METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **19**: 484–500 (2012)
Published online 8 August 2011 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.283

RMetS
Royal Meteorological Society

Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations

Pierre Pinson[a,b]* and Renate Hagedorn[b]
[a] *DTU Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark*
[b] *European Centre for Medium-range Weather Forecasts, Reading, UK*

ABSTRACT: A framework for the verification of ensemble forecasts of near-surface wind speed is described. It is based on existing scores and diagnostic tools, though considering observations from synoptic stations as reference instead of the analysis. This approach is motivated by the idea of having a user-oriented view of verification, for instance with the wind power applications in mind. The verification framework is specifically applied to the case of ECMWF ensemble forecasts and over Europe. Dynamic climatologies are derived at the various stations, serving as a benchmark. The impact of observational uncertainty on scores and diagnostic tools is also considered. The interest of this framework is demonstrated from its application to the routine evaluation of ensemble forecasts and to the assessment of the quality improvements brought in by the recent change in horizontal resolution of the ECMWF ensemble prediction system. The most important conclusions cover (1) the generally high skill of these ensemble forecasts of near-surface wind speed when evaluated at synoptic stations, (2) the noteworthy improvement of scores brought by the change of horizontal resolution, and, (3) the scope for further improvements of reliability and skill of wind speed ensemble forecasts by appropriate post-processing.

KEY WORDS probabilistic forecasting; skill; calibration; benchmark; scores; observational uncertainty; energy application

Received 12 November 2010; Revised 24 May 2011; Accepted 20 June 2011

1. Introduction

One of the major recent breakthroughs in meteorological prediction comes from the transition from point to probabilistic forecasting (Palmer, 2000; Gneiting, 2008). This phenomenon is not only observed in the meteorological literature, since probabilistic forecasts are also becoming customary products in economics and finance (Abramson and Clemen, 1995; Tay *et al*., 2000; Timmermann, 2000). Regarding meteorological prediction for decision-making in the energy field, for instance, it has been demonstrated that the optimal management and trading of wind energy generation calls for probabilistic forecasts, see Matos and Bessa (2011) and Pinson *et al*. (2007a) among others. This actually follows from a more general result which is that for a large class of decision-making problems, optimal decisions directly relate to quantiles of conditional predictive distributions, as discussed by Gneiting (2011a).

Forecasts ought to be evaluated, and various frameworks exist depending upon which of the forecast's characteristics are to be highlighted. Primarily, one should make a difference between the quality and value of the forecasts, following the discussion of Murphy (1993). The former relates to the objective evaluation of intrinsic forecast performance, while the latter is based on the benefits perceived by forecast users when making decisions based on these forecasts. Even though these two concepts have often been kept apart in the forecast verification literature, their linkage has been the focus of a few studies (see for example Katz and Murphy, 1997, and references therein for the case of forecasts of weather and climate).

Forecast quality verification is also a multi-faceted problem in the sense that a large number of scores and diagnostic tools may be considered. One could, for instance, start by looking at first-order statistics such as the bias of point forecasts or the marginal reliability of probabilistic forecasts. Scores (Mean Absolute Error – MAE, Root Mean Square Error – RMSE, Continuous Ranked Probability Score – CRPS) may additionally be considered, as well as corresponding skill scores after definition of a benchmark e.g. climatology. Finally, diagnostic approaches may be based on the joint distributions of forecasts and verifications (Murphy and Winkler, 1987). Consequently the appraisal of verification statistics and scores is a subtle task, as rightly pointed out and discussed by Mason (2008).

A core aspect of forecast verification is the definition of the reference against which the forecasts are evaluated. A common practice in the meteorological research community is to employ the model analysis as such a reference, since it comprises the best estimate of the state of the atmosphere at spatial and temporal scales consistent with those of the forecasts issued by this same model.

* Correspondence to: P. Pinson, DTU Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark. E-mail: pp@imm.dtu.dk

These forecasts are then evaluated on the numerical grid of the model. While such an approach is relevant, it may not reflect the final use of the forecasts which may be needed at any location on Earth (not just for the model grid points). In that context, it may actually be more interesting and relevant to verify the forecasts jointly against analysis and against actual observations. This is recognized by research and operational weather forecasting centres such as ECMWF, which aim to give more importance to observations in their verification suite.

A few studies focusing on the evaluation of ensemble forecasting systems against observations have recently appeared in the literature, see, for example, Candille *et al.* (2007) and Candille and Talagrand (2008). A primary objective of the present work is to look at this problem, with the aim of evaluating the quality of the ensemble forecasts of wind speed issued over Europe by the European Centre for Medium-range Weather Forecasts (ECMWF) against analysis and actual observations while accounting for observational uncertainty. The choice for this domain and for the wind speed variable arises from the growing interest in wind energy and its short-term forecasting (Giebel *et al.*, 2003; Lange and Focken, 2005; Costa *et al.*, 2008; Smith *et al.*, 2009, among others). A subsequent objective is to illustrate the disparities that appear if performing forecast verification against analysis or against observations. A final objective is then to discuss if such additional verification results may allow ways of further improving the quality of ECMWF ensemble forecasts of wind speed to be foreseen.

The data, including forecasts, analysis, and observations are first introduced in Section 2. The forecast verification methodology accounting for observational uncertainty, as well as the time-varying climatology employed as a benchmark, is then described in Section 3. The results from the application of the forecast verification methodology against observations are subsequently gathered and commented on in Section 4. The applications considered include (1) the routine evaluation of the ensemble forecasts of wind speed over a 3 month period (from December 2008 to February 2009 – DJF09), and, (2) the assessment of the impact of the change in horizontal resolution of the ECMWF ensemble prediction system. Section 5 finally develops into a discussion of the implications of such findings, drawing conclusions and perspectives for future work.

2. Data

2.1. Setup for the verification experiment, observations and analysis

The domain chosen for this study is Europe, while the forecast variable of focus is near-surface (10 m) wind speed. One of the reasons for this choice is that forecast users have shown more and more interest in that variable over the last few years, owing to the significant wind power capacities operated throughout Europe.

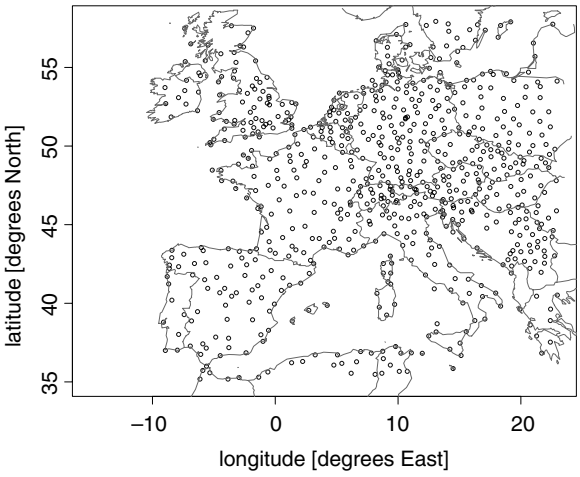

Figure 1. Map of all synoptic stations considered in this study. The domain is defined as Europe in a large sense, with longitudes in the range [−10,23] degrees East, and latitudes in the range [35,58] degrees North. ◦ synoptic stations (731).

Verification is to be performed over a set of synoptic stations located onshore throughout Europe, for which observational data is available through the Global Telecommunication System (GTS). The geographical distribution of these 731 stations can be seen from Figure 1. After inspection of the data at the various stations, 98 of these stations were discarded as having too many missing data or too long periods of suspicious behaviour in the recorded time series. No statistical methods for outlier detection have been employed. Empirical rules have been used instead, considering for instance that (1) very large spikes during a low wind speed period, or, (2) periods of more than 2 days with the same recorded wind speed values (being non-zero), were to be seen as suspicious. The interest of considering wind speed near-surface observations from synoptic stations on land in this study is that such measurements are not used in the production of the model analysis (Uppala *et al.*, 2005). An example historical reason for that relates to the heterogeneity in the representativeness of these observations in view of the very coarse spatial and temporal resolution of the model. One could then expect to see more disparities between verification results obtained if verification is performed against the analysis or against the actual observations. Local thermal and topographic effects may additionally step in and magnify the aforementioned disparities.

A station-oriented view of the verification problem is proposed: instead of considering averaging all verification scores for stations within a grid cell, all forecasts and analysis are interpolated at the stations, and the scores calculated for each of the stations individually. This idea of averaging *per* grid cell (or for larger areas) has been employed and explored for the case of precipitation: see Ghelli and Lalaurette (2000) or Pappenberger *et al.* (2009) for example. This approach would also introduce some form of filtering of the observations, and is not desirable in the present case.

Some may say that the representativeness issue, i.e. the fact that using raw observations is not consistent with the temporal and spatial scales the model aims at resolving, is not accounted for. The users of the forecasts, however, are not interested in the spatial and temporal scales of the model. They only want the best forecasts for the given locations of their choice.

The ECMWF analysis data have a temporal resolution of 6 h, while wind speed observations at synoptic stations over Europe most often have an hourly temporal resolution. This difference will be accounted for in the verification exercise, in order to be consistent with the forecasts which are described in the following. When verification scores calculated against observations or against analysis are compared, this will be done for time points for which both observations and analysis (and obviously forecasts) are available.

2.2. Wind speed forecasts

The wind speed forecasts used as input to this verification study are some of the operational products at ECMWF. Attention is given to ensemble forecasts of 10 m wind speed, with the possibility of extracting some single-valued forecasts from the ensembles, following a methodology that will be described in a further paragraph. The forecast length considered is of 6 days, corresponding to the lead times of interest to the wind energy sector. Note that the 6 day lead time also corresponds to a change in the temporal resolution of the ensemble forecasts, with forecast output being coarser for further lead times i.e. with a temporal resolution of 6 h.

The operational configuration of the ensemble forecasting system for lead times up to 6 days ahead and for the European domain is briefly summarized. Ensemble forecasts are issued twice a day at 0000 and 1200 UTC, with a horizontal resolution of about 50 km (corresponding to a spectral truncation at wave number 399) and a temporal resolution of 3 h. Operational ensemble forecasts with such a horizontal resolution were issued until the 25 January 2010. From the 26 onwards, this horizontal resolution has been changed to about 33 km, corresponding to a spectral truncation at wave number 639. Over a period spanning November 2009 to January 2010, 187 forecast series are available from the operational forecasting system with the two horizontal resolutions. These will permit application of the proposal verification framework for the assessment of the impact of the change in horizontal resolution on the quality of ensemble forecasts of near-surface wind speed.

The methodology employed for generation of the ECMWF ensemble forecasts is well documented and a number of publications can be pointed at for its various components. For a general overview, see Palmer (2000). It is not the objective here to discuss competing methodologies for the generation of ensemble forecasts or more generally of probabilistic forecasts of meteorological variables. A comparison with other global ensemble prediction systems can be found in (for example) Buizza *et al*. (2005). The ECMWF ensemble predictions aim at representing uncertainties in both the knowledge of the initial state of the atmosphere and in the physics of the numerical model used for integrating these initial conditions. For the former uncertainties, singular vectors are employed, the core methodology being extensively described by Leutbecher and Palmer (2008). A comparison of the different methodologies for the generation of initial perturbations can be found in Magnusson *et al*. (2008). In parallel for the latter type of uncertainties, stochastic physics is employed for sampling uncertainties in the parameterization of the numerical model (Buizza *et al*., 1999; Palmer *et al*., 2005). Note that the potential structural model uncertainty is, therefore, not accounted for.

The ensemble forecasts for the 633 stations of interest are obtained by applying bilinear interpolation to the gridded model output, i.e. as a weighted combination of model outputs at the four grid points around the station. The same type of bilinear interpolation is used for downscaling the analysis data at the level of the stations. By using such a bilinear interpolation scheme the land-sea mask is not considered, and grid nodes over land and sea are equally weighted.

3. Verification methodology

3.1. Time-varying climatologies as a benchmark

Verifying forecasts against a benchmark is a common practice. A benchmark has the characteristics of being a reference method, of being computationally cheap to implement, and ideally model-free. The typical benchmark in the verification of probabilistic and ensemble forecasting in meteorology is climatology. Roughly, climatology is based on all available observations over a long period of past observations, the distribution of which serves as a predictive density for any lead time $t + k$. This benchmark is difficult to outperform for longer-term lead times, typically further than 5–6 days for near-surface variables, though quite easy to outperform for short-term forecasts, say for lead times less than a day. At these shorter scales, persistent forecasts issued based on the last available measurements become the most competitive benchmark. Note that only climatology will be considered here since focus will mainly be on the medium-range (1–6 days).

Even though climatology is recognized as the central benchmark in the verification of meteorological forecasts, some concerns are also raised regarding the possibility of misinterpreting forecast verification results (Hamill and Juras, 2006). It may indeed be possible that the observed skill of a forecast system when evaluated against climatology is artificially good simply due to a drift between the reference climatology and the state of the stochastic process of interest. The discussion by Hamill and Juras (2006) implies that climatologies may (or should) be seen as time-varying, with the best estimate of climatologies permitting to minimize potential misinterpretation

of forecast verification results. Following that remark, Jung and Leutbecher (2008) have proposed an approach to the computation of time-varying climatologies, which is revisited here. Note that the approach of Jung and Leutbecher (2008) has led to the computation of the climatologies routinely used at ECMWF for the verification of ensemble forecasts against analysis. Following a similar argument, only skill scores representing improvements over the climatology benchmark will be compared, for climatologies calculated based on observations. This is because considering climatologies based on the model analysis means that forecasts would then be evaluated against benchmarks with different dynamic characteristics, hence potentially leading to misinterpretation.

Denoting by $\{x(t,s)\}_t$ the time-series of wind speed measurements being a sequence of observations for the related stochastic process $\{X(t,s)\}_t$ at the location s. Measurements are available over a period ranging from $t=0$ to $t=N$ for the number of locations considered in this study. Since talking about climatologies, N is supposed to be very large due to availability of several years if not decades of data. The core idea of time-varying climatologies is that climatologies should be defined for each hour of the year, or at least for each time of the year for which measurements are available, through smoothing the high-frequency temporal features in the recorded time series. This is in order to retain the diurnal and seasonal variations in wind speed. Since considering observations instead of analysis data in the case of Jung and Leutbecher (2008), more variability and high-frequency features are to be expected.

For convenience, let us introduce the operator v which gives the calendar date (defined in terms of the year y, month m, day d and hour h) for the absolute time t, while v^{-1} performs the opposite operation:

$$\{y,m,d,h\} = v(t), \qquad t = v^{-1}\{y,m,d,h\} \tag{1}$$

The methodology for deriving climatologies is based on kernel density estimation (KDE), an overview of which can be found in Silverman (1986). The basic idea is to attach a kernel to each of the available measurements, and to consider the time-varying climatologies as a weighted mixture of these kernels.

For simplicity, Gaussian kernels are employed here, which for a measurement $x(t,s)$ is defined as:

$$K_\sigma\,(x - x(t,s)) = \frac{1}{\sigma_x\sqrt{2\pi}} \exp\left\{\frac{(x-x(t,s))^2}{2{\sigma_x}^2}\right\} \tag{2}$$

with σ_x the standard deviation of the Gaussian density defining the bandwidth of the kernel. Such kernels are censored at 0, however, in order to be consistent with the fact that wind speed must be greater than or equal to 0. Both x and σ_x are in m s^{-1}, while K_σ is non-dimensional since defining a probability density related to a given wind speed observation. This yields:

$$K_\sigma^+(x - x(t,s)) = \begin{cases} K_\sigma(x - x(t,s)), & x > 0 \\ \Phi\left(-\dfrac{x(t,s)}{\sigma_x}\right), & x = 0 \end{cases} \tag{3}$$

where Φ is the cumulative distribution function of a standard Gaussian random variable $\mathcal{N}(0,1)$. Censored kernels put a probability mass on 0 for low and null wind speed values, being a function of the observation itself and on the chosen kernel standard deviation.

For any time of the year, the climatological distribution $\overline{F}_x$ of wind speed is then defined as a weighted mixture of kernels for the same hour of the current and neighbouring days of all years in the dataset, and for the same location. In mathematical terms this is written as:

$$\overline{F}_x(\{m,h,d\},s) = \frac{1}{N_y \sum_j w_j} \sum_y \sum_j w_j K_\sigma^+ \left(x - x(v^{-1}\{y,m,d+j,h\},s)\right) \tag{4}$$

with N_y the number of years used for producing the climatology, and with w_j a discounting factor permitting to give less weight to days that are further from the day of interest. This discounting factor is also chosen to be given by a Gaussian kernel:

$$w_j = K_{\sigma_d}(j) \tag{5}$$

Since Gaussian kernels do not have a compact support, the sum over js in Equation (4) involves an infinite number of elements. In practice since the weight defined by K_{σ_d} becomes very low for $|j|$ large, say for $|j| > 5\sigma_d$, one can limit the sum over a window of size $10\sigma_d$ around the point in time of interest. The other sum is over all N_f years in the dataset.

In practice here, the data employed as input to the calculation of time-varying climatologies consists of $N_y = 29$ years of wind speed measurements recorded with a temporal resolution of 3 h, for the 633 (validated) meteorological stations. The temporal resolution of 3 h is chosen in order to be consistent with the temporal resolution of the ensemble forecasts. These 29 years range from 1981 to 2009. For some of the stations the length of the dataset may be shorter since recording started after 1981. Also, after basic cleaning of the datasets, that is, based on simple rules and not on advanced statistical approaches, some data may be missing or considered as invalid (i.e. negative wind speeds or wind speeds greater than 60 m s^{-1}). The weights in Equation (4) can easily permit to account for these aspects, by setting w_j to 0 if measurements are missing or considered as invalid. The two bandwidths σ_x and σ_d are selected in order to be consistent with the climatologies based on the analysis derived at ECMWF. This yields $\sigma_x = 1$ m s^{-1} and $\sigma_d = 20$ days so that seasonal cycles are revealed while higher frequency fluctuations are smoothed. These bandwidth values could be further refined based on the various rules available in the statistical literature, or alternatively on a cross-validation exercise. Finally, since these climatologies have a nonparametric form, it is necessary to define them in terms of quantiles with various nominal proportions. These nominal proportions are chosen to span the whole unit range with 0.05 increments and

with a finer description of the tails end, i.e. yielding a set of nominal proportions in {0.01, 0.02, 0.05, 0.1, ..., 0.9, 0.95, 0.98, 0.99}. The mean and standard deviation values of all climatological distributions are also recorded. Climatologies are employed in their probabilistic form since some of their characteristics (mean and median) as well as full densities will be necessary for calculating the various scores for this benchmark, subsequently yielding skill scores for the ensemble forecasts.

As an illustration, Figure 2 depicts an example of a time-varying climatology for the meteorological station at Kastrup airport (Copenhagen) in Denmark for the months of April, May and June. This climatology has a strong diurnal pattern in the mean wind speed, while it also exhibits longer-term variations in the form of a seasonal trend. These dynamics at various temporal scales can also be observed for the various quantiles of the climatology, with for instance a reduction of the maximum wind speeds from April to June. The low frequency of occurrence of calm periods, even at night, is very site-specific. Such a frequency of calm periods is substantially higher for stations located in less windy areas like in central Germany for instance. It is finally worth noting that time-varying climatologies may be refined in the future by accounting for both rounding and measurement uncertainties in recorded wind speed values.

3.2. Scores and diagnostic tools

A fairly common approach to the verification of ensemble forecasts is employed here. Following arguments in a number of publications, focus is given to both reliability and sharpness of the ensemble forecasts of wind speed. In parallel, since for a large number of applications forecast users may still prefer to use point forecasts instead of ensemble or more generally probabilistic forecasts, an evaluation of a few point forecasts that may extracted from the ensembles is also performed. Especially, in view of the discussion by Gneiting (2011b), the mean and median of ensemble forecasts are specific point forecasts which aim at minimizing a Root Mean Square Error (RMSE) and a Mean Average Error (MAE) criterion, respectively. This is since the expectation of a probabilistic distribution is to minimize a quadratic loss function (as for the RMSE), while the median of that same distribution is to minimize any symmetric linear loss function (as for the MAE). Finally, the bias is generally assessed when considering the ensemble mean as the point forecast to be extracted from the ensembles.

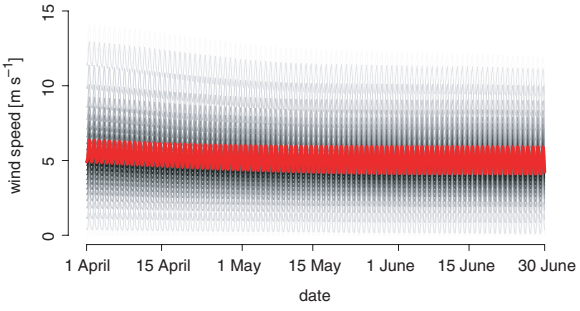

Figure 2. Example of a time-varying climatology of 10-m wind speed for the meteorological station of Copenhagen Kastrup airport in Denmark for the months of April, May and June. — Mean, — Quantiles (nominal proportions in {0.01, 0.02, 0.05, 0.1, 0.2, ..., 0.9, 0.95, 0.98, 0.99}). This figure is available in colour online at wileyonlinelibrary.com/journal/met

For a specific location s, $\{\hat{x}^j(t+k|t,s)\}_j$ denotes the set of 51 ensemble members (i.e. the control forecast and the 50 perturbed ones) issued at time t for the lead time $t+k$. $\hat{x}^j(.)$ is the jth ensemble member. The notations $\tilde{x}(t+k|t,s)$ and $\overline{x}(t+k|t,s)$ are used for the median and mean of the ensembles, respectively. The scores mentioned above are then simply given for each lead time k as:

$$\text{bias}(k,s) = \frac{1}{N_f}\sum_{t=1}^{N_f} x(t+k,s) - \overline{x}(t+k|t,s) \qquad (6)$$

and,

$$\text{MAE}(k,s) = \frac{1}{N_f}\sum_{t=1}^{N_f} |x(t+k,s) - \tilde{x}(t+k|t,s)| \qquad (7)$$

and,

$$\text{RMSE}(k,s) = \left(\frac{1}{N_f}\sum_{t=1}^{N_f}(x(t+k,s) - \overline{x}(t+k|t,s))^2\right)^{\frac{1}{2}} \qquad (8)$$

where N_f is the number of forecasts over the verification period.

Turning attention towards the probabilistic skill of the ensemble forecasts, it is appropriate to evaluate it with proper skill scores such as the Continuous Ranked Probability Score (CRPS) for instance. The expression for the calculation of the CRPS for the lead time k is:

$$\text{CRPS}(k,s) = \frac{1}{N_f}\sum_{t=1}^{N_f}\int_x \left(\hat{F}(x;t+k|t,s) - \mathbf{1}\{x > x(t+k,s)\}\right)^2 \mathrm{d}x \qquad (9)$$

where $\hat{F}(x;t+k|t,s)$ is the cumulative distribution function of the set of ensemble forecasts $\{\hat{x}^j(t+k|t,s)\}_j$, while the Heaviside function $1\{x > x(t+k,s)\}$ represents a perfectly sharp and calibrated probabilistic forecast which would have predicted a probability mass on the actual observation $x(t+k,s)$. In the present case, $\hat{F}(x;t+k|t,s)$ is given by linear interpolation through the ensemble members: for a set of 51 exchangeable members, the jth member defines the quantile of $\hat{F}(x;t+k|t,s)$ with nominal proportion $\alpha_j = \frac{j}{52}$.

Corresponding skill scores are obtained by comparing for each lead time the error criteria calculated for the ensemble forecasts and for the climatology benchmark. Single-valued forecasts are extracted from climatologies in a similar fashion as for ensemble forecasts. In other words, the bias and RMSE criteria are calculated for the ensemble mean, while the MAE criterion relies on the median of climatology predictive densities. Skill scores are then defined as

$$\text{Score}(k, s) = 1 - \frac{\text{Score}(k, s)}{\text{Score}_0(k, s)} \tag{10}$$

where 'Score' can be the 'bias', 'MAE', 'RMSE' and 'CRPS' error criteria given the above, while Score_0 is the value of such a criterion if calculated for the time-varying climatology benchmark described in Section 3.1. The resulting skill scores would, therefore, be denoted by 'Sbias', 'SMAE' or 'SRMSE' for instance. One may also obtain spatially averaged scores and skill scores by calculating the average over s of the scores and skill scores introduced above.

Particular focus should be given to ensemble forecast reliability. Reliability refers to the correspondence of empirical and nominal proportions of ensemble forecasts. In contrast recalibration relates to the post processing of ensemble forecasts in order to improve their reliability. Probabilistic reliability is visually assessed here based on PIT diagrams, being a cumulative version of Probability Integral Transform (PIT) histograms, as used and discussed by Pinson *et al*. (2010) and Marzban *et al*. (2011) for instance. Such PIT diagrams allows for straightforward visual comparison of the empirical proportions of the ensemble members against the nominal ones. Indeed, for a set of 51 exchangeable members, the nominal proportion of the jth member is $\alpha_j = \frac{j}{52}$, meaning that there should be a probability of $\frac{j}{52}$ that the observed wind speed lies below that ensemble member. PIT diagrams are therefore based on the indicator variable $\xi^j(t, k, s)$, defined as:

$$\xi^j(t, k, s) = \mathbf{1}\{x(t + k, s) < \hat{x}^j(t + k|t, s)\} \tag{11}$$

and its sample mean (over time, locations, potentially lead times). Indeed for the jth ensemble member with nominal proportion $\alpha_j = \frac{j}{52}$, the empirical (or observed) proportion $\hat{\alpha}_j(k, s)$ is estimated as:

$$\hat{\alpha}_j(k,\ s) = \text{E}[\xi^j(t,\ k,\ s)|k,\ s] = \frac{1}{N_f} \sum_t \xi^j(t,\ k,\ s) \tag{12}$$

PIT diagrams consequently depict α_j *versus* $\hat{\alpha}_j$ for all (51) ensemble members. Note that the potential effect of sampling and of the interdependence (spatial and/or temporal) in the forecast-verification pairs is disregarded here. It could be accounted for in the future by using or extending the methods described by Bröcker and Smith (2007), Marzban *et al*. (2011) and Pinson *et al*. (2010).

3.3. Accounting for observational uncertainty

One of the reasons why observations are often not favoured in verification studies is their underlying uncertainty, along with their representativeness. This is especially true for near-surface variables, e.g. wind speed and precipitation, for which observational uncertainty is known to be non-negligible, while surface effects introduce additional noise to what the numerical models aim at resolving. That representativeness issue is not accounted for here since having a station-oriented view of the forecast verification problem. Somehow a forecast user will not assess competing forecasting approaches conditional to the model capabilities, but uniquely based on verification scores and statistics for the location(s) of interest.

Observational uncertainty can be accounted for during the forecast verification process. One may distinguish between the various sorts of observational uncertainties as in Pappenberger *et al*. (2009) and potentially consider the interdependence structure (either in time or in space, or both) in the forecast errors (Candille *et al*., 2007). Various approaches may be employed for the case of the verification of ensemble forecasts, including the perturbed ensemble and observational probability proposals of Candille and Talagrand (2008). The approach followed here is of the observational probability type: the uncertainty in the observations is represented by transforming them into random variables. Their impact on scores and diagnostics is then quantified using a Monte Carlo approach similar to that of Pappenberger *et al*. (2009).

Two origins to the uncertainty in wind speed observations are considered: rounding and measurement errors. It is assumed that gross errors originating from reporting, transmission or archiving can be easily cleaned out, or that observations in that case would be seen as missing. Measurement errors come from the measuring devices themselves. They can be assumed to be Gaussian, spatially and temporally uncorrelated, with a mean μ corresponding to a systematic error and a variance σ_e^2 for the actual measurement uncertainty. μ and σ_e^2 could be defined for each station independently, but for simplicity they will be uniquely defined here. This writes:

$$e_m(t, s) = \mathcal{N}(\mu, \sigma_e^2) \tag{13}$$

In parallel, rounding errors come from the procedure of rounding measured wind speed to the closest integer (in m s^{-1}), the common practice when reporting near-surface wind speed measurements. Rounding errors can then be assumed to follow a uniform distribution around the reported value:

$$e_r(t, s) = U\left[-\frac{1}{2}, \frac{1}{2}\right] \tag{14}$$

To summarize, if writing $X(t, s)$ the random variable for the wind speed at time t and location s, and $x(t, s)$ the

reported value, $X(t,s)$ is given by the sum of $x(t,s)$ with the above two random variables:

$$X(t,s) = (x(t,s) + e_m(t,s) + e_r(t,s))\mathbf{1}\{x(t,s) + e_m(t,s) + e_r(t,s) \geq 0\} \quad (15)$$

with $\mathbf{1}\{x \geq 0\}$ indicating a censoring of the random variable at 0 since wind speed is a non-negative quantity. Given a reported wind speed value, and the measurement error characteristics μ and σ_e^2, the density of $X(t,s)$ can be obtained from a simple convolution operation. For simplicity, μ is assumed to be 0 in the following, translating to having unbiased measurements.

Subsequently, in the spirit similar to the Generalized Likelihood Uncertainty Estimation approach employed by Pappenberger *et al.* (2009), a form of Monte Carlo simulation can be used for assessing the impact of observational uncertainty on scores and diagnostics. Based on the modelled densities of observations at each point in time and in space, one can draw a number M of potential actual wind speed values $x^{(i)}(t,s)$, $i = 1, \ldots, M$, and calculate for each i the various scores and diagnostics defined in the above paragraph. This is done by plugging the drawn values $x^{(i)}(t,s)$ in the various formula of Equations (7)–(11). It will then result in empirical distributions of scores (MAE, etc.), corresponding skill scores (SMAE, etc.), but also of PIT diagrams. Indeed, in contrast to the case of Candille and Talagrand (2008), it is possible by this approach to build a set of PIT histograms or of their cumulative version in the form of PIT diagrams. This is since the set of 'actual' observations drawn from the modelled densities are then of the same nature than the predicted ensemble elements.

It should finally be noted that such a Monte Carlo approach can be highly computationally expensive. Deriving analytical expressions for the distributions of some of the simplest scores may be possible. For the case of the bias, one could use known formulae for the distribution of the sum of Gaussian variables and for the sum of Uniform variables, possibly non-identical (Mitra, 1971; Bradley and Gupta, 2002). They could be extended to the case of the MAE, based on limiting assumptions. For scores such as the RMSE and CRPS the mathematical developments would become quite technical and show the difficulty of deriving closed-form solutions. All these aspects related to the impact of observational uncertainty on the distribution of scores are discussed in Appendix A. A similar remark goes for the case of PIT histograms and diagrams. For these reasons, the computationally-intensive method described above is preferred. The fact that computational costs may lead to some limitations has also been mentioned by Candille and Talagrand (2008).

4. Application results

Two test case applications are considered, corresponding to what may be done in research and operational forecasting centres such as ECMWF. On the one hand, forecast verification is performed on a routine basis, with various scores and diagnostics reported every quarter of a year for instance. On the other hand specific verification exercises are carried out prior to an operational upgrade of the forecasting system, in order to assess the extent of expected improvements. The verification framework discussed above is applied in both cases, but with different objectives. In the first case, besides the actual routine verification it is aimed at commenting on the discrepancies between verification performed against analysis and against observations for near-surface wind speed. The impact of observational uncertainty on the routine scores that would be calculated and reported in such routine verification exercises is also illustrated and discussed. In the second case, the objective is mainly to assess the improvements brought in by the upgrade of an operational forecasting system for near-surface wind speed, at the various European stations.

4.1. Routine evaluation of ensemble forecasts

The first application case consists of the routine evaluation of the ECMWF ensemble forecasts of wind speed over the quarter December 2008, January and February 2009 (DJF09) with focus on Europe. An extensive set of maps and summary graphs have been produced for the various scores and diagnostics, depending upon lead times and possibly location. The verification suite allows for the definition of a set of stations of interest, hence permitting to look at forecast verification for a given station, on a country-by-country basis, or for a pre-defined region. Owing to the quantity of results that may be generated, only a subset of the most interesting results will be shown and commented here. The effect of observational uncertainty is disregarded in the first stage. It is then dealt with in Section 4.1.3.

4.1.1. Scores at stations

As a first illustrative example, the map of CRPS values at the various European synoptic stations for 10 m wind

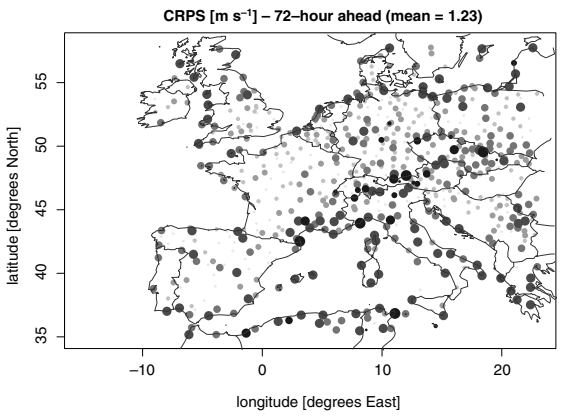

Figure 3. Map of the CRPS values calculated against observations at all synoptic stations in the case-study (633) for 72 h ahead forecasts. The mean CRPS is of 1.23 m s^{-1}. • 0.17<y<0.86; • 0.86<y<1.02; • 1.02<y<1.20; • 1.20<y<1.56; • 1.56<y<2.59; • y>2.59.

speed ensemble forecasts and for the lead time of 72 h ahead is shown in Figure 3. These CRPS values are calculated based on reported wind speed observations at the stations, hence without considering observational uncertainty. Let us explain how the results are displayed there. In view of the distribution of scores (CRPS and others) being quite skewed, it has been decided to divide such distributions in a number of equally populated classes, except for the 'extreme' score values. The 5% maximum score values represent the last one of these classes, somehow covering outlier stations. The five other classes represent equally populated classes of CPRS values for the 95% remainder of the stations, hence containing each 19% of the scores data.

Most of the highest score values are for stations located in the Alps region and in coastal areas. This could be expected since near-surface local effects are difficult to resolve at the fairly coarse resolution (50 km) of the ECMWF ensemble prediction system at the time. On a general basis, though these CRPS values are low, being below 2.59 m s^{-1} for 95% of the stations. They are even extremely low (below 1.2 m s^{-1}) for more than half of the stations. As a reference, the mean wind speed over all of these stations at this period was of 3.76 m s^{-1}, while the mean wind speed was below 6.99 m s^{-1} for 95% of these stations. It happens that for some of the stations even though the data collected were deemed acceptable since their dynamical behaviour appeared realistic, a comparison with the forecast dynamics showed that almost no correlation existed between the forecasts and measurements. Consequently, the various scores calculated at these sites appeared to be independent of the lead time. Such situations may originate from a low quality of observational data which hence could be discarded if refining the analysis. It could also be explained by a questionable quality of the ensemble forecasts, for instance due to local effects not represented in a model with such a coarse spatial resolution.

In parallel, Figure 4 depicts the disparities between the CRPS values (for the same lead time) calculated against analysis and observations at the various stations. It is in practice calculated as the difference between the CRPS values calculated against observations and against analysis. Positive values are for scores values being larger if calculated against observations than if calculated against analysis. The sorting into different classes is performed in a manner similar to the above. It appears that scores calculated against observations can actually be lower than scores calculated against analysis. It happens here for 10% of the stations. One of the potential reasons stems from the impact of observational uncertainty on the scores calculated against actual observations at the various stations. This impact will be further examined below. However, the inspection of a large number of plots with the forecasts along with corresponding analysis and observations actually revealed that for most of these stations, the forecasts really looked like they better matched the observations than the analysis.

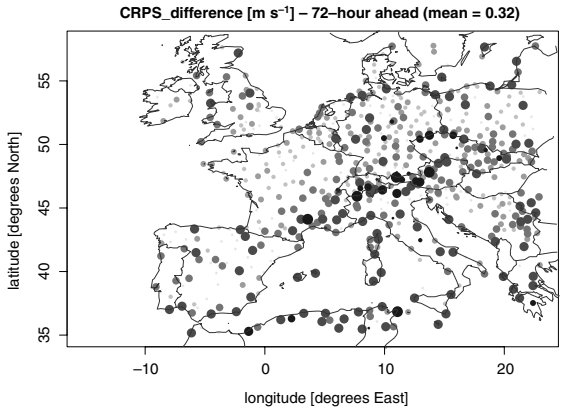

Figure 4. Map of the difference in CRPS values (when calculated against analysis or observations) at all synoptic stations in the case-study (633), for 72 h ahead forecasts. The overall mean is of 0.32 m s^{-1}. · $-1.5<y<0.17$; · $0.17<y<0.32$; · $0.32<y<0.48$; · $0.48<y<0.75$; · $0.75<y<1.68$; · $y>1.68$.

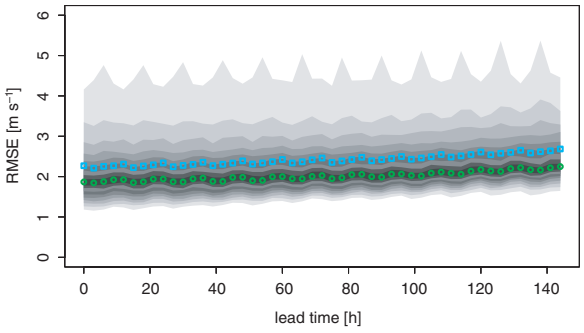

Figure 5. Distribution of RMSE score values (of the ensemble mean) for the 633 stations, as a function of lead time. □ Mean, ○ Median, ■ Central score intervals with coverage in {10,20,. . .,90}%. This figure is available in colour online at wileyonlinelibrary.com/journal/met

Generally, the results for the remaining 90% of the stations are consistent with intuitive expectations, i.e. revealing that scores calculated against observations tend to be higher than if calculating against the analysis. For 85% of the stations the discrepancies are up to 1.68 m s^{-1}, which is quite high in view of the CRPS values shown in Figure 3. Similar results have been observed when considering other forecast verification measures such as bias, MAE, RMSE, and the corresponding skill scores. A final aspect that can be looked at is the distribution of score values for all stations. As an example, Figure 5 depicts the distribution of RMSE values as a function of the lead time. These distributions are represented by a set of intervals centred on the median, and with increasing proportions (from 10 to 90%), in addition to the median and mean score values. Owing to the positive skewness of these distributions, the mean values are larger than the median ones. Mean RMSE values increase from 2.1 m s^{-1} for the first lead time to 2.4 m s^{-1} for 6 day ahead forecasts. This is because for 90% of the stations considered RMSE values may range between 1.1 and 5.2 m s^{-1} depending on the station and

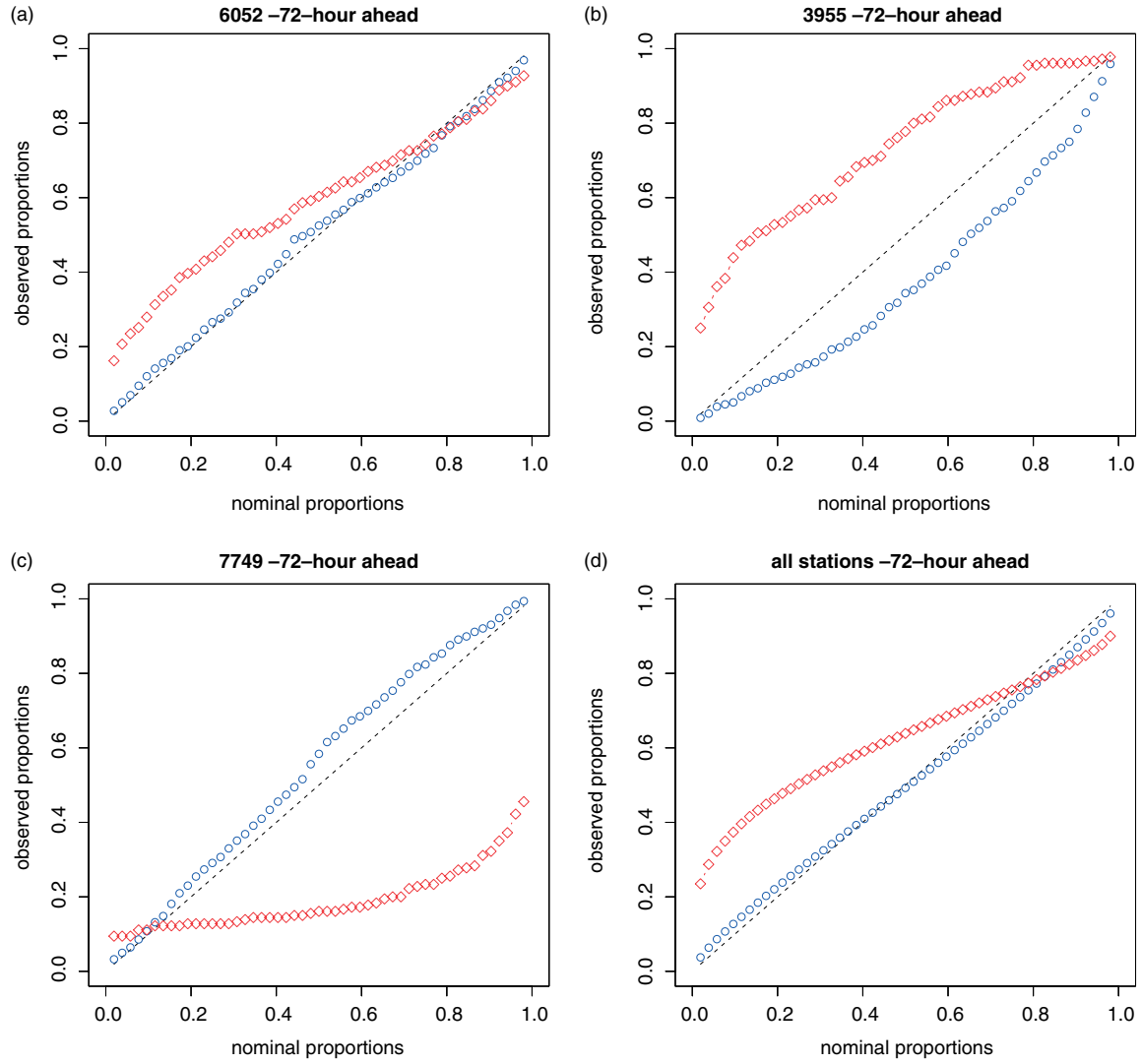

Figure 6. PIT diagrams for the reliability assessment of 72 h ahead ensemble forecasts. These diagrams are for Thyboron station in (a) Denmark, (b) Cork airport, (c) Cap Béar station in France and (d) finally all stations altogether. ○ *vs* analysis, ◇ *vs* observations. This figure is available in colour online at wileyonlinelibrary.com/journal/met

lead time. This type of representation of score distributions can be very informative for having an overview of the performance of a forecasting system over a large set of stations of interest. The periodic nature of the RMSE curves is linked to the diurnal cycles in the wind speed magnitude, the amplitude of such periodicities varying throughout Europe. To identify better the effect of the diurnal cycle on verification statistics, one may refine the analysis performed here by verifying forecasts depending on the time of the day (instead of the lead time), or by making a difference between forecasts issued at 0000 and 1200 UTC.

4.1.2. Reliability of ensemble forecasts

A crucial aspect of ensemble forecast verification related to their probabilistic reliability, for which significant disparities are expected if evaluated against analysis or against observations. For that reliability assessment, the PIT diagrams in the form of cumulative PIT histograms are employed (see Section 3.2 or Pinson *et al*., (2010) for further details). The impact of observational uncertainty on these PIT diagrams will not be discussed, since it has been found to be very limited. This might be explained by the fact that perturbed observations randomly fall between different ensemble members, but without altering much the counts over the evaluation period. Sampling or serial correlation effects on reliability statistics, as discussed by Bröcker and Smith (2007) and Pinson *et al*. (2010), could also be considered in the future. Their effect on the uncertainty of reliability statistics is expected to be larger than that of observational uncertainty.

Example PIT diagrams are gathered in Figure 6 for the stations of Thyboron in Denmark, Cork airport in Ireland, Cap Béar in France, as well as for all 633 stations altogether. These reliability assessment results are for 72 h ahead forecasts. In a fashion similar to other scores, the verification suite allows for the assessment of reliability at single stations or at pre-defined groups of stations, for a given lead time or for groups of lead time, thus permitting to focus certain geographical areas

and certain forecast ranges. When assessing reliability for various lead times and against the analysis, a fairly well known result about the ECMWF ensemble forecasts is observed, which is that they tend to be significantly under-dispersive in the short range and more reliable for the medium range.

The three PIT diagrams for single stations in Figure 6 are representative of the typical results observed over the routine verification study. The average case is similar to what is observed at Thyboron station in Denmark: a very good reliability if evaluated against analysis, while this reliability can be seen as significantly lower if assessed against observations. The ensemble forecasts appear to be slightly under-dispersive but well centred in probability when seeing the analysis as the reference. This is while ensemble forecasts appear to overestimate proportions, especially in the lower part of the ensembles, when employing observations as the reference. If differentiating lead times, these reliability issues appear to be more pronounced for the first 2 days, then improving for further lead times, consistent with what is observed when verifying ensembles against the analysis.

In parallel for (near-) coastal stations such as Cork airport, or stations located in areas with specific local wind regimes such as Cap Béar, reliability statistics obtained against the analysis already inform about a lack of sufficient reliability, while the picture clearly worsens if reliability is evaluated against observations. Similar comments can be made for the case of the Alps region. Depending on cases, clear under- or overestimation of probabilities was observed when assessing reliability against observations. For a station such as Cork, this may be since the model forecasts stronger winds as if Cork was at sea. In contrast, for a site such as Cap Béar the very specific acceleration of local wind regimes such as Tramontane and Vent d'Autan may be overlooked by the model, explaining a systematic underestimation of winds. Note that appropriate recalibration against observations would correct for this lack of reliability and hence improve overall skill scores.

To summarize the disparities in the reliability assessment *versus* analysis and observations, a quantity is defined based on the integrated absolute difference between the two reliability curves. This quantity naturally takes values in [0,1] and is referred to as reliability disparity (RD). It is low in the case of Thyboron in Figure 6, while being very high in cases such as Cork and Cap Béar (in this same figure). A map summarizing the reliability disparity values at all stations is shown in Figure 7, for PIT diagrams based on forecasts 72 h ahead. Qualitatively, similar patterns were observed for all other lead times. The sorting of the RD values in different classes is similar to the cases of Figures 3 and 4. For the 5% most extreme values, most of the corresponding stations are located either in complex coastal areas (Cornwall tip or Galicia), on small islands which are impossible for the model to resolve (e.g. Baleares), or in the Alps region. Cap Béar is one of these extreme cases. In parallel for around 40% of the stations the

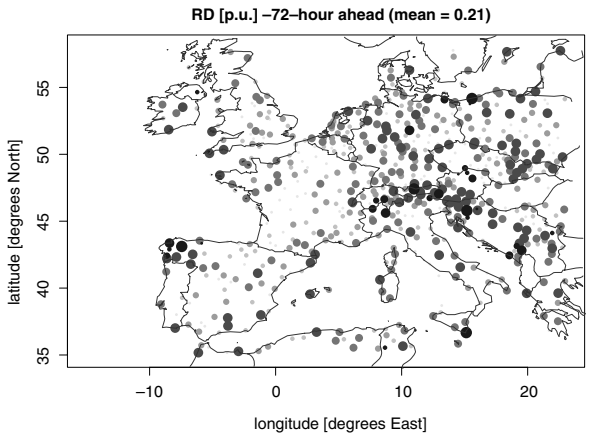

Figure 7. Map of the reliability discrepancy at all synoptic stations in the case-study (633) for 72 h ahead ensemble forecasts. · 0.03<y<0.1; · 0.1<y<0.15; • 0.15<y<0.23; • 0.23<y<0.31; • 0.31<y<0.41; ● y>0.41.

reliability disparity is fairly low (that is, below 0.15) corresponding to cases like Thyboron in Figure 6. These stations with lower disparity are spread over Europe, though a higher concentration can be observed in certain parts of France, Spain, Northern Italy, Czech Republic and Switzerland. Larger disparities tend to concentrate in northwestern parts of France (Brittany and Normandy) and Spain (Galicia), The Netherlands and northwestern regions of Germany, Austria and the Balkan region.

These results are the most surprising and interesting ones obtained from this routine forecast verification procedure. Such disparities in the reliability assessment of ensemble forecasts if considering analysis or observations as the reference were not expected beforehand. This can certainly be explained by the fact that onshore wind observations are not accounted for in the production of the analysis, and also by the significant difference in the variability of analysis and observations of wind speed. It leads to a suggestion that recalibration of near-surface wind forecasts against observations would certainly permit a significant improvement their reliability and overall skill. This should be performed in a sufficiently generic and efficient framework so that this recalibration is performed at once for the whole region, with properly identified model structures and with model parameters estimated and optimized on a site-specific basis. These models may sometimes be based on simple linear regression concepts. In other cases, they may require more advanced structures, possibly nonlinear, also accounting for potential seasonalities and serial dynamics in forecast-verification pairs.

4.1.3. Forecast quality over areas, and impact of observational uncertainty

Looking at summary verification statistics over certain areas may be particularly appealing to forecast users. In addition, while the previous results disregarded the potential impact of observational uncertainty on scores and diagnostics, it is accounted for and discussed in the

following. Due to the computational cost of the Monte Carlo method described in Section 3.3, it would be too costly to look at all 633 stations jointly over the whole Europe over periods of several months. Due to assumed spatial and temporal independence of uncertainty sources, their impact on scores greatly diminishes as the number of stations or the length of the evaluation period increases. This effect was observed to become negligible if looking at more than 100 stations over periods of more than a month (with two forecast series issued *per* day). For the case of the bias and MAE scores this can be directly supported by Appendix A, while for other scores this can be observed from computer simulations. Consider the above set-up of 100 stations and 1 month of two forecast series issued *per* day, with a standard deviation of the measurement error of $\sigma_e = 1$ m s^{-1}. In that case, the 99.7% confidence intervals for the estimated bias and MAE scores would have a width of 0.001 m s^{-1} only. The uncertainty in estimated scores would additionally decrease with more stations, longer evaluation periods or higher measurement accuracy. In view of the application in mind (wind power prediction), one can have a look instead at countries where significant wind power penetration is observed and where it is known that forecast quality is crucial for the management of wind power into the electricity network. Denmark and Ireland were consequently selected as illustrative test cases, where eight and seven validated stations can be employed, respectively.

Since no information is available about measurement accuracy at these stations, the assumption such that the standard deviation of the measurement error is $\sigma_e = 0.5$ m s^{-1}, with these measurement devices being unbiased, is formulated. This choice is supported by the review of the calibration uncertainty of state-of-the-art anemometers performed by Coquilla and Obermeier (2008). For calibrated anemometers this uncertainty may vary between 0.1 and 0.5 m s^{-1} depending on the wind speed level and the type of anemometer. It cannot be sure, however, that the measurement devices at all the synoptic stations considered are regularly calibrated. It was therefore chosen to consider 0.5 m s^{-1} as a representative value of all these stations, since representing the worst case for calibrated anemometers, and comprising a lower bound of measurement uncertainty for anemometers that are not calibrated. In the future, verification studies accounting for observational uncertainty could be refined by using up-to-date information on the quality of measurements at the various stations, or even make σ_e a function of the wind speed level. Other values for σ_e have been considered, leading to similar qualitative results. Obviously the higher σ_e gets, the larger the uncertainty on calculated scores is 200 Monte Carlo simulations are performed to estimate the uncertainty on the various scores and diagnostics performed.

As an example, Figure 8(a) and (b) depicts the CRPS as a function of the lead time for Ireland and Denmark, respectively. Each figure compares scores calculated against analysis (at the stations level), against observations, and when accounting for observational uncertainty. In that last case, the mean of the 200 Monte Carlo simulations is shown, along with 90% confidence intervals. The periodic nature of verification statistics discussed for the RMSE results of Figure 5 is also observed here for the case of the CRPS. These periodicities are directly linked to the effect of the diurnal cycle, which could be isolated if aiming at further refining the analysis. This effect does not cancel out by pooling forecasts issued at 0000 and 1200 UTC owing to the asymmetrical shape of the diurnal cycle, with a low increase during the day and a sharper drop in the evening. This effect will also be noticeable in other figures.

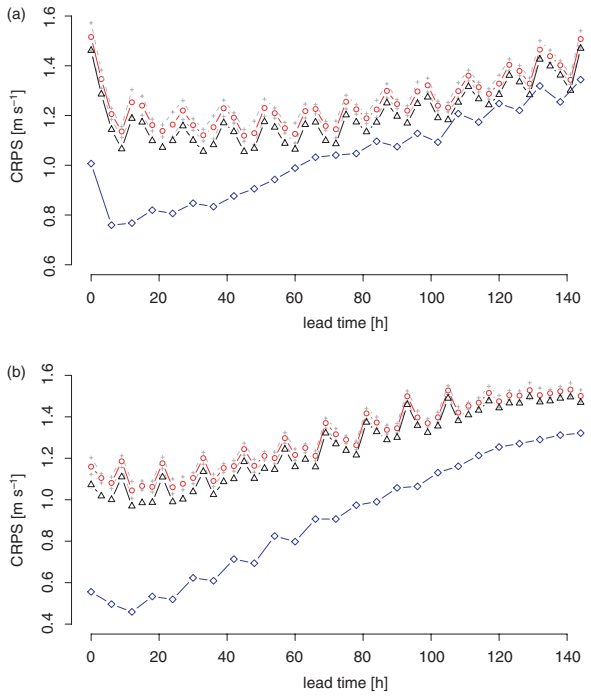

Figure 8. Comparison of the CRPS calculated as a function of lead time, as an average over stations in (a) Ireland (b) and in Denmark. It compares CRPS values calculated against analysis, against observations, with and without consideration of observational uncertainty. ◦ *vs* observations (mean), + 90% confidence intervals, ▵ *vs* observations (no unc.), ◇ *vs* analysis. This figure is available in colour online at wileyonlinelibrary.com/journal/met

For both Denmark and Ireland, there is a very large difference between CRPS scores calculated against analysis and against observations, even if the general trends are similar. For Denmark and for lead times shorter than 2 days ahead, the CRPS values calculated against observations are even twice those calculated against analysis. The mean CRPS calculated when accounting for observational uncertainty is significantly higher than if not. It even falls outside of the 90% confidence intervals. These results illustrate the discussion of Appendix A, where it is explained that accounting for observational uncertainty would generally inflate the values of certain error criteria e.g. MAE, RMSE and CRPS (but not the bias). By deconstructing the distribution of RMSE scores (see Equations (A.8)–(A.10)), it is shown that the

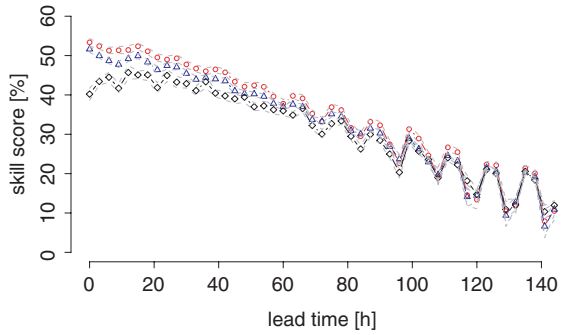

Figure 9. Skill scores giving improvements with respect to climatology over Denmark for the MAE, RMSE and CRPS scores, when accounting for observational uncertainty. Both the mean and 90% confidence intervals are represented, for each of the skill scores. ◦ SMAE, △ SRME, ⋄ SCRPS, - - - - 90% confidence intervals. This figure is available in colour online at wileyonlinelibrary.com/journal/met

mean score obtained when accounting for observational uncertainty is necessarily larger than if not. Figure 8(a,b) illustrates the fact that such a result also holds for the CRPS. The picture is different if looking at reliability though. The general deviations from perfect reliability for Denmark and Ireland are similar to those depicted in Figure 6 for the 'all stations' case. The impact of observational uncertainty is so limited that the PIT diagrams drawn for all 200 Monte Carlo simulations cannot really be distinguished. It seems that perturbations of recorded measurements globally do not significantly change the counts serving to determine the reliability of ensemble forecasts.

Finally, looking at some of the skill scores of Equation (10), calculated against time-varying climatologies, while accounting for observational uncertainty, and based on the MAE, RMSE and CRPS, plotted as a function of the lead time (Figure 9). The results are depicted for Denmark only (as the average for the eight stations), the results for Ireland being fairly similar.

The general pattern is similar to what would be observed if evaluating skill scores based on the analysis as the reference for verification. The skill (with respect to climatology) consistently decreases with the lead time, with the small subtlety of the CRPS skill score being stable for the first 36 h before starting to decrease. This is certainly due to the lack of sufficient spread of the ensembles at early lead times, since the quality of the ensemble mean and median (that is, in terms of MAE and RSME) is higher. As for the scores depicted in Figure 8(a) and (b), the impact of observational uncertainty (for the chosen value of σ_e) is limited owing to spatial and temporal dampening effect. One therefore expects that if calculating and analysing skill scores or score improvements over the whole set of European stations (as will be done in the following section), observational uncertainty would not be an issue. Interestingly, the skill scores remain positive over the whole forecast length. Periodicity in their evolution from day 4 and onwards can be observed. This can certainly be explained by the fact that the time-varying climatologies account for diurnal effects, making them more or less difficult to outperform depending on the time of day for further lead times. This effect is negligible for shorter lead times since the skill of ECMWF ensemble forecasts is highly significant while correctly capturing diurnal effects.

4.2. Evaluation of the impact of the change of horizontal resolution

The second application case relates to the assessment of the impact of the recent change of horizontal resolution (from 50 to 33 km) of the ECMWF ensemble prediction system (see Section 2.2) on the skill of ensemble forecasts of near-surface wind speed. For that purpose two versions of the ECMWF operational forecasting system were running in parallel for a targeted experiment over a period of almost 3 months. This experiment yielded 187 forecast series issued over a period starting from 3 October 2009 and ending on the 26 January 2010. Their starting times are 0000 UTC and 1200 UTC. No forecasts are available between 4 and 23 November 2009. This type of experiment allows assessing the improvements brought by the new version of the system before its actual start of operation. Such improvements are usually looked at with the analysis as a reference, and by focusing on upper-air variables (e.g. Z500). Emphasis is placed instead on a near-surface variable while seeing observations as the reference. It is foreseen that an increase in horizontal resolution yields improvements in forecast quality for near-surface winds.

Maybe the most important aspect is the improvement of overall scores, calculated for all stations, hence giving an overview of potential improvements over Europe. They are given in Figure 10 as a function of lead time, and expressed as a percentage of the scores obtained for the coarser resolution. These improvements are based on the bias, MAE, RMSE of point forecasts extracted from the ensembles, as well as on the CRPS of ensemble forecasts. Note that the bias improvement is in terms of its magnitude and therefore calculated as a decrease in its absolute value. As mentioned above, the potential effect of observational uncertainty is not considered, firstly owing to computational costs, and also since for

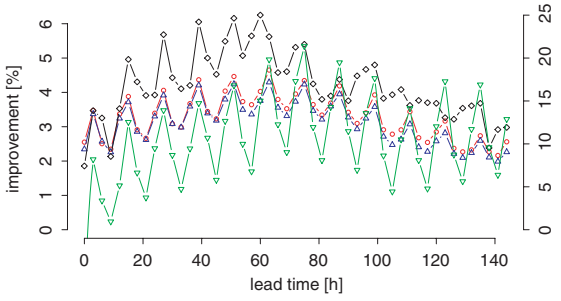

Figure 10. Global improvement of scores over Europe. The left axis scale is for the MAE, RMSE and CRPS scores, while the right one is for the bias. ◦ MAE, △ RSME, ⋄ CRPS, ▽ bias. This figure is available in colour online at wileyonlinelibrary.com/journal/met

an average over such a large number of stations it is expected to be negligible.

All improvements are positive over the forecast range considered, up to 6 days ahead. They are between 2 and 4% for the MAE and RMSE scores, while ranging from 3 to 6% for the CRPS. In view of the number of forecast series and stations involved, these improvements can be seen as noteworthy. They are even more substantial for the bias, being up to 22% for 3 day ahead forecasts. In parallel, the periodicity present for all scores (though especially for the bias, which then affect other scores) show that the change of resolution also impacted the way local diurnal effects are captured by the models. The maximum improvements for all scores are reached in the early medium range, that is, between 2 and 3 days ahead. Finally, it is interesting to see that improvements in the CRPS are larger than improvements for the more deterministic scores MAE and RMSE (since relying on point forecasts only). A potential explanation can be that the forecast quality improvements are not only related to the better ability of ensemble forecasts to target observations, and to a higher sharpness, but also originate from a better calibration.

Consequently, that point was further investigated by assessing the change in the reliability of the ensemble forecasts induced by the increased spatial resolution. This is done based on an alternative presentation of the PIT diagrams of Figure 6, which focuses on the probabilistic bias of ensemble forecasts (Pinson *et al*., 2007b; Marzban *et al*., 2011). The probabilistic bias is mathematically defined as the difference between observed and nominal proportions of ensemble members. It corresponds visually to the distance between the reliability curve and the ideal diagonal case in the plots of Figure 6. Another alternative would be to draw these PIT diagrams on the probability paper in the spirit of Bröcker and Smith (2007).

Example results are gathered in Figure 11, for the example case of 72 h ahead ensemble forecasts. Qualitatively similar results were obtained for other lead times. For a large number of stations, the situation is similar to that shown for Amsterdam Schipol and Cork airports. It consists of a substantial improvement of probabilistic reliability for the finer resolution forecasts. For some other stations e.g. Thyboron in Denmark, however, probabilistic reliability actually seems to be worse for the forecasts with finer horizontal resolution. This is also the case for some of the stations where the worst reliability statistics were observed in Section 4.1.2, such as Cap Béar in the South of France. Such a result is counter-intuitive since one would expect that more local regimes e.g. coastal effects may be better captured by increasing resolution. This may well also depend upon the physics behind the models instead. When looking at all stations altogether, the improvement in ensemble forecast reliability seems to exist, though being small.

5. Conclusions and discussion

The question of verifying ensemble forecasts against observations has been the focus of this work, with emphasis on the ECMWF ensemble prediction system and the European region. The main motivation behind this work is to argue for the proposal of verification frameworks that permit to develop a critical view of the quality of ensemble forecasts with respect to both analysis and observations. While it is fair to verify forecasts against the analysis since this one is made consistent in space and in time with the forecasts, it is also crucial to see how a forecasting system performs against actual observations. This certainly matters to the forecast users who would consider verification against observations as informing about the real quality of the forecasting system. These forecast users today have energy-related activities (e.g. wind power producers, traders, transmission system operators), are involved in airport traffic control or ship routing, etc. This approach to verification is surely also of interest to modellers and forecasters in order for them to further identify and characterize weaknesses of their forecasting approaches, for instance at the occasion of a system upgrade like the increase in spatial resolution considered here.

The disparity between verification *versus* observations and model analysis originally comes from the difference in spatial and temporal scales of these references. The model analysis is obviously consistent with the spatial and temporal scales of the model forecasts, since being based on the same numerical model. Very local effects, for example thermal or those induced by the topography, are therefore not accounted for in the model analysis while being present in the dynamics of the observations. In addition here, the fact that the near-surface wind observations on land are used in the production of the model analysis may magnify these disparities. They were observed to be substantial, their magnitude being almost comparable to the score values themselves (for the CRPS). It was also explained and shown that accounting for the effect of observational uncertainty would make the scores even worse. The study performed may be refined, if more information about measurement uncertainty at each and every station can be obtained. In the case where spatial and temporal independence of rounding and measurement errors is a safe assumption, the effect on average scores calculated over large period of times and areas is much less pronounced. It may then allow formulating robust conclusions about the comparative performance of competing forecasting approaches e.g. in the case of different horizontal resolutions here. Further work in that direction may allow issuing guidelines on the treatment of observational uncertainty depending upon the magnitude of measurement error, as well as the spatial and temporal scales involved.

In parallel, it is while focusing on reliability that the disparities between verification against analysis and observations are the most patent. The smooth characteristics of the analysis there contrast with the potentially

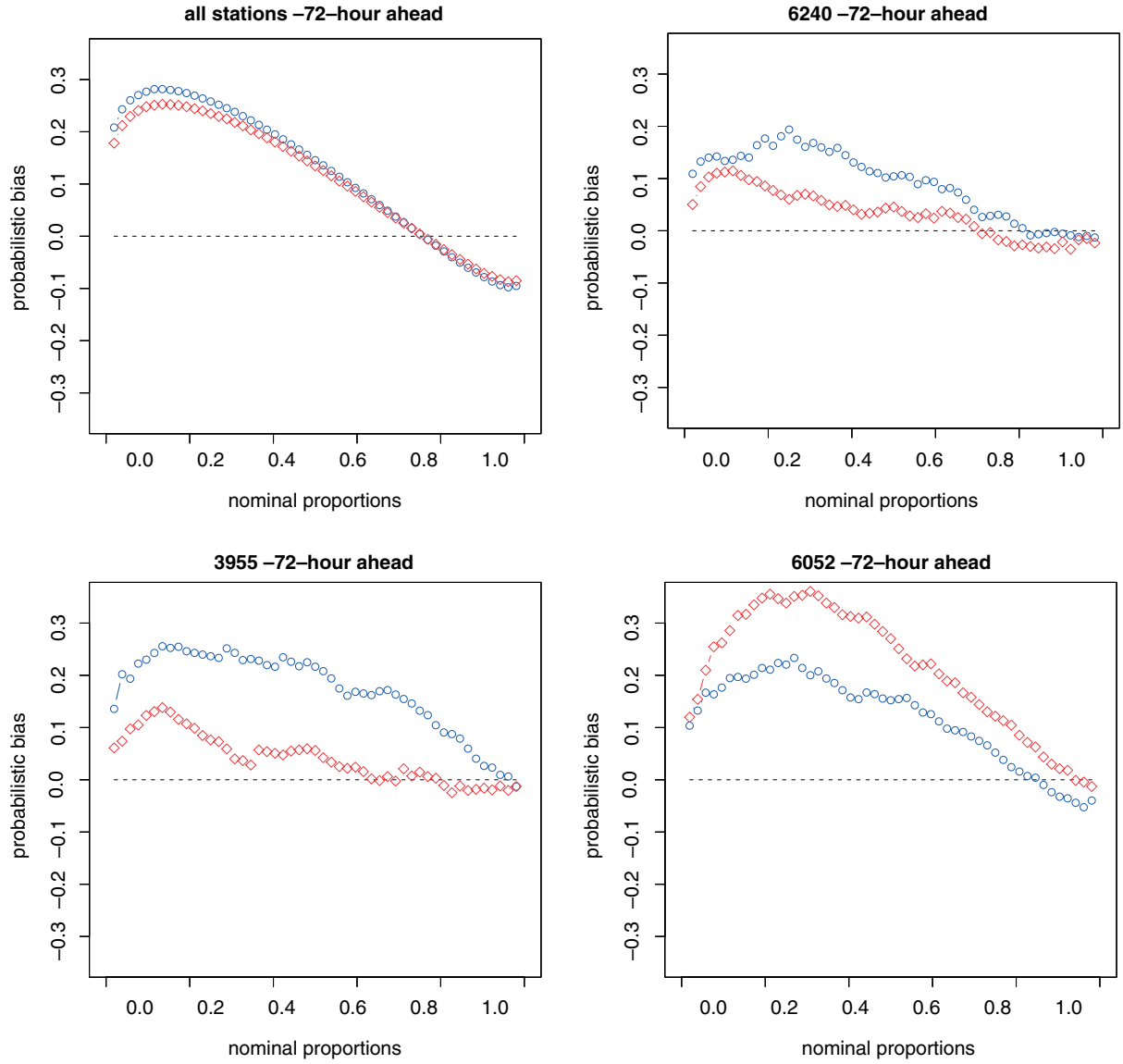

Figure 11. Impact of the change of horizontal resolution on the reliability of 72 h ahead ensemble forecasts at stations. The diagrams depicts the probabilistic bias of ensemble forecasts as a function of the nominal proportions of the ensemble members. (a) They are for all stations, (b) Amsterdam Schipol airport, (c) Cork airport and (d) finally Thyboron station in Denmark. ◦ Coarser resolution, ◇ finer resolution. This figure is available in colour online at wileyonlinelibrary.com/journal/met

strong fluctuations in observations, and consequently yields totally different reliability statistics. The ensembles tend to overestimate observed wind speeds on a general basis. For certain sites with strong local regimes though, one retrieves a more intuitive result such that ensembles significantly underestimate wind speed. The impact of observational uncertainty on the PIT diagrams was found to be minimal. This may originate from the proposal of employing a method of the observational probability type. As discussed by Candille and Talagrand (2008), perturbing ensembles in a manner consistent with observational uncertainty may allow to better account for the impact of observational uncertainty on reliability assessments. A thorough investigation of these aspects should be performed in the near future to further support results from the reliability evaluation of ensemble forecasts of wind speed against observations.

Besides its main message, this work has allowed a number of practical conclusions to be obtained from the application of this verification framework. The most important ones relate to (1) the generally good quality of ensemble forecasts of wind speed over Europe, (2) the noteworthy improvement of scores brought by the change of horizontal resolution in the system, and, (3) the scope for further improvements of reliability and skill of wind speed ensemble forecasts. Regarding that last point, let us mention a comparable study on ensemble forecasting of near-surface wind speed reported by Thorarinsdottir and Gneiting (2011b) for the North-West Pacific region of North America. Ensemble forecasts of 10 m wind speed were there issued based on the University of Washington Mesoscale Ensemble (UWME) system. For an evaluation period covering the whole calendar year of 2008, the CRPS of ensemble forecasts for lead

times up to 2 days ahead were shown to improve dramatically when employing appropriate recalibration techniques. These results support the expectations such that significant improvement in the reliability and overall skill of ECMWF ensemble forecasts (verified against observations) could be achieved with appropriate post-processing techniques.

These various conclusions are of particular relevance for various meteorological applications based on wind speed forecasts. This is particularly valid for the wind power application, for which it is known that forecast accuracy greatly impacts the cost of managing wind power production while being critical for the overall electricity networks safety.

Acknowledgements

The work presented has been partly supported by the European Commission under the SafeWind project (ENK7-CT2008-213740), which is hereby acknowledged. Acknowledgements are due to Paul Poli and Mark Rodwell at ECMWF for their help with the data. The authors are also grateful to Martin Leutbecher, Florian Pappenberger and Anna Ghelli at ECMWF as well as Robin Girard at Mines ParisTech for general discussion on forecast verification and for their comments on earlier version of that manuscript. Three anonymous reviewers are finally acknowledged for their relevant comments on an earlier version of this manuscript.

Appendix A. On the distributions of some scores when accounting for observational uncertainty.

In this appendix the distributions of some of the scores that may be employed for wind speed forecast verification are discussed. These distributions only account for observational uncertainty. Sampling uncertainty is not considered, though it could be fairly easily additionally accounted for. It is explained how some of the score distributions can be derived analytically, while it cannot be the case for some others. This motivates the use of a simulation-based approach to their estimation.

For simplicity, let us disregard the censoring of the random variable $X(t,s)$ in Equation (15). The error $e(t,s)$ around a reported measurement $x(t,s)$ is given by a sum of random variables:

$$e(t,s) = e_m(t,s) + e_r(t,s) \tag{A.1}$$

which have been defined by Equations (13) and (14). It then allows us expressing the forecast errors $\tilde{\varepsilon}(t+k|t,s)$ and $\overline{\varepsilon}(t+k|t,s)$ as the following random variables:

$$\tilde{\varepsilon}(t+k|t,s) = [x(t+k,s) - \tilde{x}(t+k|t,s)] + e_m(t+k,s) + e_r(t+k,s) \tag{A.2}$$

$$\overline{\varepsilon}(t+k|t,s) = [x(t+k,s) - \overline{x}(t+k|t,s)] + e_m(t+k,s) + e_r(t+k,s) \tag{A.3}$$

depending upon the point forecasts being defined as the median or mean of ensemble forecasts.

One remembers that the observational and rounding part of the error are independent. Spatial and/or temporal independence of the observational errors $e(t+k,s)$ is also assumed. This appears reasonable if having a diversity of measuring systems geographically spread and appropriately maintained. In that case, let us just first recall that the average of N independent Gaussian variables $Y_i \sim N(0,\sigma^2)$ is a Gaussian variable such that:

$$\frac{1}{N}\sum_{i=1}^{N} Y_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{N}\right) \tag{A.4}$$

In parallel, from the result exposed in Cramér (1946) such that the sum of N independent Uniform variables $Z_i \sim U[0,1]$ can be approximated (if N is large) by a Gaussian variable, one would obtain in the present case:

$$\frac{1}{N}\sum_{i=1}^{N} Z_i \sim \mathcal{N}\left(0, \frac{1}{12N}\right) \tag{A.5}$$

Based on the above results, for a location s and only evaluating scores over time (over N_f forecast series), the bias for the lead time k is distributed as:

$$\mathrm{bias}(k,s) \sim \mathcal{N}\left(\frac{1}{N_f}\sum_{t=1}^{N_f}[x(t+k,s) - \overline{x}(t+k|t,s)], \frac{\frac{1}{12}+\sigma_e^2}{N_f}\right) \tag{A.6}$$

In parallel in the case for which $|x(t+k,s) - \tilde{x}(t+k|t,s)| > |e(t+k,s)|$, the distribution of the MAE score accounting for observational uncertainty would similarly write:

$$\mathrm{MAE}(k,s) \sim \mathcal{N}\left(\frac{1}{N_f}\sum_{t=1}^{N_f}|x(t+k,s) - \tilde{x}(t+k|t,s)|, \frac{\frac{1}{12}+\sigma_e^2}{N_f}\right) \tag{A.7}$$

The condition expressed above seldom holds in practice. It could however be a first acceptable approximation if the magnitude of observational uncertainty is globally far smaller than that of the forecast error. If this assumption cannot be made, deriving the analytical expression of the MAE distribution becomes fairly technical owing to the presence of absolute values.

For the case of the RMSE things also get complicated due to the fact one then has to deal with products of

random variables. After a little algebra, the distribution of the RMSE can be written as:

$$\mathrm{RMSE}(k,s) \sim \mathcal{N}\left(\frac{1}{N_f}\sum_{t=1}^{N_f}[x(t+k,s)-\overline{x}(t+k|t,s)]^2, 2\sigma_\varepsilon\left[\frac{\frac{1}{12}+\sigma_e^2}{N_f}\right]\right)+\frac{1}{N_f}\sum_{t=1}^{N_f}e(t+k,s)^2 \qquad \text{(A.8)}$$

where

$$\sigma_{\overline{\varepsilon}} = \left(\frac{1}{N_f-1}\sum_{t=1}^{N_f}[x(t+k,s)-\overline{x}(t+k|t,s)]^2\right)^{\frac{1}{2}} \qquad \text{(A.9)}$$

is the standard deviation of the forecast error of the mean of the ensemble forecasts calculated based on reported observations.

The last term in Equation (A.8) involves calculating the mean of the squared distributions of observational uncertainty, which would be difficult to derive analytically. One notes however that:

$$\mathrm{E}\left[\frac{1}{N_f}\sum_{t=\mathbf{1}}^{N_f}e(t+k,s)^{\mathbf{2}}\right] = \mathrm{E}[e^{\mathbf{2}}] = \sigma_e^{\mathbf{2}} > 0 \qquad \text{(A.10)}$$

which tells us that the mean RMSE when accounting for observational uncertainty will in any case be larger than that calculated if not accounting for such observational uncertainty.

A similar problem arises when attempting to derive the distribution for the CRPS. This is since for each time step one then integrates the squared difference between the probabilistic forecast and the step function defined by the reported observation. Numerical approximation may be possible and could be the topic of further research, but globally, owing to the resulting complexity of calculation of the scores distributions and to the necessity to consider the potential censoring of observational error distributions in Equation (A.1), since wind speed cannot be negative, a simulation-based approach like that described in Section 3.3 may be seen as appropriate.

References

Abramson B, Clemen R. 1995. Probability forecasting. *International Journal of Forecasting* **11**: 1–4.

Bradley DM, Gupta RC. 2002. On the distribution of the sum of n non-identically distributed uniform random variables. *Annals of the Institute of Statistical Mathematics* **54**: 689–700.

Bröcker J, Smith LA. 2007. Increasing the reliability of reliability diagrams. *Weather and Forecasting* **22**: 651–661.

Buizza R, Houtekamer PL, Toth Z, Pellerin G, Wei M, Zhu Y. 2005. A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems. *Monthly Weather Review* **133**: 1076–1097.

Buizza R, Miller M, Palmer TN. 1999. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* **131**: 2887–2908.

Candille G, Côté C, Houtekamer PL, Pellerin G. 2007. Verification of an ensemble prediction system against observations. *Monthly Weather Review* **135**: 2688–2699.

Candille G, Talagrand O. 2008. Impact of observational error on the validation of ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society* **134**: 959–971.

Coquilla RV, Obermeier J. 2008. Calibration uncertainty comparisons between various anemometers. *Proceedings Windpower 2008, American Wind Energy Association Conference*, Houston, TX.

Costa A, Crespo A, Navarro J, Lizcano G, Madsen H, Feitosa E. 2008. A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews* **12**: 1725–1744.

Cramér H. 1946. *Mathematical Methods of Statistics*. Princeton University Press: Princeton, NJ.

Ghelli A, Lalaurette F. 2000. Verifying precipitation forecasts using upscaled observations. *ECMWF Newsletter* **87**: 9–17.

Giebel G, Kariniotakis G, Brownsword R. 2003. The state of the art in short-term prediction of wind power – a literature overview. Technical report deliverable D1.1, ANEMOS EU Project: Roskilde, Denmark. http://www.anemos-project.eu [accessed 14 July 2011].

Gneiting T. 2008. Editorial: probabilistic forecasting. *Journal of the Royal Statistical Society, Series A* **171**: 319–321.

Gneiting T. 2011a. Quantiles as optimal point predictors. *International Journal of Forecasting* **27**: 97–207.

Gneiting T. 2011b. Making and evaluating point forecasts. *Journal of the American Statistical Association* arXiv:0912.0902v2 (in press).

Hamill TM, Juras J. 2006. Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society* **132**: 2905–2923.

Jung T, Leutbecher M. 2008. Scale-dependent verification of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society* **134**: 973–984.

Katz RW, Murphy AH. 1997. *Economic Value of Weather and Climate Forecasts*. Cambridge University Press: Cambridge, MA.

Lange M, Focken U. 2005. *Physical Approach to Short-term Wind Power Prediction*. Springer: Berlin, Heidelberg.

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. *Journal of Computational Physics* **227**: 3515–3539.

Magnusson L, Leutbecher M, Källén E. 2008. Comparison between singular vectors and breeding vectors as initial perturbations for the ECMWF ensemble prediction system. *Monthly Weather Review* **136**: 4092–4104.

Marzban C, Wang R, Kong F, Leyton S. 2011. On the effect of correlations on rank histograms: reliability of temperature and wind-speed forecasts from fine-scale ensemble reforecasts. *Monthly Weather Review* **139**: 295–310.

Mason SJ. 2008. Understanding forecast verification statistics. *Meteorological Applications* **15**: 31–40.

Matos MA, Bessa RJ. 2011. Setting the operating reserve using probabilistic wind power forecasts. *IEEE Transactions on Power Systems* **26**: 594–603.

Mitra SK. 1971. On the probability distribution of the sum of uniformly distributed random variables. *SIAM Journal of Applied Mathematics* **20**: 195–198.

Murphy AH. 1993. What is a good forecast? – an essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**: 281–293.

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Monthly Weather Review* **115**: 1330–1338.

Palmer TN. 2000. Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics* **63**: 71–116.

Palmer TN, Shutts GJ, Hagedorn R, Doblas-Reyes FJ, Jung T, Leutbecher M. 2005. Representing model uncertainty in weather and climate prediction. *Annual Review of Earth and Planetary Sciences* **33**: 163–193.

Pappenberger F, Ghelli A, Buizza R, Bódis K. 2009. The skill of probabilistic precipitation forecasts under observational uncertainties within the generalized likelihood uncertainty estimation framework for hydrological applications. *Journal of Hydrometeorology* **10**: 807–819.

Pinson P, Chevallier C, Kariniotakis G. 2007a. Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Transactions on Power Systems* **22**: 1148–1156.

Pinson P, Nielsen HA, Møller JK, Madsen H, Kariniotakis G. 2007b. Nonparametric probabilistic forecasts of wind power: required properties and evaluation. *Wind Energy* **10**: 497–516.

Pinson P, McSharry P, Madsen H. 2010. Reliability diagrams for nonparametric density forecasts of continuous variables: accounting

for serial correlation. *Quarterly Journal of the Royal Meteorological Society* **136**: 77–90.

Silverman BW. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: London.

Smith J, Thresher R, Zavadil R, DeMeo E, Piwko R, Ernst B, Ackermann T. 2009. A mighty wind. *IEEE Power and Energy Magazine* **7**: 41–51.

Tay AS, Wallis KF. 2000. Density forecasting: a survey. *Journal of Forecasting* **19**: 235–254.

Thorarinsdottir TL, Gneiting T. 2010. Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society, Series A* **173**: 371–388.

Timmermann A. 2000. Density forecasting in economics and finance. *Journal of Forecasting* **19**: 231–234.

Uppala SM, Kallberg PW, Simmons AJ, Andrae U, Bechtold VD, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA, Li X, Onogi K, Saarinen S, Sokka N, Allan RP, Andersson E, Arpe K, Balmaseda MA, Beljaars ACM, Van De Berg L, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Holm E, Hoskins BJ, Isaksen L, Janssen PAEM, Jenne R, McNally AP, Mahfouf JF, Morcrette JJ, Rayner NA, Saunders RW, Simon P, Sterl A, Trenberth KE, Untch A, Vasiljevic D, Viterbo P, Woollen J. 2005. The ERA-40 reanalysis. *Quarterly Journal of the Royal Meteorological Society* **131**: 2361–3012.

