**Associate Editor Decision: Publish subject to minor revisions (review by editor)**(23 Mar 2017)
by Julie Lundquist
Comments to the Author:
The paper is in a very good shape, with a few caveats and corrections as noted below. These should be straightforward to correct.

**1. Please rewrite the sentence at the top of page 6, "The simulation and spin-up length ranged from 1 hour and 7 simulation to continuously running for the full year." What is the "1 hour and 7 simulation" intended to mean?**
Revised the sentence to: "The simulation and spin-up length ranged from 1 hour spin-up and 7 hours simulation to 24 hours spin-up and continuously running for the full year."

**2. Top of page 7: how much model data was rejected for being unphysical?**
The model output submitted were assumed to be quality checked by the submitter, but it was also checked by the authors for obvious non-physical or inconsistent behavior, and not used in that case. The number of models excluded was between two and four at each of the sites, but no model was excluded from all three sites. This has been clarified in the text.

**3. Top of page 8: "shear" instead of "sheer"**
Thanks. Corrected.

**4. Page 9 line 16: explicitly state that this section is looking at the mean profile over the entire year.**
Thanks. Renamed to 'Annual mean wind speed'.

**5. Page 12, line 9: finish the sentence, please.**
Sentence removed.

**6. Page 12, line 11: reference? Perhaps http://journals.ametsoc.org/doi/abs/10.1175/2010BAMS2770.1 ?**
Good point. We agree and have added the suggested reference.

**7. Page 12, line 14: please refer to previous work using these stability classifications.**
The wrong Bulk Richardson classication was shown in the table. This has been corrected. The classes are used in Gryning et al. (2007) and Mohan and Siddiqui (1998). A clarification of this has been added.

**8. Caption for Figure 7: please explain the box/whiskers: are the ends of the bars representing +- one standard deviation, the 9th and 91st percentile, the 2nd and 98th percentile, or some other choice? Could also be helpful to state the middle bar is the median.**
An explaination has been added to the figure text: "The boxes represent the 2nd and 3rd quartile. The whiskers extend to the smallest (bottom), or largest (top), value that is within 1.5 times the inter-quartile range. Samples outside this range are shown as outliers."

**9. Page 14 line 11: "contrast" instead of "contract"**
Thanks. Corrected.

**10. Page 17 line 12: please explicitly state whether MYNN is considered to be MYJ or not for this purpose. This section would also benefit from a brief (1-2 sentence) discussion of local vs non-local PBL schemes to explain why there is likely a difference between YSU and MY* schemes. And if MYNN is not included with the MYJ schemes, please explain why not.**
A short comment on closure order and local/nonlocal schemes have been added. We decided to treat MYNN and MYJ as seperate PBL schemes for this evaluation. This was done to keep group size similar, and because they do use a different formulation for length-scale limitation. The 'Other' group contains both local, nonlocal, and mixed schemes of different order of closure.

**11. Page 18 line 6: which model was the single most accurate?**
We consciously decided not to single out any model-setup. Particiants were promised that the performance of their specific model would not be revealed. Providing information relating model-setup and performance would, in some cases, make it easy to relate that to the participant.

**12. Page 21 line 3: should "No conclusive evidence" be tempered with "For these sites, no conclusive evidence was found that …" to make it clear that the answer might not be the same in complex terrain?**
Good point. The suggested moderation was added to the text.

**13. Page 21 line 7: please clarify what "many more runs" means. I think you mean more runs where the PBL scheme and the grid resolution is held constant but the lead-time is varied?**
The comment is more general. Future studies with a larger sample of (model) setups will provide a better statistical basis to describe the sensitivies related to model components.

# An intercomparison of mesoscale models at simple sites for wind energy applications

Bjarke T. Olsen, Andrea N. Hahmann, Anna Maria Sempreviva, Jake Badger, and Hans E. Jørgensen
DTU Wind Energy, Frederiksborgvej 399, 4000 Roskilde, Denmark
*Correspondence to:* Bjarke Tobias Olsen (btol@dtu.dk)

**Abstract.**

Understanding uncertainties in wind resource assessment associated with the use of the output from Numerical Weather Prediction (NWP) models is important for wind energy applications. A better understanding of the sources of error reduces risk and lowers costs. Here, an intercomparison of the output from 25 NWP models is presented for three sites in Northern Europe characterized by simple terrain. The models are evaluated using a number of statistical properties relevant to wind energy and verified with observations. On average the models have small wind speed biases offshore and aloft ($<4\%$), and larger biases closer to the surface over land ($>7\%$). A similar pattern is detected for the inter-model spread. Strongly stable and strongly unstable atmospheric stability conditions are associated with larger wind speed errors. Strong indications are found that using a grid spacing larger than 3 km decreases the accuracy of the models, but we found no evidence that using a grid spacing smaller than 3 km is necessary for these simple sites. Applying the models to a simple wind energy offshore wind farm highlights the importance of capturing the correct distributions of wind speed and direction.

## 1 Introduction

Numerical Weather Prediction (NWP) models are increasingly being used in wind energy applications, e.g. wind power resource mapping and site assessment, for planning and developing wind farms, power forecasting, for electricity scheduling, maintenance of wind farms, and energy trading on electricity markets. In site assessment, NWP models are commonly part of the model chain used to estimate the Annual Energy Production (AEP) and are responsible for a large part of the uncertainty of this estimate.

The extensive use of NWP models, and the vast customization-space of each model, means that a strong demand exists for quantification of a) the overall model uncertainties, and b) the sensitivity of the uncertainties to the choice of sub-components and parameters. Understanding the sensitivities and uncertainties of the NWP model output can reduce their associated risks, and improve decision making. Model users aware of the sensitivity of individual model components will be able to optimize the model setup for specific applications.

In the following, the NWP models will be referred to as "mesoscale" models, signifying that they partly resolve atmospheric phenomena in the mesoscale range, defined as the range of horizontal length scales from about one to several hundreds of kilometers (Orlanski, 1975).

A common way to assess NWP model uncertainties is to use an ensemble approach, where a number of parallel model runs, referred to as ensemble members, are run with slightly perturbed initial conditions (Warner, 2004). The magnitude of the perturbations is typically limited by the uncertainty associated with the particular perturbed variable, in the expectation that the ensemble of solutions will cover the solution-space arising from the uncertainties of the input parameters. Ensemble-based techniques are used for many meteorological application, including: precipitation forecasting (Gebhardt et al., 2011; Bowler et al., 2006), wind power production forecasting (Constantinescu et al., 2011). However, one would not expect that the ensembles of any particular modeling system fully represent the uncertainties of another modeling system. This was also demonstrated in the DEMETER project (Development of a European multi-model Ensemble for seasonal to inTERannual climate prediction) (Palmer et al., 2004), where a multi-model ensemble approach, consisting of a number of different modeling systems, each split into a number of ensembles, provided a better representation of the overall uncertainties than any single model ensemble.

Mesoscale model uncertainties in wind speed near the ground are particularly sensitive to some model components, e.g. the choice of Planetary Boundary Layer (PBL) scheme, the spin up and simulation time, and the grid spacing. In the last couple of decades these sensitivities have been studied in great detail. Vincent and Hahmann (2015), Draxl et al. (2014), and ~~Hahmann et al. (2014)~~ Hahmann et al. (2015b) studied the sensitivities of the Weather Research and Forecasting (WRF) model (Skamarock et al., 2008) in offshore and coastal areas in Northern Europe. Vincent and Hahmann (2015) studied the effect of grid nudging, spin-up time, and simulation time, on near-surface and upper PBL wind speed variance. They showed that: 1) spatial smoothing is observed when nudging is used, but the impact is small in the lower part of the atmosphere, and 2) nudged longer simulation times (11 days) only have slightly lower variance than short simulations (36 hours), which makes longer simulations appropriate for climatological wind energy studies. Draxl et al. (2014) studied the ability of the WRF model to represent the wind speed and wind shear profiles at a Coastal site in Denmark using seven different PBL schemes. They showed that the Yonsei University (YSU) (Hong et al., 2006) scheme represents the profiles best for unstable atmospheric stability conditions, while the Asymmetric Convective Model version 2 (ACM2) (Pleim, 2007b), and the Mellor-Yamada-Janjic (MYJ) (Janjić, 1994) PBL schemes had more realistic profiles for neutral and stable conditions respectively. Using the WRF model for wind resource assessment, ~~Hahmann et al. (2014)~~ Hahmann et al. (2015b) showed that the choice of PBL scheme and spin up time has the greatest impact on the simulated mean wind speed for a number of offshore sites, while the number of vertical levels, and the source of initial conditions had a smaller impact.

Several studies have investigated the WRF model sensitivities in regions of complex terrain. Carvalho et al. (2012) studied the sensitivities related to the choice of initialization frequency, grid nudging, and suite of Surface Layer (SL) scheme, PBL scheme, and Land Surface Model. They observe that using grid nudging and frequent starts (every second day) gives the best agreement for wind speed with several masts located in complex terrain in Portugal. Carvalho et al. (2012) and García-Díez et al. (2013) found a seasonal dependency of the optimal suite of SL-PBL-LSM for simulating PBL winds and temperature. Carvalho et al. (2014b) investigated the sensitivities related to the SL and PBL scheme in WRF model at both land and offshore sites in and near Portugal. They showed that the PX SL scheme (Pleim, 2006) combined with the ACM2 PBL scheme (Pleim, 2007b) gave the smallest errors for wind speed, and wind energy production estimates, across the sites, while the QNSE-

QNSE (SL-PBL) scheme (Sukoriansky et al., 2005) gave smaller errors for offshore sites. In a similar study Gómez-Navarro et al. (2015) analysed the sensitivities of the WRF model to the choice of PBL scheme, and grid spacing, in complex terrain in Switzerland. They found that using a modified version of the YSU PBL scheme, that account for effects of unresolved topography (Jiménez and Dudhia, 2012), in combination with the smallest grid spacing (2 km), and Analysis Nudging, gave the best agreements with measurements during a number of wind storms. Carvalho et al. (2014a) studied the sensitivities of simulating the local wind resource with the WRF model at several masts in Portugal, to the choice of data set used for initial and boundary conditions. They show that using the ERA-Interim reanalysis data set (Simmons et al., 2007) gave the smallest errors, compared to NCEP (National Centers for Environmental Prediction) R2 (Kanamitsu et al., 2002), CFSR (Saha et al., 2010), FNL, and GFS data sets, as well as the NASA (National Aeronautics and Space Administration) MERRA data set (Rienecker et al., 2011).

Sensitivities to the choice of modeling system have also been studied for wind energy applications. Horvath et al. (2012) compared the MM5 (Grell et al., 1994) and WRF models for a site in west-central Nevada characterized by complex terrain. Both models were run in a grid nesting setup from 27 kilometers to 333 meters grid spacing, and the near surface wind was compared to wind observations from several 50-meter tall towers. The study showed that the WRF-derived winds were in better agreement with mean wind speed observations, but thermally-driven flows were overestimated in both intensity and frequency. Hahmann et al. (2015a) compared two downscaling methodologies: the KAMM-WAsP (Badger et al., 2014) and WRF Wind Atlas ~~(Hahmann et al., 2014)~~ (Hahmann et al., 2015b) methods, both based on a model chain approach between a NWP model and a linearized flow microscale model, for a number of mast sites in South Africa. The study showed that the WRF-based method gave smaller biases than the KAMM-based approach, which underestimated the wind speeds.

Community-driven model intercomparison projects provide an opportunity to study both model uncertainties, and sensitivities to model components. In the last decade, several intercomparison projects have been successfully carried out based on model output submitted by modelers from the wind energy community. The Bolund experiment (Bechmann et al., 2011) was an intercomparison of flow models, from simple linearized flow models to Computational Fluid Dynamics (CFD) models. The models were compared to measurements around the small island Bolund in Denmark. The Comparison of Resource and Energy Yield Assessment Procedures (CREYAP; Mortensen et al., 2015) was an intercomparison of energy yield assessment procedures based on four case-studies. The study revealed a large spread amongst the different procedures, and highlighted the need for further studies into the uncertainties associated with the models themselves. A similar intercomparison of NWP models is attractive for a number of reasons. First, it offers an opportunity for model developers, model users, and stake-holders, to get a better understanding of the model uncertainties. Secondly, a collaborative intercomparison project, which utilizes model data crowd sourced from the wind energy community, increases the scalability of the study compared to traditional sensitivity studies, by distributing the workload and computational cost among participants. Finally, if sufficient meta-data is collected, it offers a unique insight into the "common-practices" in mesoscale modeling within the wind energy community.

In this paper, a blind intercomparison of the output from 25 different NWP simulations is presented for three locations in Northern Europe. The study is based on model output submitted by the modeling community to an open call for model data for a benchmarking exercise co-organized by the European Wind Energy Association (EWEA, now WindEurope) and

the European Energy Research Alliance, Joint Programme Wind Energy (EERA JP WIND). The three chosen sites represent some of the simplest terrains: offshore, inland near the coast and inland in flat terrain, where the smoothing of the terrain representation is not an issue. The three sites have quality observations from tall meteorological masts with many heights. The main objectives of this study are: 1) To highlight and quantify the uncertainties of the models and serve as motivation for future analysis of model uncertainties. 2) To identify model setup decisions that have an impact on the model performance. The models are evaluated using simple metrics relevant to wind energy applications.

The structure of the paper is as follows. In sect. 2 we present a detailed description of the methodology used, including a description of the three study sites and the models used by the participants. Sect. 3 presents the intercomparison results, and finally sect. 4 contains the summary and conclusions of the study.

## 2 Methodology

### 2.1 Sites and observations

Three sites with quality measurements from tall meteorological masts with different terrain characteristics were chosen for this study: (1) FINO3, an offshore mast in the North Sea, (2) Høvsøre, a land mast near the Danish west coast, and (3) Cabauw, a land mast in the Netherlands. The mast locations are shown in Fig. 1, and the coordinates and characteristics of each site are provided in Table A1. Long-term measurements are available from each of the masts, but a single year (2011) was selected as the study period due to its excellent data availability.

FINO3 ~~(Fabre et al., 2014a)~~ (Fabre et al., 2014b) is a marine platform located in the North Sea 80 kilometers off the coast of Denmark, with a meteorological mast reaching 120 m above mean sea level (AMSL). We used measurements at 40, 60 and 90 m AMSL in this study. The Høvsøre (Peña et al., 2014) mast is located about 2 km east of the coastline in western Jutland, Denmark. Apart from the sharp surface roughness change at the coastline, and the presence of a small coastal escarpment, the surrounding terrain is homogeneous and flat. We used measurements at 10, 40, 60, 80, 100 m at this site. The Cabauw mast ~~(Ulden and Wieringa, 1996)~~ (Ulden and Wieringa, 1995) is located 40 km inland near the small towns of Cabauw and Lopik in the Netherlands. The surroundings are flat and characterized by fairly homogeneous agricultural fields, although with patches of forest and buildings. Here we used measurements at 10, 20, 40, 80, 140, and 200 m.

Figure 2 shows availability of wind speed observations for 2011 at the three meteorological masts. At Cabauw, the data was gap-filled by simple interpolation as the missing values were few (less than 2% missing data per month) and the gaps short. The time series from the two other sites were not gap-filled.

At FINO3, the wind speed measurements at three of the heights, 50, 70, 90 m, are a combination of the measurements from three anemometers at three separate booms 120° apart. This procedure minimizes the effects of the mast flow distortion. At the other two heights, 40 and 60 m, only one anemometer is available, and the wind measurements are therefore susceptible to flow distortion. Thus, instead of using the single-anemometer data from 40 and 60 m, the measurements from 50 and 70 m were vertically interpolated in log height to 40 and 60 m. This assumes that the errors due to interpolation and extrapolation are much smaller than those caused by mast flow distortion.


**Figure 1.** Map of Northern Europe with the three site locations used in the model intercomparison: (1) FINO3, in the North Sea. (2) Høvsøre, Denmark. (3) Cabauw, The Netherlands.

### 2.2 Submission procedure and models

EWEA issued an open call for data and the submission procedure consisted of a template spreadsheet and a questionnaire downloadable from the EWEA website. The participants filled the spreadsheet with the time series of the required variables at each location and height. The questionnaire contained details about the setup of the modeling system used. The participants returned the spreadsheet to EWEA, whom passed it on to the authors in an anonymized version.

The requested model variables were hourly wind speed and direction, air temperature, and atmospheric stability. The questionnaire asked about the modeling setup, i.e. the model code and version, the surface and planetary boundary layer schemes, the Land Surface Model (LSM), the grid nests size(s) and spacing(s), the vertical levels, the land use data, the length of the simulation and spin-up time, as well as the source of the initial and boundary conditions. The participants were also asked to comment on any additional modifications made to the model, including assimilation, ensemble or other methods used.

Table A3 lists the various groups participating in the exercise. It includes representatives from private companies, universities, research centres, and meteorological institutes. Table A4 summarizes the models and the different model setup options used. The WRF model is by far the most commonly used model in the study, with 18 out of 25 models (Table A4). The Noah LSM was the most common LSM used, and the Era-Interim Reanalysis the most common source of boundary and initial conditions. The PBL scheme used and the source of land cover data were more varied amongst the participants. Most models used a maximum simulation length of less than 100 hours, including the spin-up time (most typically 12 hours spin-up and 36 hours

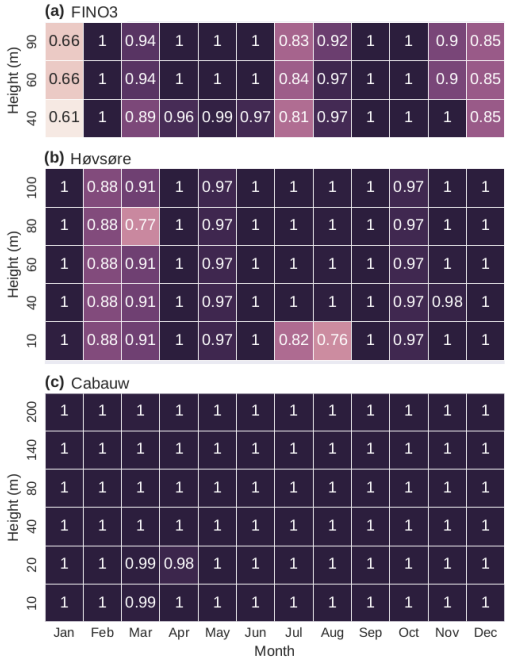


**Figure 2.** Availability of wind speed and direction observations for (a) FINO3, (b) Høvsøre, and (c) Cabauw given as the fraction of completeness for each month of the year 2011 for each height.

of total simulation). The simulation and spin-up length ranged from 1 hour spin-up and 7 ~~simulation to~~ hours simulation to 24 hours spin-up and continuously running for the full year.

For reference, wind time series from the ERA-Interim reanalysis (Dee et al., 2011) were included in the comparisons whenever possible. The ERA-Interim reanalysis data set is a global data set based on extensive assimilation of surface and upper-air observations. The data is available on a grid spacing of about 80 km in the horizontal with 60 vertical levels, with values at approximately 10, 34, 69, 118, 187 and 275 m above the model surface. We used bilinear interpolation to interpolate to the sites coordinates, and linear interpolation in the vertical. The data set is available in 6 hour intervals, thus linear interpolation in time was used to obtain hourly samples.

### 2.3 Statistical methods

This study is based on direct comparison between the observations and model output at collocated positions, as well as intercomparison of the modelled output. The sampling frequency for the study was chosen to be one hour. For the observation data this means hourly mean values; for the mesoscale models the inter-hourly variation is small, so instantaneous values were used. To ensure temporal consistency between observations and modelled output, instances of missing data from the observations were removed from the modeled output. Furthermore, to get consistent vertical profiles, only instances where all heights for a

particular mast had available data were used. The model output submitted were assumed to be quality checked by the submitter, but it was also checked by the authors for obvious non-physical or inconsistent behavior, and not used in that case. The number of models excluded was between two and four at each of the sites, but no model was excluded from all three sites.

**Inter-model mean and inter-model variations**

The emphasis of this study is on the wind speed, $u$, and wind direction, as they are the most important variables for wind energy applications. In the following, a subscript $m$ signifies the temporal mean of a variable, i.e. $u_m$ is the temporal mean wind speed. This is not to be confused with the mean value of the model-ensemble, also referred to as the inter-model mean, which is denoted with a tilde. For example, the mean of the model-ensemble for the temporal mean wind speed is denoted $\tilde{u}_m$, and calculated as:

$$\tilde{u}_m = \frac{1}{N}\sum_{i}^{N} u_{m,i} \tag{1}$$

Here $i$ is the model index, and $N$ is the total number of models. Likewise, it is useful to define its standard deviation:

$$\tilde{\sigma}_{u_m} = \sqrt{\frac{1}{N}\sum_{i}^{N}(u_{m,i} - \tilde{u}_m)^2}, \tag{2}$$

which is the standard deviation of the inter-model variation between the temporal model means. Since $\tilde{u}_m$ and $\sigma_{u_m}$ are both sensitive to outliers, we used the following procedure:

1. Calculate $\tilde{u}_m$ and $\tilde{\sigma}_{u_m}$
2. Remove models whose mean $|u_{m,i} - \tilde{u}_m| > 3.5\,\tilde{\sigma}_{u_m}$
3. Recalculate $\tilde{u}_m$ and $\tilde{\sigma}_{u_m}$ with the new subset of models

The value of $3.5\,\tilde{\sigma}_{u_m}$ was chosen somewhat arbitrarily to ensure that only "extreme" outliers were removed. The procedure included only models with output available at all the heights, to ensure a vertically consistent profile of the mean and its variation. Typically, only one or two models were removed by this criteria.

**Coefficient of variation**

Variations in wind speed often scale with the mean wind speed. Thus, to allow for intercomparison of wind speed variation intensity across vertical levels we define the coefficient of variation, $C_{v,u}$. It is defined as the ratio of the standard deviation and the mean, $\sigma_u/u_m$, and is a unit-less measure of the relative variation at the sampling time scale. At timescales of seconds it is known as the turbulence intensity, but in this case, with a sampling frequency of one hour, it represents the intensity of variations of synoptic- and mesoscale weather phenomena.

**Wind speed shear exponent**

To diagnose the wind ~~sheer~~ shear in the boundary layer, we use the wind ~~sheer~~ shear exponent, $\alpha$, which uses the wind speed $u_1$ and $u_2$ at two heights $z_1$ and $z_2$, given by the expression:

$$u_2 = u_1 \left( \frac{z_2}{z_1} \right)^{\alpha} \tag{3}$$

In the surface layer $\alpha$ is strongly influenced by the surface roughness and the atmospheric stability. By comparing the modelled to the measured $\alpha$ it is thus possible gain insights into how the model captures these effects.

**Error metrics**

The Root Mean Squared Error (RMSE) and the Normalized RMSE (NRMSE) were used as error metrics to obtain single value measures of the error across heights at a site. The RMSE and NRMSE are defined as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \left( x_j^M - x_j^O \right)^2}, \tag{4}$$

$$NRMSE = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \left( \frac{x_j^M - x_j^O}{x_j^O} \right)^2}, \tag{5}$$

for a set of $n$ modelled values $x_j^M$ and observed values $x_j^O$. The RMSE was used for variables that do not scale with height in the surface layer, e.g. wind speed shear exponent; the NRMSE was used for variables that do scale with height, e.g. wind speed.

## 2.4 Wind energy application

To investigate the errors associated with the use of each model in wind energy applications, we performed a simple wind resource assessment exercise, using both measurements and modelled time series at FINO3.

A typical approach to resource assessment is to run a mesoscale model for a number of years, followed by a downscaling process where the wind-climate statistics obtained from the mesoscale model are used as input to a microscale model (Badger et al., 2014; Hahmann et al., 2015a). In simple terrain, the microscale model usually consists of a flow model like the one used by the Wind Applications and Analysis Program (WAsP). WAsP uses a linearised flow model based on Jackson and Hunt (1975). The procedure in WAsP consists first of an upscaling, where local effects from variations in orography, surface roughness, and obstacles, are removed from the wind-climate statistics. This is referred to as "generalisation" of the wind climate, which makes it representative for a larger area than the site specific wind climate. The size of this area depends on the complexity of the surface roughness, and orographic variations in that area. To obtain a site-specific wind climate at a new site in this area, the generalised wind climate is downscaled by "reversing" the generalization process, i.e. by introducing the site-specific effects of orography, surface roughness, and obstacles of the new site.

Given the wind-climate and the turbine power curve, the expected power output can be calculated for any site. Since the participants in this intercomparison were not requested to submit the model-specific orography and roughness maps near each site, it is not possible to go through the generalization procedure, and subsequent downscaling process at the inland sites. However, for the offshore site FINO3 there are no effects of orography, and the differences in roughness between the models can be assumed to be negligible. Therefore, we can use the raw model output at this site to estimate the wind resources estimated by each of the models, without the generalisation procedure.

We performed the wind resource exercise at 90 m at FINO3, assuming first a single Vestas V80 turbine at the site, and then repeated for the wind farm of Horns Rev, which is a 80 turbine wind farm located near FINO3. The resource estimations for the wind farm includes the simple wake parametrization present in the WAsP model, which was used to estimate the power losses.

## 3 Results

### 3.1 Mean quantities and distributions

The following subsection is dedicated to the general performance of the models, and their ability to capture the mean and the distributions of a number of wind-related quantities. As previously stated, the goal is to highlight the weaknesses of the models to encourage further analysis of model sensitivities.

#### 3.1.1 ~~Mean~~ Annual mean wind speed

Figure 3 shows the vertical profiles of mean wind speed ($u_m$) at the three sites. At FINO3 (Fig. 3a), most Mesoscale Models (MMs) underpredict $u_m$ at all heights. However, the bias on average is less than $0.27\,\mathrm{ms}^{-1}$ ($\sim 2.8\%$). This is a small bias compared to that of the ERA-Interim data, which shows a larger bias than all the mesoscale models. The inter-model variance $\tilde{\sigma}_{u_m}$ at FINO3 is 2.7–3.1% of the inter-model mean, and decreases with height. That is the lowest combined inter-model variance of any of the three sites.

At Høvsøre (Fig. 3b), the MMs generally have small wind speed biases above 10 m. The error of the inter-model mean of the models is smaller than $\pm 0.16\,\mathrm{ms}^{-1}$ ($\sim 1.9\%$), and the inter-model variance is 3.0–5.2%, decreasing with height, which is low compared to the biases at the other site on land (Fig. 3c). At 10 m, most MMs overpredict the mean wind speed. The inter-model mean has a positive bias of $0.54\,\mathrm{ms}^{-1}$ ($\sim 8.4\%$). The largest inter-model variance is also seen at 10 m (7.8%). The ERA-Interim also overpredicts the mean wind speed at 10 m, with a larger bias than $\tilde{\mathrm{u}}_{\mathrm{m}}$. Above 10 m, ERA-Interim has smaller errors, but the shape of the profile is not well captured. Signs of a "kink" in both the observed and modelled profiles are present, which could indicate the transition from the low surface roughness of the sea to the higher surface roughness inland.

At Cabauw (Fig. 3c), most of the MMs overpredict $u_m$. Only one of the models and the ERA-Interim shows a significant underprediction, and in the case of the reanalysis, this underestimation increases with height. The overprediction by the rest of the MMs varies in magnitude, but the average of the models, excluding the outliers, is in the range 4–9% across the different

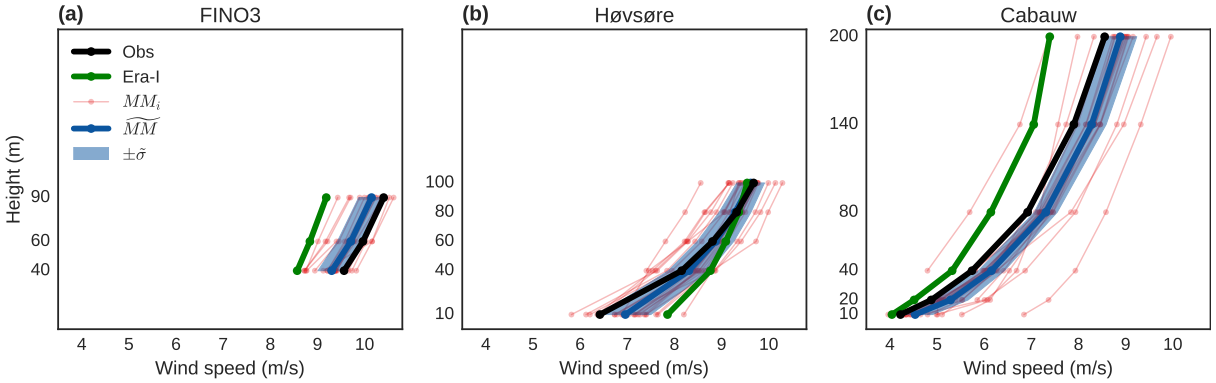


**Figure 3.** Vertical profiles of mean wind speed ($u_m$) at the three sites for: the observations (black), the ERA-Interim data set (green), the Mesoscale Models $MM_i$ (red), and the inter-model mean $\widetilde{MM}$ (blue line) and its standard deviation $\pm\tilde{\sigma}$ (blue shade).

heights. The largest relative errors are at the lowest levels. The inter-model variance ($\tilde{\sigma}_{u_m}$) at Cabauw varies between 3.3–8.1% across the different heights, and is largest at the lowest levels. The decrease of wind speed bias with height was also observed by Jiménez et al. (2016), whom associated this with excessive turbulent mixing, which may be caused by a misrepresentation of the surface roughness length.

### 3.1.2 Frequency distribution of wind speed

Figure 4 shows that, on average, the MMs capture the wind speed distributions well compared to the observations. The only exception is a slight shift towards higher wind speeds at Cabauw, corresponding to the positive bias in mean wind speed observed in Fig. 3. The ERA-Interim data set captures the distribution well at Høvsøre, but it has distributions that are shifted towards lower wind speeds at FINO3 and Cabauw, corresponding to the bias in Fig. 3.

### 3.1.3 Distribution of wind direction

Figure 3 shows that the MMs generally capture the mean wind speed well, this is also true for the wind direction distributions, commonly called "wind roses". The distributions are split into 15° sectors at heights of either 80 or 90 meters. Figure 5 also shows that the models are in good agreement. In all three sites the MMs capture the distribution better than the reanalysis data. At all sites, but most markedly at Cabauw, the ERA-Interim distribution is rotated clockwise relative to the distribution from the observations and MMs. This rotation might result in a different wind farm layout if its power is optimized according to the wind roses from MMs or the ERA interim.

### 3.1.4 Annual wind speed cycle

Figure 6a shows the monthly distribution of the mean wind speed for the MMs, and the measurements. Apart from a few models outside the 3x quartile range, most models capture the diurnal cycle well. Interesting, the figure also reveals that both

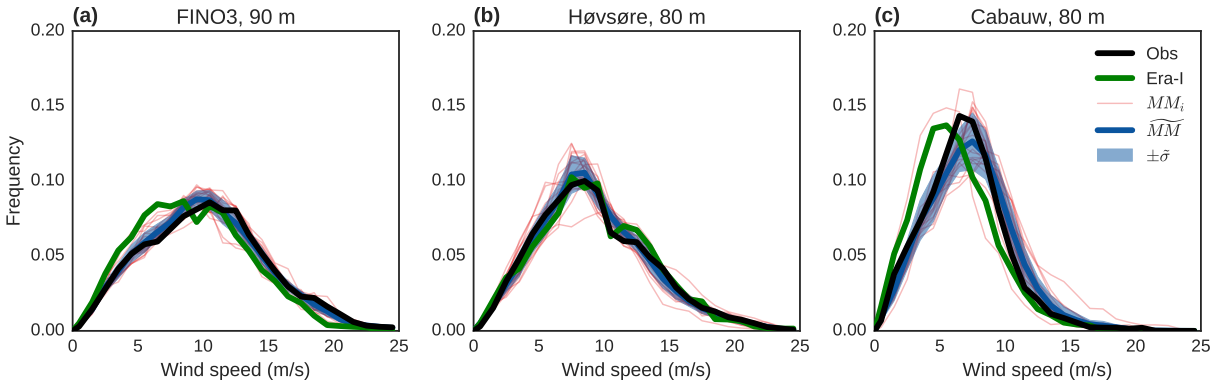


**Figure 4.** Wind speed distributions at the three sites (FINO3 at 90 m, Høvsøre at 80 m and Cabauw at 80 m), for: the observations (black), the ERA-Interim data set (green), the Mesoscale Models $MM_i$ (red), and the inter-model mean (blue line) and its standard deviation $\widetilde{MM} \pm \tilde{\sigma}$ (blue shade).

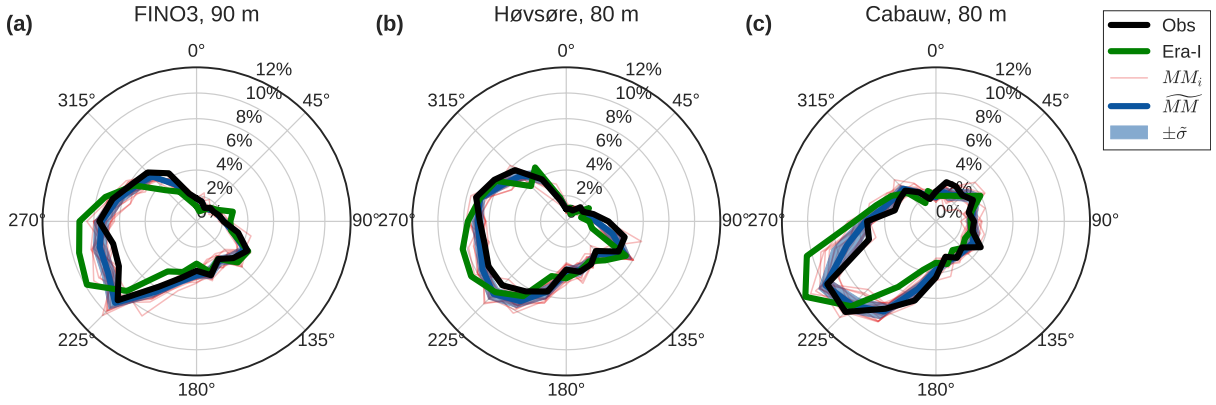


**Figure 5.** Wind direction distributions at the three sites (FINO3 at 90 m, Høvsøre at 80 m and Cabauw at 80 m), based on 24 sectors, for: the observations (black), the ERA-Interim data set (green), the Mesoscale Models $MM_i$ (red), and the inter-model mean $\widetilde{MM}$ (blue line) and its standard deviation $\pm\tilde{\sigma}$ (blue shade).

the overestimation by the models at Cabauw and the underestimation at FINO3, seen in Fig. 3, is evenly distributed throughout the year. At Høvsøre, a mix of under- and overestimations are observed.

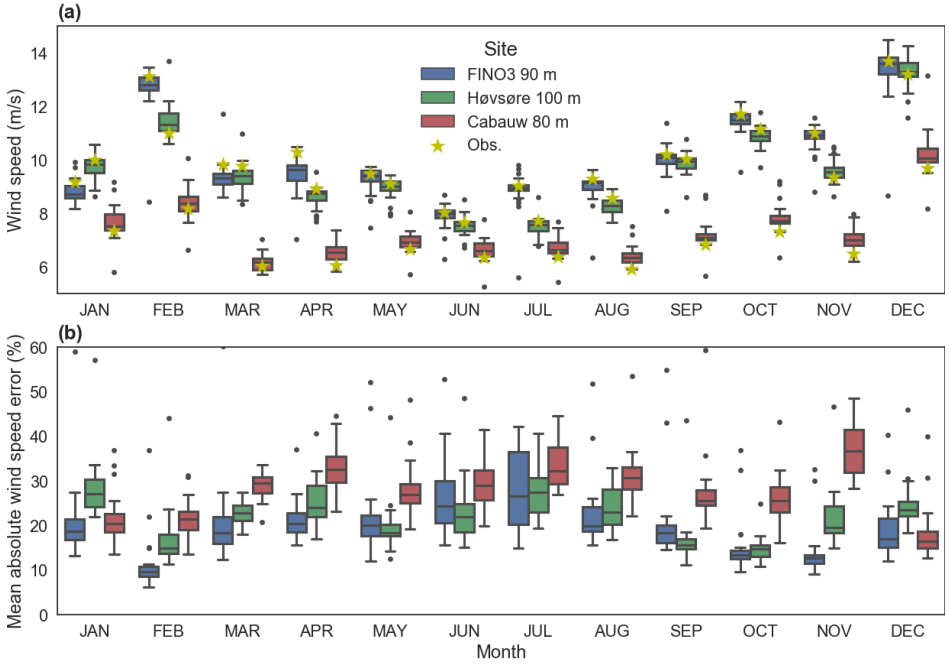


**Figure 6.** (a) Monthly distributions of mean wind speed for the MMs (boxplots) and observations (star), at each location (colors). (b) Monthly distributions of the models for the Mean Absolute Error (MAE) for wind speed at each location (colors). The boxes represent the 2nd and 3rd quartile. The whiskers extend to the smallest (bottom), or largest (top) value that is within 1.5 times the inter-quartile range. Samples outside this range are shown as outliers.

Figure 6b shows the monthly distribution of the Mean Absolute Error (MAE) for wind speed for the MMs. Summer and spring are generally associated with larger deviations between the modeled and observed wind speeds. It is well established that fall and winter weather in Northern Europe is governed by large-scale planetary and synoptic weather phenomena, that is well captured by mesoscale models. During spring and summer, meso- and thermally induced phenomena (e.g. sea breezes and convection) have a larger impact on the flow, which is more difficult for the models to correctly capture. The lowest MAE is observed at FINO3 in February, October and November with most MAE values near 10%. The largest MAE are in November at Cabauw (values in the range 30–45%). ~~For June and July FINO3 shows~~

### 3.1.5 Effect of atmospheric stability

It is generally acknowledged that non-neutral atmospheric stability conditions pose one of the greatest challenges for MMs (Fernando and Weil, 2010). To study the performance of the models in different stability regimes, the stability parameters supplied for each model (inverse Obukhov length or bulk Richardson number) were used to group the hourly samples into

five stability classes based on Gryning et al. (2007) and Mohan and Siddiqui (1998), shown in Table A2. Because the models represent atmospheric stability in different ways, the number of samples in each stability group varies for the different models. However, the number of samples in each group was never below 150 hours (out of 8760 hours), and it was more than 400 in most cases. The MAE for wind speed was calculated for each of groups and for all models. The results are shown in Fig. 7.

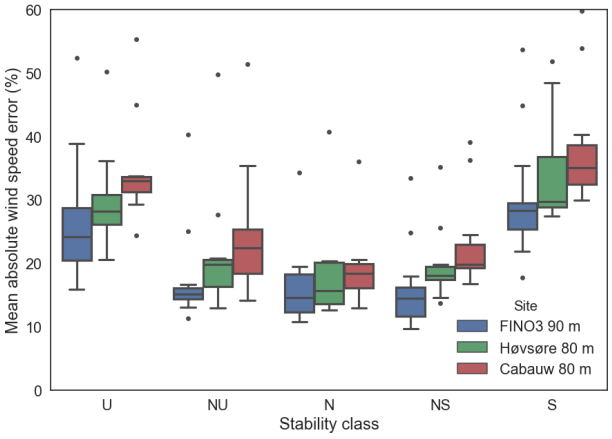


**Figure 7.** Distribution of Mean Absolute Error (MAE) for wind speed at the three sites for five stability classes: ~~Very~~ Unstable (~~VU~~U), ~~Unstable~~ Near-Unstable (~~U~~NU), Neutral (N), ~~Stable~~ Near-Stable (~~S~~NS), ~~Very~~ Stable (~~VS~~S). See definitions in Table A2. The boxes represent the 2nd and 3rd quartile. The whiskers extend to the smallest (bottom), or largest (top) value that is within 1.5 times the inter-quartile range. Samples outside this range are shown as outliers.

At all three sites, the smallest deviations between modelled and measured wind speeds are found when the models perceive the surface layer stability from unstable (U) to stable (S). The MAE in these cases typically range from $10\%$ to $35\%$, with just a few models outside of 3x quartile range. The largest deviations are found when the models estimate very stable conditions (VS) or very unstable conditions (VU) (typical values in the range 15–45% MAE). The site where the largest errors are found is Cabauw, and the smallest is FINO3. This is in agreement with the results in section 3.1.

### 3.1.6 Coefficient of variation of wind speed

Figure 8 shows the mean coefficient of variation ($C_{v,u}$) for wind speed, at the three sites. At FINO3, the average of the MMs $\tilde{C}_{v,u}$ is similar to the observations, with a bias of less than $1\%$ at all three heights. Ignoring one "outlier", the inter-model variance ranges between 3.0% and 3.5% at the three heights. The "outlier", which shows much lower values, is a consequence of the low variance for that model compared to the other models. It was removed by the filtering method described in sect. 2.3 when calculating the mean of the models ($\tilde{C}_{v,u}$) and the inter-model variance ($\tilde{\sigma}_{C_{v,u}}$). The ERA-Interim data set also captures the magnitude of $C_{v,u}$ well.

At Høvsøre, $C_{v,u}$ decreases with height for both the observations and most of the MMs. The inter-model mean of the models ($\tilde{C}_{v,u}$) agrees well with the observations, but underestimates it by about 2%. The ERA-Interim data set does not capture this

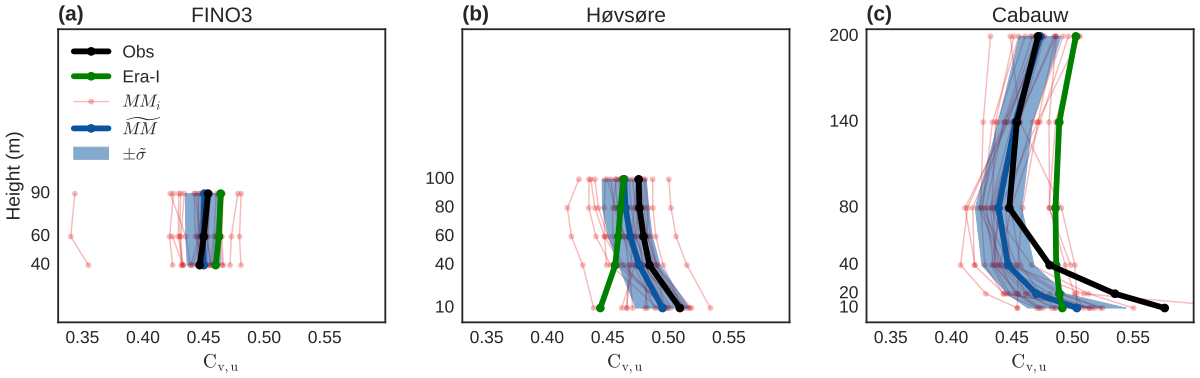


**Figure 8.** Vertical profiles of the coefficient of variation for wind speed $C_{v,u}$ at the three sites, for: observations (black), ERA-Interim (green), the mesoscale models $MM_i$ (red), and the mesoscale models mean and inter-model variance $\widetilde{MM} \pm \tilde{\sigma}$ (blue).

behavior, and instead shows an increase with height. At the highest levels, however, it reaches the average of the models and the observed values. The spread of the MMs ($\sigma_{C_{v,u}}$) is slightly higher than at FINO3 (3.6–4.4%), and is highest at the lowest levels.

At Cabauw, $C_{v,u}$ at 10 m is the largest value found across all sites. Above 10 m a sharp drop-off is found up to 80 m, where is starts to slowly increase up to 200 m. Most of the MMs capture this behavior, which is reflected in the mean of the models ($\tilde{C}_{v,u}$). However, the models underestimate the magnitude and the drop-off of $C_{v,u}$ at the lowest levels with a bias up to 12% at 10 and 20 m. Above 80 m the models agree with the observations. The ERA-Interim data set is nearly constant with height, and underestimates $C_{v,u}$ below 40 m, and overestimates it above. The inter-model variance ($\tilde{\sigma}_{C_{v,u}}$) of the MMs is largest at the lowest levels, 8.0% at 10 m, and gradually decreases to less than 4% at 200 m.

**Effect of upstream conditions on the variation of wind speed**

The coastal site Høvsøre and the offshore site FINO3 is used to investigate whether there is a dependency o the coefficient of variation for wind speed (shown in Fig. 8) on upstream surface conditions. With a nearby coastline aligned north-south, Høvsøre represents the case with anisotropic surface roughness conditions: westerly winds comes from the sea (onshore flow), and easterly winds from land (offshore flow). In ~~contract~~contrast, the offshore site FINO3 has isotropic upstream surface roughness. To study the differences, the coefficients of variation were binned according to four wind direction sectors, each spanning 90 degrees: north, east, south, and west. The values for the east and west sectors were then extracted and analyzed. Figure 9 shows the profiles of $C_{v,u}$ for the two wind directions at FINO3 and Høvsøre.

At FINO3, the coefficient of variance is almost constant with height and slightly lower for easterly winds than for westerly flow. This is true for both models and observations. The sample size for easterly winds is smaller, about half, than for westerly flow. However, both sample sizes are large (N > 1000), so the influence from sample sizes is expected to be small. The average

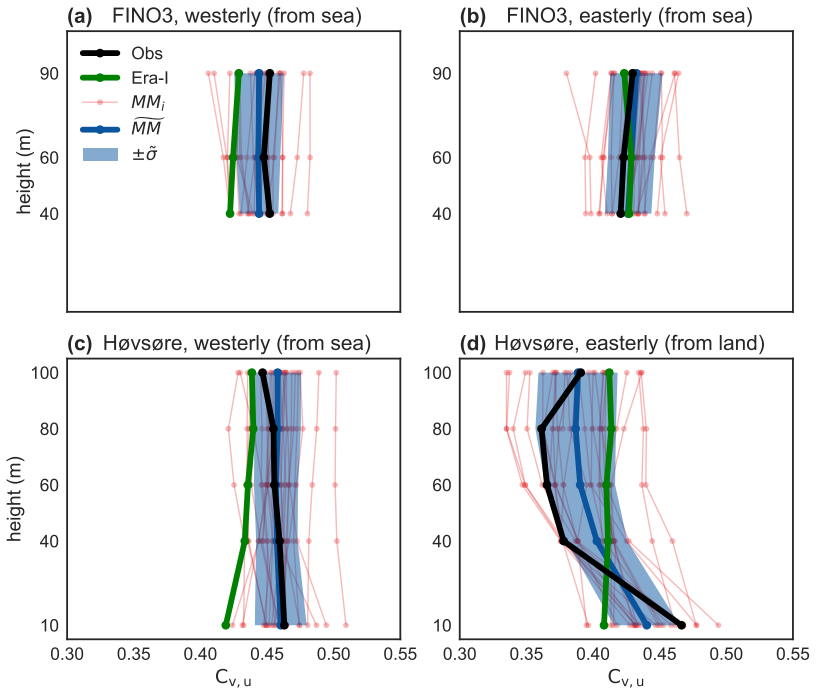


**Figure 9.** Coefficient of variation for wind speed $C_{v,u}$ for easterly (top) and westerly (bottom) winds at FINO3 (left) and Høvsøre (right), for: the observations (black), the ERA-Interim data set (green), the MMs $MM_i$ (red), and the mesoscale models mean and inter-model variance $\widetilde{MM} \pm \tilde{\sigma}$ (blue).

of the MMs captures the observed behavior well for both westerly and easterly winds, and the inter-model variance is similar for the two sectors. The ERA-Interim agrees better with the observations during easterly flow at FINO3.

At Høvsøre, the coefficient of variation is larger for westerly than for easterly winds. Easterly winds show larger coefficients of variation at 10 m than higher up. The reduction of $C_{v,u}$ with height up to 40 m for easterly flow is underestimated by most of the mesoscale models, and completely missed by the ERA-Interim data set. For westerly winds, the mean of the models and the observations agree, but is underestimated by ERA-interim.

The dependence on height of $C_{v,u}$ is only present at Høvsøre for easterly winds, and points to the influence of upstream surface conditions on the variation. The observed pattern is captured by the MMs, but the models show a more "smooth" vertical transition than do the observations. The ERA-Interim does not capture the pattern.

### 3.1.7 Distribution of wind speed shear exponent

Figure 10 shows the distributions of wind speed shear exponent ($\alpha$) for each of the three sites calculated between 40 and 80 or 40 and 90 m. Under neutral atmospheric stability conditions and isotropic surface roughness, a sharp distribution centered around a single value is expected. This means that for an offshore sites such as FINO3, the spread in shear exponent comes

primarily from variations in atmospheric stability. With this in mind, the distributions show that most MMs capture the stability well at the site. The ERA-Interim data set does not capture the strongest shear situations well. This can be easily explained by the low data frequency (6 hours).

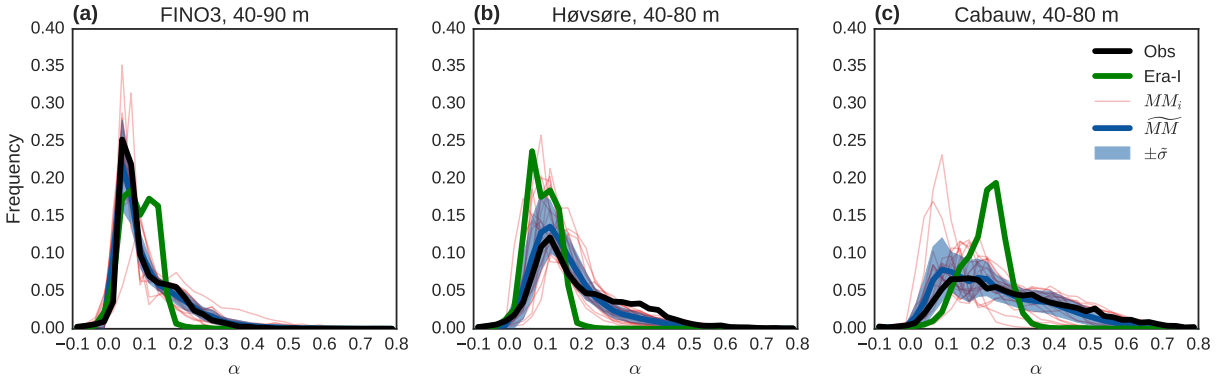


**Figure 10.** Frequency of occurrence of the shear exponent ($\alpha$) at the three sites, for: the observations (black), the ERA-Interim data set (green), the Mesoscale Models $MM_i$ (red), and the inter-model mean (blue line) and standard deviation $\widetilde{MM} \pm \tilde{\sigma}$ (blue shade).

At Høvsøre and Cabauw, the distributions of $\alpha$ reflect the combined effect of both the non-homogenous upstream surface roughness, and the variations in atmospheric stability. At the coastal site, the wind speed profile changes depending on whether the fetch is from land or from the sea, which is also reflected in the distribution of $\alpha$ ~~(Hahmann et al., 2014)~~(Hahmann et al., 2015b). Figure 10 also shows that while the shear distributions are generally also well captured at Høvsøre and Cabauw, a slight shift towards lower values is observed at both sites. This points to an underestimation of the surface roughness, a misrepresentation of the atmospheric stability, or a combination of the two. Just like at FINO3, the ERA-Interim data set does not capture the weak and strong shear cases at Høvsøre and Cabauw.

### 3.2 Relating performance to model setup

To identify what model setup choices lead to better model performance, the statistics of each model across all heights are reduced to just two values at each site: NRMSE for wind speed ($\text{NRMSE}_u$) and RMSE for wind speed shear exponent ($\text{RMSE}_\alpha$). The shear exponent was calculated between pairs of nearby levels, e.g. at FINO3 two values were calculated, one between 40 and 70 m, and one between 70 and 90 m. The $\text{RMSE}_\alpha$ was then calculated as described in section 4 between modelled and observed values of the shear exponent across all height-pairs.

Figure (11) shows $\text{NRMSE}_u$ and $\text{RMSE}_\alpha$ for all MMs at all three sites. It shows, similarly to section 3.1, that the models generally have smaller mean wind speed and mean shear exponent errors at the offshore site FINO3. But, as previously shown, errors are larger near the surface, and the three levels used at FINO3 is at 40 m and above, unlike Høvsøre and Cabuaw where levels below 40 m are included.

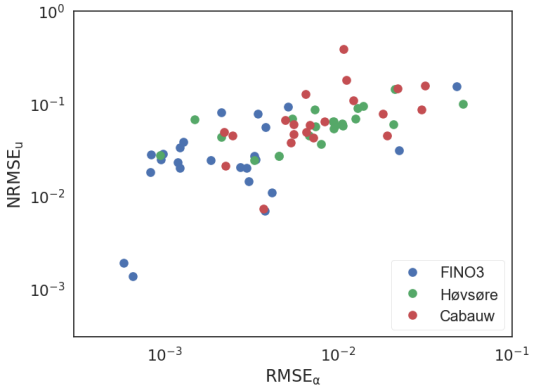


**Figure 11.** RMSE for wind speed shear exponent (RMSE$_\alpha$) versus Normalized RMSE for wind speed (NRMSE$_u$) at the three sites.

The models were then grouped according to specific model components. Given the range of setup choices that influence the model performance, large groups were needed to obtain useful statistics. With this in mind, three setup options were chosen for analysis: PBL scheme, grid spacing, and simulation lead-time, and statistics of NRMSE$_u$ and RMSE$_\alpha$ were computed for each group. The choice of groupings was based mainly on two criteria: 1) it was possible to form groups with at least six members in each group. 2) each of the options were highlighted in the literature as being important for model performance ~~(Hahmann et al., 2014; Gómez-Navarro et al., 2015; Carvalho et al., 2012; Draxl et al., 2014)~~(Hahmann et al., 2015b; Gómez-Navarro et a
Several other setup options were considered: MM, LSM, land cover, spin-up time, and data set used for initial and boundary conditions, but either it was not possible to group them in a meaningful way, or they were deemed of too little importance based on previous studies. Models missing information about particular setup options, or missing output at some heights, were excluded from this analysis.

### 3.2.1 PBL scheme

The PBL scheme in a MM ensures an accurate representation of thermodynamic and kinematic structures of the lower troposphere (Cohen et al., 2015). Two important characteristics of the PBL schemes are their order of closure and whether mixing happens through a local or a nonlocal process. Equations describing turbulent motion of order $n$ contains terms of order $n+1$. The order of closure describes the highest order of equations included, higher orders are parametrized. In local schemes variables are only affected by adjacent cells, while nonlocal schemes relate changes to gradients in the whole PBL column (Cohen et al., 2015).

To study the influence of the PBL schemes used, the MMs were split into three groups: YSU, MYJ, and Other. The statistics of NRMSE$_u$ and RMSE$_\alpha$ for these groups are shown in Table A5. The YSU group consists of six models that used the YSU PBL scheme (Hong et al., 2006), which is a first order nonlocal scheme. The models in this group span a range of grid spacings and lead-times, but models with larger than average grid spacing and longer than average lead-times dominate the group. The

MYJ group contains six models that used the MYJ PBL scheme (Janjić, 1994), ~~most of them~~ which is a 1.5 order local scheme, most of the models use a short lead-time limit, and a grid spacing that is close to the average for the MMs in this study. The last group labeled 'Other' contains nine models that used a mix of different PBL schemes (see Table A4), with different order of closures and a mix of local and nonlocal formulations. These models have a wide representation of different grid spacings and lead-times.

At FINO3, the group consisting of models not using either the YSU or MYJ PBL schemes generally have smaller wind speed errors; even though the group also contains the model with the largest $\mathrm{NRMSE}_u$. The models using the MYJ PBL scheme have smaller wind shear exponent errors, and on average also smaller wind speed errors than YSU. But the median model in the YSU and MYJ groups have similar wind speed errors.

At Høvsøre, the three groups have very similar mean wind speed error-statistics, with YSU showing only slightly smaller errors. However, for wind shear exponent the models in the YSU group have the smallest errors, both on average and for the median model. Draxl et al. (2014) studied similar error-statistics at Høvsøre for the WRF model run with a number of different PBL schemes during October 2009. They, unlike this study, found that MYJ gave slightly smaller errors than YSU. However, Draxl et al. (2014) used a version of the YSU scheme with a bug that was corrected in WRF version 3.4.1 ~~(Hahmann et al., 2014)~~(Hahmann et al., 2015b).

At Cabauw, the YSU group has smaller errors than the other groups for both wind speed and wind shear exponent, but the errors for the median model in the YSU and MYJ groups are quite similar. The single most accurate model is found in the 'Other' group, but that group as a whole has larger errors.

### 3.2.2 Grid spacing

A mesoscale model should be able to explicitly resolve smaller and smaller phenomena as the grid spacing is decreased. Skamarock (2004) illustrated that the effective resolution of the WRF model is approximately seven times the grid spacing used. However, mesoscale models, as the name suggests, have been developed to simulate the 'meso'-scale, they are often not capable of simulating weather at scales that lie between the micro- and mesoscale, i.e. between approximately 100 and 2000 m. To study the importance of the grid spacing, the models were ranked by grid spacing, similar to table A4. The models were then split into three groups: Fine, Moderate, and Coarse. The Fine group consists of seven models that all have a grid spacing below 3 km. The Moderate group consists of eight models at exactly 3 km, and the Coarse group consists of six models above 3 km. The Fine group contains models that are well distributed in terms of PBL schemes and simulation lead-time. The Moderate models also has a good representation of different PBL schemes and lead-time limits, but the MYJ PBL scheme and short lead-times are most common. The Coarse group contains no models using the MYJ PBL scheme, and half of the models use a short lead-time.

Table A6 shows the statistics for $\mathrm{NRMSE}_u$ and $\mathrm{RMSE}_\alpha$. At FINO3, the Fine group has the smallest wind speed errors. For the wind shear exponent, the smallest error is found in the Coarse group, but, on average, the Fine and Moderate groups have smaller errors. At Høvsøre, the Fine and Moderate groups have similar errors for both wind speed and shear exponent. However, the model with the smallest shear exponent error is found in the Coarse group. At Cabauw, the Moderate group

shows the smallest errors for both metrics, followed by the Fine group. But, just as for Høvsøre, the model with the smallest $\text{RMSE}_\alpha$ is found in the Coarse group.

### 3.2.3 Simulation time

As the solution in mesoscale models is integrated forward in time, the uncertainties associated with the errors in the initial conditions increase (Yoden, 2007). This can cause the model solution to drift away from the true solution. Furthermore, amplification errors can reduce the variance, which reduces the accuracy of the model in a statistical sense. To study the influence of the simulation time on the model performance, the models were ranked and split into three groups: Short, Medium, and Long. The Short group consists of nine models with a lead-time below 48 hours. Four models in the group use the MYJ scheme, and one the YSU scheme. The Short group has a good representation of models with different grid spacings. The Medium group includes eight models with a lead-time between 48 and 335 hours. The group has a good representation of different PBL schemes and grid spacing. The Long group consists of seven models with a lead-time limit above 335 hours. Five of the models use the YSU PBL scheme, and most of the models use a larger than average grid spacing.

Table A7 shows the errors statistics for the three simulation-time groups. At FINO3, the median model from the Short group has the lowest $\text{NRMSE}_u$ and $\text{RMSE}_\alpha$, but because one model has large errors, the lowest mean errors are found in the Medium group. The Medium group has smaller errors across all metrics compared to the Long group.

At Høvsøre, the Short and Long groups have similar error statistics for wind speed, and both measures are lower than those for the Medium group. For $\text{RMSE}_\alpha$ the median model from the Short group has the smallest error, while, on average, the errors are smallest in the Medium group.

At Cabauw, the smallest errors for both wind speed and shear exponent are, on average, found in the Long group, while the median model with the smallest errors are in the Short group. It is worth noting that five of the seven models in the Long group use the YSU PBL scheme, and in section 3.2.1 the models using the YSU PBL scheme were shown to have smaller errors at Cabauw, so it cannot be ruled out that the small errors in the Long group at Cabuaw is related to the over representation of the YSU scheme and not the simulation length.

## 3.3 Wind energy application

As described in sect. 2.4, the output from the mesoscale models was applied to a simple wind energy exercise. The 90-m wind resource of a Horns Rev wind farm was estimated using the output from the various MMs at FINO3. Figure 12 shows the errors for four metrics: 1) error in mean wind speed $u_m$, 2) error in mean power density $P_m$, 3) error in mean power density using a single power curve $P_{m,pc}$, and 4) error in the mean power density of a wind farm of 80 turbines $P_{m,wf}$, including wake effects.

Figure 12 shows that the majority of the models have less than $\pm 5\%$ error in mean wind speed. The errors are mostly underestimations, and, in a few cases, severe underestimation of more than $10\%$ (outside the scale of the Figure). For the mean power density, the spread of the models is, as expected, much larger due to the "third power" dependence on the wind speed. However, when the power density is calculated using a turbine power-curve, where the highest wind speeds ($>14\,\text{m}\,\text{s}^{-1}$) are less important, the inter-model variance is comparable that for mean wind speed. For the wind farm case, where the power

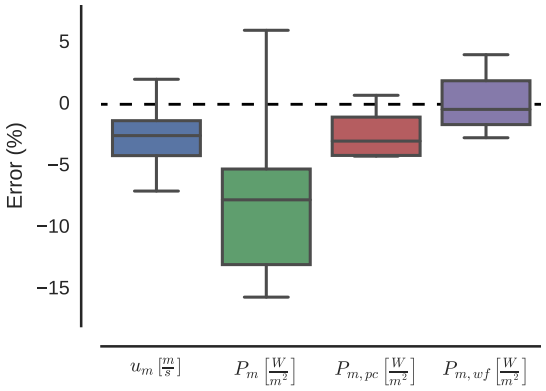


**Figure 12.** Distributions of errors from the model's output at 90 m at FINO3 for the following errors: 1) the mean wind speed $u_m$ (blue), 2) the power density $P_m$ (green), 3) the power density with an implied power curve $P_{m,pc}$ (red), and 4) the averaged power density of a wind farm including the same implied power curve as 3) and the wake effects (purple). Outliers are not shown; the most extreme ones are $-25\%$ for $u_m$, $-60\%$ for $P_m$, $-37\%$ for $P_{m,wf}$, and $-35\%$ for $P_{m,pc}$

density depends on the wind direction distribution, because of the wake losses, the variance is comparable in size to that of the mean wind speed and $P_{m,pc}$, and most models have errors smaller than $\pm 2\%$. The improvement seen for $P_{m,wf}$ is caused by the under estimation of the wake effects by most models, leading to a relative increase in mean power density, off-setting the underprediction from the modelled wind speed distribution. However, the relative effect of over- or underprediction the wake effects may just as well enhance the total power density errors, given slightly different wind direction distributions.

## 4 Summary and conclusions

The mesoscale models in this study are able to reproduce well the observed mean wind speed profiles, and the distributions of wind speed. At FINO3 and above 10 meters at Høvsøre, the average of the models has a bias of 3% or less. The largest mean wind speed biases (7–9%) are found at the lowest levels at Høvsøre and Cabauw. Similarly, the MMs were able to reproduce the relative variations of wind speed well in most cases (Fig. 8), but underestimated the relative variations at the lowest levels at Cabauw. A simple analysis of the impact of upstream surface roughness conditions on the relative wind speed variations, suggested that the models may be misrepresenting the surface characteristics (Fig. 9), which could be a misrepresentation of either the landuse classification, the conversion of landuse classes into surface roughness lengths, or in the PBL scheme. This problem highlights the need for: 1) further analysis of the representativeness of the surface characteristics in mesoscale models, and 2) downscaling the mesoscale results using a coupled microscale model to capture subgrid-scale influence from variations in orography and surface roughness. The modeled distributions of the wind direction showed only minor differences compared to the observed ones.

For future benchmarking exercises, our study shows that the focus should be on the model representation of surface characteristics, such as orography and landuse, and their associated surface roughness. An attempt was made here to include these details, but because only a subset of the participants supplied this information, it was not feasible. Further studies could also benefit from including more land masts with low to moderate complexity, where capturing the surface characteristics is important, but still manageable by mesoscale models.

The impact of choosing specific model sub-components was studied in some detail. To allow this, the output from the models was reduced to two metrics at each site, one related to the wind speed bias (NRMSE for wind speed), and one related to the shape of the wind speed profile (RMSE for wind speed shear exponent). The models were then separated into large groups according to their model setup for three setup choices: PBL scheme, grid spacing, and simulation lead-time. At FINO3, the grouping revealed that the models using the MYJ PBL scheme had smaller wind speed and shear exponent errors than those that use the YSU scheme. At Høvsøre and Cabauw, the opposite was true. However, the differences between the two groups were not significant and the median model from the two groups had similar errors. Grouping the models according to grid spacing showed that the models with 3 kilometer grid spacing or smaller had lower errors than the group with the largest grid spacings. ~~No~~ For these sites, no conclusive evidence was found that reducing the grid spacing below 3 kilometers results in smaller errors. For simulation lead-time, the median model from the group with short lead-times had the smallest errors at all sites, with the exception of the shear exponent error at Høvsøre. However, no significant difference between the mean of the groups were found, which suggests that the PBL scheme and grid spacing may be of greater importance for the performance at these sites. Future studies should include many more runs to provide more robust statistics, which can provide a basis for "best-practice" guidelines for wind energy applications using NWP models.

Last, we used the observed and modelled time-series for a classical wind energy application, the estimation of power production at a hypothetical wind farm at FINO3. The power production, including wake losses, was estimated for both a single turbine and for a wind farm, using a standard power curve. The exercise showed that while a large spread exists between the modeled power density, it is reduced when the power is calculated using a power-curve. It also showed the importance of accurately estimating the wind direction distribution, since a small deviation in the distributions might induce large changes in the power production, because of its sensitivity to the wind farm layout.

## 5 Data availability

The output data from the mesoscale models have been submitted to the European Wind Energy Association (EAWE) for the mesoscale benchmarking study under an agreement that ensures that individual participants are anonymized in the reported results, and that the model output was not publicly shared. The measurements from the meteorological masts FINO3, Høsøre, and Cabauw are provided by the data owners under an agreement of not sharing the data with any third party.

*Competing interests.* The authors declare that they have no conflict of interests

*Acknowledgements.* We would like to thank the three anonymous reviewers for constructive criticism. Their feedback elevated the level of the paper. Funding from the EU and the Danish Energy Agency through the project EUDP 14-II, ERA-NET Plus - "New European Wind Atlas" is greatly appreciated. The authors would also like to thank the European Wind Energy Association (EWEA) for organizing this mesoscale benchmarking study, the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU), and the Project Management Jülich (PTJ) for sharing the FINO3 mast data, and the Cabauw Experimental Site for Atmospheric Research (CESAR) for making the measurements from the Cabauw mast freely available online (www.cesar-labaratory.nl). Furthermore, we would like to thank the Test and Measurement section of DTU for providing the Høvsøre mast data. Finally, we would like to thank all the modeling groups that submitted output for this intercomparison. This study would not be possible without their contributions.

## References


Arino, O., Bicheron, P., Achard, F., and Latham, J.: The most detailed portrait of Earth, ESA Bull-Eur. Space, http://www.esa.int/esapub/bulletin/bulletin136/bul136d_arino.pdf, 2008.

Badger, J., Frank, H., Hahmann, A. N., and Giebel, G.: Wind-Climate Estimation Based on Mesoscale and Microscale Modeling: Statistical–Dynamical Downscaling for Wind Energy Applications, J. Appl. Meteorol. Clim., 53, 1901–1919, doi:10.1175/JAMC-D-13-0147.1, http://journals.ametsoc.org/doi/abs/10.1175/JAMC-D-13-0147.1, 2014.

Bechmann, A., Sørensen, N. N., Berg, J., Mann, J., and Réthoré, P. E.: The Bolund Experiment, Part II: Blind Comparison of Microscale Flow Models, Bound-Lay. Meteorol., 141, 245–271, doi:10.1007/s10546-011-9637-x, 2011.

Bossard, M., Feranec, J., and Otahel, J.: CORINE Land Cover Technical Guide - Addendum, European Environment Agency, pp. 1–105, http://www.eea.europa.eu/publications/COR0-landcover, 2000.

Bowler, N. E., Pierce, C. E., and Seed, A. W.: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP, Q. J. Roy. Meteor. Soc., 132, 2127–2155, doi:10.1256/qj.04.100, http://doi.wiley.com/10.1256/qj.04.100, 2006.

Carvalho, D., Rocha, A., Gómez-Gesteira, M., and Santos, C.: A sensitivity study of the WRF model in wind simulation for an area of high wind energy, Environ. Modell. Softw., 33, 23–34, doi:10.1016/j.envsoft.2012.01.019, http://linkinghub.elsevier.com/retrieve/pii/S1364815212000382, 2012.

Carvalho, D., Rocha, a., Gómez-Gesteira, M., and Silva Santos, C.: WRF wind simulation and wind energy production estimates forced by different reanalyses: Comparison with observed data for Portugal, Appl. Energ., 117, 116–126, doi:10.1016/j.apenergy.2013.12.001, http://dx.doi.org/10.1016/j.apenergy.2013.12.001, 2014a.

Carvalho, D., Rocha, a., Gómez-Gesteira, M., and Silva Santos, C.: Sensitivity of the WRF model wind simulation and wind energy production estimates to planetary boundary layer parameterizations for onshore and offshore areas in the Iberian Peninsula, Appl. Energ., 135, 234–246, doi:10.1016/j.apenergy.2014.08.082, http://linkinghub.elsevier.com/retrieve/pii/S0306261914008939, 2014b.

Champeaux, J. L., Masson, V., and Chauvin, F.: ECOCLIMAP: a global database of land surface parameters at I km resolution, Meteorol. Appl., 12, 29–32, doi:10.1017/S1350482705001519, http://doi.wiley.com/10.1017/S1350482705001519, 2005.

Cohen, A. E., Cavallo, S. M., Coniglio, M. C., Brooks, H. E., Cohen, A. E., Cavallo, S. M., Coniglio, M. C., and Brooks, H. E.: A Review of Planetary Boundary Layer Parameterization Schemes and Their Sensitivity in Simulating Southeastern U.S. Cold Season Severe Weather Environments, Weather Forecast., 30, 591–612, doi:10.1175/WAF-D-14-00105.1, http://journals.ametsoc.org/doi/10.1175/WAF-D-14-00105.1, 2015.

Constantinescu, E. M., Zavala, V. M., Rocklin, M., Lee, S., and Anitescu, M.: A computational framework for uncertainty quantification and stochastic optimization in unit commitment with wind power generation, IEEE Transactions on Power Systems, 26, 431–441, doi:10.1109/TPWRS.2010.2048133, 2011.

Cox, P. M., Betts, R. A., Bunton, C. B., Essery, R. L. H., Rowntree, P. R., and Smith, J.: The impact of new land surface physics on the GCM simulation of climate and climate sensitivity, Clim. Dynam., 15, 183–203, doi:10.1007/s003820050276, 1999.

Cuxart J, Bougeault P, R. J. L.: A turbulence scheme allowing for mesoscale and large-eddy simulations, Q. J. Roy. Meteor. Soc., 126, 1–30, doi:10.1256/qj.03.151, http://onlinelibrary.wiley.com/doi/10.1002/qj.49712656202, 2000.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Mor-

crette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. Roy. Meteor. Soc., 137, 553–597, doi:10.1002/qj.828, 2011.

Draxl, C., Hahmann, A. N., Peña, A., and Giebel, G.: Evaluating winds and vertical wind shear from Weather Research and Forecasting model forecasts using seven planetary boundary layer schemes, Wind Energy, 17, 39–55, doi:10.1002/we, http://onlinelibrary.wiley.com/doi/10.1002/we.1608/full, 2014.

Fabre, S., Stickland, M., Scanlon, T., and Oldroyd, A.: Measurement and simulation of the flow field around the FINO 3 triangular lattice meteorological mast, Journal of Wind, http://www.sciencedirect.com/science/article/pii/S0167610514000804, 2014a.

Fabre, S., Stickland, M., Scanlon, T., Oldroyd, A., Kindler, D., and Quail, F.: Measurement and simulation of the flow field around the FINO 3 triangular lattice meteorological mast, J. Wind Eng. Ind. Aerod., 130, 99–107, doi:10.1016/j.jweia.2014.04.002, 2014b.

Fernando, H. J. S. and Weil, J. C.: Whither the stable boundary layer?, B. Am. Meteorol. Soc., 91, 1475–1484, doi:10.1175/2010BAMS2770.1, 2010.

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., and Huang, X.: MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets, Remote Sens. Environ., 114, 168–182, doi:10.1016/j.rse.2009.08.016, http://dx.doi.org/10.1016/j.rse.2009.08.016, 2010.

Garbarino, R., Struzeski, T., and Casadevall, T.: US Geological Survey, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.404.5834, 2002.

García-Díez, M., Fernández, J., Fita, L., and Yagüe, C.: Seasonal dependence of WRF model biases and sensitivity to PBL schemes over Europe, Q. J. Roy. Meteor. Soc., 139, 501–514, doi:10.1002/qj.1976, 2013.

Gebhardt, C., Theis, S., Paulat, M., and Ben Bouallègue, Z.: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries, Atmos. Res., 100, 168–177, doi:10.1016/j.atmosres.2010.12.008, 2011.

Gómez-Navarro, J. J., Raible, C. C., and Dierer, S.: Sensitivity of the WRF model to PBL parametrisations and nesting techniques: Evaluation of wind storms over complex terrain, Geosci. Model Dev., 8, 3349–3363, doi:10.5194/gmd-8-3349-2015, 2015.

Grell, G., Dudhia, J., and Stauffer, D.: A description of the fifth-generation Penn State/NCAR mesoscale model (MM5), NCAR Technical note, pp. 1–121, 1994.

Gryning, S. E., Batchvarova, E., Brümmer, B., Jørgensen, H., and Larsen, S.: On the extension of the wind profile over homogeneous terrain beyond the surface boundary layer, Bound-Lay. Meteorol., 124, 251–268, doi:10.1007/s10546-007-9166-9, 2007.

Hahmann, A. N., Vincent, C. L., Peña, A., Lange, J., and Hasager, C. B.: Wind climate estimation using WRF model output : method and model sensitivities over the sea, Int. J. Climatol., doi:10.1002/joc.4217, 2014.

Hahmann, A. N., Lennard, C., Badger, J., Vincent, C. L., Kelly, M. C., Volker, P. J. H., Refslund, J., Lennard, C., Badger, J., Vincent, C. L., Kelly, M. C., Volker, P. J. H., and Refslund, J.: Mesoscale modeling for the Wind Atlas of South Africa (WASA) project, DTU Wind Energy, No. 0050, 80 pp, doi:10.13140/RG.2.1.3735.6887, 2015a.

Hahmann, A. N., Vincent, C. L., Peña, A., Lange, J., and Hasager, C. B.: Wind climate estimation using WRF model output: Method and model sensitivities over the sea, Int. J. Climatol., 35, 3422–3439, doi:10.1002/joc.4217, 2015b.

Hong, S.-Y., Noh, Y., and Dudhia, J.: A New Vertical Diffusion Package with an Explicit Treatment of Entrainment Processes, Mon. Weather Rev., 134, 2318–2341, doi:10.1175/MWR3199.1, http://journals.ametsoc.org/doi/abs/10.1175/MWR3199.1?prevSearch=&searchHistoryKey=, 2006.

Horvath, K., Koracin, D., Vellore, R., Jiang, J., and Belu, R.: Sub-kilometer dynamical downscaling of near-surface winds in complex terrain using WRF and MM5 mesoscale models, J. Geophys. Res-Atmos., 117, 1–19, doi:10.1029/2012JD017432, 2012.

Jackson, P. S. and Hunt, J. C. R.: Turbulent wind flow over a low hill, Q. J. Roy. Meteor. Soc., 101, 929–955, doi:10.1002/qj.49710143015, 1975.

Janjić, Z. I.: The Step-Mountain Eta Coordinate Model: Further Developments of the Convection, Viscous Sublayer, and Turbulence Closure Schemes, Mon. Weather Rev., 122, 927–945, doi:10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2, 1994.

Janjić, Z. I.: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model, NCEP office note, 2002.

Jiménez, P. A. and Dudhia, J.: Improving the representation of resolved and unresolved topographic effects on surface wind in the wrf model, J. Appl. Meteorol. Clim., 51, 300–316, doi:10.1175/JAMC-D-11-084.1, 2012.

Jiménez, P. A., de Arellano, J. V. G., Dudhia, J., and Bosveld, F. C.: Role of synoptic- and meso-scales on the evolution of the boundary-layer wind profile over a coastal region: the near-coast diurnal acceleration, Meteorology and Atmospheric Physics, 128, 39–56, doi:10.1007/s00703-015-0400-6, 2016.

Kallberg, P.: The HIRLAM level 1 system, Documentation manual, 1989.

Kallos, G., Nickovic, S., and Papadopoulos, A.: The regional weather forecasting system SKIRON: An overview, Proceedings of the symposium on regional weather prediction on parallel computer environments, 1997.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S. K., Hnilo, J. J., Fiorino, M., and Potter, G. L.: NCEP-DOE AMIP-II reanalysis (R-2), B. Am. Meteorol. Soc., 83, 1631–1643+1559, doi:10.1175/BAMS-83-11-1631, 2002.

Lean, H. W., Clark, P. a., Dixon, M., Roberts, N. M., Fitch, A., Forbes, R., and Halliwell, C.: Characteristics of High-Resolution Versions of the Met Office Unified Model for Forecasting Convection over the United Kingdom, Mon. Weather Rev., 136, 3408–3424, doi:10.1175/2008MWR2332.1, http://journals.ametsoc.org/doi/abs/10.1175/2008MWR2332.1, 2008.

Lock, a. P., Brown, a. R., Bush, M. R., Martin, G. M., and Smith, R. N. B.: A New Boundary Layer Mixing Scheme. Part I: Scheme Description and Single-Column Model Tests, Mon. Weather Rev., 128, 3187–3199, doi:10.1175/1520-0493(2000)128<3187:ANBLMS>2.0.CO;2, 2000.

Loveland, T. R. and Belward, a. S.: The IGBP-DIS global 1km land cover data set, DISCover: First results, Int. J. Remote Sens., 18, 3289–3295, doi:10.1080/014311697217099, 1997.

Mohan, M. and Siddiqui, T. A.: Analysis of various schemes for the estimation of atmospheric stability classification, Atmos. Environ., 32, 3775–3781, doi:10.1016/S1352-2310(98)00109-5, 1998.

Moigne, P. L. and Boone, A.: SURFEX scientific documentation, http://geosci-model-dev.net/6/929/2013/gmd-6-929-2013-supplement.pdf, 2009.

Mortensen, N. G., Nielsen, M., and Jørgensen, H. E.: Comparison of Resource and Energy Yield Assessment Procedures 2011-2015 : What have we learned and what needs to be done?, In Proceedings of the EWEA Annual Event and Exhibition 2015 European Wind Energy Association (EWEA)., pp. 1–10, 2015.

Nakanishi, M. and Niino, H.: An Improved Mellor-Yamada Level-3 Model: Its Numerical Stability and Application to a Regional Prediction of Advection Fog, Bound-Lay. Meteorol., 119, 397–407, doi:10.1007/s10546-005-9030-8, 2006.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res., 116, 1–19, doi:10.1029/2010JD015139, http://doi.wiley.com/10.1029/2010JD015139, 2011.

Noilhan, J. and Mahfouf, J. F.: The ISBA land surface parameterisation scheme, Global Planet. Change, 13, 145–159, doi:10.1016/0921-8181(95)00043-7, 1996.

Orlanski, I.: A rational subdivision of scales for atmospheric processes, B. Am. Meteorol. Soc., 56, 527, http://www.citeulike.org/group/17501/article/12086670, 1975.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Délécluse, P., Déqué, M., Díez, E., Doblas-Reyes, F. J., Feddersen, H., Graham, R., Gualdi, S., Guérémy, J. F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., Marletto, V., Morse, A. P., Orfila, B., Rogel, P., Terres, J. M., and Thomson, M. C.: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), B. Am. Meteorol. Soc., 85, 853–872, doi:10.1175/BAMS-85-6-853, 2004.

Pan, H. L. and Mahrt, L.: Interaction between soil hydrology and boundary-layer development, Bound-Lay. Meteorol., 38, 185–202, doi:10.1007/BF00121563, 1987.

Peña, A., Floors, R., and Gryning, S. E.: The Høvsøre Tall Wind-Profile Experiment: A Description of Wind Profile Observations in the Atmospheric Boundary Layer, Bound-Lay. Meteorol., 150, 69–89, doi:10.1007/s10546-013-9856-4, 2014.

Pielke, R. A., Cotton, W. R., Walko, R. L., Tremback, C. J., Lyons, W. A., Grasso, L. D., Nicholls, M. E., Moran, M. D., Wesley, D. A., Lee, T. J., and Copeland, J. H.: A comprehensive meteorological modeling system-RAMS, Meteorology and Atmospheric Physics, 49, 69–91, doi:10.1007/BF01025401, 1992.

Pleim, J.: A Simple, Efficient Solution of Flux–Profile Relationships in the Atmospheric Surface Layer, J. Appl. Meteorol. Clim., 45, 341–347, doi:10.1175/JAM2339.1, 2006.

Pleim, J. E.: A combined local and nonlocal closure model for the atmospheric boundary layer. Part I: Model description and testing, J. Appl. Meteorol. Clim., 46, 1383–1395, doi:10.1175/JAM2539.1, 2007a.

Pleim, J. E.: A Combined Local and Nonlocal Closure Model for the Atmospheric Boundary Layer. Part II: Application and Evaluation in a Mesoscale Meteorological Model, J. Appl. Meteorol. Clim., 46, 1396–1409, doi:10.1175/JAM2534.1, http://journals.ametsoc.org/doi/abs/10.1175/JAM2534.1, 2007b.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G. K., Bloom, S., Chen, J., Collins, D., Conaty, A., Da Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M., and Woollen, J.: MERRA: NASA's modern-era retrospective analysis for research and applications, J. Climate, 24, 3624–3648, doi:10.1175/JCLI-D-11-00015.1, 2011.

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y. T., Chuang, H. Y., Juang, H. M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Van Den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J. K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C. Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G., and Goldberg, M.: The NCEP climate forecast system reanalysis, B. Am. Meteorol. Soc., 91, 1015–1057, doi:10.1175/2010BAMS3001.1, 2010.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France convective-scale operational model, Mon. Weather Rev., 139, 976–991, doi:10.1175/2010MWR3425.1, http://journals.ametsoc.org/doi/abs/10.1175/2010MWR3425.1, 2011.

Simmons, A., Uppala, S., Dee, D., and Kobayashi, S.: ERA-Interim: New ECMWF reanalysis products from 1989 onwards, ECMWF newsletter, 110(110), 25–35., 2007.

Skamarock, W. C.: Evaluating Mesoscale NWP Models Using Kinetic Energy Spectra, Mon. Weather Rev., 132, 3019–3032, doi:10.1175/MWR2830.1, 2004.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, X.-Y., Wang, W., Powers, J. G., Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. G Duda, X.-Y. Huang, W. Wang, and Powers, J. G.: A Description of the Advanced Research WRF Version 3., Tech. rep., National Center for Atmospheric Research, Boulder, CO, USA, 2008.

Sukoriansky, S., Galperin, B., and Perov, V.: Application of a New Spectral Theory of Stably Stratified Turbulence to the Atmospheric Boundary Layer over Sea Ice, Bound-Lay. Meteorol., 117, 231–257, doi:10.1007/s10546-004-6848-4, http://link.springer.com/10.1007/s10546-004-6848-4, 2005.

Ulden, A. P. and Wieringa, J.: Atmospheric boundary layer research at Cabauw, Bound-Lay. Meteorol., 78, 39–69, doi:10.1007/BF00122486, http://dx.doi.org/10.1007/BF00122486, 1995.

Ulden, A. V. and Wieringa, J.: Atmospheric boundary layer research at Cabauw, Boundary-Layer Meteorology 25th Anniversary, http://link.springer.com/chapter/10.1007/978-94-017-0944-6_3, 1996.

Vincent, C. L. and Hahmann, A. N.: The impact of grid and spectral nudging on the variance of the near-surface wind speed, J. Appl. Meteorol. Clim., 54, 1021–1038, doi:10.1175/JAMC-D-14-0047.1, 2015.

Walko, R. and Tremback, C.: ATMET Technical Note 1, Modifications for the Transition from LEAF-2 to LEAF-3, ATMET, LLC, Boulder, Colorado 80308-2195, 2005.

Warner, T. T.: Numerical Weather and Climate Prediction, 2004.

Yoden, S.: Atmospheric Predictability, J. Meteorol. Soc. Jpn., 85B, 77–102, doi:10.2151/jmsj.85B.77, 2007.

**Table A1.** Site description, including latitude and longitude coordinates, classification of the site, and the height of the mast $z_s$ as well as the location terrain elevation relative to sea-level $z_{asl}$ and prevailing wind direction.

| Nr. | Name | Latitude [°] | Longitude [°] | Type | $z_s$ [m] | $z_{asl}$ [m] | Prev. wind direction |
|---|---|---|---|---|---|---|---|
| 1 | FINO3 | 55.195 | 7.158 | Offshore | 120 | 0 | WSW |
| 2 | Høvsøre | 56.441 | 8.151 | Coastal | 116 | 2 | WSW |
| 3 | Cabauw | 51.970 | 4.926 | Land | 213 | -1 | SW |

**Table A2.** Ranges of inverse Obukov length (1/L) and bulk Richardson number ($Ri_b$) used in the stability classification. The $1/L$ classes were used in Gryning et al. (2007) and the $Ri_b$ classes in Mohan and Siddiqui (1998). In both cases the original "very unstable" and the "stable" classes has been combined into the open-ended "stable" class. The same is true for the original "very unstable" and "unstable" classes which has been combined into the open-ended "unstable" class.

| Stability class | Class name | 1/L interval [$m^{-1}$] | $Ri_b$ |
|---|---|---|---|
| ~~VU~~ U | ~~Very unstable~~ Unstable | $1/L < -0.005$ | $Ri_b < ~~-0.2~~ -0.011$ |
| ~~U~~ NU | ~~Unstable~~ Near-Unstable | $-0.005 \leq 1/L < -0.002$ | $~~-0.2~~ -0.011 \leq Ri_b < ~~-0.05~~ -0.0036$ |
| N | Neutral | $-0.002 \leq 1/L < 0.002$ | $~~-0.05~~ -0.0036 \leq Ri_b < ~~0.05~~ 0.0072$ |
| ~~S~~ NS | ~~Stable~~ Near-Stable | $0.002 \leq 1/L < 0.005$ | $~~0.05~~ 0.0072 \leq Ri_b < ~~0.2~~ 0.42$ |
| ~~VS~~ S | ~~Very stable~~ Stable | $0.005 \leq 1/L$ | $~~0.2~~ 0.42 \leq Ri_b$ |

**Table A3.** Participants in the study in alphabetical order.

| Participant | Institution | Country |
|---|---|---|
| 3E | Company | Belgium |
| Anemos GmbH | Company | Germany |
| ATM Pro | Company | Belgium |
| CENER | Research Center | Spain |
| CIEMAT | Research Center | Spain |
| DEWI | Company | Germany |
| DTU Wind Energy | University | Denmark |
| DX Wind Technologies | Company | China |
| EMD International | Company | Denmark |
| ISAC-CNR | Research Center | Italy |
| KNMI | Meteorological institute | The Netherlands |
| Met Office | Meteorological institute | United Kingdom |
| RES ltd. | Company | United Kingdom |
| Statiol ASA | Company | Norway |
| University of Oldenburg | University | Germany |
| Vestas | Company | Denmark |
| Vortex | Company | Spain |

**Table A4.** Setup description of the 25 model setups ranked by horizontal grid spacing of the finest grid. The columns are: the model name and version (Model), the PBL scheme (PBL), the land surface model (LSM), whether nesting was used (Nest.), the horizontal grid spacing ($\Delta$), the land cover source, Simulation and spin-up time (Sim. time), and initial and boundary condition data (B.C.).

| Nr. | Model | PBL | LSM | Nest. | $\Delta$ [km] | Landcover | Sim. time [h] | B.C. |
|---|---|---|---|---|---|---|---|---|
| 1 | WRF V3.6.1[a] | Custom | - | yes | 1 | CORINE[b] | 48-24 | Era-I[c] |
| 2 | MAESTRO V15.01 | - | - | no | 1 | CORINE | - | Era-I |
| 3 | WRF V3.6.1 | MYJ[d] | Noah[e] | yes | 2 | USGS[f] | 78-6 | Era-I |
| 4 | WRF V3.3.1 | MYJ | - | yes | 2 | GlobCover[g] | 11064-24 | Era-I |
| 5 | WRF V3.5.1 | YSU[h] | Noah | yes | 2 | CORINE | 30-6 | Era-I |
| 6 | WRF V3.5.1 | YSU | Noah | yes | 2 | - | 264-24 | Era-I |
| 7 | HARMONIE V37h1.1[i] | SURFEX[j] | ISBA[k] | yes | 2.5 | ECOCLIMAP[l] | 7-1 | Era-I |
| 8 | WRF V3.6 | ACM2[m] | Noah | yes | 3 | USGS-MODIS | 84-12 | FNL[n] |
| 9 | WRF V3.4 | MYJ | Noah | no | 3 | USGS | 28-4 | Era-I |
| 10 | WRF V3.6.1 | YSU | Noah | yes | 3 | CORINE | 672-96 | CFSR[o] |
| 11 | WRF V3.0.1 | MYJ | Noah | yes | 3 | GlobCover | 36-6 | CFSR |
| 12 | WRF V3.6.1 | MYNN[p] | Noah | yes | 3 | USGS | 816-72 | Era-I |
| 13 | WRF V3.0.1 | MYJ | Noah | yes | 3 | GlobCover | 36-12 | MERRA[q] |
| 14 | WRF V3.0.1 | MYJ | Noah | yes | 3 | GlobCover | 36-12 | Era-I |
| 15 | WRF V3.1 | MYJ | Noah | yes | 3 | MODIS[r] | 54-6 | FNL |
| 16 | WRF V3.6.1 | YSU | Noah | yes | 3 | CORINE | 336-96 | CFSR |
| 17 | WRF V3.5.1 | MYJ | Noah | yes | 4 | IGBP-MODIS[s] | 264-24 | Era-I |
| 18 | UM V8.4[t] | Lock[u] | JULES[v] | yes | 4 | IGBP-MODIS | 36-6 | Era-I |
| 19 | WRF V3.5.1 | YSU | Noah | yes | 5 | USGS | 2424-24 | Era-I |
| 20 | SKIRON V6.9[w] | MYNN | OSU[x] | no | 5 | USGS | 51-3 | GFS[y] |
| 21 | WRF V3.5.1 | YSU | Noah | yes | 5 | USGS | 2424-24 | Era-I |
| 22 | WRF V3.5.1 | YSU | Noah | yes | 6 | IGBP-MODIS | 264-24 | Era-I |
| 23 | HIRLAM V6.4.2[z] | CBR[aa] | ISBA | no | 11 | USGS | 9-3 | IFS[ab] |
| 24 | RAMS V6.0[ac] | MYNN | LEAF[ad] | no | 12 | CORINE | 36-12 | IFS |
| 25 | MM5 V3[ae] | YSU | - | no | 20 | CORINE | 744-24 | MERRA |

[a]Skamarock et al. (2008) [b]Bossard et al. (2000) [c]Dee et al. (2011) [d]Janjić (2002)
[e]Niu et al. (2011) [f]Garbarino et al. (2002) [g]Arino et al. (2008) [h]Hong et al. (2006)
[i]Seity et al. (2011) [j]Moigne and Boone (2009) [k]Noilhan and Mahfouf (1996) [l]Champeaux et al. (2005)
[m]Pleim (2007a) [n]NCEP Final analysis [o]Saha et al. (2010) [p]Nakanishi and Niino (2006)
[q]Rienecker et al. (2011) [r]Friedl et al. (2010) [s]Loveland and Belward (1997) [t]Lean et al. (2008)
[u]Lock et al. (2000) [v]Cox et al. (1999) [w]Kallos et al. (1997) [x]Pan and Mahrt (1987)
[y]Global Forecast System [z]Kallberg (1989) [aa]Cuxart J, Bougeault P (2000) [ab]Integrated Forecasting System
[ac]Pielke et al. (1992) [ad]Walko and Tremback (2005) [ae]Grell et al. (1994)

**Table A5.** Statistics of NRMSE for wind speed ($NRMSE_u$) and RMSE for wind speed shear exponent ($RMSE_\alpha$) associated with the groups of PBL schemes across all heights at each site. The number of models in each group is: 6 in the "YSU", 6 in the "MYJ", and 9 in the "Other" group. The smallest value for each metric is in **bold**.

**FINO3**

| Metric | PBL | Mean | Median | St.d. | Min | Max |
|---|---|---|---|---|---|---|
| $NRMSE_u$ | YSU | 0.047 | 0.029 | 0.028 | 0.018 | 0.091 |
| | MYJ | 0.032 | 0.029 | **0.011** | 0.020 | **0.055** |
| | Other | **0.028** | **0.014** | 0.045 | **0.001** | 0.154 |
| $RMSE_\alpha$ | YSU | 0.029 | 0.019 | 0.034 | 0.004 | 0.116 |
| | MYJ | **0.010** | **0.010** | **0.007** | 0.004 | **0.025** |
| | Other | 0.057 | 0.019 | 0.120 | **0.003** | 0.396 |

**Høvsøre**

| Metric | PBL | Mean | Median | St.d. | Min | Max |
|---|---|---|---|---|---|---|
| $NRMSE_u$ | YSU | **0.061** | **0.058** | 0.037 | **0.024** | 0.144 |
| | MYJ | 0.063 | 0.064 | **0.013** | 0.045 | **0.090** |
| | Other | 0.062 | 0.059 | 0.026 | 0.027 | 0.100 |
| $RMSE_\alpha$ | YSU | **0.035** | **0.018** | 0.029 | **0.005** | 0.087 |
| | MYJ | 0.049 | 0.044 | **0.011** | 0.030 | **0.061** |
| | Other | 0.086 | 0.051 | 0.100 | 0.027 | 0.365 |

**Cabauw**

| Metric | PBL | Mean | Median | St.d. | Min | Max |
|---|---|---|---|---|---|---|
| $NRMSE_u$ | YSU | **0.058** | **0.049** | **0.033** | 0.021 | **0.127** |
| | MYJ | 0.066 | 0.053 | 0.037 | 0.038 | 0.146 |
| | Other | 0.124 | 0.086 | 0.106 | **0.007** | 0.389 |
| $RMSE_\alpha$ | YSU | **0.025** | **0.022** | **0.007** | 0.018 | **0.036** |
| | MYJ | 0.045 | 0.023 | 0.036 | 0.020 | 0.117 |
| | Other | 0.064 | 0.075 | 0.036 | **0.015** | 0.113 |

**Table A6.** Statistics of NRMSE for wind speed ($NRMSE_u$) and RMSE for wind speed shear exponent ($RMSE_\alpha$) associated with the group model grid spacing across all heights at each site. The number of models in each group is: 7 in "Fine", 8 in 'Moderate', and 6 in 'Coarse'. The smallest value for each metric is in **bold**.

**FINO3**

| Metric | Grid spacing | Mean | Median | St.d. | Min | Max |
|---|---|---|---|---|---|---|
| $NRMSE_u$ | Fine | **0.024** | **0.020** | **0.015** | **0.001** | **0.055** |
| | Moderate | 0.037 | 0.027 | 0.025 | 0.007 | 0.080 |
| | Coarse | 0.044 | 0.025 | 0.046 | 0.002 | 0.154 |
| $RMSE_\alpha$ | Fine | **0.013** | 0.013 | **0.008** | 0.005 | **0.025** |
| | Moderate | 0.015 | **0.011** | **0.008** | 0.004 | 0.028 |
| | Coarse | 0.067 | 0.019 | 0.121 | **0.003** | 0.396 |

**Høvsøre**

| Metric | Grid spacing | Mean | Median | St.d. | Min | Max |
|---|---|---|---|---|---|---|
| $NRMSE_u$ | Fine | 0.057 | **0.057** | 0.026 | **0.024** | 0.093 |
| | Moderate | **0.054** | **0.057** | **0.012** | 0.027 | **0.064** |
| | Coarse | 0.075 | 0.068 | 0.034 | 0.028 | 0.144 |
| $RMSE_\alpha$ | Fine | **0.040** | **0.040** | 0.021 | 0.015 | 0.076 |
| | Moderate | 0.047 | 0.048 | **0.010** | 0.030 | **0.060** |
| | Coarse | 0.088 | 0.055 | 0.109 | **0.005** | 0.365 |

**Cabauw**

| Metric | Grid spacing | Mean | Median | St.d. | Min | Max |
|---|---|---|---|---|---|---|
| $NRMSE_u$ | Fine | 0.086 | 0.064 | 0.056 | **0.007** | 0.178 |
| | Moderate | **0.048** | **0.046** | **0.015** | 0.021 | **0.078** |
| | Coarse | 0.146 | 0.107 | 0.115 | 0.049 | 0.389 |
| $RMSE_\alpha$ | Fine | 0.052 | 0.030 | 0.036 | 0.016 | 0.117 |
| | Moderate | **0.031** | **0.021** | **0.017** | 0.020 | **0.066** |
| | Coarse | 0.063 | 0.060 | 0.041 | **0.015** | 0.113 |

**Table A7.** Statistics of NRMSE for wind speed ($NRMSE_u$) and RMSE for wind speed shear exponent ($RMSE_\alpha$) associated with each group of simulation lead-time across all heights at each site. The number of models in each group is: 9 in the 'Short', 8 in the 'Medium', and 7 in the 'Long'. The smallest value for each metric is in **bold**.

**FINO3**

| Metric | Sim. length | Mean | Median | St.d. | Min | Max |
|---|---|---|---|---|---|---|
| $NRMSE_u$ | Short | 0.032 | **0.020** | 0.044 | **0.001** | 0.154 |
| | Medium | **0.028** | 0.025 | **0.014** | 0.007 | **0.055** |
| | Long | 0.051 | 0.031 | 0.028 | 0.025 | 0.091 |
| $RMSE_\alpha$ | Short | 0.052 | **0.010** | 0.122 | **0.003** | 0.396 |
| | Medium | **0.016** | 0.016 | **0.006** | **0.003** | **0.025** |
| | Long | 0.029 | 0.022 | 0.036 | 0.004 | 0.116 |

**Høvsøre**

| Metric | Sim. length | Mean | Median | St.d. | Min | Max |
|---|---|---|---|---|---|---|
| $NRMSE_u$ | Short | **0.058** | 0.059 | 0.023 | **0.024** | 0.100 |
| | Medium | 0.070 | 0.068 | **0.016** | 0.044 | **0.093** |
| | Long | 0.062 | **0.057** | 0.039 | 0.027 | 0.144 |
| $RMSE_\alpha$ | Short | 0.081 | **0.044** | 0.102 | 0.018 | 0.365 |
| | Medium | **0.044** | 0.056 | **0.023** | 0.009 | **0.076** |
| | Long | 0.046 | 0.048 | 0.025 | **0.005** | 0.087 |

**Cabauw**

| Metric | Sim. length | Mean | Median | St.d. | Min | Max |
|---|---|---|---|---|---|---|
| $NRMSE_u$ | Short | 0.088 | **0.058** | 0.108 | **0.007** | 0.389 |
| | Medium | 0.103 | 0.097 | 0.058 | 0.043 | 0.178 |
| | Long | **0.068** | 0.064 | **0.035** | 0.021 | **0.127** |
| $RMSE_\alpha$ | Short | 0.046 | **0.021** | 0.038 | **0.015** | 0.113 |
| | Medium | 0.058 | 0.054 | 0.038 | 0.018 | 0.117 |
| | Long | **0.031** | 0.025 | **0.012** | 0.020 | **0.052** |