WIND ENERGY
Wind Energ. (2015)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/we.1826

RESEARCH ARTICLE

An economic impact metric for evaluating wave height forecasters for offshore wind maintenance access

V. M. Catterson[1], D. McMillan[1], I. Dinwoodie[1], M. Revie[2], J. Dowell[1], J. Quigley[2] and K. Wilson[2]

[1] Institute for Energy and Environment, University of Strathclyde, Glasgow G1 1XW, UK
[2] Department of Management Science, University of Strathclyde, Glasgow G1 1XW, UK

ABSTRACT

This paper demonstrates that wave height forecasters chosen on statistical quality metrics result in sub-optimal decision support for offshore wind farm maintenance. Offshore access is constrained by wave height, but the majority of approaches to evaluating the effectiveness of a wave height forecaster utilize overall accuracy or error rates. This paper introduces a new metric more appropriate to the wind industry, which considers the economic impact of an incorrect forecast above or below critical wave height boundaries. The paper describes a process for constructing a value criterion where the implications between forecasting error and economic consequences are explicated in terms of opportunity costs and realized maintenance costs. A comparison between nine forecasting techniques for modeling and predicting wave heights based on historical data, including an ensemble aggregator, is described demonstrating that the performance ranking of forecasters is sensitive to the evaluation criteria. The results highlight the importance of appropriate metrics for wave height prediction specific to the wind industry and the limitations of current models that minimize a metric that does not support decision-making. With improved ability to forecast weather windows, maintenance scheduling is subject to less uncertainty, hence reducing costs related to vessel dispatch, and lost energy because of downtime. Copyright © 2015 John Wiley & Sons, Ltd.

KEYWORDS
wave height forecasting; forecast value; evaluation metrics; offshore wind

Correspondence
V. M. Catterson, Institute for Energy and Environment, University of Strathclyde, Glasgow G1 1XW, UK.
E-mail: v.m.catterson@strath.ac.uk

Received 25 April 2014; Revised 13 November 2014; Accepted 22 November 2014

1. INTRODUCTION

Development of offshore wind farms has gained significant momentum in recent years, with proposed sites being considered at increased distances from shore and at greater water depths to exploit favorable wind conditions. The number and size of individual turbines are also trending upwards, reducing the overall cost of energy. This results in maintenance becoming more logistically challenging and access having a critical impact on downtime and lost revenue.[1]

While multiple factors affect maintenance planning (including traditional concerns of spares management and crew availability), wave height is fundamental to offshore access, particularly for crew transfer. This paper considers the need for accurate wave height forecasting as part of offshore maintenance planning and identifies key criteria for assessing the suitability of a given modeling technique. Standard evaluation criteria such as root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) are not sufficiently user-specific to separate adequate from inadequate models, since the timing and size of some errors are more critical than others. These 'standard' statistical measures of accuracy do not capture the positive or negative economic impact that the forecast has on decision-making.

Murphy[2] identified three criteria to judge the quality of forecasts, namely consistency, quality and value. Consistency refers to the level of agreement between a forecaster's beliefs and their specified forecast. Quality refers to the statistical metrics between forecasts and realizations. Value relates to the added value to the decision maker using the forecasts. Of these three criteria, the most difficult to quantify is value because of the complexity of the problem and commercial

sensitivity.[3] This paper develops a metric for explicating the relationship between forecasts and value relevant to the decisions being supported by the forecaster, namely offshore wind maintenance.

As a case study, nine wave height forecasters developed using 7 years of historical wind and wave data from the Forschungsplattformen in Nord- und Ostsee FINO 1 platform in the North Sea[4] are evaluated using our metric and standard metrics for quality. Based on this analysis, the ranking of forecasters is sensitive to the evaluation criteria. This paper shows that selecting a forecaster based on statistical quality metrics results in a sub-optimal forecast with respect to what the decision maker values. Specifically, in this context, using a quality metric fails to capture the trade-off between lost opportunity cost and realized maintenance costs. This analysis serves as a benchmark for improved wave height prediction and model evaluation for offshore wind maintenance planning.

The remainder of the paper is structured as follows. In Section 2, we provide an overview of current wave forecasting modeling. In Section 3, we develop an economic metric that captures the value that is relevant to offshore maintenance planners. In Section 4, we describe the different forecasting methods that we utilize to evaluate the proposed metric. In Section 5, we compare the results of each metric for the different forecasting methods illustrating that the ranking of method changes based on the metric used. In Section 6, we conclude the paper and discuss areas of future research.

2. WAVE FORECASTING STATE OF THE ART

The distance from trough to crest of a wave is generally measured by a wave buoy fitted with accelerometers. These heights are aggregated over a sample period into a measure called significant wave height, H_s, which is the mean height of the highest third of waves recorded during the period (such as an hour or a minute).[5] Significant wave height is therefore not the largest wave measured but is representative of the height of all the large waves seen by the buoy. Forecasts of significant wave height are useful for various applications such as prediction of bubble, turbulence or air–sea drag,[5] in addition to offshore maintenance planning.

Traditional approaches to wave height modeling require detailed information about the location and geography of the site in question, in order to develop a physical model of wave energy spectra.[6] Capturing the physical processes that create and disperse waves within a region requires a detailed wind model, since wind is one of the strongest influences on wave height.[6] The complexity of the model linking wave height to wind and the computational resources currently needed to perform this kind of forecasting make a full physical model an undesirable choice for a wind farm operator.

An alternative to developing a site-specific physical model is to access wave forecast data from public sources, such as the National Oceanographic and Atmospheric Administration's WAVEWATCH III forecasts, which are published online every 12 h. This type of forecast predicts wave patterns over large areas of ocean, so local features specific to a given wind farm may not be well represented. Smaller-scale models such as Simulating Waves Nearshore can be coupled with larger-scale models to forecast at high resolutions and have been demonstrated along coastlines.[7] Since physical models have been shown to outperform statistical data-driven models at longer time horizons (over 6 h),[8] this approach may be particularly suited to planning of long maintenance actions.

In comparison to physical models, data-driven models are often computationally simple, and forecasts can be generated within seconds. Training and model validation require up-front effort, but since there is no direct knowledge of the site required, the same technique can be applied to multiple sites by simply retraining the model. The disadvantage of a data-driven approach is that training depends on a set of representative historical data. But the site of a wind farm will have gone through many years of analysis and assessment prior to operational maintenance, so data relating to wind speed and wave height are often readily available.

Because of the strengths and weaknesses of each approach, research continues into physical modeling,[9,10] data-driven modeling [including the use of artificial neural networks (ANNs)[11,12] and comparison with autoregression[13] or Kalman Filters[14]] and hybrid models.[15,16] However, these studies use standard measures of model accuracy such as RMSE, without considering the nature of scheduling offshore wind farm access. This paper considers the unique properties of the application, including key wave thresholds that limit access to certain vessel types, which means that RMSE or MAE does not adequately capture the success or failure of a given model.

As a result, this paper identifies useful criteria for evaluating a wave height forecaster in the specific context of scheduling maintenance for offshore wind turbines. An economic impact metric is proposed, based on these criteria. To benchmark the new metric against the standard RMSE metric, nine data-driven models* are evaluated using an openly accessible data set for the North Sea. Some conclusions are drawn about the relative merits of the evaluation metrics for this application.

*The economic impact metric can also be applied to physical models, but the creation and test of a physical model are out of scope in this paper.

3. MODEL EVALUATION CRITERIA

Statistical measures of errors such as RMSE, MAE and MAPE do not reflect the economic consequences of inaccurate forecasts and as such provide inappropriate criteria for forecast selection to support decision-making. Maintenance offshore is restricted by the access constraints of service vehicles. Different operations require different vessels, from access transfer boats for transporting crew and small items to specialized field support or mobile jack-up vessels in cases of major failures. The upper limit on wave height constraints is that of a helicopter, since health and safety regulation mean that crew can only be offshore while a helicopter sea rescue is possible.

Wind farm operators do not own vessels but hire those required on a day-rate basis. Therefore, minimizing the hire duration of maintenance vehicles is critical to controlling operations and maintenance costs.

3.1. Forecasting weather windows

Importantly, wave height restrictions apply for the duration of the mobilization from shore, termed a weather window. Completion of a maintenance action requires a weather window to exist, but planning of maintenance relies on the forecasting of weather windows.

Forecasting of weather windows has been studied in the context of offshore construction projects,[17] where the prediction of a weather window, which does not hold, can result in project delays. A risk reduction strategy is to accompany each weather window forecast with probabilities of the window being breached by large and moderate margins.[18] However, maintenance planning has different challenges to construction. A construction project generally assumes that a large vessel is installed on-site until the conclusion of the project and a breached weather window has a negative impact, i.e. it delays operations. For maintenance, a correctly predicted breached weather window holds no penalty; instead, it is a forecast at odds with its realization, which holds negative consequences.

Specifically, an inaccurate forecast, which predicts continuously low wave heights, could result in a vessel being chartered; then the trip being aborted before maintenance is complete, effectively wasting the cost of the day hire of the vessel and the opportunity cost of the crew assigned to the site. Conversely, an inaccurate forecast of high wave heights could result in a lost opportunity to return a turbine to service.

This paper concentrates on crew transfer vessels (CTVs), used for transporting crew and tools for common maintenance operations offshore. The day rate for chartering a CTV is typically £1750[1] and the operational limit is a wave height of 1.5 m. In practice, a very urgent repair may warrant the charter of a more expensive vessel, which can withstand greater wave heights, but this paper focuses on the common case of CTVs only. While a range of forecasting techniques is available for maintenance planning, in practice an ad hoc unstructured method led by the ship captain is typically used to make the final decision of whether or not to commence the mobilization.[19]

3.2. A new economic metric

Drawing on the specifics of the application area, this paper introduces a metric for evaluating the strength of a wave-forecasting model for offshore maintenance access. Traditional methods of comparing accuracy using error, such as RMSE, MAE or MAPE, do not take account of the requirements for accurate forecasting over the period of a full journey and weight all errors equally. This could lead to a favorable evaluation in cases where the practical impact on vessel charter is poorer than other techniques.

Specifically, for a true wave height above an access threshold of 1.5 m, any predicted height above this is equally good, and any prediction below it is equally bad[†]. The impact of an incorrect low prediction is an aborted trip, regardless of whether the error is 0.1 or 1 m. There is no negative impact from a prediction on the correct side of the height limit, even if the prediction contains a large error. Therefore, the height forecast can be treated as a classification, either above or below the constraint limit of 1.5 m for a typical access vessel. This means standard classification accuracy tools can be used to evaluate model performance.

Each forecast is assigned a label of True Above (*TA*) or True Below (*TB*) (where prediction and actual height are on the same side of the boundary), *FA* (where the prediction is above the limit and the actual height is below) or *FB* (where the prediction is below and the actual height is above the limit). These labels can be used to generate standard measures of classifier accuracy, such as a confusion matrix or figures for sensitivity and specificity.

[†]Note that in practice, this threshold is likely to have some flexibility, based on attitude and preferences of the vessel captain. For the purposes of this evaluation, we assume that there is no directional bias to the application of this threshold. As such, we assume that small errors above and below the threshold will cancel each other out.

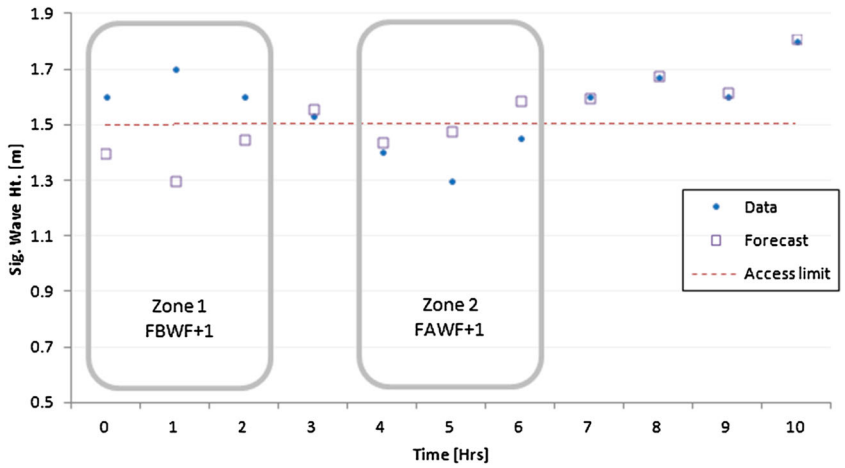

Figure 1. Summary of weather window forecast error contributions toward *FAWF* or *FBWF*.

The number of consecutive accurate forecasts is also of great importance. If the maintenance task round trip time (Δt, defined as travel time, crew transfers and work done) is scheduled to be 3 h, all forecasts within the 3 h duration window should be *TB* height predictions in order to be considered a successful forecast sequence. If even a single *FA* prediction occurs in conjunction with such a real access period, this will result in the trip being aborted and rescheduled, resulting in a lost opportunity to carry out valuable reactive maintenance. The annual frequency of such events is described by false above frequency (*FAF*). The result of these individual hourly forecast events impacts on the accuracy of the weather window forecasts; this is measured via false above window frequency (*FAWF*).

Conversely, a sequence of three consecutive 'below' forecasts, any one of which is *FB*, corresponding to a false prediction of a weather window, will result in a CTV being dispatched in error—with implications for fuel and vessel costs. The annual frequency of such events is described by false below frequency (*FBF*). In terms of weather window impact, this is measured via false below window frequency (*FBWF*).

These two possibilities are illustrated in Figure 1, which summarizes these two situations (Zone 1: *FBWF* increases by 1 during the false window $t = 0, 1, 2$ and Zone 2: *FAWF* increases by 1 because of a single *FA* forecast at $t = 6$ during a real window at $t = 4, 5, 6$). The main assumptions when utilizing these metrics are that for both metrics (*FAWF* and *FBWF*), the decision to dispatch a vessel is taken solely on the basis of the quantitative forecast. Further, it is assumed that there is always at least one turbine in the wind farm on outage, waiting to be restored to service.

A simple algorithm (Figure 2) compares the measured and forecast data to identify the position and frequency of *FA* and *FB* predictions for each look-ahead time. Note that the 1 h-ahead prediction of a 3 h long window requires the 1 h-ahead, 2 h-ahead and 3 h-ahead significant wave height forecasts for the first, second and third hours of the window respectively. The case study results presented later consider predictions of 3 h windows at greater horizons, up to the 8 h-ahead, 9 h-ahead and 10 h-ahead forecast window. Following from these metrics, the wind farm operator can place an economic value on a forecasting approach by calculating the cost of False Above (*CFA*) and the cost of False Below (*CFB*) predictions during a time horizon.

For CTV operations, there are two general implications of a bad forecast. First, the incurred annual cost of False Above events (C_{FA}) is worked out on the basis that for each event, a CTV is not dispatched even though a suitable weather window does exist. Here, the economic impact is lost revenue from not producing energy (C_G) from the wind turbine. The required inputs are turbine capacity (*CAP*), cost of energy (C_{en}), average capacity factor (*CF*) and duration of weather window (Δt).

$$C_{FA} = FAWF.C_G(\Delta t, CAP) \tag{1}$$

where $C_G(\Delta t, CAP) = \Delta t.CAP.CF.C_{en}$, and we assume that $CAP = 5MW$, $C_{en} =$ £120/MWh, $CF = 0.5$ and $\Delta t = 3$ h.

Second, the incurred cost of a False Below event (C_{FB}) is defined as a situation in which the CTV is dispatched, but in reality weather does not allow the operation to be completed in the time allocated. Here, the incurred costs are fuel (C_F, based on an assumed speed of 15 knots and distance to wind farm of 25km) and the cost of the contracted CTV time (C_{ctv}).

$$C_{FB} = FBWF.(C_F(\Delta t) + C_{ctv}(\Delta t)) \tag{2}$$

We assume that $C_{ctv}(\Delta t) =$ £220/hr, $C_F(\Delta t) =$ £50/hr and $\Delta t = 3$ h. Combining equations (1) and (2) gives an overall economic forecasting metric (*EFM*):

t discrete time index
k forecast look-ahead time

$d(t)$ Binary wave data
$f^{(k)}(t)$ Binary forecast wave data k hours ahead.
1=wave height below limit, 0=above limit.

$D(t)$ Binary window data
$F^{(k)}(t)$ Binary forecast window data k hours ahead
1=window start, 0=no window start
Initially: $D(t), F^{(k)}(t) = 0$ for all t.

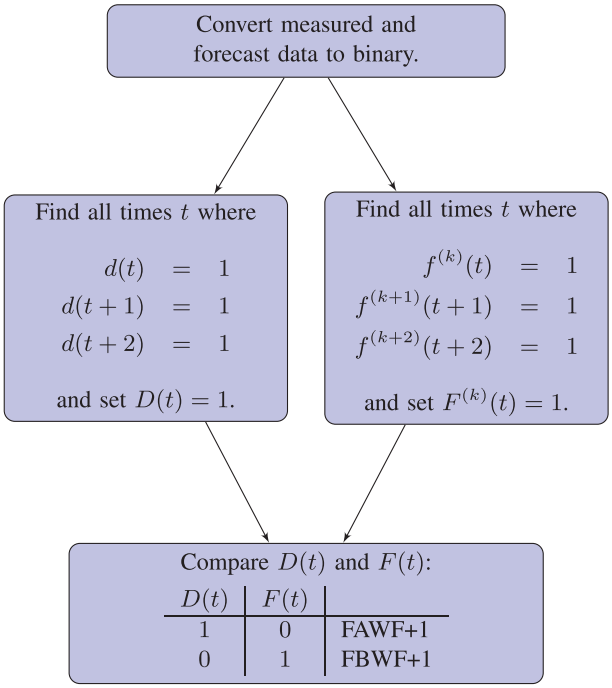

Figure 2. Flow chart of window counting procedure implemented in MATLAB.

$$EFM = C_{FB} + C_{FA} \tag{3}$$

Note that the aforementioned metric assumes that a CTV is dispatched whenever a weather window is available. Discussions with operators and service providers suggest that for large wind farms (e.g. greater than 100 turbines), this assumption is reasonable.

4. METHODOLOGIES

This study aims to assess the effectiveness of the *EFM* for ranking forecasters for offshore maintenance planning. A suite of data-driven wave height prediction techniques was built, and their performance was assessed using the new and standard metrics. This section focuses on the methodologies for constructing the suite of forecasting models.

Nine techniques from three broad classes of model were compared. First, simple smoothing approaches such as Exponential Smoothing and Trigg and Leach were investigated. Second, statistical models such as Autoregressive moving average (ARMA), Splines, dynamic linear models (DLMs) and Markov chains were used. Finally, artificial intelligence models such as support vector machines (SVMs), ANN and ensemble learning models were used. These classes were chosen in order to assess whether different types of model would better capture the wave regime.

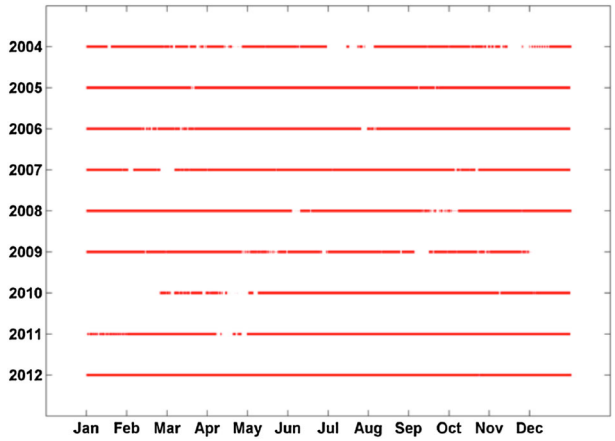

Figure 3. Monthly wave data quality by year, showing gaps.

The same training and test data were applied to each model. The aim was to predict wave height at multiple forecast horizons, ranging in 1 h time steps from 1 h ahead up to 10 h inclusive. No restriction was placed on how many historical time steps were used to generate a prediction, and this was optimized per model. The full training regime and parameter optimization for each model are described as follows. After training, each model was used to generate wave height predictions for the test set. The new metric outlined in Section 3 was calculated for these predictions and compared against RMSE.

4.1. FINO research data

The data used in this study were captured from the FINO 1 offshore research platform situated in the North Sea 45 km off the German coast.[4] This area is marked for development of future offshore wind farms and is therefore a suitable site for model testing. It also represents the highest quality, longest measured data set publicly available.

Both mean wind speed and significant wave height for 1 h periods are available from the platform. For this study, wind speed and wave height between 00:00 1 January 2006 and 00:00 1 December 2012 were used. The training set was limited to 2 years of data (2006 and 2007), while the test set was from 00:00 1 January 2008 until the end of the data set. An earlier study showed that varying the length of the training set had no significant impact on model accuracy for this data.[20]

The overall quality of the data for the study period is 89.82% (Figure 3). Gaps in the data were filled using a simple cubic spline interpolation. The quality of data and simplistic gap filling process have the potential to influence the accuracy of each forecasting technique to different degrees. Therefore, the forecasting techniques were also applied to the longest continuous period of data between May 2011 and July 2012, in order to quantify this influence. The average percentage difference in RMSE between forecasting over the entire data set and subset was 6%. Critically for this study, the ranking order of forecasters was preserved in both cases. For clarity and repeatability, the use of the entire data period and simple interpolation methodology was adopted for the analysis in this paper.

4.2. Persistence

The simplest method of forecasting is to assume that current conditions predict future conditions exactly. For wave height forecasting, the persistence approach assumes that the wave height at time t is the same as the wave height at time $t + n$, for some point n hours in the future.

This technique does not make use of wind speed or any historical wave height data apart from the most recent measurement. The persistence model is considered to be the baseline that other techniques in this study are compared against.

4.3. Exponential smoothing

One of the most basic, yet fundamental, models in forecasting is the exponential smoothing model first proposed by Brown[21] and developed by Holt *et al.*[22] The model is widely used by industry, partially because of its simplicity (computationally and intuitively)[23] and partially because of its accuracy. Computationally, even large data sets can be processed quickly and simply within standard spread sheet packages. Intuitively, the model simply weights past data to predict future data. Studies have demonstrated that despite its simplicity, there is sometimes little difference in accuracy between exponential smoothing and other forecasting techniques such as ARMA.[24,25]

The model assumes that the data are stationary and non-trending. The following formula is used for the exponential smoothing model: $S_t = \alpha X_{t-1} + (1-\alpha)S_{t-1}$ where S_t is the prediction for time step t, X_{t-1} is the observed value in time step $t-1$ and α is a smoothing factor. Typically, small values of α are chosen, usually between 0.1 and 0.3. Where the model does not perform well for these values, typically a more complex model is adopted.[26] A value of α close to 1 places greater emphasis on recent data. Note that a value of $\alpha = 1$ is equivalent to the persistence model.

Predictions were made for the test set for different values of α. Exponential smoothing only provides predictions for $t+1$. Using the aforementioned formula, we calculated S_{t+i} using α_i, $i = 0, \ldots 9$. The R[27] linprog package was used to minimize the RMSE for each α_i bounded between 0 and 1. From this analysis, $\alpha_i = 1$, i.e. persistence.

4.4. Trigg and Leach

Trigg and Leach[28] extended the simple exponential smoothing approach by including a dynamic smoothing constant that adjusts the performance of the forecasting by either increasing or decreasing the weight applied to historical data depending on the local stability of the series being forecast. The performance of the forecasts is measured through a tracking signal:

$$\textit{Tracking Signal} \approx \frac{\textit{smoothed error}}{\textit{smoothed absolute error}} \tag{4}$$

As such, the tracking signal must be between -1 and 1 as the absolute error must be at least as large in magnitude. The sign of the signal provides insight into direction of bias, and the magnitude provides insight into the extent of the bias. A signal near 1 or -1 indicates a systematic over or under estimation, while a value close to 0 indicates unbiased forecasting. The use of exponential smoothing for these assessments allows the analyst to apply more weight to recent observations.

The absolute value of the tracking signal is then used to provide an adaptive exponential smoothing constant. A value close to 0 implies that the series is 'currently' stationary and as such more weight can be applied to recent history. A value close to 1 implies that the series tends to be either increasing or decreasing and as such more weight should be applied to the most recent observation.

4.5. Cubic spline

A number of statistical tools have been subsequently developed from their original purpose, such as regression models, to model time-series data. One such example of that are splines,[29] which have been used for modeling time-series data in different domains.[30–32] A spline aims to link data points through a simple function, the simplest being a straight line. Functions with higher degrees, such as polynomials with degree n, can also be considered.

Cubic splines have been proposed as a method for local-linear extrapolation when modeling a time series with nonlinear trend. Model predictions were made using the R forecast package. To model the time-series data, the R forecast package develops a cubic spline based on historical data and then linearly extrapolates. When predicting $t+1$, this produces an improved estimation when compared with persistence. However, using this extrapolation performs poorly for $t+i$ when $i \geq 4$. As such, we have chosen to adopt the $t+1$ prediction for all $t+i, i \geq 2$.

4.6. DLMs

Dynamic linear models[33,34] give an approach to time-series analysis in which the response, in general a vector, X_t, is assumed to move through time based on the value of an unobserved state vector θ_t. The state vector then evolves through each successive time step $t = 1, \ldots, p$. A further requirement of DLMs is that all unknowns are assumed to be normally distributed within the model. The general structure of a DLM is then

$$\begin{aligned} X_t &= F_t\theta_t + v_t\,, v_t \sim \mathrm{N}(0\,, V_t) \\ \theta_t &= G_t\theta_{t-1} + w_t\,, w_t \sim \mathrm{N}(0\,, W_t) \end{aligned}$$

and to complete the specification, $\theta_0 \sim \mathrm{N}(m_0, C_0)$, which is the prior distribution for θ_0 given a mean vector m_0 and covariance matrix C_0. The parameters v_t and w_t represent observation and evolution errors with possibly non-diagonal variance matrices V_t and W_t respectively, and F_t and G_t are the observation and evolution matrices. This flexible structure allows DLMs to take many specific forms including a simple random walk, ARMA processes, polynomials, regressions and seasonality. Complex models are built up from combinations of these simple components.

Many different combinations of various individual components were fit to the wave height data. The best fitting DLM, used in the remainder of the paper, incorporates a random walk element plus seasonal components for each quarter and the year in question as a covariate. The specific model is

$$y_t = \alpha_t + \sum_{i=1}^{4} \beta_{i,t} q_{i,t} + \sum_{j=1}^{9} \gamma_{j,t} x_{j,t} + v_t,\ v_t \sim \mathrm{N}(0, V)$$
$$\beta_{i,t} = \beta_{i,t-1} + b_t,\ b_t \sim \mathrm{N}(0, B)$$
$$\gamma_{j,t} = \gamma_{j,t-1} + c_t,\ c_t \sim \mathrm{N}(0, C)$$

where α_t is associated with the random walk, $q_{i,t}, x_{j,t}$ are indicator variables of whether observation t is in quarter i and year j respectively and $\beta_{i,t}, \gamma_{j,t}$ are the parameters associated with each of these effects. Stationary variances B and C provided an excellent fit. Updating and forecasting within the model are performed using the Kalman filter after suitable initial values are chosen for the parameters in the model. These initial values do not have an effect on the forecasts in this case as a result of the large quantity of training data.

Dynamic linear models are widely used in Bayesian time-series analysis, and the model could be easily adapted to incorporate informative prior knowledge.

4.7. ARMA model

Autoregressive (AR) approaches to describe time-series data were originally developed by Box and Jenkins[35] and since then, have been applied widely. Of particular relevance to this work, AR models have been used to describe significant wave height,[13] mean wind speeds for wind turbine power generation[36] and wind turbine maintenance.[37] The AR model, normalized to the mean, μ, of the data at time step t, X_t is

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{p} \phi_i (X_{t-i} - \mu) \tag{5}$$

where ϕ is a correlation coefficient. The model described in equation (5) is valid only for a normally distributed process. Neither annual wind speed nor significant wave heights follow a normal distribution and must therefore be transformed before equation (5) is applied to the data sets.

For significant wave heights, it is necessary to remove the fit of monthly mean and then apply a transformation on the data shown in equation (6)[35] where $T(Hs_t)$ represents the transformed series and $\hat{\mu}_{ln}$ is a Fourier series fit of seasonality of the transformed time series.

$$Y_t = T(H_{s_t}) = \ln(H_{s_t}) - \hat{\mu}_{\ln(H_{s_t})} \tag{6}$$

The required order of AR model was determined using the autocorrelation function and partial autocorrelation function and determined as 4 for wave height modeling. The determination of AR coefficients and model generation was performed using the MATLAB 2013a system identification toolbox (MathWorks Inc., Natick, MA).

For the multivariate case, predicted wave height is informed by previous observations of wave height and wind speed, and a vector AR model is used. In this case, the correlation term ϕ in equation (5) is replaced by a correlation matrix allowing for the influence of additional parameters on the wave forecast to be captured. Multivariate coefficients are determined using a multivariate least squares estimation.

4.8. Markov chain

Markov chains have been deployed to solve several problems in the wind energy literature. This framework was used by Castro Sayas and Allan[38] to model wind turbine failure rates and the influence of wind speed on reliability. In terms of forecasting applications, the work of Pinson and Madsen[39] is prominent.

This paper applies a pure Markov chain (that is memoryless and has time-homogeneous parameters) with discrete time and discrete state space to model and forecast wave height. The main criterion when setting up the chain is the bin size, which determines how the state space is partitioned. This is established first by determining the maximum value in the data set. Then, an appropriate bin size is chosen, which is specific to the variable being modeled (for example, the modeler should take into account the resolution of the original data when selecting the bin size). For the wave height forecaster, the bin size was set to 0.05 m, with a maximum value of 11 m, resulting in a 220×220 transition matrix for each forecast time horizon.

The parameter estimation process is based on the normalized frequency of transition from one state to another and the frequency balance method of Billinton and Allan[40] and is summarized as follows:

$$P_{a,b} = P(s_b, t_{k+1} | s_a, t_k), k = 1, 2, 3 \ldots N \tag{7}$$

where $P_{a,b}$ is the probability of transit from state a (s_a) to state b (s_b) and t_{k+1} is time at $k + 1$ up to a maximum number of states N.

It is noted that performance improvement for the Markov chain could theoretically be achieved by partitioning the model into seasonal or monthly models (see e.g.[41]). However, this drastically cuts down on the data available to estimate the model parameters and results in a highly sparse matrix. In this sense, the Markov chain is much more data intensive than the other forecast methods in this paper. Because of these constraints, an unpartitioned, pure Markov chain is adopted.

4.9. Neural network

The ANN is perhaps the most commonly applied intelligent system technique for nonlinear regression problems and has in the past been applied to wave height modeling.[14,20] The attraction of an ANN is that with a three-layer network comprised of simple units (neurons), any function can be approximated.[42] Each neuron performs a weighted sum of its inputs before passing the result through an activation function to produce an output. Common activation functions include the sigmoid, hyperbolic tangent and linear.

The three layers are termed the input layer (data inputs plus a bias term), the hidden layer and the output layer, and this architecture is also referred to as the multi-layer perceptron. Each layer is fully connected to the next, meaning all inputs connect to all hidden neurons and all hidden neurons connect to all output neurons.

Training is performed using a back-propagation algorithm, where the network output for sample input is compared against the target value and neuron weights are updated to minimize error. Model training was performed using the R nnet library, with weight decay of 0.0005 and the maximum number of iterations increased to 1000. A standard multi-layer perceptron architecture was used, with a sigmoid function for the hidden nodes and a linear function for the output node.

A number of networks were trained, varying by the number of inputs and number of hidden nodes in the network. Inputs ranged from one prior time step of wind and wave up to seven ($t-i, 0 \leq i \leq 6$). The number of hidden nodes ranged from 2 to 20. For each architecture, three networks were trained. This gave a total number of networks evaluated for each output as 7input steps $\times$ 19hidden layer configurations $\times$ 3 = 399.

Model predictions were generated using the predict routine in R for the test data set. Of the 399 networks trained, the one with the lowest RMSE from the test data was selected as the network for that output[‡] .

4.10. SVMs

In contrast to the ANN, which many researchers have applied to wave height prediction, the SVM has seen less study. One paper was found,[43] which uses an SVM to predict wave height on Lake Michigan.

The SVM maps input data into a space using a particular function called a kernel function.[44] Originally used for classification, the SVM learns the boundary separating one class from another with maximal distance. The kernel function aims to translate a problem that is nonlinearly separable into a feature space, which is linearly separable by a hyperplane. When used for regression, the hyperplane represents the function in feature space, rather than a classification boundary.

The SVM is parameterized through the choice of kernel function. Common functions used include linear, polynomial and radial basis function (RBF). While there is no definitive methodology for appropriate training of SVMs, there are some generally agreed best practices to follow.[45] For a problem which may be nonlinear, the RBF kernel is recommended. This is parameterized by the error cost c and the RBF width parameter γ. Standard practice is to optimize these parameters through a two-stage grid search:[45] firstly, trying pairs of c and γ values with large steps across a wide space (rough grid search) and followed by smaller-stepped pairs of values around the region of the best rough values (fine grid search).

Model training was performed using the R e1071 package, which is a wrapper to the popular libsvm implementation. The RBF was chosen as the kernel, and grid search was used to optimize γ and c within the following ranges: $\gamma \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and $c \in \{1, 10, 100\}$.

Inputs to the SVM ranged from one up to seven prior time steps ($t-i, 0 \leq i \leq 6$) of wind and wave. Model predictions were generated using the predict routine in R for the test data set. Of the trained SVMs, the one with the lowest RMSE from the test data was selected as the choice for that output.

4.11. Ensemble learning

Research suggests that ensemble forecasters, which aggregate multiple predictions together, can outperform individual models.[46,47] For physics-based forecasting, an ensemble is used to find the most probable forecast given small variations

[‡] This presents the possibility of increasing model accuracy by changing the ANN evaluation criteria, that is by altering the training of the neural network to optimize for the *EFM*. This was not done for this study, since the other modeling techniques are being used to predict wave height. Changing the predictor parameter of the ANN was thought to make it a less-valid comparison of technique performance, and therefore the ANN was also trained to predict wave height.

in model initial conditions. For the statistical models presented in this work, an ensemble was created to aggregate the predictions of each model, with the intention of improving forecast accuracy.

The nine models described previously were taken as input to the ensemble, while the aggregator itself was chosen to be an ANN. Separate ensemble models were trained for each forecasting time step. For example, the $t + 1$ ensemble takes as input the $t + 1$ predictions from the nine source models and gives as output another $t + 1$ prediction. As a result, the ensemble has no memory of prior time steps.

The ensemble ANNs had nine inputs and one output. The number of hidden nodes was varied between 2 and 14 inclusively, and three training runs were performed for each architecture. For each time step, the best architecture was chosen based on lowest RMSE.

5. RESULTS

This section presents the results of the different forecast metrics calculated for the data and models described previously. The RMSE of all models is shown in Figure 4. This suggests very similar accuracy between models, with the exception of the ensemble that significantly outperforms the others. Investigating the individual models more closely, Figure 5 shows the percentage improvement in RMSE over a persistence model. From this, we see that RMSE is able to distinguish between the predictive power of each model. Figure 6 captures the RMSE percentage improvement of the ensemble.

Repeating the analysis for MAE, similar results emerge. As MAPE is scale sensitive and wave height can take values close to 0, we discard it as an appropriate metric to benchmark the methods.

Compared with Figure 4, Figure 7 illustrates the *EFM* for 3 h weather windows for different time horizons. Note that since the *EFM* is concerned with weather windows rather than individual point forecasts, several forecast values are used. For example, to forecast a 3 h weather window 8 h ahead, the forecasts for $t + 8$, $t + 9$ and $t + 10$ are utilized. Since *EFM* represents a cost, smaller values indicate better models.

The most immediate feature of Figure 7 is the poor performance of the Markov chain model, probably because of its time-homogenous properties. All the other forecasters in some way preserve the dynamic temporal aspect of the data,

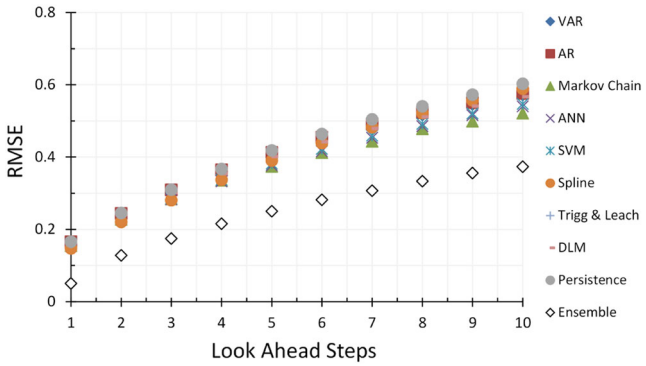

Figure 4. Absolute RMSE for all models. VAR, vector autoregressive.

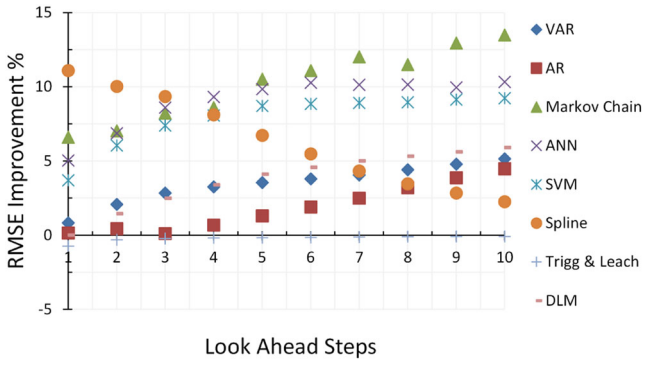

Figure 5. Percentage RMSE improvement over a persistence model. VAR, vector autoregressive.

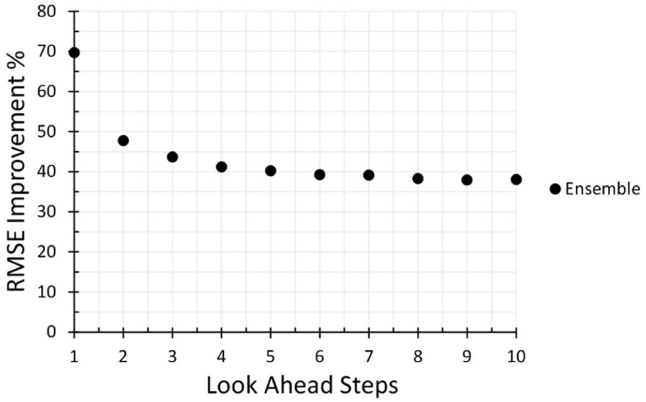

Figure 6. Percentage RMSE improvement over a persistence model for ensemble model.

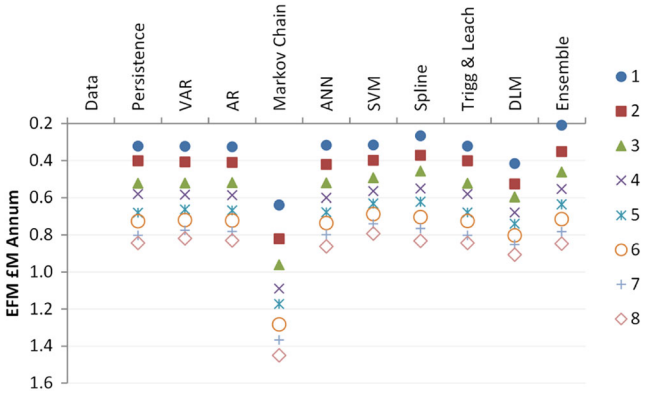

Figure 7. Distribution of *EFM* for 3 h weather windows. Legend indicates forecast horizon of 1 to 8 h ahead. VAR, vector autoregressive.

whereas the homogenous Markov chain parameters do not. This highlights the flaws in using RMSE or MAE as the forecasting benchmark as it weights bad forecasts equally over the entire range of the wave height. The *EFM* focuses on the economic impact of forecasts around the key operating point of the plant and is thus more relevant for offshore operational decisions.

The best-performing model for shorter forecast horizons (1–2 h ahead) is the ensemble. However, it can be seen that selection of the 'best' forecaster very much depends on the time horizon of the forecast. For the longest-range forecast considered in this study (8 h ahead), the SVM outperforms all other models. This is a very interesting result as inspection of Figures 5 and 6 would suggest that the ensemble forecast outperforms all of the individual forecast tools. Figure 7 illustrates that in an operational environment there is value in a deeper understanding of how individual forecasts perform under different conditions.

Figure 8 takes the extreme time horizon of the 8 h ahead forecast and provides a breakdown of the *EFM* into its constituent costs for these conditions (when forecasting a weather window of 3 h' duration). It is observed that for most cases, the economic impact of *FA* forecasts is greater than *FB* forecasts. There are two key reasons for this. First, to increase *FBWF*, a forecast must be consistently *FB* (in this case for a full 3 h). This is in contrast to the *FAWF*, where only a single *FA* during a real weather window will result in an increase in *FAWF*. The relative frequency of these events in part drives the cost distributions in Figure 8. Second, C_{FB} is driven by fuel and CTV charter costs, while C_{FA} depends on turbine and wind resource characteristics. In this case study, the costs associated with each are different, i.e. $C_{FA} = FAWF.£900$ whereas $C_{FB} = FBWF.£810$. An interesting avenue of future work would be to explore the relationship between the cost structure of the metrics and how this should be adapted depending on wind turbine unit capacity, cost of chartering the CTV and so on.

If a utility were to rank the methods using a quality metric, e.g. RMSE or MAE, or a value metric, e.g. *EFM*, a different ordering would be observed. Table I shows that the ensemble model performs best for RMSE across all time horizons. However, for the *EFM*, the ensemble is outperformed by the SVM for the longer time horizons (over 6 h). Similarly, the

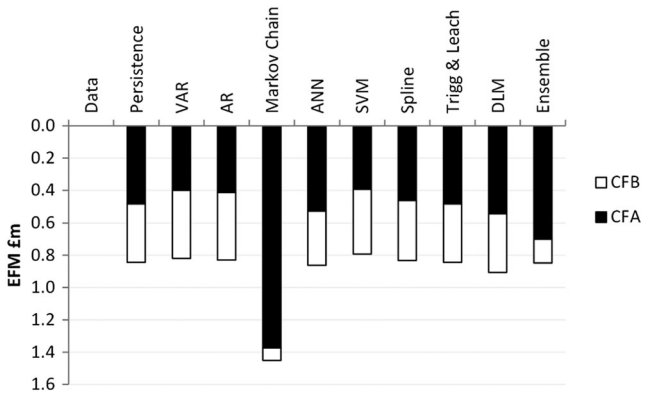

Figure 8. *CFB*, *CFA* and combined *EFM* for 3 h weather windows, for a forecast horizon of 8 h ahead. VAR, vector autoregressive.

Table I. Ranking of each method by RMSE and *EFM*.

Metric	Method	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8
	VAR	6	6	6	7	7	7	7	6
	AR	7	8	8	8	8	8	8	8
	Markov chain	3	3	4	3	2	2	2	2
	ANN	4	4	3	2	3	3	3	3
RMSE	SVM	5	5	5	5	4	4	4	4
	Spline	2	2	2	4	5	5	6	7
	Trigg and Leach	9	9	9	9	9	9	9	9
	DLM	8	7	7	6	6	6	5	5
	Ensemble	1	1	1	1	1	1	1	1
	VAR	6	5	6	5	4	4	3	2
	AR	7	6	4	6	5	5	4	3
	Markov chain	9	9	9	9	9	9	9	9
	ANN	4	7	5	7	6	7	6	7
EFM	SVM	3	3	3	3	2	1	1	1
	Spline	2	2	1	1	1	2	2	4
	Trigg and Leach	5	4	7	4	7	6	7	5
	DLM	8	8	8	8	8	8	8	8
	Ensemble	1	1	2	2	3	3	5	6

VAR, vector autoregressive.

Markov chain model performs well for the RMSE metric but performs extremely poorly for the *EFM*. Alternatively, Trigg and Leach performs very poorly for RMSE but performs better for *EFM*.

If a utility were to select one of these models to plan maintenance 8 h ahead based on RMSE, they would select the ensemble or the Markov chain if a single model is preferred. The corresponding costs of utilizing these models are £847,633 (ensemble) and £1,450,119 (Markov chain). By considering instead *EFM*, the utility would select the SVM with a cost of £792,283. This represents a saving of £55,350 to £657,836 per annum. Focusing on a quality metric rather than a value metric results in sub-optimal decision support for offshore wind farm maintenance.

6. DISCUSSION

Forecast accuracy has an important role to play in controlling offshore maintenance costs. However, standard metrics such as RMSE evaluate the absolute accuracy of a forecast, without accounting for how the forecast will be used. Since access for maintenance is predicated on the wave height access threshold, the size of a forecast error is unimportant as long as the prediction falls on the correct side of the threshold.

This paper has introduced an economic impact metric, *EFM*, for evaluating the utility of wave height forecasters for offshore access. The metric incorporates the financial cost of two types of forecast error: chartering a CTV for a false access window or lost generation when a turbine repair could have been completed. As a case study, nine data-driven models were evaluated using both RMSE and *EFM*.

The key conclusion to be drawn from the results is that the ensemble forecaster significantly outperforms all other models when evaluated using RMSE, but it is outperformed economically by the splines and SVMs at longer forecasting horizons. The economic impact of utilizing RMSE instead of *EFM* for the 8 h-ahead case study is £55,350 or more per annum. This demonstrates that the value of a wave height forecaster for offshore maintenance access depends not simply on its accuracy in absolute terms.

Future work should focus on exploring the sensitivity of the economic metric to changes in fuel costs, charter costs, turbine characteristics and so on. Some of the simplifying assumptions should be considered in more detail, such as there always being a turbine experiencing downtime. The metric assumes a reactive maintenance regime to simplify the quantification of the impact of a missed weather window; the effects of a proactive approach to maintenance would require further study. Future research could also investigate the applicability of the metric to other applications. For example, forecasting power output has asymmetric penalties and could be improved by using a similar metric over MAE or RMSE.

ACKNOWLEDGEMENTS

For data from the FINO project, we thank the BMU (Bundesministerium fuer Umwelt, Federal Ministry for the Environment, Nature Conservation and Nuclear Safety) and the PTJ (Projekttraeger Juelich, project executing organisation). We would like to thank the editor, Pierre Pinson, and the anonymous reviewers for their valuable suggestions for the paper.

REFERENCES

1. Tavner P. *Offshore Wind Turbines: Reliability, Availability and Maintenance*, Vol. 13. Inst of Engineering & Technology, 2012.
2. Murphy AH. What is a good forecast? An essay on the nature of goodness on weather forecasting. *Weather and Forecasting* 1993; **8**: 281–293.
3. Mailier PJ, Jolliffe IT, Stephenson DB. Assessing and reporting the quality of commercial weather forecasts. *Meteorological Applications* 2008; **15**(4): 423–429.
4. Bundesministerium fuer Umwelt, Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, Germany. FINO project. (Available from: http://fino.bsh.de/) (Accessed 1st December 2014).
5. Andreas EL, Wang S. Predicting significant wave height off the northeast coast of the United States. *Ocean engineering* 2007; **34**(8): 1328–1335.
6. Janssen PAEM. Progress in ocean wave forecasting. *Journal of Computational Physics* 2008; **227**(7): 3572–3594.
7. Booij N, Ris RC, Holthuijsen LH. A third-generation wave model for coastal regions: model description and validation. *Journal of Geophysical Research* 1999; **104**(C4): 7649–7666.
8. Reikard G, Pinson P, Bidlot J-R. Forecasting ocean wave energy: the ECMWF wave model and time series methods. *Ocean Engineering* 2011; **38**(10): 1089–1099.
9. Ma G, Shi F, Kirby JT. Shock-capturing non-hydrostatic model for fully dispersive surface wave processes. *Ocean Modelling* 2012; **43-44**: 22–35.
10. Allard R, Rogers E, Martin P, Jensen T, Chu P, Campbell T, Dykes J, Smith T, Choi J, Gravois U. The US Navy coupled ocean-wave prediction system. *Oceanography* 2014; **27**(3): 92–103.
11. Deo MC, Naidu CS. Real time wave forecasting using neural networks. *Ocean Engineering* 1998; **26**: 191–203.
12. Jain P, Deo MC. Real-time wave forecasts off the western Indian coast. *Applied Ocean Research* 2007; **29**(1): 72–79.
13. Soares CG, Ferreira AM, Cunha C. Linear models of the time series of significant wave height on the southwest coast of Portugal. *Coastal Engineering* 1996; **29**(1): 149–167.
14. Zamani A, Azimian A, Heemink A, Solomatine D. Wave height prediction at the Caspian Sea using a data-driven model and ensemble-based data assimilation methods. *Journal of Hydroinformatics* 2009; **11**(2): 154–164.
15. Woodcock F, Greenslade DJM. Consensus of numerical model forecasts of significant wave heights. *Weather and Forecasting* 2007; **22**: 792–803.
16. Durrant TH, Woodcock F, Greenslade DJM. Consensus forecasts of modeled wave parameters. *Weather and Forecasting* 2009; **24**: 492–503.
17. Hall JN. Use of risk analysis in North Sea projects. *International Journal of Project Management* 1986; **4**(4): 217–222.
18. Trbojevic VM, Bellamy LJ, Brabazon PG, Gudmestad T, Rettedal WK. Methodology for the analysis of risks during the construction and installation phases of an offshore platform. *Journal of Loss Prevention in the Process Industries* 1994; **7**(4): 350–359.

19. Feuchtwang J, Infield D. Offshore wind turbine maintenance access: a closed form probabilistic method for calculating delays caused by sea state. *Wind Energy* 2012.
20. Dinwoodie I, Catterson VM, McMillan D. Wave height forecasting to improve off-shore access and maintenance scheduling. *IEEE Power and Energy Society General Meeting* 2013: 1–5. DOI: 10.1109/PESMG.2013.6672438.
21. Brown RG. *Statistical Forecasting for Inventory Control*. McGraw-Hill, 1959.
22. Holt CC, Modigliani F, Muth JF, Simon HA. *Production Planning, Inventories, and Workforce*. Prentice Hall: New York, 1960.
23. Gardner ES. Exponential smoothing: the state of the art. *Journal of forecasting* 1985; **4**(1): 1–28.
24. Makridakis S, Hibon M, Moser C. Accuracy of forecasting: an empirical investigation. *Journal of the Royal Statistical Society. Series A (General)* 1979: 97–145.
25. Makridakis S, Andersen A, Carbone R, Fildes R, Hibon M, Lewandowski R, Newton J, Parzen E, Winkler R. The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting* 1982; **1**(2): 111–153.
26. Montgomery DC, Johnson LA, Gardiner JS. *Forecasting and Time Series Analysis*. McGraw-Hill: New York, 1990.
27. R Development Core Team. R: a language and environment for statistical computing, R foundation for statistical computing, Vienna, Austria, 2005. (Available from: http://www.R-project.org), ISBN 3-900051-07-0 (Accessed 1st December 2014).
28. Trigg DW, Leach AG. Exponential smoothing with an adaptive response rate. *OR* 1967; **18**(1): 53–59.
29. De Boor C. *A practical guide to splines*, Applied Mathematical Sciences, vol. 27. Springer-Verlag: New York, 1978.
30. Harvey A, Koopman SJ. Forecasting hourly electricity demand using time-varying splines. *Journal of the American Statistical Association* 1993; **88**(424): 1228–1236.
31. Cai Z, Fan J, Yao Q. Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* 2000; **95**(451): 941–956.
32. Lin CJ, Chen HF, Lee TS. Forecasting tourism demand using time series, artificial neural networks and multivariate adaptive regression splines: evidence from Taiwan. *International Journal of Business Administration* 2011; **2**(2): 14.
33. Harrison J, West M. *Bayesian Forecasting & Dynamic Models*. Springer, 1999.
34. Congdon P. *Bayesian Statistical Modelling*. Wiley, 2007.
35. Box GEP, Jenkins GM. *Time Series Analysis Forecasting and Control*. McGraw-Hill: San Francisco, London, Holden-Day, 1970.
36. Hill D, McMillan D, Bell K, Infield D, Ault GW. Application of statistical wind models for system impacts. *Proceedings of the 44th International Universities Power Engineering Conference (UPEC)*, 2009; 1–5.
37. McMillan D, Ault GW. Condition monitoring benefit for onshore wind turbines: sensitivity to operational parameters. *IET Renewable Power Generation* 2008; **2**(1): 60–72.
38. Castro Sayas F, Allan RN. Generation availability assessment of wind farms. *IET Generation, Transmission and Distribution* 1996; **143**(5): 507–518.
39. Pinson P, Madsen H. Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models. *Journal of Forecasting* 2012; **31**(4): 281–313.
40. Billinton R, Allan RN. *Reliability Evaluation of Engineering Systems*. Plenum press: New York, 1983.
41. Hagen B, Simonsen I, Hofmann M, Muskulus M. A multivariate Markov weather model for O&M simulation of offshore wind parks. *10th Deep Sea Offshore Wind R&D Conference (DeepWind 2013 Energy Procedia 35)*, 2013; 137–147.
42. Cybenko G. Approximation by superpositions of a Sigmoidal function. *Mathematics of Control, Signals, and Systems* 1989; **2**(4): 303–314.
43. Mahjoobi J, Adeli Mosabbeb E. Prediction of significant wave height using regressive support vector machines. *Ocean Engineering* 2009; **36**(5): 339–347.
44. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Statistics and Computing* 2004; **14**: 199–222.
45. Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification, 2010. (Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm) (Accessed 1st December 2014).
46. Barai SV, Reich Y. Ensemble modelling or selecting the best model: many could be better than one. *AI EDAM* 1999; **13**(5): 377–386.
47. Clemen RT. Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting* 1989; **5**(4): 559–583.