasicchio, F. D′Alessandro (2013) *Multivariate approach to design coastal and offshore structures*, In: Conley, D.C., Masselink, G., Russell, P. E. and O′Hare, T.J. (eds.), Proceedings 12th International Coastal Symposium, Plymouth, England, Journal of Coastal Research, Special Issue No.65, pp. 386-391, DOI: 10.13140/2.1.4954.5604.

E. Vanem (2016) *Joint statistical models for significant wave height and wave period in a changing climate*, Marine Structures, 49, (2016), 180-205, DOI: http://dx.doi.org/10.1016/j.marstruc.2016.06.001.

T. Bedford, R.M. Cooke (2001) *Probability density decomposition for conditionally dependent random variables modeled by vines*, Annals of Mathematics and Artificial Intelligence, 32, 1-4, 245–268, DOI: https://doi.org/10.1023/A:101672590.

T. Bedford, R.M. Cooke (2002) *Vines — a new graphical model for dependent random variables*, Annals of Statistics, 30, 4, 1031–1068, DOI: 10.1214/aos/1031689016.

H. Joe (1996) *Families of m-variate distributions with given margins and m(m−1)/2 bivariate dependence parameters*, Lecture Notes Monograph Series, vol. 28, Institute of Mathematical Statistics, Hayward, CA, pp. 120–141, DOI: 10.1214/lnms/1215452614, URL: https://projecteuclid.org/euclid.lnms/1215452614.

M. Ma, S. Song, L. Ren, S. Jiang, J. Song (2013) *Multivariate drought characteristics using trivariate Gaussian and Student t copulas*, Hydrological Processes, 27, 8, 1175–1190, DOI: https://doi.org/10.1002/hyp.8432.

X. Wang, M. Gebremichael, J. Yan (2010) *Weighted likelihood copula modeling of extreme rainfall events in Connecticut*, Journal of Hydrology, 390, 1–2, 108–115, DOI: https://doi.org/10.1016/j.jhydrol.2010.06.039.

G. Wong, M.F. Lambert, M. Leonard, A.V. Metcalfe (2010) *Drought analysis using trivariate copulas conditional on climatic states*, Journal of Hydrologic Engineering, 15, 2, 129–141, DOI: https://doi.org/10.1061/(ASCE)HE.1943-5584.0000169.

G. Poulomi, M. Reddy (2013) *Probabilistic assessment of flood risks using trivariate Copulas*, Theoretical and Applied Climatology, 111, 1-2, 341–360, DOI: https://doi.org/10.1007/s00704-012-0664-4.

P. Deheuvels (1979) *La fonction de dépendance empirique et ses propriétés: Un test non paramétrique d'indépendance*, Bulletin de la classe des sciences, Acad. Royale de Belgique, 65, 6, 274–292.

R. de Melo e Silva Accioly, F.Y. Chiyoshi (2004) *Modeling dependence with copulas: a useful tool for field development decision process*, Journal of Petroleum Science and Engineering, 44, 2004, 83– 91, DOI: 10.1016/j.petrol.2004.02.007.

B. Vaz de Melo Mendes, R. Martins de Souza (2004) *Measuring financial risks with copulas*, International Review of Financial Analysis, 13, 2004, 27– 45, DOI: 10.1016/j.irfa.2004.01.007.

W. Hu, Y. Min, Y. Zhou, Q. Lu (2017) *Wind power forecasting errors modelling approach considering temporal and spatial dependence*, Journal of Modern Power Systems and Clean Energy, 5, 3, 489-498, DOI: 10.1007/s40565-016-0263-y.

T. Schmidt (2006) *Coping with copulas*, In: Rank, J. (Ed.), Copulas-From Theory to Applications in Finance, 1st Edition, 350 pages, Incisive Media Risk Books, London, UK, pp. 3–34, ISBN: 190433945X.

Q. Zhang, M. Xiao, V.P. Singh (2015) *Uncertainty evaluation of copula analysis of hydrological droughts in the East River basin, China*, Global and Planetary Change, 129, 2015, 1-9, DOI: https://doi.org/10.1016/j.gloplacha.2015.03.001.

M. Abramowitz and I. A. Stegun, (Eds.) (1972) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th printing, New York: Dover, 1972, URL: http://people.math.sfu.ca/~cbm/aands/abramowitz_and_stegun.pdf.

W. Feller (1945) *The Fundamental Limit Theorems in Probability*, Bull. Amer. Math. Soc. 51, no. 11, 800-832, URL: https://projecteuclid.org/euclid.bams/1183507419.

W. Feller (1968) *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd edition, New York: Wiley, 528 pages, ISBN: 978-0-471-25708-0.

W. Feller (1971) *An Introduction to Probability Theory and Its Applications*, Vol. 2, 3rd edition, New York: Wiley, ISBN: 9780471257097.

O. Kallenberg (1997) *Foundations of Modern Probability*, New York: Springer-Verlag, ISBN: 978-0-387-95313-7

J. W. Lindeberg (1922) *Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung*, Mathematische Zeitschrift, Vol. 15, Issue 1, 211-225, DOI: https://doi.org/10.1007/BF01494395.

M. R. Spiegel (2003) *Theory and Problems of Probability and Statistics*, Schaum S Outline Series, New York: McGraw-Hill, ISBN: 9780070586109.

H. F. Trotter (1959) *An Elementary Proof of the Central Limit Theorem*, Arch. Math, Vol. 10, Issue 1, pp. 226-234, DOI: https://doi.org/10.1007/BF01240790.

S. L. Zabell (1995) *Alan Turing and the Central Limit Theorem*, Amer. Math. Monthly, Vol. 102, Issue 6, pp. 483-494, DOI: https://doi.org/10.1080/00029890.1995.12004608.

Fulop, Sean A., and Kelly Fitz (2006) *Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications*, Journal of the Acoustical Society of America, Vol. 119 (1), January 2006, pp. 360–371, DOI: https://doi.org/10.1121/1.2133000.

Auger, François, and Patrick Flandrin (1995) *Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method*, IEEE® Transactions on Signal Processing, Vol. 43 (5), May 1995, pp. 1068–1089, DOI: 10.1109/78.382394.

Schuster, Arthur (1898) *On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena*, Terrestrial Magnetism. 3 (1): 13–41, DOI: 10.1029/TM003i001p00013.

Steven W. Smith (1999) *The Scientist and Engineer's Guide to Digital Signal Processing*, 2nd Edition, California Technical Publishing , 1999, ISBN: 0-9660176-7-6, ISBN: 0-9660176-4-1, URL: http://www.analog.com/en/education/education-library/scientist_engineers_guide.html.

Tayfun, Aziz (1980) *Narrow-band nonlinear sea waves*, Journal of Geophysical Research, 85 (C3), pp. 1543–1552, DOI: 10.1029/jc085ic03p01548.

B. McWilliams, M.M. Newmann and D. Sprevak (1979) *The Probability Distribution of Wind Velocity and Direction Wind Engineering*, Vol. 3, No. 4, pp. 269-273, Published by: Sage Publications, Ltd, URL: http://www.jstor.org/stable/43749150.

W. Li, J. Isberg, R. Waters, J. Engström, O. Svensson and M. Leijon (2016) *Statistical Analysis of Wave Climate Data Using Mixed Distributions and Extreme Wave Prediction*, Energies, Vol. 9, pp. 396, DOI: https://doi.org/10.3390/en9060396.

J. Mathisen, E. Bitner-Gregersen (1990) *Joint distributions for significant wave height and wave zero-up-crossing period*, Applied Ocean Research, Vol. 12, Issue 2, pp. 93-103, DOI: https://doi.org/10.1016/S0141-1187(05)80033-1.

K.E. Taylor (2001) *Summarizing multiple aspects of model performance in a single diagram*, J. Geophys. Res., Vol. 106, pp. 7183–7192, DOI: https://doi.org/10.1029/2000JD900719.

S. Bastola, C. Murphy, and J. Sweeney (2011) *The role of hydrological modelling uncertainties in climate change impact assessments of Irish river catchments*, Adv. Water Resour., 34 (5), 562–576, DOI: https://doi.org/10.1016/j.advwatres.2011.01.008.