Adaptive calibration of (u,v)wind ensemble forecasts

Pinson, Pierre

Published in:
Royal Meteorological Society. Quarterly Journal

Link to article, DOI:
10.1002/qj.1873

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Pinson, P. (2012). Adaptive calibration of (u,v)wind ensemble forecasts. Royal Meteorological Society. Quarterly Journal, 138(666), 1273-1284. DOI: 10.1002/qj.1873

RMetS
Royal Meteorological Society

Adaptive calibration of (u,v)-wind ensemble forecasts

Pierre Pinson
DTU Informatics, Technical University of Denmark and European Centre for Medium-range Weather Forecasts, Reading, UK

*Correspondence to: P. Pinson, Technical University of Denmark, DTU Informatics, Richard Petersens Plads 305(212), DK-2800 Kongens Lyngby, Denmark. E-mail: pp@imm.dtu.dk

Ensemble forecasts of *(u,v)*-wind are of crucial importance for a number of decision-making problems related to e.g. air traffic control, ship routeing and energy management. The skill of these ensemble forecasts as generated by NWP-based models can be maximised by correcting for their lack of sufficient reliability. The original framework introduced here allows for an adaptive bivariate calibration of these ensemble forecasts. The originality of this methodology lies in the fact that calibrated ensembles still consist of a set of (space–time) trajectories, after translation and dilation. In parallel, the parameters of the models employed for improving the stochastic properties of the generating processes involved are adaptively and recursively estimated to accommodate smooth changes in the process characteristics and to lower computational costs. The approach is applied and evaluated based on the adaptive calibration of ECMWF ensemble forecasts of *(u,v)*-wind at 10 m above ground level over Europe over a three-year period between December 2006 and December 2009. Substantial improvements in (bivariate) reliability and in various deterministic/probabilistic scores are observed. Finally, the maps of translation and dilation factors are discussed. Copyright © 2012 Royal Meteorological Society

Key Words: ensemble prediction; probabilistic calibration; bivariate processes; recursive estimation; near-surface wind

Received 22 April 2011; Revised 14 October 2011; Accepted 29 November 2011; Published online in Wiley Online Library 3 January 2012

Citation: Pinson P. 2012. Adaptive calibration of (u, v)-wind ensemble forecasts. *Q. J. R. Meteorol. Soc.* **138**: 1273–1284. DOI:10.1002/qj.1873

1. Introduction

After the tremendous advances in the development of ensemble forecasting methodologies (Palmer, 2000; Gneiting and Raftery, 2005), ensembles and more generally probabilistic forecasts of meteorological variables are increasingly considered as a crucial input to a number of socially relevant decision-making problems. Out of the potential variables of interest, probabilistic forecasts of near-surface winds are becoming increasingly popular. This is partly due to the needs for accurate forecasting of wind power generation (from the short to medium range). Nielsen *et al.* (2006), Pinson and Madsen (2009) and Taylor *et al.* (2009) provide reviews of the methods for ensemble-based forecasting of wind power. It has been shown that the optimal management and trading of wind energy generation calls for probabilistic forecasts (e.g. Pinson *et al.*, 2007a; Matos and Bessa, 2010; Meibom *et al.*, 2011, among others). From a more general point of view, probabilistic forecasts of near-surface winds can be of great value for decision-making problems related to sailing, ship routeing, air traffic control, etc. This statement is supported by theoretical results indicating that, for a large class of decision-making problems, optimal decisions directly relate to quantiles of conditional predictive densities (Gneiting, 2011).

As is often the case for forecasts directly taken as output from physical models, ensemble forecasts of near-surface winds tend to be biased. For probabilistic forecasts, this defect consists of their lack of sufficient (probabilistic) reliability: they generally are under-dispersive. With that in mind, various approaches to the bias-correction and calibration of ensemble forecasts of wind speed (Sloughter *et al.*, 2010; Thorarinsdottir and Gneiting, 2010) and direction (Bao *et al.*, 2010) have been described. The bivariate view of wind speed and direction was in parallel touched upon by Gneiting *et al.* (2008) when discussing the skill evaluation of multivariate probabilistic forecasts. Our aim here is to further develop the calibration of ensemble forecasts of (u, v)-wind in a multivariate framework. In contrast with the argument of Wilks (2002), we are not looking at fitting probability distributions based on the ensembles. The output of our calibration methodology consists of ensemble forecasts similar in nature to the uncalibrated ones, though with improved stochastic properties. This approach is motivated by the fact that a number of decision-support systems using wind probabilistic forecasts as input need ensembles (in other words, trajectories) instead of predictive densities (Meibom *et al.*, 2011; Morales *et al.*, 2010). Indeed, by fitting probability distributions for each point in space and in time, individually, the physical spatio-temporal structure of ensemble members gets lost.

The methodology developed in the present article is mainly inspired by the approach described in Pinson and Madsen (2009) for the adaptive kernel dressing of ensemble forecasts in a univariate framework, but also by the ideas described by Sloughter (2009) for probabilistic forecasting of (u, v)-wind using Bayesian model averaging (BMA). The present proposal is developed in a multivariate Gaussian framework, with the idea of correcting the first- and second-order moment properties of the ensemble forecasts. The main innovations brought in by this approach are that:
(i) u and v wind components are jointly considered, instead of focusing on wind speed and direction individually;
(ii) the output of the models are ensemble forecasts of the same nature as the input ones, not predictive densities; and
(iii) the model parameters are considered as time-varying, while being adaptively and recursively estimated in a rigorous Maximum-Likelihood (ML) framework.

This article is structured as following. The rationale and models for calibration are introduced in the first stage by describing the bivariate Gaussian framework for (u, v)-wind, our approach to calibration, as well as the underlying mean and variance models. The question of the adaptive and recursive estimation of the model parameters is subsequently dealt with. We derive the general updating formulae for the various models in a ML framework while giving the exact expressions to be used for every set of model parameters. The methodology is finally applied for the calibration of European Centre for Medium-range Weather Forecasts (ECMWF) ensemble forecasts of (u, v)-wind (at 10 m above ground level) over Europe and over a three-year period. The original and calibrated ensemble forecasts are evaluated in a bivariate framework, focusing on the deterministic skill of the ensemble mean, as well as the reliability and skill of ensemble forecasts. The article ends with conclusions and perspectives regarding future work.

2. Rationale and models for calibration

Before we get into a description of our proposed methodology, related models and estimation methods, it is important to mention that the model analysis is considered as the target and hence is used as a reference for calibration. Calibration against observations at particular sites may rely on similar approaches and models, though the scope of their application would be different. Indeed we consider here that, before forecasts may be communicated to their potential users, a requirement for the meteorological centre issuing the forecasts is to ensure that its ensemble forecasts are calibrated with respect to its own target, i.e. its own analysis.

2.1. *The general bivariate Gaussian framework*

Instead of looking at wind speed and direction individually, wind is modelled as a bivariate process, i.e. in terms of its zonal and meridional components, denoted u and v, respectively. While physical reasons hint at the fact these two components of the wind are inter-related, their interdependence is illustrated in Figure 1 based on 48 h ensemble forecasts from ECMWF over the period December 2005–February 2006. This figure depicts a map of average values of the coefficient of determination R^2, showing the general level of interdependence between these u and v

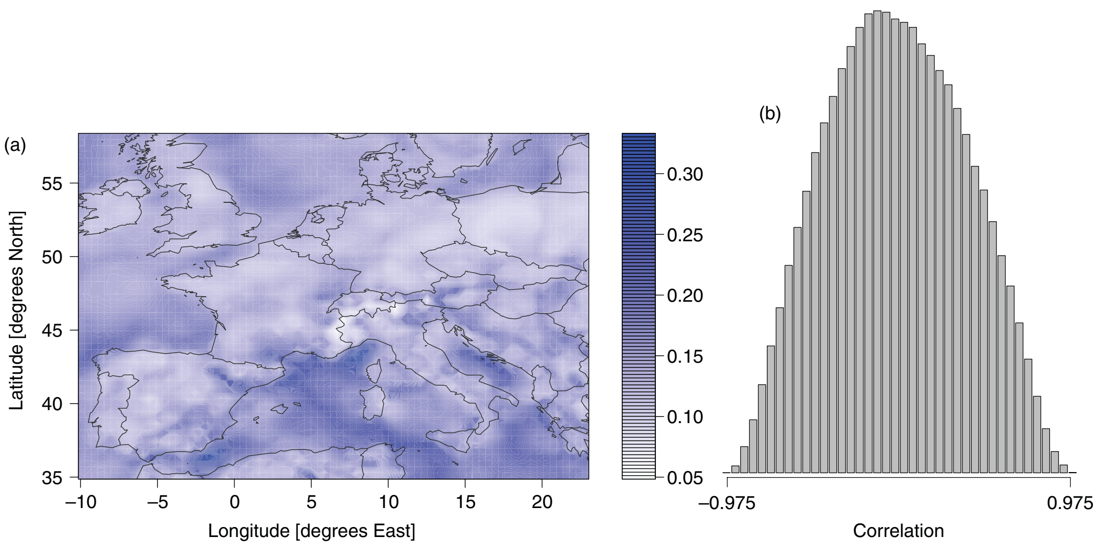

Figure 1. Interdependence of u and v wind components of ensemble forecasts, based on the 48 h ECMWF ensemble forecasts over DJF2006. (a) is a map of the average R^2 values between u and v over all forecast series, and (b) is the distribution of the (u, v) correlation for all locations and all forecast series. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

components, as well as the distribution of the (u, v)-correlation values for all locations and all forecast series over that period. While the level of correlation between u and v components may well vary in time and in space, it appears that they are significantly interrelated overall, especially in zones with specific wind regimes e.g. the Tramontane in the French Gulf of Lion. This justifies the proposal of bivariate approaches to the recalibration of (u, v)-wind ensemble forecasts, instead of the more classical univariate approaches.

For a given location s, $\mathbf{y}_{s,t} = [u_{s,t} \; v_{s,t}]^\top$ is the observed wind vector at time t. It is assumed that (u, v)-wind is distributed bivariate Gaussian, possibly after transformation. We denote by $\mathbf{Y}_{s,t} = [U_{s,t} \quad V_{s,t}]^\top$ the bivariate random variable for the wind vector at time t and for location s. Subsequently,

$$\mathbf{Y}_{s,t} \sim \mathbb{N}_2 \left(\boldsymbol{\mu}_{s,t}, \boldsymbol{\Sigma}_{s,t}\right), \tag{1}$$

where $\boldsymbol{\mu}_{s,t} = [\mu_{u,s,t} \quad \mu_{v,s,t}]^\top$ is a two-dimensional vector giving the wind vector expectation at time t and location s while $\boldsymbol{\Sigma}_{s,t}$ is the variance–covariance of the random variable:

$$\boldsymbol{\Sigma}_{s,t} = \begin{bmatrix} {\sigma_{u,s,t}}^2 & \rho_{s,t}\,\sigma_{u,s,t}\,\sigma_{v,s,t} \\ \rho_{s,t}\,\sigma_{u,s,t}\,\sigma_{v,s,t} & {\sigma_{v,s,t}}^2 \end{bmatrix}. \tag{2}$$

In the above, $\sigma_{u,s,t}$ and $\sigma_{v,s,t}$ are the standard deviation of the random variable at time t and location s along the u and v components, respectively, while $\rho_{s,t}$ is the (u, v)-correlation, controlling the anisotropic shape of bivariate distributions. Comprehensive illustrations of using bivariate Gaussian distributions for predicting (u, v)-wind can be found in Gneiting *et al.* (2008), for instance.

Similarly, for the ensemble forecasts of (u, v)-wind, for given lead time k, we write

$$\widehat{\mathbf{y}}^{(j)}_{s,t|t-k} = \left[\widehat{u}^{(j)}_{s,t|t-k} \; \widehat{v}^{(j)}_{s,t|t-k}\right]^\top \tag{3}$$

as the jth member of a set of m ensemble wind forecasts, issued at time $t - k$ for the current time t (hence with k denoting the forecast horizon) for the location s. Since in the following we will focus on each location and lead time individually, and in order to ease notations, we do not employ subscripts that would indicate the lead time k and location s (unless necessary). One should for instance remember that $\widehat{\mathbf{y}}^{(j)}_t$ actually corresponds to $\widehat{\mathbf{y}}^{(j)}_{s,t|t-k}$, the same being valid for the analysis as well as individual u and v components.

It is also assumed that ensemble forecast members sample a bivariate Gaussian distribution, possibly after transformation. An exploratory analysis of ECMWF ensemble forecasts and analysis data available showed that the bivariate Gaussian assumption was suitable for both ensemble forecasts and forecasts errors of the ensemble mean, with no transformation being necessary. This contrasts with the analysis of Sloughter (2009), who observed for the University of Washington Mesoscale Ensemble (UWME) system that a power transformation should be applied to the wind speed for the ensemble forecasts of (u, v)-wind to be more Gaussian.

For the purpose of the derivations to be performed, let us recall here the density function for bivariate Gaussian random variables:

$$f(\mathbf{y}) = \frac{1}{2\pi\sigma_u\sigma_v\sqrt{1-\rho^2}} \times \exp\left[\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{u-\mu_u}{\sigma_u}\right)^2 + \left(\frac{v-\mu_v}{\sigma_v}\right)^2 - \frac{2\rho(u-\mu_u)(v-\mu_v)}{\sigma_u\sigma_v}\right\}\right]. \tag{4}$$

In the remainder of this article and under this bivariate Gaussian assumption, the variables related to the ensemble forecasts will be denoted with a hat symbol, as $\widehat{\mu}_u$ and $\widehat{\sigma}_u$ for the expectation and standard deviation of the u component of their generating process. In parallel, a star symbol will denote all quantities linked to the calibrated ensemble forecasts e.g. $\widehat{\mu}^*_u$ and $\widehat{\sigma}^*_u$ for the corrected expectation and standard deviation related to the u component.

2.2. *The rationale behind calibration*

Proposals for the calibration of ensemble forecasts can already be found in the work of Wilks (2002) and of Buizza *et al.* (2003) among others, based on fitting probability distributions to the set of ensemble members. Since bivariate Gaussian variables are fully characterised by their mean vector and covariance matrix, fitting appropriate probability distributions would translate to the estimation of their mean and covariance, for instance in a ML framework. This would comprise a generalisation of the univariate case, as considered by Vannitsem and Hagedorn (2011) for the specific case of wind speed and by Gneiting *et al.* (2005) in a more general case.

Since we aim to conserve the original nature of the ensemble forecasts, we introduce a little twist to these approaches by avoiding the direct fitting of distributions. Our proposal is instead to concentrate on the underlying generating processes for the ensemble forecasts and for the error of the ensemble mean for every lead time, in order to introduce a two-dimensional translation and dilation of the sets of ensemble forecasts. This reduces to proposing models for the mean and variance of the bivariate Gaussian densities, then yielding translation and dilation factors. The translation corresponds to the (bivariate) bias correction of the ensemble mean, while the dilation translates to the variance correction of the (unbiased) ensemble forecasts along the u and v dimension. The potential correction of the (u, v)-correlation is not considered since it would be difficult to apply it without having to resample from the generating processes, which is exactly what we aim to avoid.

Let us illustrate the rationale behind our proposal for the calibration of ensemble forecasts of (u, v)-wind based on Figure 2. On Figure 2(a), the original and calibrated sets of ensemble members issued for a given location s, at a given time t and for a specific lead time k, are shown. The corresponding generating processes for these two sets of ensemble members are represented in Figure 2(b). The mean/mode of the generating process is displaced towards higher magnitude of u and v components, while its variance is increased–actually more along the u dimension than along the v one. The difference between the two means of these generating processes gives the translation factors to be applied in the first stage to the ensemble forecasts, while the ratio between variances (along the u and v dimensions,

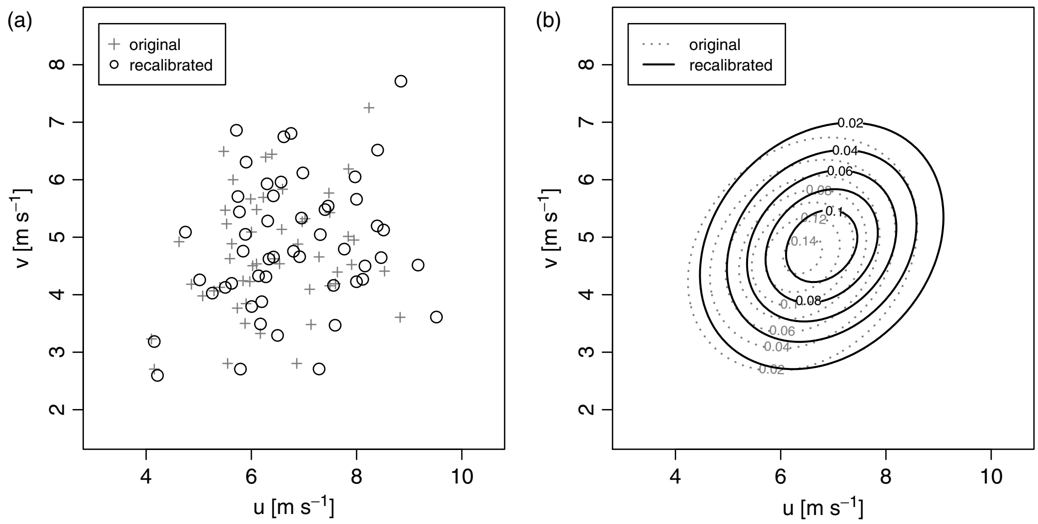

Figure 2. Illustrative example of the calibration of ensemble forecasts of (u, v)-winds, considering a translation and dilation of the set of ensemble members. (a) original and calibrated set of ensembles after derivation of translation and dilation factors, and (b) the generating processes for both original and calibrated ensembles. The translation and dilation factors are derived so as to improve the correspondence of the stochastic properties of the generating processes for both ensemble forecasts and observations.

individually) yields the dilation factors to be applied in the second stage.

In more mathematical terms, at time t, location s and for lead time k, by writing $\boldsymbol{\tau} = [\tau_u \;\; \tau_v]^\top$ and $\boldsymbol{\xi} = [\xi_u \;\; \xi_v]^\top$ as the translation and dilation factors, respectively, this yields

$$\widetilde{\mathbf{y}} = \overline{\mathbf{y}} + \tau, \tag{5}$$

$$\mathbf{y}^{(j)*} = \widetilde{\mathbf{y}} + \mathrm{diag}(\xi)\left(\mathbf{y}^{(j)} - \overline{\mathbf{y}}\right), \; j = 1, \ldots, m, \tag{6}$$

where $\overline{\mathbf{y}} = [\overline{u} \; \overline{v}]^\top$ is the bi-dimensional ensemble mean, $\mathbf{y}^{(j)*}$ $(j = 1, \ldots, m)$ are the members of the calibrated ensembles, and $\mathrm{diag}(\xi)$ is a matrix of zeros with the elements of ξ on its diagonal.

2.3. *Obtaining the translation and dilation factors*

Since placing ourselves in a bivariate Gaussian framework, the translation and dilation factors are obtained based on models for the mean and variance of the generating processes for both the ensemble forecasts and the errors of the ensemble mean. An exploratory analysis indicated that a set of linear models would be sufficient. Even though we do not use appropriate subscripts, all models below are defined for each lead time and point in space, individually. The stochastic characteristics of the ensemble forecasts necessarily evolve as a function of these two variables.

In the first stage, the translation factor is deduced from linear bivariate models used to correct the mean of the generating process,

$$\mu_u^* = \boldsymbol{\theta}_u^\top \mathbf{x}, \qquad \mu_v^* = \boldsymbol{\theta}_v^\top \mathbf{x}, \tag{7}$$

where $\mathbf{x} = [1 \;\; \overline{u} \;\; \overline{v}]^\top$ with $\overline{u}$ and $\overline{v}$ the mean of ensemble forecasts of u and v wind components for that lead time, while $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$ are vectors of model parameters. They are generically referred to as mean models. Employing such a bivariate approach implies a bi-dimensional view of the translation in the (u, v)-plane. We generically write $\boldsymbol{\theta}$ for the set of model parameters for mean correction. Note that one could potentially use additional explanatory variables in model (7), like high-resolution deterministic forecasts for instance. Subsequently, the translation factor $\boldsymbol{\tau}$ is given by

$$\tau_u = \boldsymbol{\theta}_u^\top \mathbf{x} - \overline{u}, \qquad \tau_v = \boldsymbol{\theta}_v^\top \mathbf{x} - \overline{v}. \tag{8}$$

For the case of the variance of the generating process, it has been observed that a univariate scaling along the u and v axes would be sufficient. The chosen solution was then to employ two linear models for σ_u and σ_v individually,

$$\sigma_u^* = \exp(\boldsymbol{\gamma}_u)^\top \mathbf{z}_u, \qquad \sigma_v^* = \exp(\boldsymbol{\gamma}_v)^\top \mathbf{z}_v, \tag{9}$$

where $\mathbf{z}_u = [1 \;\; \sigma_u]^\top$ and $\mathbf{z}_v = [1 \;\; \sigma_v]^\top$, with σ_u and σ_v being sample estimates of the standard deviations of the ensemble forecasts along the u and v dimensions, respectively. $\boldsymbol{\gamma}_u$ and $\boldsymbol{\gamma}_v$ are the corresponding bi-dimensional vectors of model coefficients. The models in (9) are referred to as variance models (even though they actually concentrated on standard deviations instead). We generically write $\boldsymbol{\gamma}$ for the set of model parameters for variance correction. We take the exponential of these coefficients to ensure that the coefficients applied to the various explanatory variables are always positive, so that the resulting standard deviations are in turn constrained to be positive. Finally, the dilation factors along the u and v dimension are obtained as

$$\xi_u = \frac{\sigma_u^*}{\sigma_u}, \qquad \xi_v = \frac{\sigma_v^*}{\sigma_v}. \tag{10}$$

Even though the calibration of the ensemble forecasts is based on the translation and dilation factors, the actual parameters to be estimated are those of the above mean and variance models, i.e. $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$.

3. Adaptive and recursive estimation of the model parameters

Based on these parametric assumptions for the generating processes of forecasts and errors of the ensemble mean, we propose to estimate the parameters of the models introduced in section 2.3 through a ML approach. It is thus aimed at maximising the likelihood of the observed wind vectors, given the calibrated probabilistic forecasts resulting from the model. More specifically, the method is a Recursive Maximum Likelihood (RML) approach, with exponential forgetting of past observations. An advantage of such a proposal is that only the last available set of forecasts and measurements is employed at a given time t for updating the model parameters. It hence allows for significant lowering of computational costs compared to the more traditional batch estimation methods, e.g. using a moving window of 3 months for estimating model coefficients. Another advantage brought in by the exponential forgetting is the ability for the model parameters to smoothly evolve, as a reaction to changes in the joint forecast–observation process characteristics. These changes may originate from changes in the wind dynamics e.g. due to seasonalities, but also from changes in the forecasting system, like at the occasion of a change of model physics or of a change of horizontal/vertical resolution.

3.1. *General aspects of the RML estimation*

The method described below is inspired by that of Pinson and Madsen (2009), which was introduced for the calibration of ensemble forecasts of wind power. In a RML estimation paradigm, the estimate of the model parameters is defined at a given time t as that which minimises the objective function

$$S_t(\boldsymbol{\theta},\boldsymbol{\gamma}) = -\frac{1}{n_\lambda}\sum_{i=1}^{t-k}\lambda^{t-k-i}\ln\left\{L(\mathbf{y}_i;\boldsymbol{\theta},\boldsymbol{\gamma})\right\}, \qquad (11)$$

where $\lambda \in (0,1)$ is the forgetting factor allowing for adaptivity in time (by giving less weight to older observations), and n_λ is the effective number of observations, $n_\lambda = (1-\lambda)^{-1}$, used for normalising the objective function. The value of λ is typically slightly below 1. In parallel, the term $L(\mathbf{y}_i;\boldsymbol{\theta},\boldsymbol{\gamma})$ denotes the likelihood of observing the wind vector $\mathbf{y}_i$ in view of the generating process for the calibrated ensemble forecasts issued at time $i-k$ for lead time i, given the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$:

$$L(\mathbf{y}_i;\boldsymbol{\theta},\boldsymbol{\gamma}) \;=\; P[\mathbf{y}_i|\boldsymbol{\theta},\boldsymbol{\gamma}] = \widehat{f}^*(\mathbf{y}_i), \qquad (12)$$

where $\widehat{f}^*$ is the (bivariate Gaussian) density function for the generating process, as expressed in (4). In the following, we denote by $\widehat{\boldsymbol{\theta}}_t$ and $\widehat{\boldsymbol{\gamma}}_t$ the estimate of the model parameters at time t.

The interest of this ML estimation method is that minimising the objective function of (11) is equivalent to minimising the logarithmic scoring rule known as ignorance (Roulston and Smith, 2002) (here for bivariate probabilistic forecasts). Ignorance considers a trade-off between reliability and sharpness of the probabilistic forecasts. It is also a proper scoring rule which ensures that a lower value of the score indeed corresponds to a higher skill of the probabilistic forecasts. As a consequence, recursively minimising the objective function in (11) will permit us to obtain ensemble forecasts with a generating process having maximised skill, given the ensemble forecasts used as input and the chosen model for translation and dilation. The minimisation of other proper scores like the Continuous Rank Probability Score (CRPS) could be considered instead, as discussed by Gneiting *et al.* (2007). The derivation of a similar recursive estimation scheme would be not be possible however, since it would require us to perform batch estimation on a sliding window of recent data.

Since having at time t two vectors of model parameters $\widehat{\boldsymbol{\theta}}_{t-1}$ or $\widehat{\boldsymbol{\gamma}}_{t-1}$ to be updated, they are taken care of one after the other (starting first with the translation ones). When dealing with one vector of parameters, the other one is considered fixed. By generally writing $\widehat{\boldsymbol{\nu}}$ for these model parameters (thus being $\widehat{\boldsymbol{\theta}}$ or $\widehat{\boldsymbol{\gamma}}$), the RML estimation is derived as follows. From the formulation of the ML estimation problem given above, a corresponding recursive estimation procedure can be derived by applying the method described by Madsen (2007). Indeed, the basis for derivation of such recursive procedure is to employ a Newton–Raphson step for expressing the estimate $\widehat{\boldsymbol{\nu}}_t$ as a function of the previous estimate $\widehat{\boldsymbol{\nu}}_{t-1}$,

$$\widehat{\boldsymbol{\nu}}_t = \widehat{\boldsymbol{\nu}}_{t-1} - \frac{\nabla_{\boldsymbol{\nu}} S_t(\widehat{\boldsymbol{\nu}}_{t-1})}{\nabla^2_{\boldsymbol{\nu}} S_t(\widehat{\boldsymbol{\nu}}_{t-1})}. \qquad (13)$$

We follow the reasoning of Pinson and Madsen (2009) for the developments below. One first deduces from (11) that

$$S_t(\widehat{\boldsymbol{\nu}}_{t-1}) = \lambda S_{t-1}(\widehat{\boldsymbol{\nu}}_{t-1}) - \frac{1}{n_\lambda}\ln\left\{L(\mathbf{y}_t;\widehat{\boldsymbol{\nu}}_{t-1})\right\}, \qquad (14)$$

which can then be used for deriving recursive formulae for the calculation of $\nabla_{\boldsymbol{\nu}} S_t$ and $\nabla^2_{\boldsymbol{\nu}} S_t$. Indeed, for $\nabla_{\boldsymbol{\nu}} S_t$ we write

$$\nabla_{\boldsymbol{\nu}} S_t(\widehat{\boldsymbol{\nu}}_{t-1}) = -\frac{1}{n_\lambda}\frac{\nabla_{\boldsymbol{\nu}} L(\mathbf{y}_t;\widehat{\boldsymbol{\nu}}_{t-1})}{L(\mathbf{y}_t;\widehat{\boldsymbol{\nu}}_{t-1})}, \qquad (15)$$

since $\boldsymbol{\nu}_{t-1}$ is assumed to be the optimal estimate at time $t-1$, thus minimising the objective function S_{t-1} and yielding $\nabla_{\boldsymbol{\nu}} S_{t-1}(\widehat{\boldsymbol{\nu}}_{t-1}) = 0$. In a similar manner, by assuming that L is almost linear around the optimal estimate*, a recursive formula for the Hessian of the objective function can be written as

$$\nabla^2_{\boldsymbol{\nu}} S_t(\widehat{\boldsymbol{\nu}}_{t-1}) = \lambda\nabla^2_{\boldsymbol{\nu}} S_{t-1}(\widehat{\boldsymbol{\nu}}_{t-1}) + \frac{1}{n_\lambda}\frac{\nabla_{\boldsymbol{\nu}} L(\mathbf{y}_t;\widehat{\boldsymbol{\nu}}_{t-1})\left\{\nabla_{\boldsymbol{\nu}} L(\mathbf{y}_t;\widehat{\boldsymbol{\nu}}_{t-1})\right\}^\top}{L^2(\mathbf{y}_t;\widehat{\boldsymbol{\nu}}_{t-1})}. \qquad (16)$$

Then, by defining the information vector

$$\mathbf{h}_t = \frac{\nabla_{\boldsymbol{\nu}} L(\mathbf{y}_t;\widehat{\boldsymbol{\nu}}_{t-1})}{L(\mathbf{y}_t;\widehat{\boldsymbol{\nu}}_{t-1})} \qquad (17)$$

and the estimate of its inverse covariance matrix

$$\mathbf{R}_t = \nabla^2_{\boldsymbol{\nu}} S_t(\widehat{\boldsymbol{\nu}}_{t-1}), \qquad (18)$$

one obtains from (13)–(16) the two-step scheme for updating the $\boldsymbol{\nu}$-estimate at time t, i.e.

$$\widehat{\boldsymbol{\nu}}_t \;=\; \widehat{\boldsymbol{\nu}}_{t-1} + \frac{1}{n_\lambda}\mathbf{R}_t^{-1}\mathbf{h}_t, \qquad (19)$$

$$\mathbf{R}_t \;=\; \lambda\mathbf{R}_{t-1} + \frac{1}{n_\lambda}\mathbf{h}_t\mathbf{h}_t^\top. \qquad (20)$$

*This assumption is common for iterative estimation methods where Taylor expansions of the objective function are used and with the steps of the updating procedure of limited magnitude.

The interest of the recursive estimation scheme can be observed from the above formulae: only the latest available information is used at time t for updating the model parameters. This recursive estimation scheme is initialised by setting all model parameters so that the procedure of translation and dilation corresponds to an identity transformation. In parallel, the initial inverse covariance matrix $\mathbf{R}_t$ (for $t=0$) can be filled in with zero values. Obviously, such a matrix cannot be inverted as would be necessary for updating model parameters with (19). The approach to be taken then consists of using (20) for updating $\mathbf{R}_t$ only as long as $\mathbf{R}_t$ is non-invertible, and then starting to use (19) when this stage is reached eventually.

3.2. *Obtaining the moments of the calibrated ensemble-generating process*

When the analysis of wind vectors $\mathbf{y}_t$ is made available at time t, it is then used for updating the parameters of the mean and variance models (described in section 2.3) for the ensemble-generating processes for every lead time and every location (that is, grid nodes). Analysed wind vectors are then compared with all past forecasts made for time t, hence with varying lead time k. Subsequently for a given lead time and location, it is first necessary to compute the moments of the calibrated generating process based on the last available set of the corresponding model parameters, i.e. from time $t-1$. Firstly, based on (7), one has

$$\mu^*_{u,t} = \widehat{\boldsymbol{\theta}}^\top_{u,t-1}\mathbf{x}_t\,, \quad \mu^*_{v,t} = \widehat{\boldsymbol{\theta}}^\top_{v,t-1}\mathbf{x}_t\,, \tag{21}$$

with $\mathbf{x}_t = [1\ \ \bar{u}_t\ \ \bar{v}_t]^\top$. Similarly, based on model (9) for σ_u and σ_v, one obtains

$$\sigma^*_{u,t} = \exp(\widehat{\boldsymbol{\gamma}}_{u,t-1})^\top \mathbf{z}_{u,t}\,, \quad \sigma^*_{v,t} = \exp(\widehat{\boldsymbol{\gamma}}_{v,t-1})^\top \mathbf{z}_{v,t}\,, \tag{22}$$

with $\mathbf{z}_{u,t} = [1\ \ \sigma_{u,t}]^\top$ and $\mathbf{z}_{v,t} = [1\ \ \sigma_{v,t}]^\top$.

3.3. *RML estimation for the translation factors*

Based on the general developments of section 3.1, the information vector $\mathbf{h}_t$ has to be calculated at time t when a new set of wind vector observations $\mathbf{y}_t$ is made available. In practice, the basic quantity to be computed is the gradient of the likelihood with respect to relevant parameters.

Consider first the parameters $\boldsymbol{\theta}_u$, and after a little algebra, the information vector to be employed for the updating of $\boldsymbol{\theta}_u$ at time t is

$$\mathbf{h}_t = \begin{bmatrix} 1 \\ u_t \\ v_t \end{bmatrix} \otimes \frac{1}{\sigma^*_{u,t}\left(1-\rho_t^2\right)} \left(\frac{u_t - \mu^*_{u,t}}{\sigma^*_{u,t}} - \rho_t \frac{v_t - \mu^*_{v,t}}{\sigma^*_{v,t}} \right), \tag{23}$$

with $\otimes$ denoting the fact that all elements of the vector on the left-hand side are multiplied by the scalar value on the right-hand side. In the above, the updated moments $\mu^*_{u,t}$, $\mu^*_{v,t}$, $\sigma^*_{u,t}$ and $\sigma^*_{v,t}$ are obtained using (21) and (22), that is, based on the last available model parameters $\widehat{\boldsymbol{\theta}}_{u,t-1}$, $\widehat{\boldsymbol{\theta}}_{v,t-1}$, $\widehat{\boldsymbol{\gamma}}_{u,t-1}$ and $\widehat{\boldsymbol{\gamma}}_{v,t-1}$ at time $t-1$.

In parallel, for the case of $\boldsymbol{\theta}_v$, this same information vector is written as

$$\mathbf{h}_t = \begin{bmatrix} 1 \\ u_t \\ v_t \end{bmatrix} \otimes \frac{1}{\sigma^*_{v,t}(1-\rho_t^2)} \left(\frac{v_t - \mu^*_{v,t}}{\sigma^*_{v,t}} - \rho_t \frac{u_t - \mu^*_{u,t}}{\sigma^*_{u,t}} \right). \tag{24}$$

These expressions can then be plugged into (17) and subsequent equations for the updating of the mean model parameters.

3.4. *RML estimation for the dilation factors*

In the second stage we look at the model parameters for the variance models related to dilation factors. Considering first the parameters $\boldsymbol{\gamma}_u$, the information vector to be employed at time t is

$$\begin{aligned} \mathbf{h}_t = {} & \mathrm{diag}\{\exp(\widehat{\boldsymbol{\gamma}}_{u,t-1})\} \begin{bmatrix} 1 \\ \sigma_{u,t} \end{bmatrix} \otimes \frac{1}{\sigma^*_{u,t}} \\ & \times \left\{ \frac{u_t - \mu^*_{u,t}}{\sigma^*_{u,t}(1-\rho_t^2)} \left(\frac{u_t - \mu^*_{u,t}}{\sigma^*_{u,t}} - \rho_t \frac{v_t - \mu^*_{v,t}}{\sigma^*_{v,t}} \right) - 1 \right\}, \end{aligned} \tag{25}$$

where $\mathrm{diag}\{\exp(\widehat{\boldsymbol{\gamma}}_{u,t-1})\}$ is a diagonal matrix with the elements of $\exp(\widehat{\boldsymbol{\gamma}}_{u,t-1})$ on its diagonal. In a symmetric manner, the information vector for updating the parameters $\boldsymbol{\gamma}_v$ at time t is written as

$$\begin{aligned} \mathbf{h}_t = {} & \mathrm{diag}\{\exp(\widehat{\boldsymbol{\gamma}}_{v,t-1})\} \begin{bmatrix} 1 \\ \sigma_{v,t} \end{bmatrix} \otimes \frac{1}{\sigma^*_{v,t}} \\ & \times \left\{ \frac{v_t - \mu^*_{v,t}}{\sigma^*_{v,t}(1-\rho_t^2)} \left(\frac{v_t - \mu^*_{v,t}}{\sigma^*_{v,t}} - \rho_t \frac{u_t - \mu^*_{u,t}}{\sigma^*_{u,t}} \right) - 1 \right\}. \end{aligned} \tag{26}$$

Similarly to the above, these expressions are to be plugged into (17) and subsequent equations for the updating of the variance model parameters.

4. Applications and test cases

4.1. *Test case and available data*

The ensemble forecasts of (u, v)-wind at 10 m above ground level originate from the operational ensemble forecasting system at ECMWF. The forecast length considered is 5 days, corresponding to the lead times of interest for most of the decision-making problems involving wind forecasts. Another reason for this choice of a limited range of lead times is that we do not expect calibration to be necessary or bringing substantial benefits for further lead times. The domain chosen for this study is Europe–defined here by longitudes between –10 and 23°E, and latitudes between 35 and 58°N. This domain covers a rectangular latitude–longitude grid with $S = 80 \times 57 = 4560$ grid nodes. Future work may consider the possibility of evaluating the approach proposed here over the whole globe, in order to assess its interest under various climates.

Data including ensemble forecasts and the related model analyses from ECMWF have been collected over a period spanning December 2006 to December 2009. These ensemble forecasts are issued twice a day at 0000 UTC 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. But since the model analysis which is seen as a reference has a temporal resolution of 6 h only, we consider this coarser temporal resolution in the present study.

The methodology employed for the generation of the ECMWF ensemble forecasts is well documented and a number of publications can be cited for its various components. Palmer (2000) provides a general overview. It is not our objective 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 e.g. Buizza *et al.* (2005). The ECMWF ensemble predictions aim to represent uncertainties in both the knowledge of the initial state of the atmosphere and in the physical parametrisation 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 parametrisation of the numerical model (Buizza *et al.*, 1999; Palmer *et al.*, 2005). Note that the potential structural model uncertainty is therefore not accounted for.

4.2. Model configuration and estimation set-up

From the available data, two periods are defined: the first one for identification (and initial training) of the statistical models, and the second one for evaluating what the performance of these models may be under operational conditions. The first year of data is employed as the training set, exactly covering the months from December 2006 to November 2007. The remainder of the dataset (December 2007 to November 2009) is used for out-of-sample evaluation of the reliability and skill of the ensemble forecasts of (u, v)-wind, before and after calibration. We do not use the month of December 2009 for the out-of-sample forecast evaluation since we focus on complete quarters only.

Over the training period, a part of the data is used for one-fold cross-validation (the last 6 months), in order to select an optimal forgetting factor for the various models involved in the translation and dilation of the ensemble forecasts. Actually, instead of considering the forgetting factor itself, it is preferable to use the corresponding effective number of observations $n_\lambda = 1/(1-\lambda)$. This allows one to better appraise the size of the equivalent 'sliding window' in the adaptive estimation of the dynamic model parameters, such as that considered by Gneiting and Raftery (2005) and Hering and Genton (2009) for instance. The selection of optimal values for the model structure and parameters n_λ is done in a trial-and-error manner, by evaluating the results obtained from a set of different set-ups. For more information on cross-validation, we refer to Stone (1974). The criterion to be minimised over the cross-validation set is the Energy Score, *ES*. This score comprises a multivariate generalisation of the more common Continuous Ranked Probability Score (CRPS). It is a proper skill score already employed by Gneiting *et al.* (2008) for the evaluation of density forecasts of (u, v)-wind. It will also be considered as a lead score for the out-of-sample evaluation of the skill of our calibrated ensemble forecasts. For a given location s, at a given time t for a lead time k, the value of *ES* for a set of ensemble forecasts $\{\widehat{\mathbf{y}}^{(j)}_{s,t+k|t}\}_j$ with corresponding observation $\mathbf{y}_{s,t+k}$ is obtained as

$$ES_{s,t,k} = \frac{1}{m}\sum_{j=1}^{m} ||\mathbf{y}^{(j)}_{s,t+k|t} - \mathbf{y}_{s,t+k}|| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} ||\mathbf{y}^{(i)}_{s,t+k|t} - \mathbf{y}^{(j)}_{s,t+k|t}||\,, \quad (27)$$

where $||.||$ denotes the Euclidean norm. Over an evaluation set of N time steps, the *ES* value for a given lead time k and for a given grid node is given by

$$ES_k = \frac{1}{N}\frac{1}{S}\sum_{t=1}^{N}\sum_{t=1}^{S} ES_{s,t,k}\,. \quad (28)$$

Based on this cross-validation exercise, it was found that an optimal value for the forgetting factor would be $\lambda = 0.996$, corresponding to a equivalent number of observations $n_\lambda = 250$ (or in other words 125 days).

4.3. Deterministic skill of the ensemble mean

Our aim in the first stage is to assess to what extent the deterministic skill of some single-valued forecast extracted from the ensembles is improved through recalibration. We concentrate on the ensemble mean since this is the most common single-valued forecast extracted from a set of ensemble forecasts. The ensemble mean is expected to minimise a quadratic loss function because it is an estimate of the conditional expectation of the stochastic process. Following the argument of Gneiting (2011), the criterion of choice for the evaluation of these deterministic forecasts is the bivariate Root Mean Square Error (bRMSE) for lead time k, calculated as

$$\text{bRMSE}_k = \left(\frac{1}{N}\frac{1}{S}\sum_{t=1}^{N}\sum_{t=1}^{S} ||\overline{\mathbf{y}}_{s,t+k|t} - \mathbf{y}_{s,t+k}||^2\right)^{\frac{1}{2}}, \quad (29)$$

where $\overline{\mathbf{y}}_{s,t+k|t}$ is the mean of the ensemble forecasts issued at time t for time $t+k$, and at the grid node s.

The bRMSE criterion is evaluated as a function of the lead time for the original and calibrated ensemble forecasts, for the eight quarters of the evaluation period (DJF2008, MAM2008, ..., SON2009). The bRMSE of the ensemble forecasts after bivariate calibration is also compared to that of ensemble forecasts calibrated in a univariate fashion, i.e. for the u and v components, independently. The improvement in bRMSE is quantified as the percentage of the bRMSE of the original ensemble forecasts. The results obtained are gathered in Figure 3. Note that similar curves would be obtained if skill scores were considered instead, i.e. by normalising the bRMSE of the ensemble mean by that of a benchmark like climatology. This is because the normalisation of the scores for both sets of forecasts would cancel out.

Improvements in the bRMSE criterion are positive for all lead times considered. Even though the level of improvement

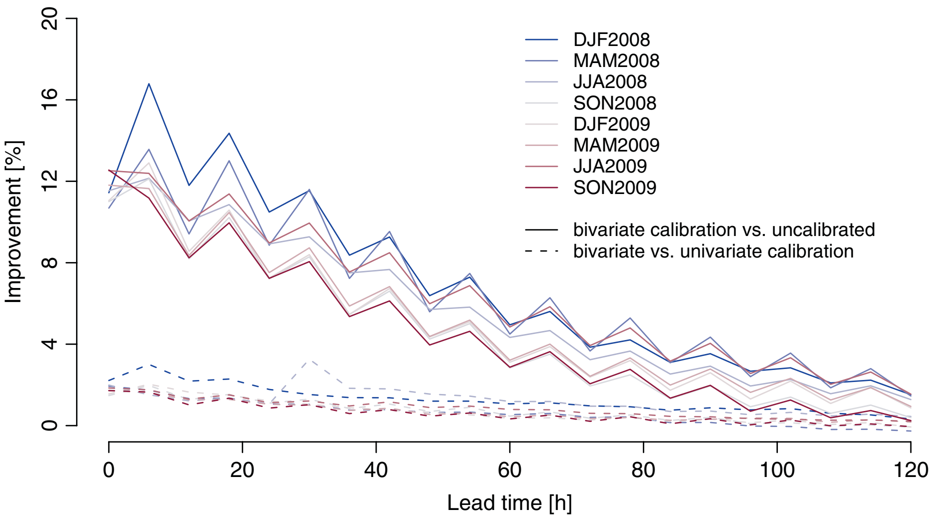

Figure 3. Improvements in the bRMSE score as a function of the lead time, computed for various quarters over 2008 and 2009. Positive improvements are for a higher skill of the ensemble mean extracted from the calibrated ensemble forecasts. Comparison is made between no calibration and bivariate calibration, and between univariate and bivariate calibration. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

varies among the various quarters, it appears to be consistent and fairly independent of the season. For the overall bivariate calibration methodology, the trend is that this improvement diminishes with the lead time, being above 5% up to the 3-day lead time, then slowly fading out as the lead time increases. It is substantial for the short to early-medium range, even reaching 10–15% for the first day. The main contributor to this improvement is certainly the translation of the ensemble forecasts, since this allows correction for the (bivariate) bias of the ensemble mean. In parallel, the benefits of jointly calibrating the u and v components can be seen from the positive improvement when going from univariate to bivariate calibration. Even though of lower magnitude (1% on average over the period and over all lead times), it clearly contributes to enhancing the skill of the ensemble mean.

4.4. *Reliability and skill of ensemble forecasts of* (u, v)*-wind*

The above improvements in the bRMSE of the ensemble mean are certainly non-negligible, though they do not reveal to what extent the calibration procedure allows for an improvement of the stochastic characteristics of the ensemble forecasts. For that purpose, it is necessary to look at improvements in the *ES* instead, since this is a proper score that evaluate the full (bivariate) predictive densities sampled by the ensembles. The improvements are evaluated in a similar manner to the bRMSE; they are a function of the lead time for the various quarters, while they give the percentage improvement of the *ES* when going
(i) from uncalibrated to bivariate calibration, and
(ii) from univariate calibration of the u and v components to bivariate calibration.
The results are depicted in Figure 4.

The observed pattern is similar to the case of the bRMSE score for the ensemble mean. For the overall calibration methodology the *ES* improvement diminishes with the lead time, while being substantial for the first 2–3 days (between 5 and 25%) and then fading out for further lead times. The contribution of going from univariate to bivariate calibration is also fairly limited (1% on average over the period and over all lead times), though always positive. In parallel, the improvement in the probabilistic score does not seem to be dependent upon the season. This may well be an advantage of the adaptive and recursive estimation of the model parameters. Remember that the optimal forgetting factor as determined through the cross-validation exercise corresponds to an equivalent window of 125 days ($\sim$ 4 months). It therefore allows for a smooth tracking of potential changes in the need for translation and dilation of the ensemble forecasts.

This improvement in skill obviously originates from the complete calibration of the ensemble forecasts, through translation and dilation. Consequently we aim at verifying the bivariate reliability of the ensemble forecast of (u, v)-wind before and after calibration. Bivariate reliability of these ensemble forecasts can be assessed thanks to multivariate rank histograms such as those described and discussed by Gneiting *et al.* (2008). They consist of a simple multivariate generalisation of the common rank histogram, even though their determination may appear slightly technical. We will not get into detail on how to produce these bivariate rank histograms since they are thoroughly described by Gneiting *et al.* (2008) as being as easy to read and interpret as their univariate counterparts. As an example of our evaluation of the reliability of ensemble forecasts of (u, v)-wind before and after calibration, we show in Figure 5 rank histograms for 24 h and 48 h forecasts for the JJA2008 quarter. This period is chosen arbitrarily since the results for other quarters have been found to be qualitatively and quantitatively similar. These two lead times are chosen since obviously the reliability improvements through calibration are more evident for shorter lead times when considering this type of ensemble forecast.

From a qualitative point of view for all lead times and quarters considered, the under-dispersiveness of the ensemble forecasts of (u, v)-wind appears to be corrected thanks to the calibration method proposed. The bivariate rank histograms after calibration look fairly flat, even though there seem to be too many events observed outside the coverage of the ensemble forecasts (corresponding to the last bin on the right). This may be explained by the fact that the distribution of forecast errors from the ensemble mean is not perfectly bivariate Gaussian, with a tail being slightly

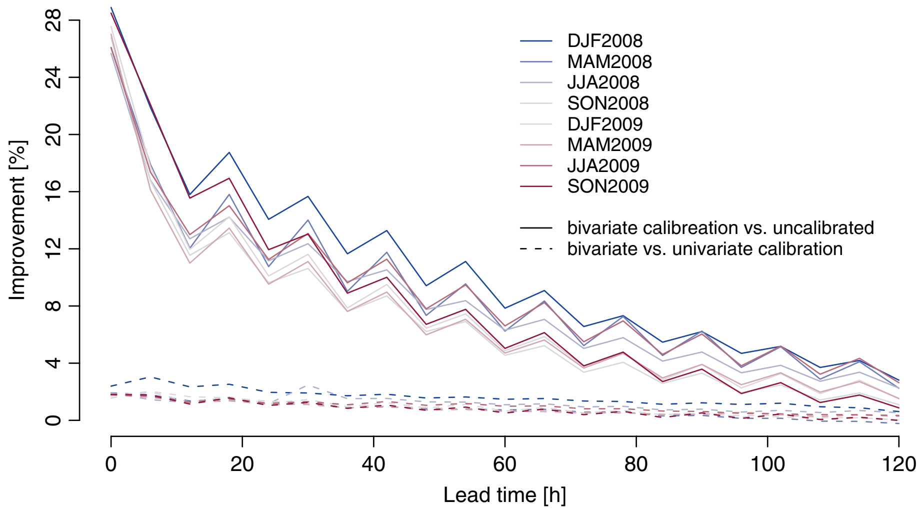

Figure 4. Improvements in the Energy Score as a function of the lead time, computed for various quarters over 2008 and 2009. Positive improvements indicate a higher skill of the calibrated ensemble forecasts. Comparison is made between no calibration and bivariate calibration, and between univariate and bivariate calibration. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

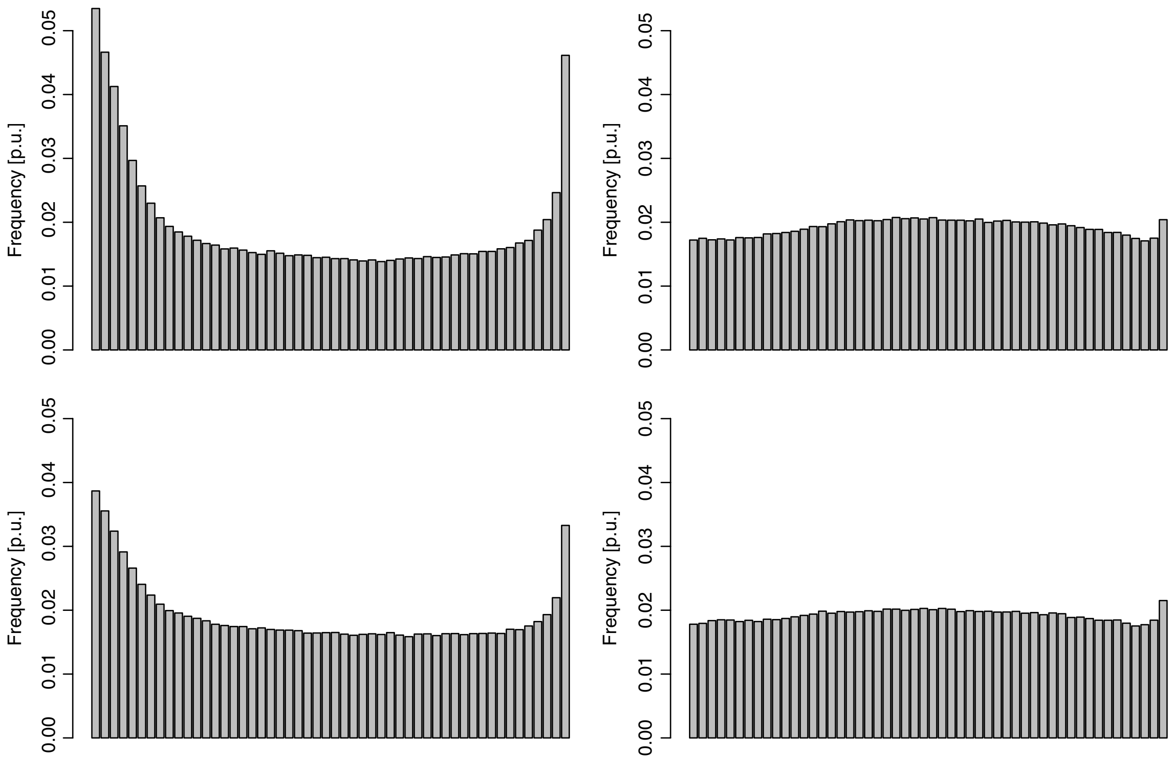

Figure 5. Bivariate rank histograms for the evaluation of the reliability of ensemble forecasts (a) before and (b) after calibration for a lead time of 24 h. (c, d) are as (a, b), but for 48 h lead time.

fatter than in the bivariate Gaussian distribution due to high winds originating from the south to west directions.

4.5. Example calibration results and maps of calibration factors

After evaluating the comparative skill and reliability of the ensemble forecasts of (u, v)-wind before and after calibration, we illustrate here some example results from our proposal calibration method. We place emphasis on a set of ensemble forecasts issued for a specific location, which is the Horns Rev wind farm in Denmark (55.51°N, 7.87°E), on 8 February 2009 at 0000 UTC. One of the reasons for carrying out a bivariate calibration of the (u, v)-wind is that it is then straightforward to derive wind speed and direction ensemble forecasts.

Consider as an illustrative example the ensemble forecasts of wind speed, before and after calibration, as presented in Figure 6. The same range of wind speed values is used for both plots, i.e. from 0 to 18 m s^{-1}, in order to ease comparison. This example illustrates the way the set of ensemble members is translated and dilated. The reduction in the bRMSE criterion discussed above also implies that on average the ensemble mean of wind speed has a lower RMSE. These evaluation results are not shown here in order to be concise. The ensemble mean appears to match more closely the observations for the calibrated

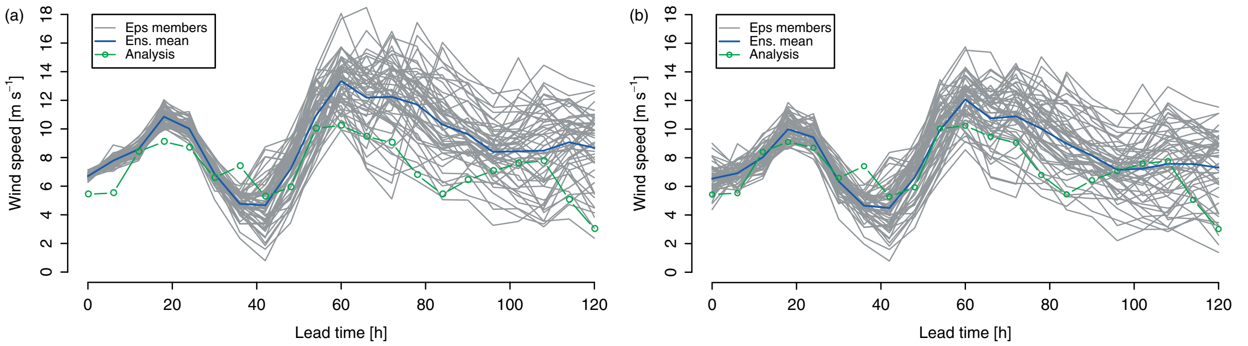

Figure 6. Ensemble forecasts of wind speed (a) before and (b) after calibration. These forecasts are extracted from the archive of (u, v)-wind forecasts, at the location of the Horns Rev wind farm in Denmark. They were issued on 8 February 2009 at 0000 UTC. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

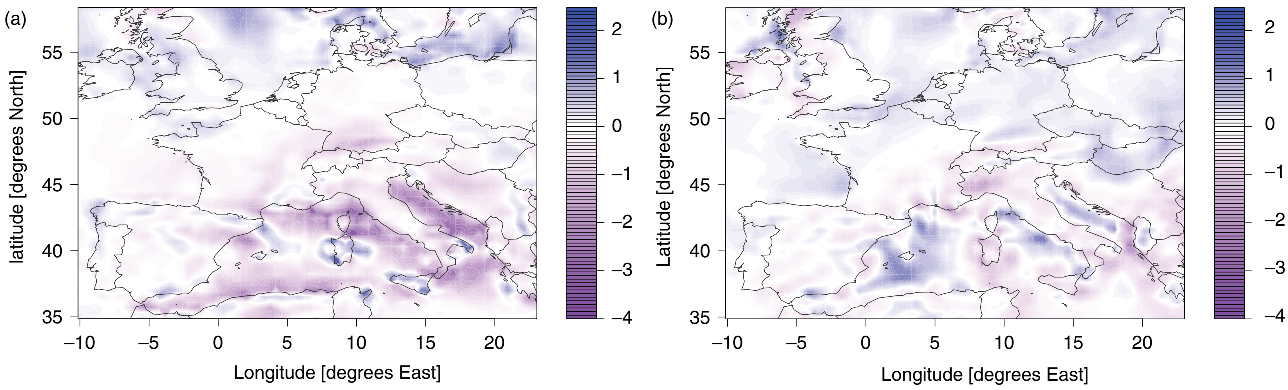

Figure 7. Maps of translation factors over Europe on 31 December 2009 and for 48 h forecasts for for (a) u and (b) v components. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

ensembles especially for the early range. Remember that the ensemble mean of (u, v)-wind is linearly shifted with respect to u and v. In terms of the corresponding wind speed, this does not lead to a linear bias correction of the ensemble members, owing to the nonlinear relationship between (u, v) and corresponding wind speed. The same goes for the dilation of the set of ensemble members which, despite the simplicity of the correction applied in u and v, implies a nonlinear correction of the dispersion of the ensemble forecasts of wind speed. Similar comments could be formulated for the case of wind direction. In Figure 6 the dispersion of ensembles is increased through calibration for short lead times, say until around 40 h ahead, and then reduced or kept at a similar level for longer ones. This figure nicely illustrates the importance of performing calibration based on different models and parameters for every lead time individually, since the deficiencies of the NWP-based ensemble forecasts are known to significantly evolve with the forecast horizon. Interestingly and maybe counter-intuitively for most practitioners, these deficiencies are stronger for short lead times while vanishing for lead times further than 3–4 days ahead.

We now look at some maps of translation and dilation factors obtained from the methodology. The values of these factors are a function of u and v themselves, and the lead time, in addition to varying in time and in space. Their maps are consequently fairly dynamic and those presented here should be seen as illustrative examples only. The maps shown in Figure 7 are for the translation factors along the zonal and meridional components, obtained at the end of the dataset on 31 December 2009 at 0000 UTC and for the 48 h lead time. In parallel, Figure 8 is for the related dilation factors (also for zonal and meridional components) at the same date and for the same lead time. The variations in all these factors are smooth in space, even though the magnitude of these variations is substantial. This smoothness lets us believe that, instead of estimating model parameters for each grid point individually, one may consider in the future proposing and estimating a spatial model of these parameters. This would have the consequence of clearly decreasing the costs of their estimation. As a reference for the empirical study in this article, it took around 12 min (on a recent laptop) to update the model parameters at all grid points (4560) and for all lead times (25) every time a new set of ensemble forecasts was considered.

We mentioned that the magnitude of the variations of translation and dilation factors was substantial. In terms of translation on that day and for that lead time, the u-component of ensemble forecasts was mainly increased over the north of the European domain, while being substantially decreased for near-coastal areas around the Mediterranean Sea. In contrast, the correction of the v-component is more mixed, with weaker corrections overall, while more (positive or negative) corrections were carried out over Ireland and Scotland, over the Mediterranean Sea and Northeast Europe. The maps of dilation factors for the u and v components appear more alike with a slightly higher average level for the meridional component. The values of dilation factors vary between 0.8 and 6.25, though most of their values are below 2.5. The range of values for these maps have hence been constrained to $[0.8, 3]$ to better highlight contrast in the most usual range of variation for the dilation factors. The darkest areas are for locations where dilation factors are equal to or above 3. It is no surprise that for this lead time the

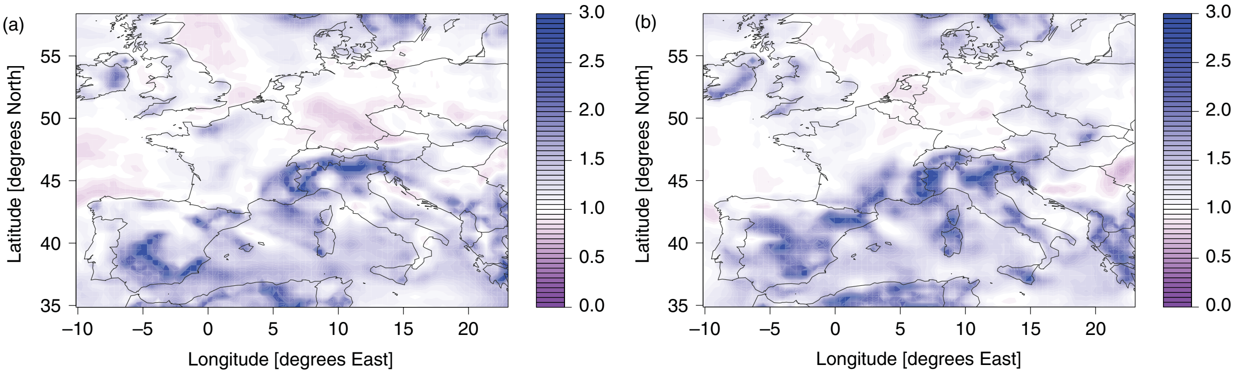

Figure 8. Maps of dilation factors over Europe on 31 December 2009 and for 48 h forecasts for (a) u and (b) v components. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

dispersion of ensemble forecasts is mainly increased over the whole domain. There are, however, a few areas over which the dispersion is actually decreased–they include Germany, the Netherlands, as well as parts of the North Sea and of the Atlantic Ocean.

5. Conclusions and discussion

The original approach described for the calibration of ensemble forecasts of (u, v)-wind relies on an adaptive and recursive estimation of the parameters of mean and variance models in a ML framework. It has the advantage of having solid theoretical foundations while being computationally cheap. This is because model parameters are updated based on the last forecasts and analysis only, each time this analysis is made available. It contrasts with the idea of having sliding windows for estimation, for which a complete batch estimation would need to be performed at every time step. The advantage here is that this calibration smoothly accommodates changes in the ensemble forecast reliability characteristics, e.g. due to seasons and upgrades in the operational forecasting system. Here the forgetting factor controlling the speed of adaptation was picked from a cross-validation exercise. In the future, we may consider dynamic forgetting factors in order to better adapt to successive phases when quasi-steady state of the dynamics of the joint forecast-verification process are followed by abrupt changes in such dynamics (e.g. originating from operational upgrades in the forecasting system). Examples of dynamic approaches to the definition of forgetting factors include Leung and So (2005) and Paleologu *et al.* (2008) among others.

Our calibration approach relies on a translation and dilation of the sets of ensemble forecasts of (u, v)-wind, in turn based on linear models allowing improvement of the stochastic characteristics of the ensemble generating process. It was shown that this approach led to substantial improvements of deterministic scores for the ensemble mean and of probabilistic scores for the ensembles themselves. All scores and diagnostics considered were defined within a bivariate framework, i.e. based on the bivariate RMSE of the ensemble mean, Energy Score and bivariate rank histograms for the ensemble forecasts. This improvement of probabilistic scores originates from the correction for the lack of sufficient reliability of the original ensemble forecasts of (u, v)-wind. As in Sloughter (2009), it has previously been considered that original data should be transformed to a bivariate Gaussian framework, with linear models for the mean and variance. The use of such a transformation was not deemed necessary in the present study based on ECMWF data. Consideration of more advanced models for the calibration of ensemble forecasts of (u, v)-wind may be the topic of further research. From the empirical work performed, we do not expect that substantial further improvements in forecast skill and reliability of the ensemble forecasts would be obtained, while computational costs would increase rapidly. An important observation from the empirical work and the maps of translation and dilation factors is that model parameters may certainly be represented by a spatial model, since they exhibit smooth variations in space. The simplification resulting from using a spatial model would also contribute to lowering computational costs. Note also that, since the calibrated ensemble forecasts are kept as (space–time) trajectories, it would be of particular interest in the future to evaluate them as trajectories instead of calculating scores and employing a diagnostic approach for each grid point and lead time independently. This may be revealing if the space–time structure of the original ensemble forecasts is affected or improved.

From a more general point of view, we explained that a prime assumption of our work relates to the role of a meteorological forecast provider to calibrate ensemble forecasts with respect to its own target. One should understand, however, that for forecast users the actual target may be different and may depend upon the intended application e.g. in the case of local observations of wind speed and direction at a wind farm. Hence, even if ensembles are calibrated with respect to their own target, further calibration may be necessary before ensemble/probabilistic forecasts are to be used in decision making.

Acknowledgements

The work presented has been partly supported by the European Commission under the SafeWind project (ENK7-CT2008-213740), which is hereby acknowledged. The authors are grateful to Roberto Buizza, Renate Hagedorn, Martin Leutbecher, Florian Pappenberger and Mark Rodwell at ECMWF for general discussion on calibration. Two anonymous reviewers are also to be acknowledged for their various comments and suggestions.

References

Bao L, Gneiting T, Grimit EP, Guttorp P, Raftery AE. 2010. Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Mon. Weather Rev.* **138**: 1811–1821.

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. *Mon. Weather Rev.* **133**: 1076–1097.

Buizza R, Miller M, Palmer TN. 1999. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **131**: 2887–2908.

Buizza R, Richardson DS, Palmer TN. 2003. Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man's ensembles. *Q. J. R. Meteorol. Soc.* **129**: 1269–1288.

Gneiting T. 2011. Quantiles as optimal point predictors. *Internat. J. Forecasting* **27**: 197–207.

Gneiting T, Raftery AE. 2005. Weather forecasting with ensemble methods. *Science* **310**: 248–249.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102**: 359–378.

Gneiting T, Raftery AE, Westwald AH III, Goldman T. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**: 1098–1118.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *J. R. Statist. Soc. series B* **69**: 243–268.

Gneiting T, Stanberry L, Grimit EP, Held L, Johnson NA. 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* **17**: 211–235.

Hering AS, Genton MG. 2009. Powering up with space-time wind forecasting. *J. Amer. Statist. Assoc.* **105**: 96–104.

Leung S-H, So CF. 2005. Gradient-based variable forgetting factor RLS algorithm in time-varying environment. *IEEE Trans. Signal Processing* **53**: 3141–3150.

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. *J. Comput. Phys.* **227**: 3515–3539.

Madsen H. 2007. *Time Series Analysis*. Chapman and Hall/CRC: London.

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

Matos MA, Bessa R. 2010. Setting the operating reserve using probabilistic wind power forecasts. *IEEE Trans. Power Systems* **26**: 594–603.

Meibom P, Barth R, Hasche B, Brand H, Weber C, O'Malley M. 2011. Stochastic optimization model to study the operational impacts of high wind penetrations in Ireland. *IEEE Trans. Power Systems* **26**: 1367–1379.

Morales JM, Conejo AJ, Pérez J. 2010. Short-term trading for a wind power producer. *IEEE Trans. Power Systems* **25**: 554–564.

Nielsen HA, Nielsen TS, Madsen H, Giebel G, Badger J, Landberg L, Sattler K, Voulund L, Tøfting J. 2006. From wind ensembles to probabilistic information about future wind power production – results from an actual application. In *Proceedings of 2006 IEEE PMAPS Conference; Probabilistic Methods Applied to Power Systems*, Stockholm, Sweden.

Paleologu C, Benesty J, Ciochina S. 2008. A robust variable forgetting factor Recursive Least-Squares algorithm for system identification. *IEEE Signal Processing Lett.* **15**: 597–600.

Palmer TN. 2000. Predicting uncertainty in forecasts of weather and climate. *Reports Prog. Phys.* **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. *Ann. Rev. Earth Planetary Sci.* **33**: 163–193.

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

Pinson P, Madsen H. 2009. Ensemble-based probabilistic forecasting of wind power at Horns Rev. *Wind Energy* **12**: 137–155.

Roulston M, Smith LA. 2002. Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.* **130**: 1653–1660.

Sloughter JM, Gneiting T, Raftery AE. 2010. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. American Statist. Assoc.* **105**: 25–35.

Sloughter JM. 2009. 'Probabilistic weather forecasting using Bayesian model averaging'. PhD final presentation, Dept. of Statistics, University of Washington, Seattle, USA.

Stone M. 1974. Cross-validation and assessment of statistical predictions (with discussion). *J. R. Statist. Soc. Series B* **36**: 111–147.

Taylor JW, McSharry P, Buizza R. 2009. Wind power density forecasting using ensemble predictions and time-series models. *IEEE Trans. Energy Conv.* **24**: 775–782.

Thorarinsdottir TL, Gneiting T. 2010. Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *J. R. Statist. Soc. Series A* **173**: 371–388.

Vannitsem S, Hagedorn R. 2011. Ensemble forecast post-processing over Belgium: comparison of deterministic-like and ensemble regression approaches. *Meteorol. Appl.* **18**: 94–104.

Wilks DS. 2002. Smoothing forecast ensembles with fitted probability distributions. *Q. J. R. Meteorol. Soc.* **128**: 2821–2836.

