Probing NWP model deficiencies by statistical postprocessing

Rosgaard, Martin Haubjerg; Nielsen, Henrik Aalborg; Nielsen, Torben S.; Hahmann, Andrea N.

Published in:
Quarterly Journal of the Royal Meteorological Society

Link to article, DOI:
10.1002/qj.2705

Publication date:
2016

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
The objective in this article is twofold. On one hand, a Model Output Statistics (MOS) framework for improved wind speed forecast accuracy is described and evaluated. On the other hand, the approach explored identifies unintuitive explanatory value from a diagnostic variable in an operational numerical weather prediction (NWP) model generating global weather forecasts four times daily, with numerous users worldwide. The analysis is based on two years of hourly wind speed time series measured at three locations; offshore, in coastal and flat terrain, and inland in complex topography, respectively. Based on the statistical model candidates inferred from the data, the lifted index NWP model diagnostic is consistently found among the NWP model predictors of the best performing statistical models across sites.

Key Words: NWP model development; statistical forecasting; general linear modelling; Model Output Statistics; backward elimination; backward stepwise selection; Bayesian Information Criterion; wind energy scheduling

1. Introduction

Modern-day studies of the atmosphere heavily rely upon data produced by numerical weather prediction (NWP) models and the study of systematic deviations of weather simulations from the observed atmosphere contributes to the continued improvement of NWP model formulations, which in turn better enables researchers to make candid statements on issues relating to climate and weather. Many forecast users worldwide depend critically on the accuracy of operational NWP models, with implications for a diverse range of applications from natural disaster risk assessment to efficient integration of renewable energy sources in the electrical grid.

As the system of equations describing the spatio-temporal continuum of observed atmospheric dynamics must be discretised in practice to enable temporal integration such that future atmospheric states can be inferred, Judd et al. (2008) argue that NWP models can be viewed as occupying a phase space of much lower dimensionality than the observed atmosphere. Leith (1974) distinguishes the subspace of inadequate representation of atmospheric dynamics on numerically unresolved scales as a separate issue, further connoted e.g. by Holton and Hakim (2013) into key physical process classes pertaining to 1) radiation, 2) clouds and precipitation, and 3) turbulent mixing and exchange, respectively. Within the insight gained from the pioneering work of Courant et al. (1928, 1967), among others, Charney (1947, 1949; 1949) recognised, from the very outset of NWP on electronic computers, the need to filter solutions to the coupled nonlinear partial differential equations describing the temporal evolution of the atmosphere. Inadequate representation of dynamics on the subgrid scale in NWP models, as well as filtering of solutions to the governing equations for atmospheric dynamics, may spur systematic deviation between simulated time series for atmospheric quantities and corresponding observations. Furthermore, each numerical discretisation scheme has implications for prediction accuracy and numerical stability, e.g. Bénard (2003) finds stability properties of time integration schemes commonly employed for NWP to be highly dependent on fundamental choices, such as the vertical coordinate.

This article concerns detection of systematic coupling of various NWP modelled quantities to atmospheric flow prediction, and to this end outlines a postprocessing methodology for assessing explanatory value within a Model Output Statistics (MOS) framework, cf. Glahn and Lowry (1972). Systematic error in NWP models is usually quantified in terms of forecast variable bias, which is formally defined as the expectation of the forecast error; see e.g. Dee and Da Silva (1998) for a description of a theoretical framework for bias estimation. In practice, bias is estimated as the average error for a sufficiently long historical dataset of NWP model simulations and observations, while Dee (2005) relaxes the definition somewhat by accepting running-mean error approaches for short time periods as estimators of bias. E.g. Hacker and Rife (2007) adopt this definition in constructing and evaluating a running-mean algorithm for estimation of systematic errors for near-surface temperature on a limited area NWP model (LAM) grid. For another LAM, Sweeney and Lynch (2011) compare running-mean window lengths spanning the previous 3 and 28 days, respectively, to a Kalman (1960) filter approach, and a weighted combination of the three, as statistical postprocessing methods for wind speed NWP data. In a subsequent study, Sweeney et al. (2013) increase the resolution of the LAM domains and evaluate performance of seven different postprocessing approaches, including linear regression (MOS), running-mean bias correction, Kalman filtering, artificial neural networks, as well as forecasting techniques involving wind direction dependency, diurnal effects, and mean and
variance nudging. They find none of these superior, and that it is difficult to improve performance beyond simple running-mean bias correction. Stensrud and Yussouf (2005) find bias correction forecast performance competitive or better than MOS for their NWP multimodel ensemble approach, using data from the previous 12 days for bias correction, though Cheng and Steenburgh (2007) arrive at general MOS superiority in their comparison to running-mean and Kalman filter techniques, using data from 145 observation stations scattered across the western United States of America, except during a persistent cold air pool event in February 2005 where the Kalman filter approach is found to perform better. In the study by Hart et al. (2004) MOS forecasts based on coarsely resolved LAM data outperform high-resolution data from the same LAM configuration. Thorarinsdottir and Gneiting (2010) introduce wind speed as response variable in extension to the nonhomogeneous Gaussian regression forecasting approach first suggested by Gneiting et al. (2005), and find substantial improvement relative to reference forecasts.

Pinson (2012) proposes another single-model NWP ensemble MOS approach for wind forecasting, namely adaptive and recursive Maximum Likelihood (ML) estimation of parameters in bivariate models for wind component ensemble mean and variance, and demonstrates better deterministic and probabilistic scores upon applying the method on a three-year dataset. Wilks (2015) describes a framework for multivariate ensemble MOS using empirical copulas and examines how recently proposed variations of these nonparametric functions linking multivariate distribution functions to their constituent univariate marginal distributions perform when the method is applied for probabilistic forecast wave forecasting. Statistical postprocessing that accounts for multivariate joint forecast distributions is not pursued here, albeit this would be a natural extension of the study presented in the event the problem extension or reformulation at hand warrants a multivariate approach.

The article is organised as follows. Section 2 outlines the dataset on which the general linear modelling framework presented in Section 3 is applied. In particular, Section 3 specifies the underlying rationale for each step in the construction of statistical model candidates. Section 4 describes the findings, and a discussion of results and conclusions follow in Section 5.

2 Data

Kanamitsu (1989) describes an early version of the Global Forecast System (GFS), one in a handful spectral NWP models used to generate global forecasts operationally, developed and maintained by the National Centers for Environmental Prediction (NCEP) in the United States of America. Four GFS forecast initialisations were available daily on a 0.5° global grid during the time period studied, downloadable with a few hours delay relative to forecast initialisation. Since the model upgrade 27 July 2010 the time integration of the governing equations was carried out at horizontal resolution T574, which roughly corresponds to a 27 km horizontal distance between computational grid points, though 14 January 2015 the horizontal resolution of operational GFS forecasts was upgraded to T1534 (13 km), and the number of vertical model layers increased from 64 to 90. Forecasts at the locations studied in this work are operational T574L64 GFS datasets spanning October 2012 through September 2014, and hence derive from the largely static model formulation available in a realtime setting during the time period 5 September 2012, where a land surface model bug fix was implemented, up till 14 January 2015. Results presented in Section 4 thus reflect forecast improvement achievable in practice in an operational setting. A two-year dataset is also considered e.g. in Cassola and Burlando (2012); Sweeney et al. (2013); two studies related to this work.

Geographical locations of the three wind farm sites studied are sketched in Figure 1. The available wind speed reference data underlying results presented in Section 4 consists of the spatial average of 5 minute instantaneous wind speed time series recorded at the wind turbine nacelles of the Horns Rev, Rejsby Hede and Stor-Rotliden wind farms, respectively, further averaged to hourly values. That is; the hourly wind speed forecast reference datasets represent \( \sim 1 \times 2 \text{ km}^2 \) (Rejsby Hede) and \( \sim 4 \times 5 \text{ km}^2 \) (Horns Rev and Stor-Rotliden) horizontal wind speed fields, respectively. An overview of the dataset is given in Table I.

Forecasts wind components and temperature 80 m above ground level (AGL) are used for Horns Rev as this is near the height above mean sea level of the wind turbine anemometers recording the wind speed, while 100 m AGL forecasts are relevant for comparison to Stor-Rotliden data. Rejsby Hede wind turbine hub-height is 45 m AGL so 10 m AGL and 80 m AGL wind speed forecasts are vertically interpolated to this height based on the logarithmic wind profile relation according to Monin (1970), assuming neutral atmospheric stability conditions and vanishing.

Table I. Statistics for wind speed measurement time series used for model training and evaluation at Horns Rev (HR), Rejsby Hede (RH), and Stor-Rotliden (SR). Wind farm acronyms are followed by the year number, separated by slash. Year 1: October 2012 till September 2013. Year 2: October 2013 till September 2014. Complete pairs refers to the percentage of datapoints for which both observation and forecast is nonmissing. Other quantities are in m s\(^{-1}\). Wind speed time series mean is denoted \( \bar{\mu} \), the median \( \bar{\sigma} \), and standard deviation \( \bar{\sigma} \). The usual meteorological bias sign convention applies, i.e. forecast minus observation.

<table>
<thead>
<tr>
<th>Site/ year</th>
<th>Complete pairs [%]</th>
<th>OBSERVATIONS</th>
<th>FORECASTS</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \bar{\mu} )</td>
<td>( \bar{\sigma} )</td>
<td>( \bar{\sigma} )</td>
<td>( \bar{\sigma} )</td>
</tr>
<tr>
<td>HR/1</td>
<td>96.0</td>
<td>8.55</td>
<td>8.25</td>
<td>3.74</td>
</tr>
<tr>
<td>HR/2</td>
<td>97.7</td>
<td>9.30</td>
<td>9.07</td>
<td>4.25</td>
</tr>
<tr>
<td>RH/1</td>
<td>92.4</td>
<td>5.87</td>
<td>5.35</td>
<td>3.19</td>
</tr>
<tr>
<td>RH/2</td>
<td>95.4</td>
<td>6.46</td>
<td>5.98</td>
<td>3.51</td>
</tr>
<tr>
<td>SR/1</td>
<td>93.5</td>
<td>6.55</td>
<td>6.46</td>
<td>2.39</td>
</tr>
<tr>
<td>SR/2</td>
<td>92.6</td>
<td>6.67</td>
<td>6.43</td>
<td>2.76</td>
</tr>
</tbody>
</table>

This article is protected by copyright. All rights reserved.
zero plane displacement (the terrain is flat cropland with a few scattered patches of trees and houses nearby). Wind direction and temperature are linearly interpolated. Forecast data from the closest four computational grid points is bilinearly interpolated horizontally to the coordinates of the three wind farm centres, respectively. GFS predictions of pressure, the friction velocity at the surface, and measures of air mass rotation in the atmospheric boundary layer and the rest of the troposphere are included in the analysis, as well as heat fluxes, different indicators of atmospheric stability, and variables relating to humidity and precipitation. Also, a scalar estimate of the local geopotential height gradient at 700 mb, quantified by the sum of gradient magnitudes in the two horizontal dimensions, is included thus coupling simulated geostrophic wind to the surface wind modelled. In this respect, the geopotential height anomaly – here understood to be the deviation from the average over time, cf. e.g. Yuan et al. (2008) – should as well potentially contribute explanatory value locally. The surface pressure tendency is closely coupled to this quantity, cf. e.g. Steenburgh and Holton (1993), and is used in the model development, alas the available GFS output is temporally coarse.

3. Method

The MOS approach outlined in this section is based on general linear modelling which involves mathematically relating one (univariate modelling), two (bivariate) or several (multivariate) dependent variable(s) mathematically to one or more independent variables assumed to explain variation observed for the dependent variable(s). The terminology of denoting a dependent variable the model response is adopted here, as employed e.g. in Thorarinsdottir and Gneiting (2010) as well as in Hastie et al. (2009); Olsson (2001); Madsen and Thyregod (2010). Alternatively, a dependent variable is also known as predictand in the literature, e.g. in Termonia and Deckmyn (2007); Wilks (2011); Warner (2011), or outcome in Pawitan (2001), while an explanatory variable is usually termed regressor, denoting continuous data only, or predictor which is a broader term also denoting categorical data. In contrast to the case of deterministic models, such as weather and climate models, statistical models account for the possibility that a relationship is not perfectly as modelled by allowing for unexplained (stochastic) response variation \( \varepsilon \) for which independence across observations and a distribution is typically assumed for model coefficient estimation and inference purposes. Deviation between statistical model predictions from corresponding observations are usually termed model residuals, denoted \( \varepsilon \) in this article.

In the present context, the basic premise is that measured wind speed can be univariately modelled as linearly dependent on additive forecast variable interaction terms. That is; in a multiple linear regression framework the response is related to a sum of predictor terms, each with an unknown coefficient to be estimated analytically by the method of ordinary least squares (OLS) based on a sufficiently large dataset of response and predictor variables. Statistically modelled quantities are in upper case to distinguish these stochastic variables from their corresponding realisations, which are in lower case. The study is based on the general linear model for hourly wind speed \( W \).

\[
W = w_{LLS}(x) + \varepsilon, \quad \text{with} \quad \Theta(x) = \begin{bmatrix} 1 & x^T \end{bmatrix} \theta, \hspace{1cm} (1)
\]

where \( w_{LLS} \) is forecasted wind speed after linear correction cf. the prediction equation Eq. (2), \( \varepsilon \) denotes the model error, \( \theta \) is a column vector containing the coefficients to be estimated based on NWP data and measurements \( w \) for the sites studied, cf. Section 2, and \( x \) is a column vector containing quantities derived from NWP modelled quantities. Termonia and Deckmyn (2007) propose to infer viable predictors based on direct inspiration from NWP formulation and subgrid scale parameterisations, yet in this work the choice of \( x \) elements in Eq. (1) is more vaguely guided by physical intuition, thus placing the MOS predictor range considered here somewhere in between the mechanistic-inspired strategy of Termonia and Deckmyn (2007) and purely empirical (data mining) statistical modelling, cf. e.g. Pawitan (2001) p. 5 on the respective merits of scientific formula and empirical models. Physical characteristica of \( x \) element candidates are briefly outlined last in Section 2; n.b. \( x \) is understood to represent continuous as well as derived categorical (binary only) variables, coded as dummy variables cf. e.g. Olsson (2001) pp. 12–21. The central predictor of Eq. (1), \( w_{LLS} \), denotes forecasted wind speed after removal of bias and scaling errors, as recommended by Fritsch et al. (2000) and employed e.g. in Woodcock and Engel (2005); Engel and Ebert (2007, 2012), i.e. \( w_{LLS} \) data points are linear least square (LLS) predictions

\[
w_{LLS} = \begin{bmatrix} 1 & w \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}, \hspace{1cm} (2)
\]

with raw forecasted wind speed \( w \), cf. e.g. Sweeney et al. (2013) from which the LLS acronym is adopted in this article. With all additive terms proportional to \( w_{LLS} \) the Eq. (1) modelled wind speed approaches zero when \( w_{LLS} \) approaches zero. While Eq. (2) formally does not guarantee \( w_{LLS} \) to be non-negative, it is unlikely that such values will occur with reasonable NWP model forecasts since both forecast and observation are positive, as well as positively correlated.

In contrast to the binning approach employed in Sweeney et al. (2013), forecasted wind direction is here implemented in Eq. (1) as Fourier expansion terms.

\[
f(w_d) = \sum_{k=1}^{N} a_k \cos(kw_d) + b_k \sin(kw_d), \quad w_d \in (-\pi, \pi) \hspace{1cm} (3)
\]

up to 5th order; \( N = 5 \), to approximate wind speed dependence on a periodic function of forecasted wind direction \( w_d \). Parameters \( a_k \) and \( b_k \) are estimated as \( \theta \) coefficients in Eq. (1), i.e. \( \cos(kw_d) \) and \( \sin(kw_d) \) are \( x \) elements, respectively.

3.1. Forecast performance assessment

One of the two key purposes of the study is to explore the extent to which NWP model forecasts can be postprocessed empirically to advance efficient operational integration of wind energy in the electrical grid, while the other concerns interpretation of the results to infer NWP model deficiencies. Madsen et al. (2005) discuss two commonly employed scalar metrics for wind power prediction accuracy assessment, namely the mean absolute error (MAE) and the root mean square error (RMSE), each addressing slightly different aspects of forecast accuracy. As the expansion of the OLS acronym suggests, the procedure for estimation of \( \theta \) elements in Eq. (1) involves minimisation of the squared residuals \( \varepsilon^2 \). Hence, the modelling framework is tuned towards RMSE minimisation. Compared to RMSE, the MAE accuracy metric is less sensitive to severe forecast errors and is therefore arguably more relevant if the assessment should reflect normal operational conditions for day ahead wind energy scheduling. When used for performance evaluation along with RMSE, the MAE constitutes a baseline for inference on the variation in the magnitude of errors. Since both systematic and random errors contribute to MAE and RMSE, these metrics are sensitive to NWP model bias in raw (unprocessed) forecasts. In contrast, the absolute value of the Pearson correlation coefficient (PCC) is invariant under linear transformations of one or both constituent variables and hence not sensitive to bias and scaling errors.
As mentioned, RMSE and, to a lesser extent, MAE are commonly used for wind speed and power forecast performance assessment, cf. e.g. Madsen et al. (2005); Pinson (2012); Sweeney et al. (2013); Müller (2011), though e.g. Ranaboldo et al. (2013) also include the PCC for evaluation of wind power predictions. Although rarely employed in wind energy forecasting, the inclusion of the PCC as estimator of forecast performance for the statistical models developed here – alongside MAE and RMSE – is indirectly motivated by the work of Murphy (1988). He meticulously highlights the deficiency of the PCC squared as a weather forecast assessment metric when used by itself, in that he demonstrates that forecast skill (in terms of mean square error) can be broken into a sum of the squared PCC and two bias-sensitive terms. Murphy (1988) argues that the PCC squared resembles potential skill attainable post elimination of systematic errors, while bias-sensitive performance metrics such as the RMSE and MAE describe actual performance. Due to wind speed forecast bias-removal cf. Eq. (2) as a preprocessing step for modelling by Eq. (1), RMSE and PCC will exhibit the same relative performance between Eq. (1) model candidates studied for a given site, yet between sites the balance between “potential skill indicator” PCC and actual performance in terms of MAE/RMSE approximately emphasises differences in atmospheric flow distortion at each site, a feature in part subject to the local terrain due to the distinct surroundings of the offshore, coastal and inland sites selected in the experimental design. This picture is approximate as the interpretation above is subject to viewing \( \Theta(\bar{x}) \) in Eq. (1) as an ordinary linear model coefficient, i.e. disregarding the \( x \)-dependence. It then follows from the definition of the coefficient of determination for Eq. (1) which is in fact equal to the PCC squared and usually denoted \( R^2 \) cf. e.g. Olsson (2001) p. 5 – that

\[
\sum_{i=1}^{N} e_i^2 = \left(1 - R^2\right) \sum_{i=1}^{N} (w_i - \bar{w})^2, \tag{4}
\]

in which RMSE squared times the number of time series data points represents the left hand side, and the sum on the right hand side represents the variation in wind speed measurements, i.e. \( \bar{w} \) denotes the average of wind speed measurements. The superscript \( \hat{\cdot} \) is henceforth denotation in the following.

Hence, by including PCC alongside RMSE (and MAE) in the results pertaining to wind speed, presented in Figure 6, the reader is indirectly reminded of the site statistics of Table I, namely \( \delta \). However, with regard to wind power only RMSE and MAE are applied in Table III, as is customary cf. Table I.

3.2 Model development strategy

The dataset available spans two years, cf. Section 2. Where e.g. Müller (2011); Sweeney and Lynch (2011) use moving window recursive approaches for one-year datasets, the first year of data is here reserved for model development, while data for the second year is solely used to assess performance of the final wind speed model candidates based on the structure outlined in Eq. (1) – a special case of the cross validation technique, cf. e.g. Efron and Gong (1983). This strategy substantiates out-of-sample any systematic wind speed forecast error coupling to GFS variables identified in-sample. Furthermore, all seasons are represented in both of the datasets for model development and evaluation, respectively, and the results presented are therefore more generic than if a subyearly period, sensitive to seasonal effects, had instead been analysed. The dataset is binned according to forecast leadtime groups of width equal to the time period between NWP model forecast initialisations. Data within each forecast leadtime group is then augmented such that all wind speed measurement data points occur exactly once in each group. In this way all hours of the day are represented in each forecast performance metric evaluation and inference on forecast performance leadtime dependency is thus readily assessed.

As mentioned in Section 2, forecasts are issued four times daily, and wind energy scheduling typically concerns the forecast leadtime horizon from 12 to 48 hours ahead, depending on application requirements and the latency with which forecast data can be obtained. E.g. Messner et al. (2013) consider forecast leadtime hours 12 to 48 for studying forecasted wind speed transformation to wind turbine power production, a topic briefly considered in Section 4.4. To simplify execution of the final, empirical model development step explained in Section 3.5 only two daily forecast initialisations are included in the model development phase, namely Coordinated Universal Time (UTC) midnight and noon forecast initialisations. Thus, for each site Eq. (1) moulding cf. sections 3.4 and 3.5 is carried out separately for the three sites considered and forecast leadtime hour groups 12 ↔ 24, 24 ↔ 36 and 36 ↔ 48, respectively. In this way all wind speed measurements are represented exactly once within each of the nine respective datasets, hence maximising the power and validity of the statistical procedures. For the evaluation of the NWP postprocessing framework in Section 4 all four daily GFS initialisations are employed, and forecast leadtime hour groups 00 ↔ 06, 06 ↔ 12, 12 ↔ 18, 18 ↔ 24, 24 ↔ 30, 30 ↔ 36, 36 ↔ 42, and 42 ↔ 48 are considered.

Before carrying out the Eq. (1) development steps described in sections 3.4 and 3.5 separately for the nine datasets, the GFS-derived quantities included in the analysis, cf. the last paragraph of Section 2 and the Eq. (1) description first in Section 3, are preliminarily screened for explanatory value subject to the Eq. (1) framework in Section 3.3. As noted above, the procedure outlined in sections 3.3, 3.4, and 3.5 is based on year 1 data, cf. Table I.

3.3 Preparatory screening of predictor candidates

This section describes a few diagnostic steps to be carried out for the dataset used for model structure identification and coefficient estimation (year 1; Table I) in order to preprocess the numeric predictor candidates for modelling. In accordance with the model development strategy outlined in Section 3.2, the quantification of the \( w_{LLS} \) data used in Eq. (1) model development is based on forecast data spanning leadtime hours 12 to 48. That is, Eq. (2) predictions are based on data for GFS forecast leadtime hours 12 to 48 in this section.

Following a standard regression model building approach one may choose to plot the deviation between \( w_{LLS} \) and observed wind speed \( w \) against potential additional explanatory variables, i.e. \( w_{LLS}x \) terms — n.b. that the quantity \( w - w_{LLS} \) is henceforth referred to as \( LLS \) error. However, each Eq. (1) model predictor can be written \( w_{LLS}(x - \bar{x}) + w_{LLS}\bar{x} \) of which the latter term expresses variation purely due to \( w_{LLS} \). Therefore, the LLS errors are instead plotted against \( w_{LLS}(x - \bar{x}) \). In the event a linear relation appears a plausible approximation to the underlying GFS intervariable coupling to wind speed, the corresponding Eq. (1) term is kept unchanged as a quantitative predictor. Local linearity within regimes divided by a \( w_{LLS}x \)-threshold is accommodated by adding separate linear terms that mutually excludes one-another depending on a threshold value. That is, depending on the value of \( w_{LLS}x \) one or the other term is active, always one of them and never both. Finally, if no linear relation is evident, though a regime-like distribution can be inferred from the scatter density plot, or perhaps more clearly from a histogram for the numeric values of the \( w_{LLS}x \) term, the predictor candidate is included as a categorical variable with regime-dividing values inferred from the scatter plot, cf. \( w_{LLS}x \) regime-dependent linear relationships discussed before. In case the threshold value differs considerably
for the three reference data sites studied the predictor term is discarded to keep the methodology generic across terrain types. Continuous NWP modelled quantities describe different aspects of the atmosphere via a set of coupled partial differential equations and subgrid scale parameterisations and are hence correlated to some extent. The degree of multicollinearity in the Eq. (1) model is examined by inspection of the PCC matrix for the predictor terms, which provides a rough estimate of the extent to which redundant information is included in the model. Since the PCC is only sensitive to linear relationships, the interpredictor PCC matrix is supplemented by inspection of interpredictor scatter density plots in order to infer whether systematic nonlinear relations between \( w_{LLS} \) terms are present. However, as noted e.g. by Williams et al. (2013) the adverse effects of multicollinearity are typically not severe in a prediction context, as pertains to this work.

3.4. Model reduction based on Maximum Likelihood criteria

Eq. (1) predictor terms containing the least explanatory value are removed sequentially according to a stepwise process known as backward elimination, e.g. in Wilks (2011), or backward stepwise selection in Hastie et al. (2009). The backward elimination strategy is employed as the full model is thus subject to study, in contrast to the case for forward stepwise selection, where each predictor is added and tested sequentially. Note that both techniques have advantages and drawbacks, depending on the modelling task and the dataset at hand. Cf. Section 3.3, scatter density plots of LLS error against \( w_{LLS} \) terms can indicate linear coupling for certain variables, while others may look less promising as explanatory variables in the assumed response-relationship structure Eq. (1). The objective is then to identify and exclude redundant information to avoid suboptimal and inconsistent model coefficient estimates, as discussed e.g. in Birgé (2006).

The Akaike Information Criterion (AIC), first proposed by Akaike (1974), and the Bayesian Information Criterion (BIC), Schwarz (1978), are two penalised ML criteria, both dependent on the number of coefficients to estimate and the latter also explicitly depends on the number of data points available for estimation;

\[
AIC = -2\ell(W) + 2n, \quad BIC = -2\ell(W) + n \log(N),
\]

where \( \ell(W) \) denotes the maximised value of the log-likelihood function for the model specified by Eq. (1), \( n \) is the number of coefficients \( \theta \) to be estimated, and \( N \) is the number of observations. As e.g. pointed out in Olsson (2001) p. 3, OLS estimation of Eq. (1) coefficients corresponds to ML estimation if the errors \( \varepsilon \) are assumed independent and zero-mean normal distributed, which in turn enables the AIC and BIC as Eq. (1) model reduction tools. Results discussed in Section 4.2 motivate the Gaussian-\( \varepsilon \) assumption, while the issue of finite autocorrelation in part motivates the choice of backward elimination scheme, cf. the next paragraph. The AIC and BIC are common tools for model selection within various scientific disciplines, see e.g. Kuha (2004); Alfaro and Huelsenbeck (2006); Jones (2011); Vrieze (2012), while alternatives such as bootstrap criteria, cross validation criteria, Mallow’s \( C_p \), etc. are also found in the literature, see e.g. Hastie et al. (2009); Zucchini (2000); Bultheel et al. (2013); Birgé (2006). Vrieze (2012) argues that the BIC is consistent under certain assumptions, i.e. this criterion will asymptotically select the “true” model, given that this is a subset of the full model space investigated, whereas the AIC is not consistent. Hastie et al. (2009) p. 135 notes that the BIC is more strict in that it penalises the inclusion of redundant model predictors more severely than the AIC. However, given that the full true model is not a part of the model space studied the AIC is said to be efficient in that it will asymptotically choose the model that minimises the mean square error, while the BIC is not efficient if the true model is not a subset of the investigated model space. Model selection based on the AIC is always subject to the risk of selecting too large a model (i.e. with redundant information) regardless of the number of observations in the model development dataset, while the corresponding risk when applying the BIC is very small given a sufficient number of observations. On the other hand, the risk of selecting too small a model is higher with the BIC than with the AIC, regardless of the number of observations. The merits of the two model selection criteria can roughly be summarised in the following points.

- If a false negative finding is considered more misleading than a false positive, the AIC is preferable.
- If a false positive finding is considered more misleading than a false negative, the BIC is preferable.

Inclusion of redundant information in the coefficient estimation process can lead to poor predictive ability for the Eq. (1) model candidates when applied for postprocessing of future NWP data. Furthermore, measured wind speed and NWP modelled quantities are time series of smoothly varying atmospheric dynamics aspects and as such exhibit finite autocorrelation, which implies that Eq. (1) model errors are likely not chronologically independent. Therefore the effective number of (independent) data points is lower than the actual number of data points \( N \). Since the width of confidence intervals for Eq. (1) coefficient estimates scales inversely with \( \sqrt{N} \), the test statistic used to accept/reject the null-hypothesis stating non-value for individual predictor terms is prone to attribute significance (rejection of null-hypothesis) to terms for which the test statistic would otherwise infer non-significance, subject to the effective number of independent observations. Thus, more predictor terms may survive the backward elimination process than warranted according to the assumption of independent observations. The BIC employs stricter predictor-penalisation than the AIC and the former is therefore preferred for backward elimination of Eq. (1) predictor terms. With the assumption of zero-mean Gaussian Eq. (1) errors \( \varepsilon \) with variance \( \sigma^2 \), Hastie et al. (2009) p. 233 note that the BIC can be expressed in terms of the Eq. (1) residual mean \( \tilde{\sigma} \).

\[
BIC = \frac{N\tilde{\sigma}^2}{\sigma^2} + n \log(N).
\]

Since the BIC, although more restrictive than the AIC, may yet keep too near-redundant predictor terms in Eq. (1), the model reduction process is extended with additional model selection criteria inferred from the models resulting from BIC reduction.

The BIC model reduction process allows for elimination of the standalone \( w_{LLS} \) term in Eq. (1), in which case the predictor is reinstated posthumously since the quantity corresponds directly to the model response. Similarly, not all Eq. (1) wind direction terms, cf. Eq. (3), are found significant post BIC model reduction, and all wind direction terms up to the highest Fourier expansion order surviving the BIC reduction process are reinstated if eliminated during the process.

3.5. Further model reduction based on empirical assessment

Statistical significance of a predictor identified cf. the Eq. (1) model development approach outlined in Section 3.4 is no guarantee for notable influence on model predictions. E.g. the discussed issue of model residual autocorrelation may lead to exaggerated significance attribution for some predictors, which in turn increases the risk of sustaining a predictor set subject to severe multicollinearity. Hence, the Eq. (1) model development...
This section describes the assessment of the models resulting from carrying out the Eq. (1) predictor reduction steps of Section 3 for the dataset presented in Section 2. As prescribed in Section 3.3, the exploratory data analysis outlined in Section 4.1 preprocesses the data for the backward elimination of Eq. (1) predictor terms. The model predictor reduction process described in sections 3.4 and 3.5 is executed in Section 4.2, leading to the final model candidates evaluated in sections 4.3 and 4.4.

Note that sections 4.1 and 4.2 concern only 0000 and 1200 UTC forecast initialisations pertaining to year 1 of Table I, cf. the model development strategy laid out in Section 3.2, while all four daily GFS forecast initialisations are used in sections 4.3 and 4.4.

### 4.1 Exploratory data analysis prior to statistical modelling

LLS error scatter density plots cf. Section 3.3 constitute approximate guidance as to whether a predictor candidate should be 1) excluded, 2) included as a binary categorical variable, further substantiated by predictor term histograms cf. Section 3.3, 3) as a x-regime dependent term, or 4) as an ordinary regressor. Figure 2 includes examples of GFS-derived quantities processed as categorical, two-regime regressors, and ordinary continuous regressors prior to Eq. (1) predictor reduction based on the BIC carried out in Section 4.2.

1) The thermal buoyancy, and hence atmospheric stability, in the lowest part of the boundary layer arguably depends on variables such as daily sunshine duration and downward short wave radiation. Though based on scatter density plots cf. Section 3.3 these GFS quantities are excluded in the further analysis, as inclusion as neither categorical, regime-dependent linear, nor ordinary regressor appear suitable due to fairly evenly spread-out scatter with no hint of linear trends (not shown).

2) Figure 2a indicates for Horns Rev – as for the other two sites (not shown) – that convective available potential energy (CAPE) is likely not suitable for inclusion as continuous x in an Eq. (1) type predictor term, as no linear relation is evident from the scatter density plot. However, viewing the horizontal axis as a histogram – i.e. data points per $w_{LLS}x$ bin (not shown) – further highlights a tendency already apparent in Figure 2a for values to
be either concentrated around zero or with values far exceeding zero, with approximately half the data points belonging to each binary class. From a physical point of view, whether potential energy of a parcel of air is present is one indicator of atmospheric stability, with implications for the vertical wind profile, and hence the GFS variables CAPE and the related convective inhibition (CIN) are implemented as binary categorical variables in the modelling process of Section 4.2. Forecast variables related to precipitation are indicators of water phase shift in the atmosphere, potentially associated with weather fronts, and are processed in a similar manner due to CAPE/CIN-like scatter density plot traits. Respective data point counts in the two binary factor levels are ensured to be distributed somewhere in between 30–70% and 30–50%; if one factor level is sparsely populated the Eq. (1) model coefficient estimation becomes unreliable and crashes if the design matrix M becomes degenerate, cf. Eqs. (5–6).

3) Moving on to Eq. (1) predictor terms based on regime-dependent x, Figure 2b shows the scatter density plot for latent heat flux simulated by the GFS. This variable, along with the related sensible heat flux, has a natural threshold at zero; separating condensation from evaporation for the former and vertical propagation direction for the latter.

4) Finally, the GFS-simulated lifted index vs. LLS error displayed in Figure 2c do suggest a linear relation, in contrast to the case for variables like CAPE and latent heat flux. Similarly, the scatter density plots for all GFS-derived quantities appearing in Figure 3 indicate no processing necessary, and these are therefore included as continuous x in the Eq. (1) model.

Figure 3a quantifies the linear coupling between the various quantities predicted from the GFS formulation of the governing equations for atmospheric dynamics. As all Eq. (1) predictor terms involve proportionality to LLS wind speed severe multicollinearity is expected and verified in Figure 3b, namely for temperature and mean sea level pressure. Notice that these terms are, in fact, fully correlated with LLS wind speed to two digits precision — and are hence likely not adding value to Eq. (1) prediction accuracy, though neither deteriorating model performance in-sample cf. e.g. Williams et al. (2013).

Figure 3.b quantifies the linear coupling between the various quantities predicted from the GFS formulation of the governing equations for atmospheric dynamics. As all Eq. (1) predictor terms involve proportionality to LLS wind speed severe multicollinearity is expected and verified in Figure 3b, namely for temperature and mean sea level pressure. Notice that these terms are, in fact, fully correlated with LLS wind speed to two digits precision — and are hence likely not adding value to Eq. (1) prediction accuracy, though neither deteriorating model performance in-sample cf. e.g. Williams et al. (2013).
model residuals are computed for the three sites studied. If \( \sigma_e^2 \) is constant \( \sum_{i=1}^{N} \sigma_i^2 \) vs. the time stamp will have near-constant slope. As data for the offshore site, Horns Rev, exhibits the most severe sudden \( \epsilon^2 \) increases disrupting this picture, the cumulative \( \epsilon^2 \) for this site is shown in Figure 4, along with two examples of weather events underlying the \( \epsilon^2 \)-jumps; storms. Note that removing the two 24-hour storm periods results in what looks like a steady, constant-slope development which in turn implies homoscedasticity. However, removing the 4–5 events for which cumulative \( \epsilon^2 \) exhibits sudden increase with time in Figure 4c had no impact on the final model candidates arrived at later in this section, and the impact on OLS estimates for \( \theta \) elements of Eq. (2) only influences model predictions up to the fourth decimal in the Section 4.3 analysis.

As argued for in Section 3.5, additional model reduction upon having BIC-reduced the nine Eq. (1)-type models separately based on the nine datasets described last in Section 3.2, is likely necessary to achieve consistent results. Thus quantile ranges displayed in Figure 5c, and similar plots for the other eight datasets (not shown), are used in order to better secure that out-of-sample performance (second year data) is consistent with in-sample performance (first year data). Cf. the discussion in Section 3.5 this information is applied as a tool for direct assessment of explanatory value of predictor terms remaining post backward elimination processing of Eq. (1) for the nine datasets. The \( x\theta \) histogram width, cf. Figure 5b, quantifies the extent of influence on the model response. Note that in comparison with Figure 5a the lifted index \( \hat{\sigma}_x \) histogram is wider than the model response coefficient histogram. This apparent discrepancy can be understood in light of Figure 3a, exhibiting similar traits as the corresponding PCC matrix for the Stor-Rotliden site, in that specific humidity is inversely correlated with lifted index and hence acts to counter the stronger influence of the lifted index. Thus the specific humidity histogram (not shown) is enveloped by the histogram for lifted index and acts to “contract” the variation of the lifted index. The variation histogram for temperature is narrower than that for specific humidity, cf. Figure 5c, and centered in the far negative around \(-1.6\) (not shown). The lifted index variation, mainly of positive sign, is also balanced by the 4-layer lifted index (not shown), looking like a narrower mirror image around zero – i.e. of opposite sign – of the lifted index variation histogram.

The empirical reduction steps proposed in Section 3.5 and executed here conclude with recursively leaving one predictor out of the model in order to quantify the negative impact and construct a predictor-hierarchy of explanatory value. A clear example of the adverse effect of multicollinearity arises when including both pressure at the surface and at the mean sea level in the model development. For some of the nine datasets modelled, cf. the last paragraph in Section 3.2, both appear to be dominating predictors producing nine quantile range plots cf. Figure 5c, and in the final screening predictor terms are tested according to the hierarchy-listing approach sketched above in order to rule out false signals. As a final criteria, at least 0.1 variation within the 95% \( x\theta \)-quantile range is required for one or more of the nine datasets; see the dashed vertical line in Figure 5c. When multiplying \( x\theta \) with \( w_{LLS} \) this corresponds to a 95% quantile bandwidth of \( \sim 1\) m s\(^{-1}\) for LLS wind speeds around \( \sim 10\) m s\(^{-1}\). The resulting model is denoted Model 2 in Table II of Section 4.3.

![Figure 4](image-url)  
Figure 4. Quantitative illustration of how two poor storm event forecasts for Horns Rev challenge the assumption of unchanging \( \sigma_x \), which underlies the modelling framework. The times of the year when the two storm events, each of 24 hours duration, are removed from the dataset are marked on the horizontal axis in (c).

This article is protected by copyright. All rights reserved.
4.3. Model performance assessment

In contrast to the case during model development steps carried out in sections 4.1 and 4.2, all four daily GFS initialisations; 0000, 0600, 1200, and 1800 UTC, are included in Section 4.3. Hence, with inclusion of all forecast leadtime hours up to 48 hours, the considered forecast leadtime hour groups are, cf. Section 3.2:

-0.00 → 6, 06 → 12, 12 → 18, 18 → 24, 24 → 30, 30 → 36, 36 → 42, and 42 → 48, respectively. In this section, the $\theta$ elements of the final model candidates are estimated based on year 1 data and model performance evaluated based on year 2 data, cf. Table 1. Thus the dependence of performance on the forecast leadtime is resolved twice as fine as in datasets underlying construction of the model candidates, while staying within the forecast performance assessment framework of Section 3.2.

As mentioned last in Section 4.2, the model arrived at there is denoted Model 2. A few additional model candidates are evaluated in order to nuance the performance investigation. A natural baseline is, of course, the Eq. (2) predictions representing forecasts corrected for wind speed bias (see Table 1) and scaling errors. This is the aptly named Model 1, a prerequisite for the other models in that Model 1 predictions are essential for modelling by Eq. (1). An additional perspective on performance in the present context; Model 3, could be to assess the effect of adding wind direction as is typically, if not always, done in statistical wind power prediction models. It turns out that, besides wind direction, the only other $x$ variables present in all nine models post BIC-reduction are the lifted index and hub-height temperature, the further addition of which makes Model 4. Model 5 predictions quantify the effect of leaving out the wind direction from Model

![Figure 6. Top row: Forecast performance in terms of PCC. Middle row: Forecast performance in terms of RMSE superposed by forecast performance in terms of MAE (bold type). Bottom row: Forecast performance relative to Model 1 performance in terms of RMSE superposed by PCC (dotted). (a) Horns Rev, (b) Rejsby Hede, (c) Stor-Rolfienden.](image)

<table>
<thead>
<tr>
<th>Table II. Overview of models evaluated. Wind direction, Eq. (3), includes up to four Fourier expansion terms. Wind speed, wind direction, and temperature are at wind turbine nacelle hub-height, specific humidity is at 2 m AGL. $\dagger$ superscript denotes Eq. (2) predictions, while $\ddagger$ marks an Eq. (1) type model.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model #</td>
</tr>
<tr>
<td>Model 1†</td>
</tr>
<tr>
<td>Model 2‡</td>
</tr>
<tr>
<td>Model 3†</td>
</tr>
<tr>
<td>Model 4‡</td>
</tr>
<tr>
<td>Model 5†</td>
</tr>
<tr>
<td>Model 6‡</td>
</tr>
<tr>
<td>Model 7‡</td>
</tr>
</tbody>
</table>
4, while Model 6 and 7 assess the individual contribution to Model 4 of temperature and lifted index, respectively.

Model performance results are shown in Figure 6. Recall cf. Eq. (4) that the high forecast accuracy in terms of PCC and relatively poor RMSE in Figure 6a, compared to the corresponding metric balance in Figure 6c, highlights severe fluctuations in atmospheric flow off the Danish west coast relative to corresponding weather conditions in central Sweden, cf. Table I. Leaving out the wind direction (Model 1 and 5) has relatively severe impact on performance for the two Danish sites, while Model 7 for the inland Swedish site suffers a minor penalty gradually diminishing as leadtime until Model 5 achieves Model 4 performance around 30 hours GFS integration time. This is in line with expectation as namely the coastal site is situated ~ 1–2 km from the Danish west coast and thus experience winds from the North Sea for westerly flow and winds subject to heterogenous land surface drag for easterly flow, which in turn makes direction-dependent forecast errors plausible. In fact, Model 5 (4 without wind direction) performs inferior to Model 1 (LLS wind speed) for the coastal site, i.e. the inclusion of temperature and lifted index deteriorates the out-of-sample wind speed prediction accuracy at this location. The Swedish site is situated on the top of a hill and surrounded by terrain of some heterogeneity, and hence the wind represented on the coarse ~ 27 km horizontal computational grid of the GFS is more undisrupted by the topography in the simulation than are observed surface winds. This hypothesis is supported by the fact that Model 3 quickly drops to Model 1 performance as a function of leadtime, i.e. forecasted wind direction do not contribute much explanatory value here. Note also that Model 4 appears to be on par with- or better than Model 2. And for the coastal site, in particular, it is even an advantage to further leave out temperature from Model 4 (Model 7). For the inland site models 6 and 7 start out with similar performance, though Model 6 gradually drops as the leadtime increases, while Model 7 sustains the leading position performance-wise.

Given that two or more models perform equivalently well the one with the least complexity is preferable. In that respect, the wind direction model (Model 3) appears to be the best candidate (cf. Section 4.4) for the coastal site, while slightly surpassed by Model 2 for the offshore site. Inland, Model 4 consistently outperforms the other model candidates, though approximately one day into the forecast temperature contributes very little explanatory value as Model 7 (Model 4 without temperature) performance merges with that of Model 4. Overall, predictability in terms of MAE and RMSE drops more rapidly for the offshore site than for the onshore sites and the former starts out better and ends up poorer, performance-wise. Also, the relatively high PCC for the offshore site, albeit similar MAE/RMSE across sites, may indicate that severe GFS wind speed magnitude errors influence offshore prediction to a greater extent than weather system phase errors relative to locations on land, cf. Section 3.1 and the discussion on conditional (scaling errors) and unconditional (bias) contributions to forecast skill in Murphy (1988).

4.4. Practical implications for renewable energy feasibility

Motivation for considering all available forecast data up to 48 hours ahead is given first in Section 4.3. By the same token, the four daily GFS initialisations are applied in Section 4.4 where model coefficient estimation is instead recursive and adaptive. Also, the focus on forecast leadtime resolution is relaxed in order to recast the dataset underlying the conclusion given in Section 5, and for the sake of a more summarising (clear-cut) comparison of model candidates within an alternative framework for coefficient estimation than the cross validation, cf. e.g. Efron and Gong (1983), employed in Section 4.3.

In this section explanatory value of the GFS-derived quantities summarised in Table II is quantified in terms of wind power forecast improvement for the three sites studied. Nielsen et al. (2001, 2012) describe the Wind Power Prediction Tool (WPPT) to estimate the collective electric power production of wind farm turbines, transforming kinetic energy in the moving air into electricity, for use in operational planning of wind energy on a daily basis. The stochastic simulation process in the WPPT involves wind direction dependent power curve modelling and the interested reader is referred to e.g. Lydia et al. (2014) for a recent overview of the topic. As Orths and Eriksen (2012) remark in their description of the world’s largest day ahead market for trading power, the relevant forecast leadtime span for wind energy scheduling is 12–36 hours ahead of the auction deadline. In a realtime setting the four daily GFS forecasts are available with a 5 hours delay and this is accounted for in the statistical modelling underlying results in Table III, showing relative improvement in WPPT forecasts issued each hour, based on the GFS-derived predictors underlying models 2, 4, 6 and 7 of Table II, respectively.

By virtue of the pronounced differences in the rationales underlying the two employed evaluation strategies, results of Figure 6 and Table III indicate that modelled lift index may be key to understanding formulation deficiencies in the GFS, as this variable contains systematic information to consistently reduce wind speed forecast error based on two years of hourly data. Note, for this site also, that the inclusion of predictor terms involving temperature and friction velocity, both of which exhibit severe collinearity with LLS wind speed cf. Figure 7, results in forecast performance reduction relative to the performance of the WPPT setup for GFS input. This prompts a review of the fundamental difference between model coefficient estimation techniques underlying results of Figure 6 and Table III. The sharp division of fixed training and test datasets employed to arrive at the results of Figure 6 is – as mentioned in the first paragraph of this section – replaced by recursive and adaptive coefficient estimation in the WPPT. Recursive in the sense that coefficient estimates are updated in a computationally efficient manner as wind farm measurements and forecast data becomes available, and adaptive due to exponentially decaying weight on previous data points as a function of the temporal distance to the current

<table>
<thead>
<tr>
<th>Site</th>
<th>Model #</th>
<th>Prior</th>
<th>Post</th>
<th>∆</th>
<th>Prior</th>
<th>Post</th>
<th>∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>2</td>
<td>16.14</td>
<td>0.36</td>
<td>11.18</td>
<td>0.23</td>
<td>2.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>16.50</td>
<td>0.42</td>
<td>11.14</td>
<td>0.21</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>16.28</td>
<td>0.27</td>
<td>11.24</td>
<td>0.17</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>16.23</td>
<td>0.27</td>
<td>11.24</td>
<td>0.17</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>RH</td>
<td>2</td>
<td>11.97</td>
<td>0.42</td>
<td>8.25</td>
<td>0.26</td>
<td>3.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>12.39</td>
<td>0.47</td>
<td>8.51</td>
<td>0.29</td>
<td>3.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>12.05</td>
<td>0.34</td>
<td>8.29</td>
<td>0.22</td>
<td>3.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>12.01</td>
<td>0.35</td>
<td>8.25</td>
<td>0.25</td>
<td>3.83</td>
<td></td>
</tr>
<tr>
<td>SR</td>
<td>2</td>
<td>16.12</td>
<td>0.31</td>
<td>11.90</td>
<td>0.24</td>
<td>4.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>15.50</td>
<td>0.31</td>
<td>11.66</td>
<td>0.18</td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>15.96</td>
<td>0.15</td>
<td>11.85</td>
<td>0.19</td>
<td>4.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>15.54</td>
<td>0.27</td>
<td>11.52</td>
<td>0.14</td>
<td>4.05</td>
<td></td>
</tr>
</tbody>
</table>
time stamp. In the WPPT configuration employed here, coefficient estimates for all models are based on fewer data points than in the Section 4.3 approach, and namely the poorly performing models 2 and 6 are subject to strong collinearity with the standalone LLS wind speed predictor, which in turn renders the coefficient estimation unreliable cf. Section 3.5. While all wind direction inclusive models perform similarly for Rejsby Hede, cf. Figure 6b, all GFS-derived quantities improve wind power forecasts for this coastal site relative to the standard WPPT configuration, which can roughly be viewed as representing Model 3 (the one with only LLS wind speed and wind direction). This may indicate strong seasonal dependence of the model coefficients which is accommodated for in the WPPT, though thwarted in the full-year model training dataset determining the model coefficient estimates underlying subsequent-year model predictions in Figure 6.

5 Discussion and conclusions

The purpose of a model is to provide a basis for understanding data and, namely in a weather forecasting context, to predict future measurements. The feasibility of various approximations to the governing equations for atmospheric dynamics, numerical discretisations for dynamical cores, and physical process schemes have been investigated since the dawn of the scientific discipline of computer-aided NWP 65 years ago, and new research to improve the status quo is continually carried out. As a tool for supporting this endeavour, the work presented in this article offers a methodology for improving prediction accuracy for a model quantity of interest and, potentially, indirectly diagnosing weather and climate model deficiencies.

Provided assumptions 1) and 2) below are somewhat satisfied, the method of Section 3 holds for any model based on a set of coupled hyperbolic partial differential equations. That is, the practical implications of mathematical relations for prognostic and diagnostic variables can be explored based on the methodology framed in sections 3 and 4. In particular, inference on systematic model variable coupling to an observed quantity of interest is facilitated, and quantification of explanatory value for individual predictor candidates is discussed. However, the applied univariate modelling framework is subject to the following assumptions: 1) A physically reasonable general linear model

This article is protected by copyright. All rights reserved.

relation between an observed response and simulated predictor candidates can be established, 2) approximate Gaussianity, homoscedasticity, and vanishing autocorrelation can be assumed for the statistical model error. A careful analysis of the extent to which 2) holds constitutes an essential component in such studies. For nonnormal response characteristics alternative statistical modelling techniques may be more suitable. In an ensemble forecasting context e.g. Vaniseim (2009) proposes a MOS framework for the methodology presented in this work. Furthermore, quantifying forecast uncertainty is of major interest within many disciplines, namely concerning scheduling of wind power production, cf. e.g. Thorarinsdottir and Gneiting (2010); Pinson (2012). In this context, extending the here proposed deterministic methodology to probabilistic forecasting in order to assess wind resource predictability may be achieved in terms of forecast error quantile modelling, cf. e.g. Nielsen et al. (2006), and further extended to scenario forecasts cf. Pinson et al. (2009).

With access to the source code practically all weather and climate model aspects can be investigated in the quest for model consistency with atmospheric measurements of interest, the most computation-intensive estimation step being the BIC model reduction which, for this study, took a few minutes on a mid-range Intel-based workstation. Conducting the study on NWP model forecasts rather than reanalysis data exposes forecast leadtime-dependent features – such as the here uncovered departure between explanatory value of lifted index vs. temperature for the inland site as a function of leadtime – that would otherwise not be disclosed, hence adding a temporal dimension to the NWP model deficiency analysis. Thus, one gets a peek inside the numerics engine of a complex system of discretised partial differential equations, for which less obvious numerical scheme and formulation deficiencies can be hard to detect. The models listed in Table II involve only Figure 2c-type predictors, i.e. ordinary linear regressors, indicating that it may be justifiable to skip the visual classification of a candidates for Eq. (1) in general. Hence, in light of the fact that the set of predictors recurring for all three sites considered in Figure 6 appears to contain the essential explanatory information, the method presented in this article can be automated and applied for several sites, subject to prior selection of a suitable variation-threshold cf. Figure 5c.

The results suggest that statistical modelling of forecasted lifted index and wind direction, by the approach described in Section 3 and executed in Section 4, contributes additional accuracy to GFS predictions of wind speed in all terrain types. For the offshore and coastal sites excluding the lifted index predictor achieves equivalent performance cf. Figure 6, though it is important to note that lifted index inclusion does not deteriorate the out-of-sample prediction at any leadtime. This is, however, not the case for the coastal site when excluding the wind direction predictor terms, leading to inferior performance of including lifted index and temperature relative to the pure LLS wind speed model. That wind direction in NWP models has systematic influence on wind speed forecast error is perhaps not surprising. In contrast, the fact that the NWP modelled lifted index contributes \( \sim 0.1 \text{ m s}^{-1} \) better MAE and RMSE than the pure wind direction model for complex terrain data is more interesting, namely when considering the equivalent performance of surface temperature for short leadtimes deteriorating for forecast hours beyond \( \sim 12 \) hours ahead. However, the wind power forecast results of Table III substantiate the explanatory value of the lifted index for all sites and hence hint a strong seasonal model coefficient dependency, namely for the offshore and coastal sites, both of which are in proximity to the North Sea.

The lifted index, first proposed by Galway (1956) as a predictor of latent instability (severe weather), is defined as the temperature difference between a parcel of air raised adiabatically to the...
500 mb pressure level and the value at that level. The relatively significant performance impact from including this predictor in Eq. (1) could hint moisture-dependent bias in the GFS simulation of surface layer winds, cf. e.g. Bénard et al. (2000), which in turn points toward inadequate surface and planetary boundary layer parameterisations and/or land surface scheme. The study suggests that the GFS accurately predicts the lifted index, a quantity characterising a wide vertical extent, while prediction accuracy for the highly fluctuating surface wind speed is inferior, yet systematically coupled to the lifted index. In other words, explanatory value is derived from mapping the GFS predicted lifted index covariance with measured wind speed to covariance with forecasted wind speed, respectively.

Through the relation Eq. (4) the high PCC and similar RMSE performance offshore, relative to the coastal and inland sites, express the more volatile weather conditions in the North Sea region than in central Sweden. However, the accuracy metric discrepancy may in part be attributed the inferior representation of the influence of the ocean surface on atmospheric flow. i.e. if the simulation of large scale weather systems is fairly well temporally correlated with observed dynamics (as the high PCC suggests), the severe wind speed error magnitude penalty expressed by the MAE/RMSE relative to the PCC, compared to the other sites, could adhere to NWP model lower boundary condition issues. Hence, the authors speculate that efforts toward improving the simulation of offshore surface layer dynamics in the GFS would be feasible for efficient offshore wind energy integration in the power grid, as well as for other applications depending critically on wind speed forecast accuracy at sea.

Acknowledgements
The study was funded by the Danish transmission system operator Energinet.dk, within the framework of public service obligation project PSM 10464, to promote environmentally friendly energy. The analysis and visualisation of results were generated using R (2013), Figure 1 was created using NCL (2014). Wind farm datasets were kindly provided by the major European power company Vattenfall. The authors thank Stig Mortensen and Sven Thomsen from ENFOR A/S for discussions on statistical modelling, as well as Mark Kelly and Jake Badger from DTU Wind Energy for comments on the results. Also, discussions with Joshua Hacker from NCAR helped improve the manuscript. Finally, the authors are thankful for the constructive feedback from the anonymous reviewers.

References

This article is protected by copyright. All rights reserved.


