METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **18**: 94–104 (2011)
Published online 28 July 2010 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.217

Ensemble forecast post-processing over Belgium: comparison of deterministic-like and ensemble regression methods

Stéphane Vannitsem[a]* and Renate Hagedorn[b]
[a] *Institut Royal Météorologique de Belgique, Brussels, Belgium*
[b] *European Centre for Medium-Range Weather Forecasts, Reading, UK*

ABSTRACT: A comparison of the benefits of post-processing ECMWF ensemble forecasts based on a deterministic-like and a regression technique is performed for Belgium. The former is a Linear Model Output Statistics technique (EVMOS) recently developed to allow provision of an appropriate ensemble variability at all lead times, and the latter is the Non-homogeneous Gaussian Regression, NGR. The training of the post-processing techniques is based on the reforecast dataset of ECMWF which covers a period from 1991 to 2007. The EVMOS approach is mainly providing a correction of the systematic error and does not enhance substantially the variance of the ensemble. The application of the NGR method provides an ensemble which encompasses the observations, unlike the EVMOS scheme. However, by taking into account the observational error, the analysis suggests that the ensemble based on the EVMOS post-processing scheme is also found to be consistent. This apparent contradiction is clarified and it turns out that both schemes are valuable depending on the specific purpose, the evaluation of the uncertainty of large scale flows or the downscaling of the temperature uncertainty at the level of the local observations. Copyright © 2010 Royal Meteorological Society

KEY WORDS model output statistics; non-homogeneous Gaussian regression; predictability; probabilistic forecasts; ECMWF; ERA-interim

Received 10 November 2009; Revised 9 April 2010; Accepted 26 May 2010

1. Introduction

Ensemble forecasts have been an integral part of the forecasting suite for almost 20 years at the European Centre for Medium-Range Weather Forecasts, ECMWF, and at the National Center for Environmental Prediction, NCEP (Molteni *et al.*, 1996; Toth and Kalnay, 1997). Today about 10 global ensemble prediction systems provide probabilistic forecasts on an operational basis (Bougeault *et al.*, 2009). Together with the prediction of the potential evolution of the current weather situation, they provide information on the uncertainty of a forecast issued at a specific date. Even if many modifications of the design of ensemble forecasts have allowed for an improvement of the probabilistic skill of the ensemble, some robust deficiencies are still present, amongst others the underdispersive character of the ensemble and/or the presence of systematic errors (Buizza *et al.*, 2005; Doblas-Reyes *et al.*, 2005; Bechtold *et al.*, 2009). In this context it is not easy to disentangle the respective roles of the sources of errors, which can be either from a modelling origin or from the mis-specification of the initial condition ensemble, or both.

Beyond the natural tendency in improving these forecasts by modifying the physics of the model or the way the initial ensemble is built, there is a growing interest in applying post-processing techniques to the ensembles. Two main approaches were followed. The first consists in performing a regression of a probability density on the whole ensemble or a kernel density smoothing of each ensemble member. The advantages and drawbacks of several variants of theses methods were thoroughly reviewed in Wilks (2006a), and it was found that one of the best approach is the Non-Homogeneous Gaussian Regression method (NGR) proposed by Gneiting *et al.* (2005). The second main approach is to apply the post-processing technique developed for single forecasts to the ensemble directly. This approach was at first sight discarded because the most popular post-processing technique introduced in the 1970s (see Glahn and Lowry, 1972) has the tendency to converge to the climatological mean for long lead times (e.g. Wilks, 2006b). This approach is based on a linear regression between forecasts and observations and is referred hereafter as the Linear Model Output Statistics (LMOS). Recently, a modification of this approach has been proposed which is strongly related to the ensemble kernel density smoothing (Glahn *et al.*, 2009; Unger *et al.*, 2009), which allows alleviation of this problem.

An alternative approach has also been recently proposed which consists in modifying the linear regression itself (Vannitsem, 2009). The idea behind this new approach is to incorporate in the fitting the natural

* Correspondence to: Stéphane Vannitsem, Institut Royal Météorologique de Belgique Avenue Circulaire, 3, 1180 Brussels, Belgium. E-mail: svn@oma.be

Copyright © 2010 Royal Meteorological Society

fact that the forecasts are displaying errors, an assumption not present in the LMOS approach. It has been shown that this simple consideration allows for keeping natural variability to the corrected forecast. This in turn allows its use for ensemble post-processing as has been demonstrated on a low order system (Vannitsem, 2009). This approach is hereafter referred as EVMOS (Error-in-Variables MOS).

In the present work, the latter approach is compared to the NGR method, which is believed to be one of the most performant ensemble regression techniques (Wilks, 2006a) in an operational environment. Since the decision whether a post-processing method should be implemented operationally at a national meteorological service depends on the local results for their specific area, here the respective advantages and drawbacks of both post-processing methods for Belgium, as a specific example, are analyzed. To this aim, the 2 m temperature reforecasts now operationally produced and archived at ECMWF for the calibration of the operational Ensemble Prediction System (Hagedorn, 2008; Hagedorn *et al.*, 2008) is used to train the EVMOS equations and the NGR calibration. The resulting coefficients are then applied to the operational 2 m temperature ensemble. Two sets of data will be used as reference, (1) eight synoptic stations located in Belgium, and, (2) the Eta-interim re-analyses (Simmons *et al.*, 2007).

In Section 2, the post-processing approaches are described, while the data sets are presented in Section 3. The properties of the corrected ensemble forecasts are then discussed in Section 4. The conclusions are drawn in Section 5.

2. The post-processing schemes

2.1. The EVMOS technique

One of the central difficulties in using the Linear Model Output Statistics (LMOS) technique for ensemble correction in its classical formulation (Glahn and Lowry, 1972) is related to the fact that the LMOS correction converges toward the climatological mean for long lead times. In the present section the new scheme proposed in Vannitsem (2009), in which this constraint is relaxed, is described briefly.

The Linear approach consists in building a linear combination of a set of model observables (or predictors), $V_i(t)$, $i = 1, \ldots, n$, at a certain lead time t of the forecast:

$$X_C(t) = \alpha(t) + \sum_{i=1}^{n} \beta_i(t) V_i(t) \tag{1}$$

in order to provide a corrected forecast, $X_C(t)$ (or predictand), for a specific reference variable $X(t)$. The parameters $\alpha(t)$ and $\beta_i(t)$ are estimated using a set of K past forecasts by minimizing a cost function, $J(t)$, measuring the distance between the observations and the corrected forecast. In the classical approach, the basic assumption is the fact that the input observables are supposed to be error free, but not the observations. The cost function displays, therefore, a very simple form:

$$\sum_{k=1}^{K} \left[\{\alpha(t) + \sum_{i=1}^{n} \beta_i(t) V_{i,k}(t)\} - X_k(t) \right]^2 \tag{2}$$

When this basic error-free assumption is relaxed, the cost-function for one model observable (EVMOS1) reads:

$$J(t) = \sum_{k=1}^{K} \frac{[\{\alpha(t) + \beta(t) V_k(t)\} - X_k(t)]^2}{\sigma_X^2(t) + \beta^2(t)\sigma_V^2(t)} \tag{3}$$

where $\sigma_X^2(t)$ and $\sigma_V^2(t)$ are the variance of X and V, at lead time t. After minimization, one arrives at:

$$\alpha(t) = < X(t) > - \beta(t) < V(t) > \tag{4}$$

$$\beta(t) = \sqrt{\frac{\sigma_X^2(t)}{\sigma_V^2(t)}} \tag{5}$$

For two model observables (EVMOS2), relation (3) can be easily modified by inserting in Equation (3) a linear relation of the form $V_{1,k} + \delta V_{2,k}$ in place of V_k. The minimization of this modified relation leads then to the estimation of the three parameters (see Vannitsem, 2009).

In the classical LMOS approach, one central feature is the progressive decrease of the variance of the corrected forecasts, implying that for long lead times the corrected forecasts are converging towards the climatological mean. This feature could be useful when using single forecasts because the climatological mean is better than any single forecasts when they become completely uncorrelated with the actual atmospheric situation (see Wilks, 2006b), but once the interest is placed on the evaluation of the uncertainty of the forecasts based on ensembles this feature is a drawback which is easily alleviated by using the EVMOS approach since the variance of the corrected forecasts is now equal to the variance of the observations.

2.2. The NGR method

This method is a regression technique aimed at providing probabilistic forecasts, related to the classical LMOS technique in the sense that a linear regression of the ensemble mean is performed. The regression residuals are then supposed to be normally distributed with a variance linearly related to the variance of the ensemble itself. In this case the corrected ensemble forecast is given by:

$$X_{C,j}(t) = a(t) + b(t)\overline{V}(t) + \varepsilon_j(t) \tag{6}$$

with,

$$\varepsilon_j(t) \in N\left\{0, c(t) + d(t)\sigma_{\text{ens}}^2(t)\right\} \tag{7}$$

Copyright © 2010 Royal Meteorological Society

where $\overline{V}(t)$ and $\sigma_{\text{ens}}(t)$ are the ensemble mean and spread of the ensemble at lead time t, respectively. The parameters a, b, c and d should be estimated through the minimization of a cost function. The one proposed by Gneiting *et al.* (2005), also used subsequently in several applications, is the average Continuous Ranked Probability Score defined as:

$$\overline{CRPS(F,q)} = \frac{1}{K}\sum_{i=1}^{K}\int_{-\infty}^{\infty}\{F_i(s) - H(s-q_i)\}^2\,\mathrm{d}s \qquad (8)$$

where F_i denotes the probability distribution issued for the ith observation, q_i, and $\mathrm{H}(s-q_i)$ is the Heaviside function. For a Gaussian distribution of the form (7), this cost function reduces to:

$$\overline{CRPS} = \frac{1}{K}\sum_{i=1}^{K}\sqrt{c(t)+d(t)\sigma^2_{\text{ens},i}(t)} \cdot \left[Z_i\left\{2\Phi(Z_i)-1\right\} + 2\phi(Z_i) - \pi^{-1/2}\right] \qquad (9)$$

where,

$$Z_i = \frac{q_i - \left\{a(t) + b(t)\overline{V}_i(t)\right\}}{\sqrt{c(t)+d(t)\sigma^2_{\text{ens},i}(t)}} \qquad (10)$$

and ϕ and Φ are the probability density and probability distribution of the standard Gaussian distribution. The parameters are then minimized numerically through standard techniques such as the simplex method which can be found in Press *et al.* (1996).

3. The data

One of the central problems of most post-processing techniques is the need to have a sufficiently large set of data on which the statistical relations can be trained. This imposes that the model used for the forecast should be the same as the one used for training. This has led different centres to develop long archive data sets based on the integration of the model starting from past analyses. These are known as reforecasts (Hamill *et al.*, 2006; Hamill and Wilks, 2007; Hagedorn, 2008). In the present work the 2 m temperature reforecast produced by ECMWF is used. Each week, a set of ensemble forecasts starting from the same calendar date for 17 years from 1991 to 2007 is performed. The initial state of the control forecast is selected from the ECMWF reanalyses data set (Uppala *et al.*, 2005). The number of ensemble members has been fixed to four, in addition to the control forecast. It has been shown that this number is sufficient to get a quite robust estimate of the parameters of the NGR methods presented in Section 2 (Hagedorn, 2008).

Ensemble post-processing is applied on a set of synoptic stations located in Belgium. Eight of them (Uccle, Beauvechain, Elsenborn, Florennes, Deurne, Kleine Brogel, Koksijde and Middelkerke) have been selected, in view of their long records without many data gaps. Some of these are located in the central part of the country (Uccle, Florennes, Beauvechain, Deurne). Elsenborn and Kleine Brogel are located in the eastern part of Belgium which is a more undulating region than the central part, while Middelkerke and Koksijde are located on the coastal zone. The same stations have been used in Vannitsem (2008) to evaluate the nature of the corrections provided by the classical LMOS algorithm for single (or deterministic) 2 m temperature forecasts, in the spirit of the theoretical investigation of Vannitsem and Nicolis (2008). As it turned out, the dominant correction was mainly of a systematic nature, either from a systematic bias in the initial conditions or in the model formulation. The only exception is the coastal zone for which an important correction is obtained for the variable part of the model error source.

The same analysis is also performed using the most recent ECMWF reanalysis, ERA-interim (Simmons *et al.*, 2007) as verification dataset. The purpose of this is to evaluate the role of the observational error. The investigation is done at the same grid points as the ones providing the forecasts for the synoptic stations (five grid points).

The post-processing techniques presented above are trained for each season (December, January, February; March, April, May; June, July, August; September, October, November) separately, using the reforecasts generated during the corresponding periods. The regressions are then applied to the ensemble forecasts of the four seasons covering the period from March 2008 up to February 2009. For each season there are 221 ensemble forecasts (17 years × 13 weeks) on which the post-processing techniques are trained. All the results displayed in the sequel are averaged over all stations or grid points, in order to compensate for the shortness of the verification period.

4. Results

4.1. Source of errors

Before starting the presentation of the results obtained in the context of the ECMWF data, a discussion of the role of the different sources of errors affecting the forecasts, the post-processing and the scores, is needed. One can distinguish between two different sources of errors: the initial condition and model errors. The post-processing calibration technique of ensemble forecasts aims at modifying the original ensemble in order to take into account these different sources of errors, as discussed extensively in previous works (Wilks, 2006a; Vannitsem, 2009), but one important source of errors also affects the scores used to evaluate the quality of ensembles, the observational error, as discussed in recent papers (Saetra *et al.*, 2004; Bowler, 2006; Candille and Talagrand, 2008). This observational error originates from (1) the presence of instrumental errors and (2) from the so-called representativeness errors. Instrumental errors are, of course, related to the quality of the instrument and the environmental setting in which it is placed. For the hourly temperature at

Copyright © 2010 Royal Meteorological Society

2 m this error is quite small (of the order of a few tenths of degrees at synoptic stations). The second source of error is larger and is related to the fact that the measurement station evaluates the quantity of interest on a very small scale as compared with the scale of the forecasted observable. For the ECMWF model the forecasted variable is valid for a domain of the order of a few tens of kilometres, which is of course much larger than the size of the instrument used to measure the 2 m temperature, which can vary substantially over a few tens of kilometres. It is worth emphasizing that the representativeness error is not *per se* associated with a defect of the observational technique designed to measure the local properties of the atmosphere, but stems from a mismatch between the scale of the measurement and of the model observables. This representativeness error reflects, therefore, the under-sampling of the atmospheric variability.

To evaluate the potential impact of this observational error, it is necessary to focus on the mean square error between one forecast and the observations, denoted respectively as V and $O = X_T + \varepsilon$ where ε is the observational error, including both the measurement error and the representativeness error, and X_T is the truth. Assuming that this error is independent of the forecast and of the true value, the mean square error is then given by:

$$< (V - O)^2 > = < (V - X_T)^2 > + < \varepsilon^2 > - 2 < \varepsilon > < V - X_T > \tag{11}$$

If no systematic observational errors ($<\varepsilon> = 0$) are present or if $< V > = < X_T >$, it reduces simply to the sum of the mean square error as compared to the truth and the variance of the observational error. Note that in data assimilation the standard deviation of this observational error is assumed to be of the order of 1.5 °C (e.g. Parrish and Derber, 1992; Mahfouf *et al.*, 2009).

Turning more specifically to the impact of this observational error on the corrected forecast based on the EVMOS1 scheme, the first point to make is the role of this observational error on the estimation of the parameter α and β, obtained by minimizing Equation (3) in which X is replaced by $\mathrm{O} = X_T + \varepsilon$:

$$\alpha = < X_T > + < \varepsilon > - \beta < V > \tag{12}$$

$$\beta = \sqrt{\frac{{\sigma_{X_T}}^2 + \sigma_\varepsilon^2}{\sigma_V^2}} \tag{13}$$

If the observational error is small as compared to the natural variability of the variable considered, then β is not very much affected by the presence of the observational error. Parameter α is shifted by an amount equal to the systematic observational error. Both mean and variance of the corrected forecasts are of course affected by these errors. The mean will be shifted from the reality by an amount equal to $< \varepsilon >$ and the variance by an amount equal to the variance of this observational error. The mean square error of the corrected forecast using one model observable, Y, can be easily evaluated (Equation (1)):

$$X_C = \alpha + \beta Y \tag{14}$$

where Y here refers to any observable of the deterministic or ensemble forecasts (with the same units as the nominal variable). The MSE decomposition is then:

$$\begin{aligned} < (X_C - O)^2 > = & < (Y - X_T)^2 > - \overbrace{(< X_T > - < Y >)^2}^{DC} \\ & \underbrace{+ (\beta - 1)^2 < (Y - < Y >)^2 > - 2(\beta - 1)\{C(X_T, Y) - < (Y - < Y >)^2 >\}}_{-VC} + \sigma_\varepsilon^2 \end{aligned} \tag{15}$$

where $C(X_T, Y)$ refers to the covariance between X_T and Y. The second term on the right hand side is called the drift correction (DC) and the third and fourth terms are referred to as the variability correction (VC). The main difference with the estimation, assuming that there are no observational errors, lies in the presence of the term σ_ε^2, the variance of the observational error. If known, this quantity can be removed to get a reliable estimate of the mean square error of the correction as compared to reality. In the next section, relation (15) will be estimated for the ensemble mean.

Note that this observational error does not affect explicitly the evaluation of the variance of the raw ensemble itself or of the corrected ensemble by EVMOS, except through the amplitude of the parameters α and β. In order to take into account the presence of observational errors in the estimate of MSE some perturbative techniques are usually used to increase the variance of the ensemble (Hamill, 2001; Hagedorn *et al.*, 2008), the purpose being to clarify the reliability of the ensemble taking into account the presence of observational errors. This approach is not used here. Instead, the amplitude of the observational error is removed based on Equation (15) to evaluate the quality of the post-processing (e.g. Hagedorn *et al.*, 2008). For other variables, this operation cannot necessarily be performed due to the poor knowledge of the observational error. Note that in this case, the use of other datasets such as the reanalyses can help in clarifying the role of observational error as will be shown in Section 4.

For the NGR method, the estimation of the parameters of the Gaussian distribution is highly sensitive to all sources of errors, and in particular the observational errors. The main reason lies in the fact that the estimated variance (8) is directly related to the use of real observations in Equation (10). This will be confirmed in the next section.

4.2. Application to the ECMWF ensemble forecasts

Starting with the evaluation of the correction obtained with the MOS equations discussed in Section 2, Figure 1(a) displays the mean square error evolution for the corrected forecasts based on EVMOS1 with 2 m temperature, T2m, as predictor and EVMOS2 with two

Copyright © 2010 Royal Meteorological Society

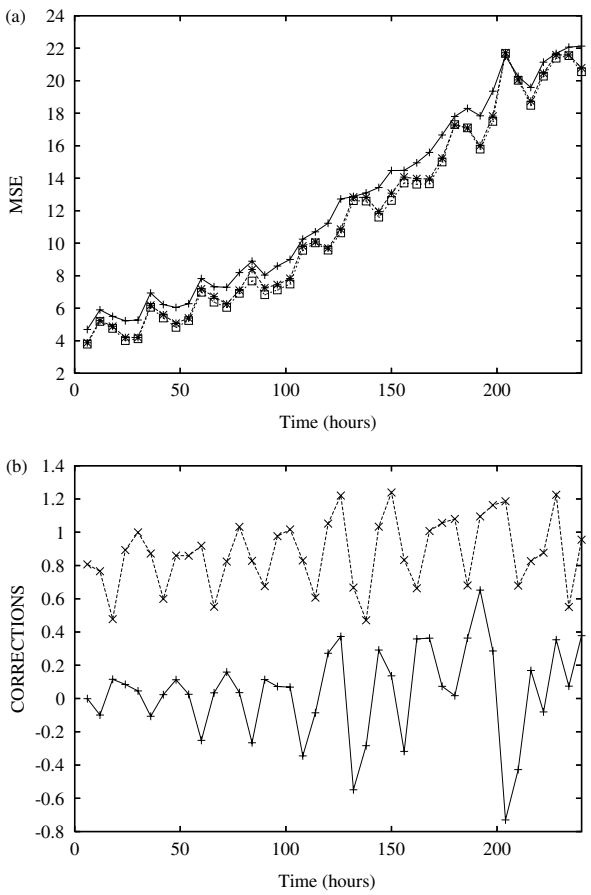

Figure 1. (a) Mean square error evolution for the raw forecasts (pluses), EVMOS1 (stars) and EVMOS2 (squares) during the training period covering 1991–2007. EVMOS2 is based on the temperature at 2 m and 850 hPa as predictors. (b) Variability (pluses) and drift (crosses) corrections provided by the EVMOS1 scheme during the same period.

observables T2m and T850 (temperature at 850 hPa), as obtained from the full reforecast training data set described in Section 3. The results are averaged over all synoptic stations and the four seasons.

A large correction is obtained with an amplitude oscillating between 0.5 and 1 °C depending on the lead time of the forecast. The largest improvement occurs at 0000 and 0600 UTC. The EVMOS2 scheme does only provide a slight improvement as compared to the EVMOS1 scheme during the training phase. Other second observables were also considered such as latent or sensible heat fluxes, but without success. As shown later, the improvement obtained with T850 disappears when considering the forecast quality of the ensemble mean computed during the verification phase. This suggests that the second observable is of little use in this specific forecast case.

The two corrections, the systematic drift (DC) and the variability (VC), are plotted in Figure 1(b) for EVMOS1, indicating that the nature of the correction is mainly of a systematic nature. This feature together with the fact that the parameter β is usually not very far from 1, indicates that a good variability is already produced by the model. This result is obviously very encouraging in terms of model quality for Belgium.

As already discussed in the previous section, the comparison of the forecasts is made with observations affected by observational errors. The variance of these errors is typically estimated to be of the order of $(1.4)^2$ or $(1.5)^2$ (see e.g. Mahfouf *et al.*, 2008; Hagedorn *et al.*, 2008). This implies that the mean square error evaluation of the corrected forecast should be shifted by about 2 °C^2 in order to get an estimate of the error as compared with the true evolution at the scale of the model. So the error of the corrected forecast oscillates between 2 and 3 °C^2 for short times. The correction obtained with the post-processing varies between 20 and 40% of the error variance for short lead times, as compared with the true atmospheric evolution.

One of the central objectives of ensemble forecasts is to provide an uncertainty information such that the observation can be considered as one state undistinguishable from the states generated by the ensemble forecasts. This feature implies that the mean square error of the ensemble mean (MSE) should be close to the variance of the ensemble (VAR), averaged over a set of initial dates. Figure 2 displays the MSE and the VAR for the different MOS schemes used here, averaged over the whole verification year (March 2008 to February 2009). Lead times at 0600, 1200, 1800 and 0000 UTC are displayed separately in order to emphasize the corrections obtained during the daily cycle. Whatever the lead time, the mean square error of the ensemble mean is smaller for the three MOS schemes as compared with the raw forecast. The correction amplitude is similar for the three schemes, indicating that they all provide similar corrections of the systematic error. It also reveals that the addition of the temperature at 850 hPa as second predictor does not provide much additional correction for the ensemble mean forecast, but seems to affect the variance of the ensemble slightly (as compared with EVMOS1).

In order to evaluate the robustness of these results in view of the small number of stations and the single verification year, a confidence interval has been estimated for each MSE and VAR curves through a bootstrap approach (Efron and Tibshirani, 1993). For each lead time, the MSE and VAR computed at the eight stations are first re-sampled through random drawings. This operation is performed for each season and then averaged to get new values for the whole year. It is then repeated 1000 times. A mean and a variance for the curves MSE and VAR can then be estimated. This bootstrap method shows that the variability around the MSE curves displayed in Figure 2 is quite small, with a standard deviation close to 0.15 for short times and 0.4 °C^2 for longer lead times, as illustrated in Figure 2(e) for EVMOS1 at 0000 UTC lead times. For the VAR curves, these standard deviations are very small at short lead times, close to 0.01, up to 0.3 °C^2 for long lead times. It indicates that the differences between the MSE and VAR curves of the corrected forecasts and the raw forecasts are quite robust, even with this small number of stations.

Copyright © 2010 Royal Meteorological Society

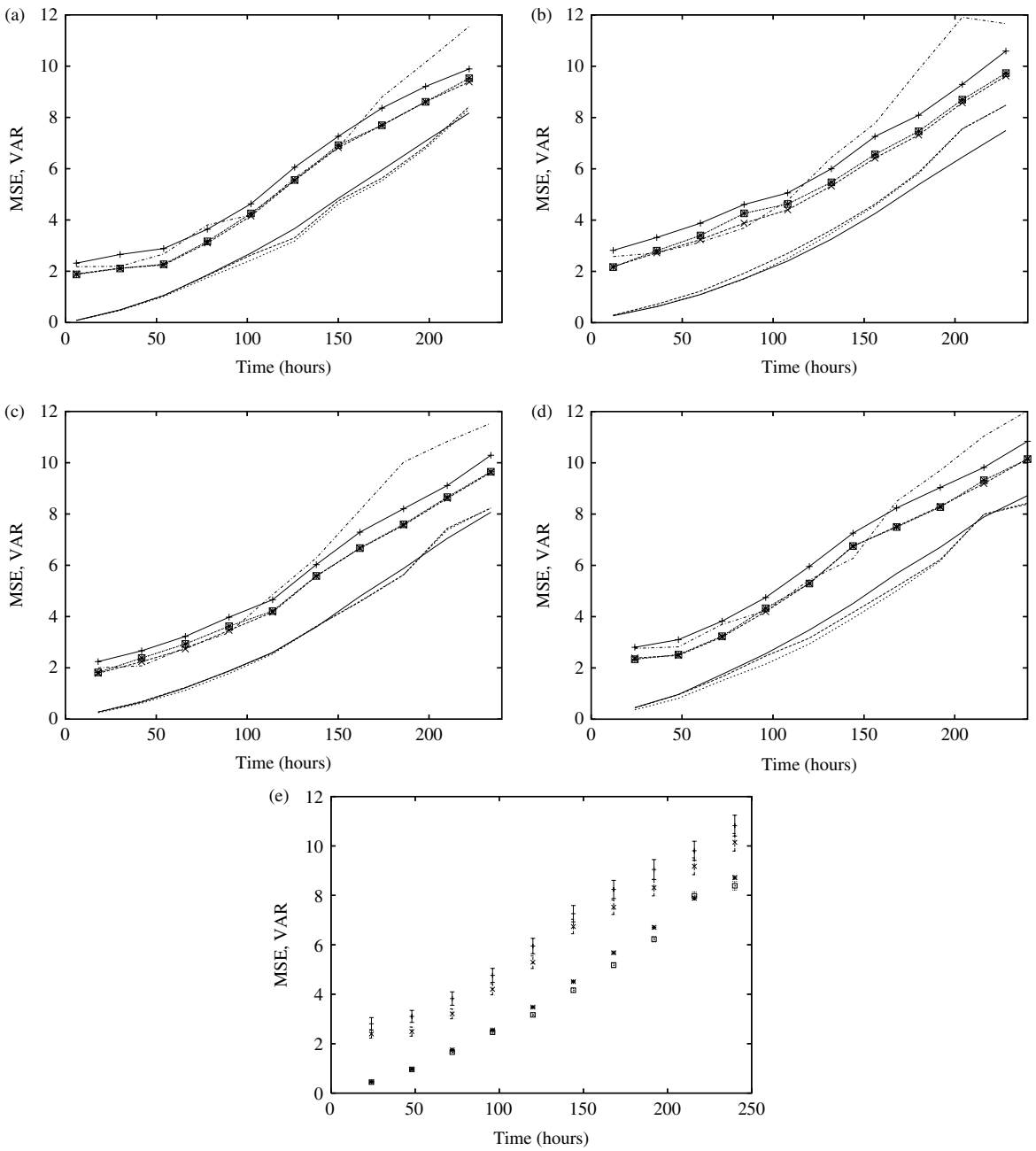

Figure 2. Evolution of the mean square error of the ensemble mean (MSE; pluses for the raw forecast, crosses for EVMOS1, stars for EVMOS2 and circles for NGR) and variance (VAR, curves without symbols; continuous line for the raw forecasts, dashed line for EVMOS1, dotted line for EVMOS2, dash-dotted line for NGR) for the raw and corrected ensemble forecasts of the period March 2008 to February 2009. The results are displayed separately for the four different forecast times of the day, (a) 0600, (b) 1200, (c) 1800, and (d) 0000 UTC. EVMOS2 is using the Temperature at 850 hPa as second predictor. (e) As in (d) for EVMOS1 and the raw forecasts only, with the error bars (one standard deviation on both sides of the averaged curves) obtained with the bootstrapping method (see text for details).

An additional important result is that for short lead times the EVMOS1 and EVMOS2 scheme do not substantially modify the variance of the ensemble, whereas the NGR method does this quite well. The NGR calibrated ensemble variance VAR is very close to the MSE, indicating that the observation can be considered as one potential member of the model ensemble. Note that the variance of the ensemble is smaller for EVMOS2, but in view of the small difference with EVMOS1 and our inability in evaluating precisely the observational error, one cannot infer at this stage what the best scheme is.

To understand this major difference between EVMOS and NGR one must come back to the discussion of Figure 1, which mentioned that the role of the observational errors is mainly of a representativeness nature. The EVMOS approach does not correct for short lead times the random part of error sources uncorrelated with the predictors (Vannitsem, 2009). The absence of correction of the observational error present when comparing the corrected large scale forecast mean and the local scale observation suggests that this error is not correlated with the predictors used. This in turn implies

Copyright © 2010 Royal Meteorological Society

this large difference between the VAR and MSE curves for EVMOS. However, if one uses relation (15), one immediately realizes that by removing the variance of the observational error ($2\,^{\circ}C^2$), one gets an MSE curve very close to the VAR curve, suggesting that indeed the ensemble is consistent, but at the scale described by the model. The NGR method, on the contrary, inflates the ensemble VAR during the minimization process in order to get a distribution close to the observational one. This VAR curve therefore becomes close to the MSE by construction.

Small differences are present between the various daily hours, in particular an increase of the variance for 1200 UTC using EVMOS (Figure 2(b)) and a decrease for 0000 UTC (Figure 2(d)). For NGR, the VAR curves are usually larger than the MSE for long lead times, suggesting a tendency of the ensemble based on NGR to be over dispersive.

Note that both techniques have been applied using the ensemble of reforecasts archived at the ECMWF (four members and a control), but an important aspect of EVMOS is that it can be trained using the control forecasts of the reforecast archives only. The corrections for the ensemble obtained in this case do not differ substantially from the ones based on the full archive, as revealed in Figure 3. In this figure, all lead times have been plotted for the winter season of the verification period. The EVMOS scheme trained using the four ensemble members + control is practically undistinguishable with the corrected forecast obtained by using the control only. This suggests that with this approach one can achieve similarly good results but with a less expensive reforecast data set.

Let us now focus on the rank histogram of the ensembles (see Wilks, 2006b). This diagram evaluates the fraction of occurrences of the observations within the ensemble. For a consistent ensemble, this histogram should be flat, otherwise it indicates some deficiencies such as an under dispersion when it has a U-shape with more occurrence of the observation outside the ensemble (see Wilks, 2006b for a detailed discussion). Figure 4 displays the rank histogram at four different lead times averaged over the whole year and the ensemble of stations considered. For short lead times, all the ensembles are under dispersive (U-shape of the histogram), except the one generated by the NGR method. This feature obviously reflects the small VAR of the ensembles of the DMO and EVMOS schemes. The NGR method is clearly much more useful than the EVMOS method (in its current form) when a reliable ensemble at the scale of the observation is needed for short lead times. For long lead times this conclusion is weakened because the rank histogram is now much flatter for the EVMOS method. These features reflect the results already summarized for the MSE and VAR, indicating that the observational error is taken into account by the NGR method.

In order to investigate the skill of specific extreme events, Figure 5 displays the Brier Skill Score (BSS) for predicting the occurrence of temperature less than $0\,^{\circ}C$

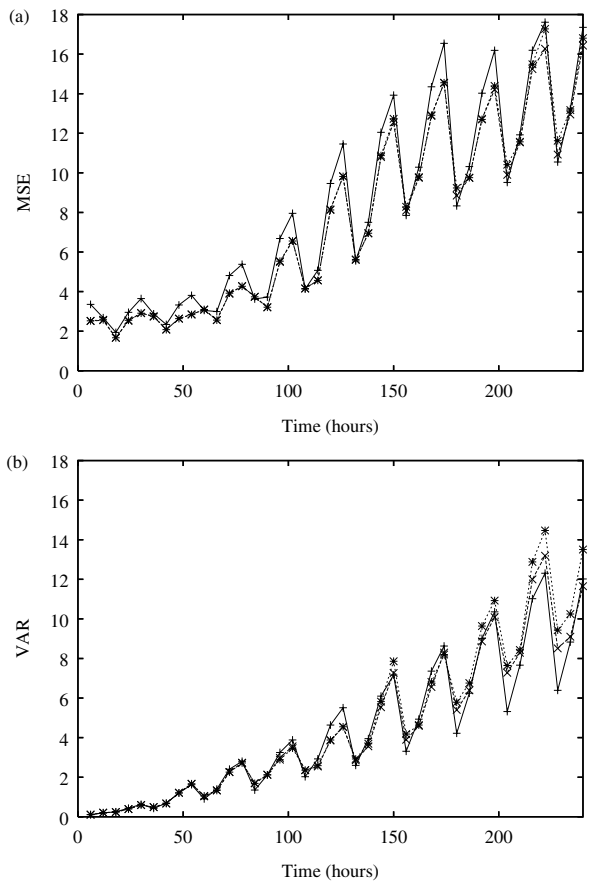

Figure 3. Evolution of the mean square error (MSE) of the ensemble mean (a) and variance (VAR) of the ensemble (b) for the raw (pluses) and EVMOS1 corrected ensemble forecasts, trained using the control forecast (crosses) or the four ensemble members + the control (stars). The period covered is winter 2008–2009.

during winter times for 0000 UTC as in Figure 2(d). Interestingly the two approaches EVMOS and NGR provide similar corrections with a large improvement of skill. For summer temperature larger than $20\,^{\circ}C$ at 1200 UTC (Figure 6), the results indicate a large improvement of skill by using one of the post-processing techniques, with at some moments a doubling of skill scores. The similarity of improvement in skill between the EVMOS and NGR approaches suggests that the observational error does not affect excessively the ensemble evaluation based on this score.

The impact of the observational error present at small scales and emphasized when discussing the results for synoptic stations is further examined by investigating the correction obtained by the various techniques when trained and verified against another reference. A suitable reference is provided by the ERA-interim reanalyses (Simmons *et al.*, 2007) covering the same period and the same region. The smaller scale at which these reanalyses are performed is twice as large as the scale of the ensemble forecast model, indicating that the scale discrepancy between the two systems (i.e. the representativeness error) is much smaller than in the comparison with the synoptic stations.

Copyright © 2010 Royal Meteorological Society

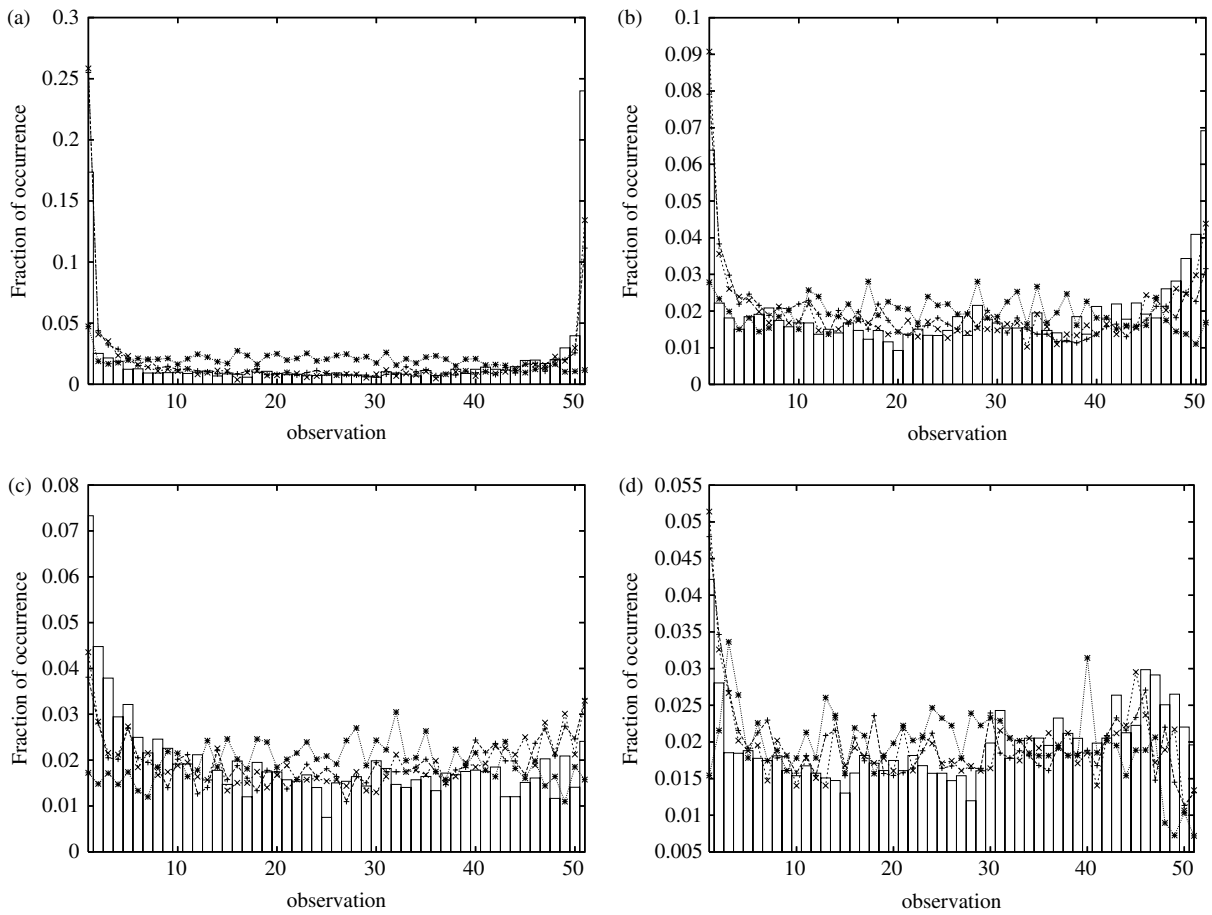

Figure 4. Rank histogram for the raw ensemble forecasts (boxes) and the corrected ensembles based on EVMOS1 (pluses), EVMOS2 with as second observable the temperature at 850 hPa (crosses) and the NGR method (stars), averaged over the whole verification year March 2008 to February 2009 and over all the stations, at 24 (a), 120 (b), 180 (c) and 240 (d) h lead time.

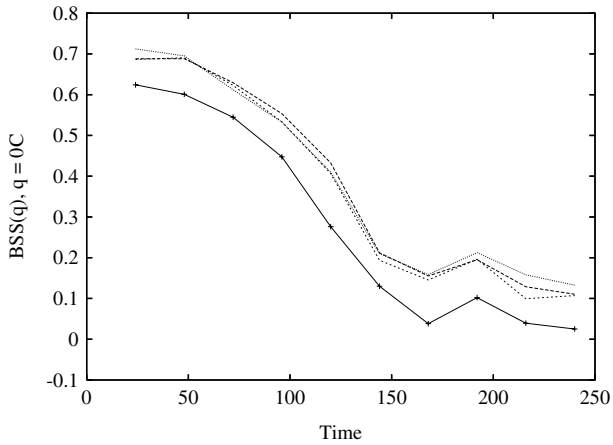

Figure 5. Brier Skill Scores for 2 m temperature probabilistic forecasts below 0 °C during Winter 2008–2009 for the raw ensemble (pluses) and the corrected ones based on EVMOS1 (long-dashed), EVMOS2 with as second observable the temperature at 850 hPa (dashed) and the NGR method (dotted), at 0000 UTC.

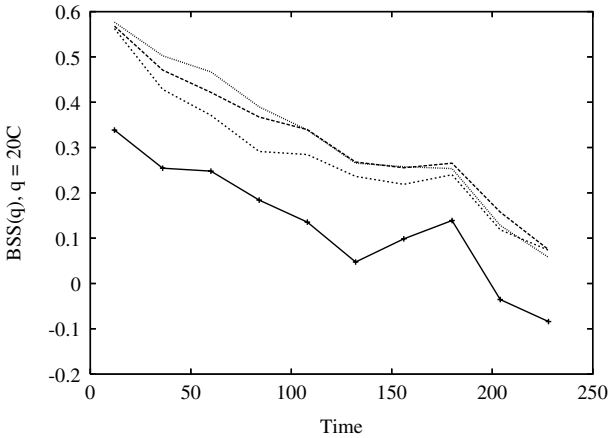

Figure 6. Brier Skill Scores for 2 m temperature probabilistic forecasts larger than 20 °C during Summer 2008 for the raw ensemble (pluses) and the corrected ones based on EVMOS1 (long-dashed), EVMOS2 with as second observable the temperature at 850 hPa (dashed) and the NGR method (dotted), at 1200 UTC.

Figure 7 displays the MSE and VAR obtained with the different techniques (DMO, EVMOS1, EVMOS2 and NGR) based on the reanalyses as reference at the same grid points as the ones used for the calibration at synoptic stations (five grid points). Clearly the VAR curves are now closer to the MSE curves for all post-processing techniques, suggesting that they all provide a good correction at these larger scales. It is also worth noting that the NGR ensemble variance is now similar to the ones obtained with the other techniques. For the Brier

Copyright © 2010 Royal Meteorological Society

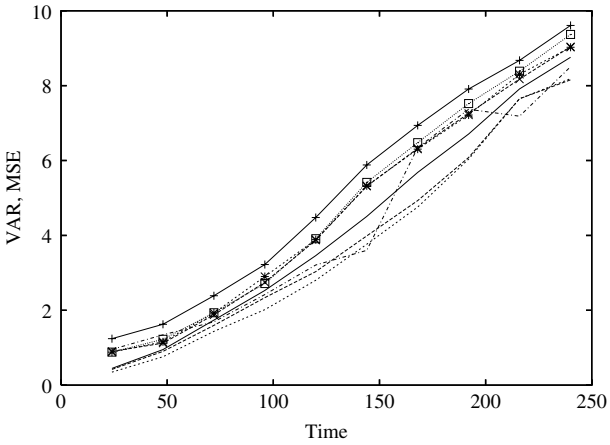

Figure 7. Mean square error evolution (MSE; pluses for the raw forecast, crosses for EVMOS1, stars for EVMOS2 and squares for NGR) of the ensemble mean and the ensemble variance (VAR; lines without symbols) for the raw ensemble (continuous line) and the corrected ensemble by the post-processing schemes (EVMOS1, long-dashed; EVMOS2, dashed; NGR, dot-dashed) for the whole year at 0000 UTC. EVMOS2 is using the Temperature at 850 hPa as second predictor. The reference used for training and verification is the ERA-interim reanalyses.

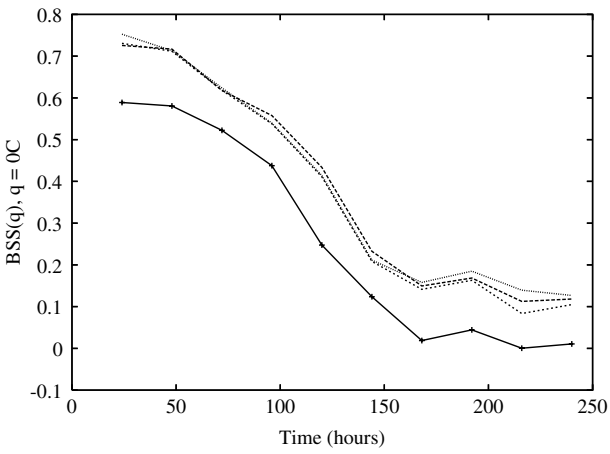

Figure 8. Brier Skill Scores for 2 m temperature probabilistic forecasts below 0 °C during winter 2008–2009 for the raw ensemble (pluses) and the corrected ones based on EVMOS1 (long-dashed), EVMOS2 with as second observable the temperature at 850 hPa (dashed) and the NGR method (dotted), at 0000 UTC. The reference is the ERA-interim reanalysis.

skill score the picture is similar to the one obtained for the synoptic stations, with a slight improvement at short lead times (Figure 8).

5. Discussion and conclusions

Two post-processing techniques of ensemble forecasts have been tested, one correcting directly each member of the ensemble referred as EVMOS, and the other, calibrating the ensemble based on a regression density estimator. The results suggest that both the EVMOS and NGR approaches are able to correct operational ensemble forecasts of the 2 m temperature and reduce considerably the distance between the ensemble mean and the observation over Belgium. The main difference between the two approaches lies in the amplitude of the corrected ensemble variance. For the EVMOS scheme, this variance is quite close to the one of the raw ensemble forecast for most lead times, while the one associated to the NGR scheme is generally much larger. This difference reflects two important aspects: (1) the EVMOS scheme does not correct any deficiency in the random distribution of the initial condition errors nor does it take the presence of observational errors into account, and, (2) the NGR method does take into account all the possible sources of errors affecting the observations. These differences do not imply, however, that one scheme is better than the other, but rather each scheme can have its own interest depending on the question at hand.

More specifically, if one would like to know whether the corrected ensemble is consistent with the truth, one has to take into account the presence of observational errors. If one subtracts this error from the MSE of the EVMOS scheme, one gets a result which is now very close to the variance of the ensemble, suggesting that the EVMOS scheme does provide a consistent ensemble, at the scales of the model dynamics. On the other hand, this observational error is already taken into account in the NGR approach indicating a good consistency at the scale of the observation itself.

To emphasize the differences between the two schemes further, Figure 9 shows two examples of operational ensemble forecast for the station of Uccle. These examples also demonstrate how an operational forecast product would change depending on the different calibration schemes. The starting dates are 4 December 2008 (Figure 9(a)) and 14 February 2009 (Figure 9(b)), respectively. For the first date (Figure 9(a)), the spread of the raw ensemble and the ensemble corrected by EVMOS1 are systematically too small, while the NGR method is indeed covering the range of potential values in which the observation is falling. On the other hand, the spread of the NGR can sometimes be too large when a useful information in a real forecast situation is needed. Such an example is provided in Figure 9(b) at short lead times. In this case, at 18 h lead time, the NGR ensemble covers a range of about 10 °C, while the raw ensemble provide already a good forecast with an uncertainty of the order of 2 °C. The EVMOS ensemble is performing even better by debiasing the raw ensemble. Another interesting feature of both methods is visible after day 3 of the forecast in the same panel (Figure 9(b)): the raw ensemble is in this case highly skewed indicating that the ensemble contains multiple regimes (or the possibility of regime transitions). This feature is preserved after the correction by EVMOS but not by the NGR method, transforming the ensemble to a Gaussian distribution. This can sometimes induce a failure of the NGR method when such regime transition or extreme events occur, as at the 4 day lead time (18 February 2009) where the NGR does not capture the event while the raw ensemble and the EVMOS

Copyright © 2010 Royal Meteorological Society

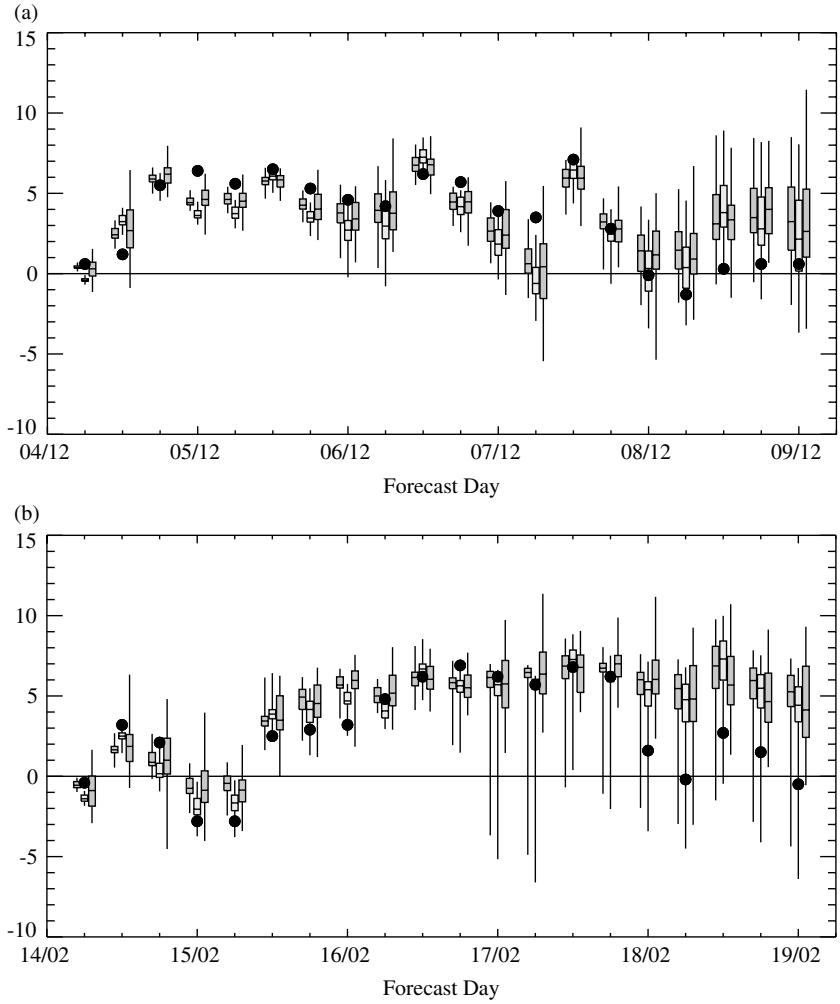

Figure 9. Ensemble forecast for two specific dates (4 December 2008 (a) and 14 February 2009 (b)) at Uccle. The black dots represent the observation. The ensemble distribution is displayed with box-and-whiskers, with 50% of the ensemble lying in the inner box and the remaining members represented by the whiskers. At each lead time, the box-and-whiskers represent, from left to right, the EVMOS1 corrected ensemble (dark grey), the raw ensemble (light grey) and the NGR corrected ensemble (dark grey).

ensemble give an indication of the potential occurrence of this actual event.

In summary, the technique should be chosen depending on the specific purpose. The EVMOS approach explored here is able to correct ensemble forecasts at the scales effectively described by the model, and its specific interest is that it can be trained on single past forecasts. The NGR method on the other hand provides information on the full uncertainty including the representativeness error and acts therefore as a probabilistic approach valid for downscaling. However its Gaussian shape can sometimes be counterproductive when raw ensembles are highly skewed.

An extension of this work through the analysis of post-processing of GFS re-forecasts (Hamill *et al.*, 2006) or TIGGE ensembles (Thorpex Interactive Grand Global Ensemble, Bougeault *et al.*, 2009) is envisioned, with emphasis of their potential combinations with the ECMWF ensemble forecasts to form a multi-model ensemble system.

Acknowledgements

We thank Tom Hamill for many fruitful discussions and the reviewers for their constructive comments. This work is partly supported by the Belgian Federal Science Policy Office under contract MO/34/020.

References

Bechtold P, Köhler M, Jung Th, Doblas-Reyes F, Leutbecher M, Rodwell MJ, Vitart F, Balsamo G. 2008. Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time scales. *Quarterly Journal of the Royal Meteorological Society* **134**: 1337–1351.

Bougeault P, Fuentes M, Raoult B, and co-authors. 2009. The THORPEX Interactive Grand Global Ensemble (TIGGE). *Bulletin of the American Meteorological Society* in press.

Bowler NE. 2006. Explicit accounting for observation error in categorical verification of forecasts. *Monthly Weather Review* **134**: 1600–1606.

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

Copyright © 2010 Royal Meteorological Society

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.

Doblas-Reyes FJ, Hagedorn R, Palmer TN. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting-II Calibration and combination. *Tellus* **57A**: 234–252.

Efron B, Tibshirani R. 1993. *An Introduction to the Bootstrap*. Chapman & Hall: New York, NY.

Glahn HR, Lowry DA. 1972. The use of Model Output Statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology* **11**: 1203–1211.

Glahn HR, Peroutka M, Wiedenfeld J, Wagner J, Zylstra G, Schuknecht B, Jackson B. 2009. MOS uncertainty estimates in an ensemble framework. *Monthly Weather Review* **137**: 246–267.

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

Hagedorn R. 2008. Using the ECMWF reforecast dataset to calibrate EPS forecasts. *ECMWF Newsletter* **117**: 8–13.

Hagedorn R, Hamill TM, Whitaker JS. 2008. Probabilistic forecast calibration using the ECMWF and GFS ensemble reforecasts. Part I: Two-Meter temperatures. *Monthly Weather Review* **136**: 2608–2619.

Hamill TM. 2001. Interpretation of Rank Histograms for verifying ensemble forecasts. *Monthly Weather Review* **129**: 550–560.

Hamill TM, Whitaker JS, Mullen SL. 2006. Reforecasts: an important dataset for improving weather predictions. *Bulletin of the American Meteorological Society* **87**: 33–46.

Hamill TM, Wilks DS. 2007. Comparison of Ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review* **135**: 2379–2390.

Mahfouf JF, Bergaoui K, Draper C, Bouyssel F, Taillefer F, Taseva L. 2009. A comparison of two off-line soil analysis schemes for assimilation of screen level observations. *Journal of Geophysical Research* **114**: D08105.

Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996. The ECMWF Ensemble Prediction System: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society* **122**: 73–119.

Parrish DF, Derber JC. 1992. The National Meteorological Center's spectral statistical-interpolation analysis system. *Monthly Weather Review* **120**: 1747–1763.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1996. *Numerical Recipes in Fortran 77*. Cambridge University Press: Cambridge.

Saetra O, Hersbach H, Bidlot J-R, Richardson D. 2004. Effects of observation errors on the statistics for ensemble spread and reliability. *Monthly Weather Review* **132**: 1487–1501.

Simmons A, Uppala S, Dee D, Kobayashi S. 2007. ERA-Interim: new ECMWF reanalysis products from 1989 onwards. *ECMWF Newsletter* **110**: 25–35.

Toth Z, Kalnay E. 1997. Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review* **125**: 3297–3319.

Unger DA, van den Dool H, O'Lenic E, Collins D. 2009. Ensemble regression. *Monthly Weather Review* **137**: 2365–2379.

Uppala SM and co-authors. 2005. The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society* **131**: 2961–3012.

Vannitsem S. 2008. Dynamical properties of MOS forecasts: Analysis of the ECMWF operational forecasting system. *Weather and Forecasting* **23**: 1032–1043.

Vannitsem S. 2009. A unified linear model output statistics scheme for both deterministic and ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society* **135**: 1801–1815.

Vannitsem S, Nicolis C. 2008. Dynamical properties of model output statistics forecasts. *Monthly Weather Review* **136**: 405–419.

Wilks DS. 2006a. Comparison of ensemble-MOS methods in the Lorenz'96 setting. *Meteorological Applications* **13**: 243–256.

Wilks DS. 2006b. *Statistical Methods in the Atmospheric Sciences*. Academic Press: New York, NY.

Copyright © 2010 Royal Meteorological Society