Sensitivity Analysis of the WRF Model: Wind-Resource Assessment for Complex Terrain

SERGIO FERNÁNDEZ-GONZÁLEZ, a MARÍA LUISA MARTÍN, b EDUARDO GARCÍA-ORTEGA, c ANDRÉS MERINO, c JESÚS LORENZANA, d JOSÉ LUIS SÁNCHEZ, e FRANCISCO VALERO, a AND JAVIER SANZ RODRIGO e

a Department of Earth Physics, Astronomy and Astrophysics II, Faculty of Physics, Complutense University of Madrid, Madrid, Spain
b Department of Applied Mathematics, Faculty of Computer Engineering, University of Valladolid, Segovia, Spain
c Atmospheric Physics Group, Instituto de Matemática Interdisciplinar, University of León, León, Spain
d Supercomputing Center of Castile and León, University of León, León, Spain
e National Renewable Energy Centre (CENER), Sarriguren, Spain

(Manuscript received 4 May 2017, in final form 6 November 2017)

ABSTRACT

Wind energy requires accurate forecasts for adequate integration into the electric grid system. In addition, global atmospheric models are not able to simulate local winds in complex terrain, where wind farms are sometimes placed. For this reason, the use of mesoscale models is vital for estimating wind speed at wind turbine hub height. In this regard, the Weather Research and Forecasting (WRF) Model allows a user to apply different initial and boundary conditions as well as physical parameterizations. In this research, a sensitivity analysis of several physical schemes and initial and boundary conditions was performed for the Alaiz mountain range in the northern Iberian Peninsula, where several wind farms are located. Model performance was evaluated under various atmospheric stabilities and wind speeds. For validation purposes, a mast with anemometers installed at 40, 78, 90, and 118 m above ground level was used. The results indicate that performance of the Global Forecast System analysis and European Centre for Medium-Range Weather Forecasts interim reanalysis (ERA-Interim) as initial and boundary conditions was similar, although each performed better under certain meteorological conditions. With regard to physical schemes, there is no single combination of parameterizations that performs best during all weather conditions. Nevertheless, some combinations have been identified as inefficient, and therefore their use is discouraged. As a result, the validation of an ensemble prediction system composed of the best 12 deterministic simulations shows the most accurate results, obtaining relative errors in wind speed forecasts that are <15%.

1. Introduction

Unlike conventional energy sources, wind speed fluctuates both spatially and temporally, causing variations in wind energy production (Deppe et al. 2013). Therefore, integration of wind power into the electric grid system requires estimates of future wind speed (Liu et al. 2011). Furthermore, electrical system operators and wind farm owners need short-term wind speed forecasts to sell the energy generated in the spot market (Costa et al. 2008). In addition, accurate estimation of future wind speed can improve the safety, reliability, and profitability of the operation of several wind farms (Staid et al. 2015; Torralba et al. 2017). Wind power fluctuations can lead to undesired deviations of frequency and voltage in the electric grid (Ellis et al. 2015). As a result, the economic viability of wind energy is linked to accurate wind speed predictions, necessitating the improvement of numerical weather prediction models (NWP).

In recent years, information about the vertical wind profile has become vital because rotor diameters can be >100 m (Draxl et al. 2014). In fact, during recent decades, wind turbine size has increased, requiring the analysis of wind shear in the rotor area (Sanz Rodrigo et al. 2013). Thus, two objectives have emerged: 1) avoidance of damage caused by extreme wind events through the use of wind forecasts, and 2) more precise planning of the locations of future wind farms through the use of wind-resource assessment (Outten and Esau 2013).

Wind speed and direction outputs can be provided by NWP. However, global atmospheric models tend to greatly underestimate wind speed, particularly over...
complex terrain (Kunz et al. 2010). Therefore, the use of mesoscale models is vital in such areas. Wind flows simulated by mesoscale numerical models are sensitive to resolution, model physics, initial and boundary conditions, and parameterizations to simulate processes at the subgrid scale (Zhong and Whitman 2008). This fact suggests that different sources of error should be considered in numerical model forecasts (Siuta et al. 2017a).

The aim of numerical weather forecasts is to estimate the future state of the atmosphere. Since forecasts are not perfect, forecasts for meteorological phenomenon should be probabilistic (Gneiting and Raftery 2007). Probabilistic forecasts assign probabilities to future event occurrences and can be derived from ensemble prediction systems (EPSs). They can be used to provide estimates of the uncertainty associated with a weather forecast (Sivillo et al. 1997). However, the number of ensemble members is restricted by computer limitations.

The Weather Research and Forecasting (WRF) Model has been commonly used during recent decades to simulate, among other phenomena, wind flow over complex terrain (Hari Prasad et al. 2017). Because of its versatility, it is possible to create an EPS by several methods. One option includes the perturbation of initial and boundary conditions, in order to estimate the uncertainty associated with model assimilation (Hamill et al. 2000). However, this technique requires a long period for the ensemble spread to increase and effectively capture forecast error (Kieu et al. 2014), even periods exceeding the lead time of the simulation. The other option involves the use of various analyses and/or reanalysis as initial and boundary conditions (Hally et al. 2014). It is also possible to construct an ensemble by combining different models or physics parameterizations, creating a multiphysical ensemble (Fujita et al. 2007). This option may be more convenient for short-range mesoscale forecasts because the ensemble spread increases faster than with initial condition perturbation (Hou et al. 2001). In fact, the main source of error in a short-term forecast stems from model physics according to Costa et al. (2008). In particular for short-term forecasts in complex terrain, the planetary boundary layer (PBL) scheme and grid length are relevant factors according to the results obtained by Siuta et al. (2017a).

Wind power forecasting needs uncertainty metrics, so a probabilistic approach is convenient (Liu et al. 2011). Because an ensemble consisting only of a combination of physical parameterizations may be under-dispersive for both flatter (Deppe et al. 2013) and complex terrain (Siuta et al. 2017b), in this work we also used distinct initial and boundary conditions to increase the spread among ensemble members. Lee et al. (2012) recommended considering several sources of uncertainty to adequately characterize the probability density function of the variable of interest. Carvalho et al. (2014) identified ERA-Interim and National Centers for Environmental Prediction Global Forecast System (NCEP–GFS) analyses as the most reliable initial and boundary condition datasets, so we selected these databases in advance. In this way, ensemble spread (defined as the absolute maximum difference between any pair of ensemble members) is expected to grow (Jerez et al. 2013). One advantage of an EPS is that forecast uncertainty may be correlated with the ensemble spread magnitude (Lee et al. 2012). In addition, the error attained from the ensemble mean is on average smaller than that from the best deterministic simulation (control forecast), as suggested by Komaromi and Majumdar (2014).

This paper is focused on wind-resource assessment over complex terrain. Because model performance varies with location, atmospheric stability, and meteorological fields (Weigel et al. 2008), a sensitivity analysis of several physical schemes and initial and boundary conditions was developed for the Alaiz mountain range in the northern Iberian Peninsula, where several wind farms are located. In this way, one of the most important findings of this paper is to discover which combinations are the most accurate and which should be rejected because of their poor results in modeling wind flow over the study area. Subsequently, an EPS is proposed that provides probabilistic information and a measure of forecast uncertainty, given that the ensemble mean of this EPS yields the most accurate wind speeds for wind energy applications. Thus, the continuous ranked probability score (CRPS) is calculated. This information can be very valuable for wind farm planners during site assessment and turbine selection (Ritter and Deckert 2017).

2. Experimental design

This work is within the New European Wind Atlas (NEWA) project (Petersen et al. 2014), which includes several field campaigns for wind measurement in different locations. The present study is focused in one of these locations, the Alaiz mountain range.

a. Study area and database

A reference meteorological mast 118 m tall was used for developing a sensitivity analysis of the WRF Model. This mast is in the Alaiz range, 15 km south-southeast of Pamplona (Fig. 1). Data from Vector A100LK-PC3 cup anemometers on the mast at 40, 78, 90, and 118 m above ground level (m AGL) were used to observe wind speed.
Temperature data at 38 and 97 m AGL were from Ammonit P6312 probes. The instrument setup and calibration meets the requirements of MEASNET (2009) for cup anemometers.

The mast was installed at an altitude of 1100 m above sea level. Northward from the Alaiz range, there is a wide valley with an altitude of 400 m at the valley bottom. The Pyrenees to the northeast of the study area favor the channeling of northerly component winds from the Cantabrian Sea, known as cierzo. This wind then reaches the Ebro Valley, which is south of the Alaiz range. At synoptic scales, this wind is produced by a high pressure center over the Cantabrian Sea and a low pressure system over the Mediterranean Sea. Further explanation about the influence of complex terrain in this dataset is given in Sanz Rodrigo et al. (2013).

Following the recommendations of Sanz Rodrigo et al. (2013), the sensitivity analysis focused only on the northerly wind component (0° ± 60°) to remove data affected by flow distortion caused by the mast (because this wind direction is perpendicular to the orientation of support booms for mast instruments) or wind farms to the south. Fifteen days in May 2015 were selected for validation. This month was chosen because of full coverage of observed data. Moreover, May is representative of the transition from winter to summer weather. Specifically, alternation of the subtropical ridge and low pressure systems over the northern Iberian Peninsula was the main large-scale feature in May 2015.

The anemometer temporal resolution is 10 min. However, a simple moving average with three values on either side of the central value was used to obtain wind values representative of the analyzed hour. In addition, the northerly wind component was filtered, resulting in a database composed of 223 h (≈60% of the study period). Static stability in the PBL was analyzed by comparing potential temperatures at 38 and 97 m AGL. Over the study period, a thermal inversion was registered during 41 h, and the PBL was categorized as unstable (layer potential temperature gradient $d\theta/dz < 0$) during 40 h. PBL static stability was considered neutral ($d\theta/dz = 0$) for 41 h, although the prevailing state (101 h) was stable ($d\theta/dz > 0$).

b. WRF Model setup

We used version 3.6.1 of the Advanced Research version of WRF (ARW), which is a three-dimensional, nonhydrostatic model widely used among the scientific community for atmospheric research. A more detailed explanation of this model is in Skamarock and Klemp (2008).

Two sets of simulations were run, differing by initial and boundary conditions. The first was initialized using the NCEP–GFS analysis with 0.25° × 0.25° global latitude–longitude grid. This database includes analysis with 3-h time steps and is included within the NCEP Climate Forecast System Reanalysis (Saha et al. 2010). The ERA-Interim reanalysis (referred to as ERA) was...
used for the second set of simulations. This database is provided by the European Centre for Medium-Range Weather Forecasts on a regular grid with 0.75° × 0.75° spacing (Dee et al. 2011) and a temporal resolution of 6 h. It includes an assimilation method with four-dimensional analysis (Dee and Uppala 2009). Since both ERA and NCEP–GFS analysis are only available with some delay, they cannot be used for operational purposes, but can be useful for wind-resource assessment and hindcasting.

Simulations of the 15 days in May 2015 were run individually. This number of case studies was selected following recommendations of similar sensitivity analysis using mesoscale models (Evans et al. 2012; Johnson and Wang 2012; Fernández-González et al. 2015; García-Ortega et al. 2017). Each simulation was initialized at 0000 UTC, with a hindcast period of 36 h. However, the first 12 h were not used in the sensitivity analysis and were considered model spinup. Therefore, the sensitivity analysis was from 1200 UTC on the day of simulation start through 1200 UTC the following day.

Four nested domains were defined following a two-way nesting strategy, with spatial resolution 27 (D01), 9 (D02), 3 (D03), and 1 km (D04). Each domain had 100 × 100 grid points in the east–west and south–north directions, and the area of each is shown in Fig. 2. There were 61 sigma levels in the vertical, with increasing resolution approaching the surface to faithfully simulate winds in the PBL. A total of 8 vertical levels were included in the first 120 m AGL (mast height), and 21 vertical levels in the first 1000 m AGL.

The WRF Model offers several options for physics schemes, enabling optimization of the model for a specific weather phenomena and study area. We tested two different radiation, surface, and PBL parameterizations to evaluate which best predicted wind speed in the Alaiz mountain range. Thus, we obtained two sets (one with each initial and boundary condition as defined above) of eight deterministic simulations by combining two radiation schemes, two surface parameterizations, and two PBL schemes. As a result, a hindcast composed of 16 members was obtained.

Wind flow over complex terrain can be altered by radiative fluxes (Evans et al. 2012). Therefore, two sets of radiation schemes were tested. First, the MM5 shortwave scheme described by Dudhia (1989) for shortwave radiation, and the Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997) for longwave radiation. In the other set of simulations, the new Goddard (NG) shortwave and longwave schemes were used (Chou et al. 2001).

The WRF Model determines surface heat, momentum, and moisture fluxes mainly via the PBL scheme, which estimates wind shear and friction at the subgrid scale. For this reason, two PBL schemes were evaluated, to determine which achieved the most realistic simulation of wind speed in Alaiz range. We used the Yonsei University (YSU) PBL parameterization (Hong et al. 2006), a first-order scheme that calculates turbulent fluxes using nonlocal eddy diffusivity coefficients. This PBL scheme requires combination with the Monin–Obukhov (MO) surface-layer scheme, which through the Monin–Obukhov similarity theory provides exchange coefficients to the surface land scheme. The Mellor–Yamada–Nakanishi–Niño (MYNN) PBL scheme (Nakanishi and Niino 2006) was also tested. This PBL scheme was combined with the MYNN surface-layer scheme, which is also based on Monin–Obukhov similarity theory. According to Hari Prasad et al. (2017), the YSU and MYNN PBL schemes simulate the most realistic local winds and boundary layer characteristics.

The Noah Land Surface Model (NLSM) was selected for surface processes (Chen and Dudhia 2001). It is a four-layer soil temperature and moisture scheme that provides data of sensible and latent heat fluxes in the PBL. In addition, the Rapid Update Cycle (RUC) land surface model (Smirnova et al. 1997) was used. This scheme incorporates soil temperature and moisture in six layers and multilayer snow and frozen soil conditions. The use of cumulus parameterizations is not recommended for high horizontal resolutions, so convective processes were resolved explicitly in domains D03 and D04. For domains D01 and D02, the new Kain–Fritsch cumulus scheme (Kain 2004) was used.
For microphysics parameterization, we selected the Thompson six-class microphysics scheme (Thompson et al. 2008). This is a single-moment scheme, acting as a double-moment scheme for cloud ice and cloud rain, estimating not only the hydrometeor mixing ratio but also hydrometeor number concentrations.

A summary of combined physics schemes used in the 16 sets of simulations are shown in Table 1. These schemes were selected after an exhaustive literature review, choosing those with the best results in similar studies (e.g., Draxl et al. 2014; Evans et al. 2012; Fernández-González et al. 2015). Initially, we assumed that all simulations had the same potential predictability, as Chmel (2008) recommends. It was decided not to compare different microphysics and cumulus schemes to avoid increased computational cost.

c. Sensitivity analysis methodology

The sensitivity analysis was done at several levels within the PBL, to evaluate model accuracy dependence on height. First, the optimal grid length of the model was determined. Subsequently, each simulation was validated individually with the aim of detecting inadequate combinations of parameterizations in the wind speed forecast for the study area. Several EPSs were evaluated to select the one that gave the lowest error relative to the observed wind speed. After that, the convenience of deterministic or probabilistic approaches was considered. Finally, model performance was tested under diverse atmospheric conditions.

Because this work focused on wind energy applications, the sensitivity analysis was applied to heights of the wind turbine hub (60–90 m AGL). Specifically, the verification of model hindcasts was done at heights 40, 78, 90, and 118 m AGL. Because of the high vertical resolution of WRF in the PBL, no vertical interpolation of the model output was needed to match measurement heights; the difference between observation and model levels was <5 m.

To evaluate temporal correlation and wind speed accuracy of the hindcasts, the following indexes were calculated.

First, to visualize systematic deviations from expected values, bias was calculated by subtracting observed wind data from model estimates. Bias values > 0 indicate model overestimation of wind speed, and values < 0 reflect underestimation:

$$bias = \frac{1}{N} \sum_{i=1}^{N} (M_i - O_i).$$

Here, $M_i$ corresponds to each forecast wind speed value with $O_i$ the corresponding observed wind speed at the same time. The term $N$ reflects the total number of hours within the study period, and $i$ represents each time step.

Similarly, mean absolute error (MAE) is the average absolute forecast error, regardless of model overestimation or underestimation:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |M_i - O_i|.$$

To compare model performance during periods with strong and weak wind speeds, mean absolute percentage error (MAPE) was calculated. Large MAPE values caused by weak winds were not expected because the weakest observed wind speed during the sensitivity analysis was >2 m s⁻¹:

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|M_i - O_i|}{O_i}.$$
Finally, temporal similarity between the hindcasts and observed wind speed was evaluated by Pearson’s product moment correlation coefficient \( r \):

\[
    r = \frac{\sum_{i=1}^{N} (M_i - \bar{M})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N} (M_i - \bar{M})^2} \sqrt{\sum_{i=1}^{N} (O_i - \bar{O})^2}}.
\]

An accuracy metric, adequate for probabilistic forecasts, was also applied: the CRPS. Indeed, the CRPS is similar to the MAE, but for probabilistic forecasts. Its units are the same than those of the observed variable, but in this paper we have decided to also show the values in percentage terms (relative CRPS) to compare episodes with distinct wind speed. According to Gneiting et al. (2005), the CRPS offers a strictly proper scoring metric, evaluating both forecast calibration and sharpness in a single score. Thus, an accurate representation of forecast uncertainty is obtained, facilitating the decision-making process (Juban et al. 2007).

In addition, a reliability diagram was used to assess forecast calibration (Wilks 2011). We have tested the wind at 90 m AGL, fixing a threshold of 10 m s\(^{-1}\), to represent turbine-rated winds in modern wind turbines (speed at which turbines begin to produce the maximum power output). A sharpness histogram is inserted within the reliability diagram to measure the width of the forecast probability distribution (Gneiting et al. 2007). The sharpness histogram indicates the frequency of hours that such a threshold is exceeded in each probability bin.

Finally, rank histograms (Hamill 2001) were used for verifying ensemble forecasts. Rank histograms are obtained by determining which ensemble forecast bin matches the associated observed wind speed value.

### 3. Results and discussion

#### a. Sensitivity analysis

1) **DEPENDENCE OF MODEL PERFORMANCE ON SPATIAL RESOLUTION**

Figure 3 summarizes estimated results of the sensitivity analysis at 40, 78, 90, and 118 m AGL for domains D02, D03, and D04 and the 16 simulations. Domain D01 was not considered in the validation because of its low horizontal resolution (grid length = 27 km). Validation of D02, D03, and D04 examined improvement in model performance according to horizontal resolution. In Fig. 3 (and Figs. 4, 6, and 7), the color scale is defined such that light grays indicate the best values of the validation indices (\( r \), bias, MAE, and MAPE), and dark grays reflect the poorest values.

With regard to grid length, sharp improvement of model accuracy was observed in the hindcast of wind speed magnitude when model resolution was increased to 1 km (D04). For instance, the bias indicates marked underestimation (>3 m s\(^{-1}\) at 40 m AGL; >2 m s\(^{-1}\) at greater heights) for 9- and 3-km grid spacing (D02 and D03), with slight underestimation (~1 m s\(^{-1}\) at 40 m AGL; ~0 m s\(^{-1}\) at 78, 89, and 118 m AGL) in domain D04 (1-km grid spacing). Similarly, MAE and MAPE results show remarkable improvement in the highest-resolution domain. MAPE decreases from >25% in domains D02 and D03 to <16% in domain D04. Nevertheless, \( r \) did not increase with horizontal resolution, indicating no improvement in temporal evolution of wind speed with increasing spatial resolution. Smaller \( r \) values in D04 may be connected with the fact that phase errors tend to increase in high-resolution simulations (Mass et al. 2002). Taking these results into account, it may be beneficial to increase the model resolution to 1 km in the study area during episodes with a northerly wind component because of remarkable improvement in the estimation of the wind flow over complex terrain.

2) **EVALUATION OF PHYSICS SCHEMES AND INITIAL AND BOUNDARY CONDITIONS FOR D04**

Because the results of the previous section demonstrate that a horizontal resolution of 1 km was the best for these case studies, a more detailed analysis of model performance at various heights is displayed in Fig. 4, only for the highest-resolution domain. There was considerable underestimation at 40 m AGL (bias from −2.4 to −0.6 m s\(^{-1}\)), and slight underestimation at 78, 90, and 118 m AGL (bias near 0 m s\(^{-1}\)).

In general, the GFS-initialized simulations underestimated wind speed more than those initialized with ERA-Interim. Simulations using the YSU PBL scheme (code ending in 1 in Fig. 4) had a better performance than those using the MYNN parameterization (which had considerable underestimation). In research carried out by Hari Prasad et al. (2017), they explained that the deeper mixed layer simulated by the MYNN scheme, in contrast to the shallow mixed layer estimated by the YSU PBL parameterization, might lead to wind speed underestimation in their study area. MAE and MAPE values were slightly better in the case of GFS members. Nevertheless, ERA-Interim simulations were better than those initialized with GFS analysis for temporal evolution because \( r \) values were larger. Therefore, an EPS developed by combining ERA-Interim and GFS analysis can be very useful for wind-resource assessment since each one obtains better results depending on the metrics evaluated.

With regard to the deterministic simulations, poor results of simulations 135 and 535 stand out for both
GFS and ERA-Interim initial and boundary conditions. Although temporal evolution was also poorer, the most worrying results were shown by MAE and MAPE values, especially at 40 m AGL. These results discourage the combination of the RUC land surface model with the MYNN PBL and surface-layer schemes. On the contrary, the best deterministic simulation was initialized with GFS analysis, the NG radiation scheme, NLSM land surface model, and MYNN PBL and surface-layer schemes (GFS 525). Although this simulation underestimated wind speed a bit more than other deterministic simulations, the correlation coefficient, MAE, and MAPE were slightly better than the others.

Finally, we also considered the option of developing a hindcast composed by different deterministic simulations. In this regard, distinct ensembles were defined, testing their respective ensemble means. The construction of ensembles by combining distinct physical...
parameterizations and initial and boundary conditions is based on the assumption that all model configurations are equally skillful. As a result, the ensemble mean tends to provide more skillful information than any individual EPS member (Grimit and Mass 2002). The reason is that the ensemble mean attenuates random errors (Warner 2011).

Values of validation indexes for distinct ensemble systems are shown in the four rightmost columns of the panels in Fig. 4. We developed an ensemble with eight simulations initialized with the GFS analysis (denoted by e.m. GFS in Fig. 4), one with eight simulations initialized with the ERA-Interim reanalysis (e.m. ERA),
one with all the simulations (e.m. 16), and one of 12 members by removing the worst four simulations (e.m. 12).

For temporal evolution, r values of e.m. 16 and e.m. 12 were >0.93, comparable only to the most accurate deterministic forecast (GFS 525). MAE and MAPE values were also improved by using e.m. 12, demonstrating the convenience of selecting an EPS but removing the poorest ensemble members (those in which the RUC land surface model and MYNN PBL scheme were used together). MAE values < 1.5 m s⁻¹ were obtained at 40 m AGL, decreasing to <1.25 m s⁻¹ at higher altitudes. Staid et al. (2015) considered errors between 1 and 2 m s⁻¹ satisfactory for wind energy applications. However, ensembles were not able to correct the slight underestimation (bias values between −0.4 and −0.1 m s⁻¹).

Following the recommendations of Lee et al. (2012), inaccurate or redundant ensemble members should be removed from the EPS in order to minimize computational cost and optimize model performance. For this reason, we consider the EPS with 12 ensemble members as the best option for estimating wind speed for our case study. In addition, the use of different initial conditions and physical parameterizations reduces overconfidence (the degree of ensemble underdispersion), producing a net gain in forecast skill (Weigel et al. 2008).

3) VALIDATION OF VERTICAL WIND PROFILE

Mesoscale models must provide accurate information about the wind profile around hub height (commonly 60–100 m AGL) to help minimize stress on the rotor (Draxl et al. 2014). Average wind speeds over the entire study period are represented in Fig. 5, both the observation and the mean of different combinations of hindcasts (all the simulations initialized with the same data source and using the same PBL scheme are represented in the same way). This figure indicates general underestimation in almost all hindcasts, revealing greater systematic underestimation in hindcasts initialized with GFS analysis. These results agree with those of Carvalho et al. (2014), who concluded that the ERA-Interim provides the most realistic initial and boundary conditions for wind-resource assessment.

There was notable wind speed underestimation at low levels in the boundary layer, with less error at higher altitudes. According to Arroyo et al. (2014), the reason may be connected to the fact that mesoscale models smooth the orography and strengthen wind gradients in the lower levels of the PBL. Our results agree with those of Hari Prasad et al. (2017), who observed greater underestimation in simulations using the MYNN PBL scheme and also showed the YSU scheme as the most accurate PBL parameterization for their study area. Although postprocessing was not the aim of the present work, our results indicate that it might mitigate the observed systematic underestimation, especially at 40 m AGL. Zhu et al. (2002) recommended statistical postprocessing in ensemble forecasts to remove systematic errors. Indeed, Siuta et al. (2017a) found up to 56% improvement in MAE values by using a simple mass-balance factor. This issue will be studied in future research because more case studies and observation locations are needed for developing accurate postprocessing.

4) WIND SPEED DEPENDENCE

Figure 6 shows the behavior of WRF physics parameterizations depending on wind speed. The entire study period was divided into hours with the observed wind speed stronger and weaker than 10 m s⁻¹. This threshold was chosen because 10 m s⁻¹ was the approximate average wind speed in the Aïhaiz mountain range over the study period.

As shown in Fig. 6, MAPE values are much better during strong wind episodes (10%–15%) than for wind speed < 10 m s⁻¹ (20%–40%). Furthermore, there is a systematic underestimation (bias < −1 m s⁻¹) for speeds > 10 m s⁻¹ at 40 m AGL. According to the results exposed in this paper, it appears that the WRF Model overestimates friction losses at low boundary layer levels, causing systematic underestimates of wind speed at 40 m AGL, especially during strong wind episodes. The poor results at the lowest levels of the PBL could be related to the lack of observational data over complex terrain on the initial and boundary condition data sources, causing unrealistic initial conditions. Another reason could be poor surface-layer physics that are applied over complex terrain, possibly connected to the use of inadequate surface-layer similarity theories. Nevertheless, we have not enough evidence to determine the exact reason of these poor results with the outcomes obtained in this paper.

In addition, bias values indicate that simulations initialized with GFS analysis performed better under weak winds, whereas those using ERA-Interim reanalysis were more accurate under strong winds. As mentioned above, simulations 135 and 535 gave the worst results. According to Fig. 6, this is mainly because of their poor performance during weak winds. In contrast, GFS 525 and GFS 531 produced the best values of r, MAE, and MAPE during weak winds. As referred to in previous section, the combination of ERA-Interim and GFS analysis in an EPS system for wind-resource assessment could be convenient since the performance of each one is higher during certain weather conditions.

5) VALIDATION FOR DIFFERENT ATMOSPHERIC STABILITY CONDITIONS

To check if WRF wind speed forecast errors are dependent on atmospheric stability, verification of the
various WRF simulations during different static stability conditions was done (Fig. 7). For temporal evolution, $r$ values indicated that the WRF Model performed better during unstable conditions. In contrast, the temporal evolution of wind speed is very unpredictable when a thermal inversion is dominant.

Wind speeds are underestimated by the WRF Model simulations during neutral and unstable conditions, especially at 40 m AGL. Nevertheless, this underestimation is reduced with thermal inversions, even changing to overestimation in simulations initialized using the GFS analysis. During stable conditions, there was underestimation at lower levels (40 and 78 m AGL), but slight overestimation is observed at upper levels (90 and 118 m AGL), especially in simulations initialized with the ERA-Interim.

MAE and MAPE were notably poorer during thermal inversions. In the bottom panels of Fig. 7, the worst results correspond again to GFS 135, GFS 535, ERA 135, and ERA 535 simulations, mainly for unstable conditions. For neutral and unstable atmospheric stability classes, there were poorer bias, MAE, and MAPE values at 40 m AGL. However, MAE was worse at 118 m AGL under thermal inversions, possibly because this altitude can be near the upper limit of the inversion layer. Thus, bias, MAE, and MAPE were better in the simulations using the YSU PBL scheme, especially during unstable conditions. These results are in agreement with those of Draxl et al. (2014).

There was no combination of parameterizations that performed better for every episode, that is, some schemes were more accurate under certain meteorological conditions, and others during other weather regimes. The
same conclusion can be applied to initial and boundary conditions. This leads us to conclude that it is appropriate to develop probabilistic forecasts so that we have different ensemble members, each being superior for a certain meteorological situation. Since it is almost impossible to know ahead of time which of the ensemble members will be the best during each episode, the use of the ensemble mean allows obtaining the lowest error when statistics are aggregated over long study periods (Warner 2011). In addition, an EPS can be used to derive probabilities of possible future wind speeds (a probabilistic forecast), thereby also giving an estimate of forecast uncertainty.

b. Probabilistic verification

A reliability diagram and a sharpness histogram were generated using a threshold of 5 m s$^{-1}$. This value was
selected for analyzing the events in which the cut-in speed is exceeded. Figure 8 shows the values for the EPSs composed by 16 and 12 members. The forecast probability is calculated by relating the number of members of the EPS that exceeds the threshold, while the observed frequency represents the relation of observations that exceeds the threshold for each forecast probability. The reliability diagram indicates that both EPSs are fairly well calibrated, since both reliability curves are close to the diagonal (black line). In addition, the reliability diagram also shows that both EPSs are characterized by overestimation (underestimation) in the lower (higher) forecast probabilities. Thus, the EPSs developed in this paper are slightly underdispersive. The overestimation in the lower forecast probabilities is lower in the EPS composed by 12 members, while the
underestimation in the higher probabilities is attenuated by using the EPS of 16 members. Therefore, the election of one or another EPS should be decided depending on the needs of the user.

In addition, a sharpness histogram was inserted in the bottom-right corner of Fig. 8. This histogram specifies how often the event threshold is in each forecast probability bin. Therefore, probabilities closer to 0 and 1 indicate sharper forecasts. In these EPSs, the majority of the forecasts obtain a probability of 0 or 1, showing a high sharpness, especially in the EPS composed of 12 members, in which the frequency of forecasts with a probability of 0 or 1 is slightly higher than in the EPS composed of 16 members. According to Siuta et al. (2017b), a high sharpness is related to underdispersive EPSs, connected to the overconfident nature of the ensemble distributions.

Finally, rank histograms for the EPSs composed by 16 and 12 members are shown in Fig. 9. These tools were used for evaluating the reliability of the EPSs. In both cases, rank histograms indicate underdispersive ensembles, since observed values fall very often at the extremes of the ensemble distribution (flatter histograms would be desirable). The fact that the rightmost bar is considerably taller than the leftmost bar indicates that the EPSs have deficiencies in the forecast of extreme high wind speed values (Hamill 2001). In addition, a not flat histogram can also indicate the existence of bias (Siuta et al. 2017b), confirming our results obtained in the sensitivity analysis developed in this paper.

c. Evaluation of case studies

Finally, two episodes were selected to describe in detail the influence of initial and boundary conditions and physical schemes on the wind speed forecast. These episodes were chosen for their distinct wind speeds and atmospheric stabilities. Both episodes meet the requirement of a northerly wind component ($\pm 60^\circ$) throughout the events.

1) 15–16 May 2015

First, we analyzed an event with strong sustained winds and statically stable PBL. This event had strong synoptic forcing. A north-northwest wind was generated...
by a high pressure center to the northwest of the Iberian Peninsula and a low pressure system over the Gulf of Lion. Figure 10 shows wind speeds > 10 m s\(^{-1}\) during the entire episode, with a relatively stable speed throughout. Stable atmospheric conditions were predominant. Under these conditions, the spread between ensemble members was small, with all deterministic simulations very accurate. This is in accord with Toth et al. (2001), who claimed that the ensemble mean is more likely to be verified when the spread is smaller, particularly at mid-latitudes where model physics are accurate (Buzzi 1997).

In fact, \(\text{MAPE} = 9.40\%\) was obtained from the ensemble mean of 12 simulations. During this episode, there are two ensemble members slightly more accurate than the ensemble mean (\(\text{MAPE} = 9.39\%\) and 8.72\% were obtained for GFS 131 and GFS 525, respectively). This is likely because the verification is only applied to a single day, which is not long enough to show the true benefits of using the ensemble mean. However, when developing the verification over the 15 days, the ensemble mean indicates a more accurate verification than any of the ensemble members.

In addition, the CRPS for this day was also calculated, obtaining a value of 0.59 m s\(^{-1}\). Considering that the average wind observed during this day in the Alaiz mountain range was of 14.5 m s\(^{-1}\), a relative CRPS of 4.0\% was obtained.

Stability of the wind component and speed is remarkable in Fig. 11 during both day (left panel) and night (right panel). Wind flow from the Cantabrian Sea reached the valley to the north of the Alaiz mountain range, intensified by orographic forcing from this relief. Subsequently, the flow was channeled into the Ebro Valley. These results indicate that the WRF Model was able to accurately predict strong wind episodes in the study area characterized by intense synoptic forcing.

2) 29–30 MAY 2015

This subsection describes the analysis of a light wind speed episode caused by weak large-scale forcing. At synoptic scale, the period 29–30 May 2015 was characterized by a weak surface pressure gradient to the north of the Iberian Peninsula, causing weak synoptic forcing of airflow reaching the Alaiz mountain range. During days characterized by a statically stable PBL, a diurnal cycle of wind speed is seen in Fig. 12, since wind speed is greater during daytime than at night because of radiative processes. Atmospheric stability was also affected by radiative processes. Clear skies favor surface heating (daytime) and radiative cooling (nighttime), causing instability during daytime and formation of a thermal inversion layer at night, respectively.

As seen in Fig. 12, all ensemble members were very similar to observed wind speeds during early hours (daytime), although the spread between ensemble members and error grew during nighttime, contemporary with the dominance of the thermal inversion. By analyzing error from the ensemble mean of the best 12 simulations, \(\text{MAPE} = 18.61\%\) was calculated for this episode, approximately twice the MAPE obtained from the strong wind speed event with atmospheric stability. As in the previous study case, there are six ensemble members (all the ensemble members initialized with the GFS analysis except simulations 135 and 535) that improve the verification of the ensemble mean during this day. Nevertheless, the ensemble mean obtains the best verification when averaging the results of the whole study period.

After calculating the CRPS, a value of 0.54 m s\(^{-1}\) was obtained. Although it is slightly inferior to those estimated in the previous case study, if we consider that the average wind observed during this day in the Alaiz mountain range was of 7.6 m s\(^{-1}\) (approximately half as the previous case), a higher relative CRPS is obtained: 7.1\%.

This is in accord with Grimit and Mass (2002), who observed that small-spread events are more predictable than those with large spread. In the same vein, McMurdie and Ancell (2014) claimed that a large ensemble spread is connected with poor predictability. Analogously, Hally et al. (2014) proved that a probabilistic forecast is less skillful during weak low-level flow.

In this episode, simulations initialized using the GFS analysis indicated a better temporal evolution (higher correlation with the observations) of wind speed during thermal inversion conditions than those initialized with ERA-Interim reanalysis. Although Stensrud and Fritsch (1994) asserted that using different model schemes might be a better option during weak large-scale forcing episodes, our results indicate that spread can also be generated by combining distinct initial and boundary conditions. Thus, the ensemble mean can retain information about anomalous wind flows (Toth 2001).

Shin and Hong (2011) suggested that the YSU PBL scheme does not adequately represent the diurnal wind cycle, but our results show a similar behavior of the MYNN and YSU parameterizations during daytime and nighttime, with the initial and boundary conditions more influential in simulating that cycle. Despite the fact that ERA-Interim reanalysis yielded better overall results, under certain weather conditions the GFS analysis appeared more valid, justifying the use of an EPS for reliable information under every atmospheric situation. Indeed, Carvalho et al. (2014) pointed out that the GFS analysis is the best alternative to the ERA-Interim reanalysis, especially for real-time forecasts because of its rapid availability (the ERA-Interim reanalysis is...
available after a two-month waiting period). These results indicate that considerable error comes from the initial and boundary conditions introduced in the model. Nevertheless, there has been improvement of data assimilation and resolution of reanalyses during recent years, increasing the accuracy of mesoscale wind forecasts and generating realistic simulations of local wind regimes in complex terrain.

The maximum spread between ensemble members was at 0500 UTC 30 May 2015, when the hindcast error is larger. Compared with the previous event, the relative spread is smaller throughout the episode, as was MAPE. This is consistent with Grimit and Mass (2002), who claimed that the magnitude of the ensemble spread is indicative of the forecast error scale. Thus, we decided to analyze the existing correlation between the observed

![Graph showing rank histograms for the EPS composed by (top) 16 and (bottom) 12 members. The x axis indicates the position within the EPS in which the observed value falls, and the y axis represents the frequency of each bin.](image-url)
error (estimated from the ensemble mean) and the ensemble spread during the whole study period (15 days). A correlation coefficient of 0.26 was obtained, with a \( p \) value of \( 2 \times 10^{-2} \), significant for a 0.05 significance level. Therefore, the null hypothesis (the two variables are uncorrelated) can be rejected. Although a similar value was obtained by Fernández-González et al. (2017), this correlation coefficient can be considered quite low, which might be related to the small dataset used in this study.

Wind flow differences between day and night are portrayed in Fig. 13. Daytime conditions are shown in the left panel, when northerly winds were prevalent. However, speeds were lighter than in the previously analyzed event because of weak synoptic forcing. Under these conditions, the flow was partially blocked by the orography and, as a result, it was channeled to the east and west of the Alaiz range because of the lower altitude, strengthening the wind in those areas. At night, the flow was irregular under a statically stable PBL, consistent with Hari Prasad et al. (2017). Without wind intensification caused by thermal gradients generated by varying insolation conditions, wind speeds decrease notably. This is especially noticeable in the valley to the north of the Alaiz range, where the wind was nearly calm, probably because of a thermal inversion layer formed by radiative cooling. On the leeward side of the Alaiz range, wind flow was channeled into the Ebro Valley.

Wind speed is less predictable under weak synoptic forcing and changing atmospheric stability, especially during thermal inversions. These results demonstrate that sensitivity to physical parameterizations and to initial and boundary conditions are case dependent, in agreement with Hally et al. (2014).

4. Concluding remarks

Mesoscale models are a useful tool in forecasts of wind speed in a given region. Prior testing of the model can be used to determine optimum WRF Model configurations; however, the best configuration is a function of location and meteorological conditions. After developing a sensitivity analysis of several initial and boundary conditions, as well as several physical schemes of the WRF Model, the following conclusions were reached:

- High horizontal resolution is useful for realistically simulating wind flow over the complex terrain present
In the study area, model accuracy of the wind speed hindcast improves with altitude increase; the poorest results were at lowest study level (40 m AGL), especially during strong wind events. The reasons for these poor results should be further analyzed in future research.

There was systematic wind speed underestimation in simulations initialized with the MYNN PBL scheme, which was partially corrected using the YSU parameterization. Throughout the study period, the deterministic simulation with the best results was initialized with the GFS analysis, NG radiation scheme, NLSM land surface model, and MYNN PBL and surface-layer schemes. In contrast, one of the main findings of this paper was to discover that simulations in which the RUC land surface model was combined with the MYNN PBL scheme gave the worst results, so its use for wind speed prediction in the Alaiz mountain range is discouraged.

Another important finding is that the verification depends markedly on wind speed. Indeed, MAPE was sharply reduced during strong wind speed episodes. Simulations initialized with the GFS analysis had superior results during low wind speed events. The results suggest that model accuracy is greater during episodes with weak synoptic forcing in the Alaiz mountain range.

It is also very important to note that the type of atmospheric stability greatly altered model performance. The poorest results were under thermal inversion conditions. Model skill increased during stable, neutral, and unstable conditions. Temporal correlation was also much greater during unstable situations.

Because model accuracy is case dependent, the development of an EPS composed of different physical schemes and initial and boundary conditions is advisable. In fact, the best results were attained using an EPS of 12 members (MAPE < 15%), from which the four ensemble members with the poorest results were removed.

The probabilistic verification developed in this paper detected high sharpness in the EPSs. However, the rank histograms show that the EPSs developed in this paper are underdispersive.

In conclusion, the results of the present research demonstrate that the WRF mesoscale model is able to realistically simulate wind flow over the complex terrain present in the study area. However, a sensitivity analysis is necessary to find the combination of physical parameterizations and initial and boundary conditions with a better performance under different atmospheric conditions. EPSs are useful because the use of the ensemble mean can improve the accuracy of deterministic prediction. Moreover, EPS use allows the user to estimate...
uncertainty associated with a given forecast, based on the dispersion of different ensemble members. The inaccuracies detected in the probabilistic verification will be treated in a future paper by including additional uncertainty in the design of the EPS, and by implementing methods for dressing the forecast distribution. Future research will also focus on quantification of the uncertainty and predictability of the ensemble proposed.
in the present work, and analysis of its applicability to other regions. In addition, a longer evaluation period would be beneficial.

Acknowledgments. This work was partially supported by research projects METEORISK (RTC-2014-1872-5), PCIN-2014-013-C07-04, PCIN-2016-080 (UE ERA-NET Plus NEWA Project), ESP2013-47816-C4-4-P, CGL2010-15930, CGL2016-78702-C2-1-R and CGL2016-78702-C2-2-R, CGL2016-81828-REDT, and the Instituto de Matemática Interdisciplinar (IMI) of the Universidad Complutense. Special thanks go to Roberto Weigand. The authors also thank the Centro Nacional de Energías Renovables (CENER) for providing the database used.

REFERENCES


Torralba, V., F. J. Doblas-Reyes, D. MacLeod, I. Christel, and M. Davis, 2017: Seasonal climate prediction: A new source of information for the management of wind energy resources.