Koninklijk Meteorologisch Instituut van België
Institut Royal Météorologique de Belgique

Toward post-processing ensemble forecasts based on hindcasts

Bert Van Schaeybroeck and Stéphane Vannitsem

2012

Wetenschappelijke en
technische publicatie
Nr 061

Uitgegeven door het
KONINKLIJK METEOROLOGISCH INSTITUUT VAN BELGIE
Ringlaan 3, B-1180 Brussel
Verantwoordelijke uitgever: Dr. D. Gellens

Publication scientifique
et technique
N° 061

Edité par
l'INSTITUT ROYAL METEOROLOGIQUE DE BELGIQUE
Avenue Circulaire 3, B-1180 Bruxelles
Editeur responsable: Dr. D. Gellens

Koninklijk Meteorologisch Instituut van België
Institut Royal Météorologique de Belgique

Toward post-processing ensemble forecasts based on hindcasts

Bert Van Schaeybroeck and Stéphane Vannitsem

2012

Wetenschappelijke en
technische publicatie
Nr 061

Uitgegeven door het
KONINKLIJK METEOROLOGISCH INSTITUUT VAN BELGIE
Ringlaan 3, B-1180 Brussel
Verantwoordelijke uitgever: Dr. D. Gellens

Publication scientifique
et technique
N° 061

Edité par
l'INSTITUT ROYAL METEOROLOGIQUE DE BELGIQUE
Avenue Circulaire 3, B-1180 Bruxelles
Editeur responsable: Dr. D. Gellens

Summary

With an operational implementation of post-processing at RMI in mind, we study possible approaches of correcting the ECMWF ensemble forecast for stations in Belgium using the ensemble hindcast data set. This data set is enlarged each week by eighteen independent five-member ensemble forecasts using the current IFS system. Therefore, the hindcasts constitute an ideal basis for the training cycle of post-processing. Combined with a forward predictor-selection procedure we propose to use a post-processing technique called error-in-variables model output statistics or EVMOS. This technique was recently proposed and is based on linear regression and suited for correcting ensemble forecasts. The corrected forecasts are produced for nine synoptic stations in Belgium.

Different factors which influence the correction quality and which we aim to optimize are the number of weeks of training data, the number of predictors and the clustering of daily training data. We also investigate the influence of the training period, that is, the period of days over which the training forecasts are initialized. More specifically, we compare a training window which is centered around the forecast day with the case where the days of training precede the forecast day. Different results for the different training periods arise due to seasonal effects. We validate the different approaches against the bias-corrected forecasts using observations at nine stations for the ten-meter zonal and meridional wind speed and the two-meter temperature and for lead times up to one week. This is performed by cross-validation for a period of fourteen weeks.

For the inland stations and for all lead times, a mean-square-error (MSE) improvement of around 1.5 m^2/s^2 and 0.8 $(°C)^2$ for wind and temperature respectively is obtained. The MSE gain for wind at the two coastal stations is lower, especially for the meridional wind. Systematic biases are negligible for wind and thus most of the EVMOS post-processing is obtained by a variability correction. For two-meter temperature, on the other hand, systematic biases dominate the EVMOS corrections evidencing the correct variability representation of the model. For forecasting the day-time (12h) two-meter temperature, the best forecast is obtained by a simple bias correction whereas EVMOS post-processing turns out most effective for predicting the night-time (0h) forecast. In order to utilize EVMOS post-processing operationally, we propose the use of three predictors, a training period of at least seven weeks and preferably a training period centered around the forecast date.

Initially more than eighty candidate predictors are considered from the hindcast data set, based on which we construct eleven additional predictors. From the set of selected predictors during validation we isolate the most prominent ones. Except for the corresponding variables, by far the most frequently-selected predictors for the ten-meter wind are North-South and East-West surface stresses as well as the boundary layer height. For the two-meter temperature, temperature at 850 hPa and maximal temperature in the last 6 hours are the most crucial predictors.

1 Introduction

The current resolution of the Ensemble Prediction System (EPS) at ECMWF is around 32 kilometer for the first ten days and 64 kilometer for the subsequent five days. Although many processes are well-represented, detailed orographic features and subgrid processes are not accurately resolved. Therefore it may be useful to consider a direct post-processing of the EPS by comparison of the past forecasts with measurements at different stations.

Post-processing a forecast consists in changing the raw model output in such a way that its error is reduced. It is mostly done based on the knowledge of error statistics coming from the comparison of existing model output with measurements or analysis. In 1972 Glahn and Lowry introduced a post-processing procedure based on linear regression, called Linear Model Output Statistics (LMOS). The procedure consists of two phases. In the first phase, called the *training phase*, one aims at finding a linear relationship between past forecast variables and corresponding observations. Then, in the *post-processing phase*, this relationship is applied to new forecasts in order to produce a corrected forecast.

Recent post-processing approaches specifically designed for correcting *ensemble* forecasts can be divided into statistical-like and deterministic-like methods. In this work we adopt the EVMOS approach (Vannitsem, 2009), which, like LMOS, is a deterministic-like approach in the sense that each ensemble member is corrected in a similar fashion. The rationale behind EVMOS is to encompass both observational and forecast errors in such a way that the first and second moments of the distribution of observations coincide with those of the predictand. Examples of statistical-like approaches include Non-homogeneous Gaussian Regression (NGR) and Bayesian Model Averaging (BMA). NGR assumes Gaussianity and modifies spread and mean of each ensemble by minimizing the CRPS score (Gneiting et al., 2005). BMA, on the other hand, creates a new ensemble distribution by superposing kernel dressings on each member. Both methods produce a new ensemble distribution from which, potentially, a new ensemble may be sampled. The statistical and deterministic methods are similar in the sense that they use model variables as predictors, usually assume Gaussian error statistics and allow for the removal of biases. A discussion of some of their differences is given in Vannitsem and Hagedorn (2010), Van Schaeybroeck and Vannitsem (2011) and Wilks (2006).

Upon post-processing the relationships obtained during the training phase are strongly dependent on the model under consideration and a substantial number of past forecasts are required for obtaining stable relationships. Such a long training set is in general not available when model changes are frequent. This problem has been partly resolved for the ECMWF ensemble prediction system (EPS): not only is the 51-member EPS generated twice daily, each Thursday eighteen 5-member EPS forecasts, known as hindcasts, are issued for the same date of all past eighteen years. As we will show in this work a few weeks of hindcast are sufficient to provide stable post-processing for the wind and temperature near the surface.

This work is structured as follows. In section 2 we give the specifications of the hindcast data set present at ECMWF while the procedures for post-processing and predictor selection are explained in sections 3 and 4. In section 5 we present the validation results for the different approaches. We focus briefly on the ensemble features of the corrected forecasts in section 6 and on the most useful predictors in section 7. Lastly in section 8 we conclude and discuss the possibility of an operational post-processing scheme at RMI.

2 Hindcast data set

The hindcast data set of ECMWF contains ensemble reforecasts based on the ERA-Interim reanalysis using the current IFS cycle with the same spatial resolution as the EPS. Each week on Thursday eighteen forecasts, all initialized at 0h, are produced for the same date for the past eighteen years.

The prerequisite for a good LMOS post-processing is a large data set which poses a serious constraint on the use of LMOS in practice. Many data points are required in order to obtain stable regression relations and thus to yield a reliable corrected forecast. For most meteorological quantities, it was concluded that a data set of 300 observation/forecast pairs for each lead time were sufficient. Therefore, a model change involves a waiting period of the order of ten months to get a sufficiently large training set. Note, however, that such a set then contains data that could be highly correlated in time, and that does not take into account seasonal effects. The existence of the ECMWF hindcast data set provides a large step forward to resolve these issues. The hindcast data reduce the time required to obtain stable regression, they contain almost no time correlations since the hindcasts are issued with one week of delay, and seasonal effects may be taken into account.

From the hindcast data set present at ECMWF we have extracted 52 surface variables, 6 variables at pressure levels 500 hPa, 850 hPa and 925 hPa, and, additionally, we constructed eleven model-based predictors, leading to a total of 81 candidate predictors. In section 7 we discuss only the few among these predictors surviving the predictor-selection procedure introduced in section 4.

3 Training Procedure

Statistical inference based on past forecasts and associated observations allows for constructing the statistical error characteristics. Assuming these features to be generic for all outputs of the model, this allows us to "correct" future model output. For example, if the mean of the error distribution of a certain variable is nonzero, it makes sense to subtract the bias from future forecasts. Another natural requirement is that the variability of the forecasts should correspond to the variabililty of the observations. For long forecast lead times, this last condition guarantees the correct climatological variance of the forecasts. In the following we sketch the classical LMOS and the new EVMOS post-processing methods, their underlying assumptions and their relation to these conditions.

3.1 Model Output Statistics and Ordinary Least Squares

The standard method of correcting deterministic forecasts is by the application of LMOS (Glahn and Lowry, 1972). This procedure uses ordinary-least-squares (OLS) regression to obtain a relation between forecasted variables which are called predictors and the corresponding observations. As a first assumption LMOS considers only the presence of observational errors. Consider a variable X, for example two-meter temperature, for which $X_{_N}$ denotes the true value and $X_{_O}$ the observed value. If we have a series of K measurements $X_{_{O,j}}$ $(j = 1, \ldots, K)$ with corresponding true values $X_{_{N,j}}$, the errors are $\varepsilon_{_{O,j}} = X_{_{N,j}} - X_{_{O,j}}$. OLS assumes the model is able to provide a perfect prediction $X_{_C}$ for $X_{_N}$ and also the errors $\varepsilon_{_O}$ to be uncorrelated and normally distributed. Therefore,

the most probable values of X_C and thus also X_N are obtained by minimizing the mean square error (MSE):

$$\mathrm{MSE} = \langle \varepsilon_O^2 \rangle = \langle (X_O - X_C)^2 \rangle. \tag{1}$$

Here the brackets denote the average over the entire data set $\langle \cdot \rangle = (1/K) \sum_{j=1}^{K} \cdot$. We still lack a relationship between the forecast estimate X_C and the uncorrected model output which consists of the predictors V_p $(p = 1, \ldots, P)$. Linear regression assumes a linear relationship between the predictand X_C and the predictors V_p. Hence, for each observation $X_{O,j}$ and associated predictor values V_{jp} $(p = 1, \ldots, P)$:

$$X_{C,j} = \alpha + \sum_{p=1}^{P} \beta_p V_{jp}.$$

A minimization of the MSE with respect to α and each component of the vector $\boldsymbol{\beta}$ governs the well-known solution (Casella and Berger, 1990):

$$\boldsymbol{\beta}_{OLS} = \left(\overline{\boldsymbol{V}}^T \overline{\boldsymbol{V}} \right)^{-1} \left(\overline{\boldsymbol{V}}^T \overline{\boldsymbol{X}}_O \right), \tag{2a}$$

$$\alpha_{OLS} = \langle X_O \rangle - \sum_{p=1}^{P} \beta_{OLS,p} \langle V_p \rangle. \tag{2b}$$

Here the matrix $\overline{\boldsymbol{V}}$ contains the elements $(\overline{\boldsymbol{V}})_{jp} = V_{jp} - \langle V_p \rangle$ while the observation vector is $(\overline{\boldsymbol{X}}_O)_j = X_{O,j} - \langle X_O \rangle$. The proportionality of the regression coefficients β_p to the correlations between observation and forecast variables $\overline{\boldsymbol{V}}^T \overline{\boldsymbol{X}}_O$ implies that they tend to zero for long lead times when, due to the chaotic nature of the system, correlations inevitably vanish. Thus, along with the convergence of the predictand X_C towards the (climatological) mean $\langle X_O \rangle$, its variability $\langle (X_C - \langle X_C \rangle)^2 \rangle$ vanishes and the aforementioned condition of equality of variance of observation and corrected forecast cannot be satisfied. Furthermore, when applying LMOS on all members of a certain ensemble, the corrected-ensemble spread vanishes and becomes an incorrect predictor of the forecast skill.

3.2 Error-in-Variables Model Output Statistics (EVMOS)

The problem at hand can be resolved by a reconsideration of the underlying assumption of OLS which is the mere presence of observation errors (Vannitsem, 2009). As well known, forecasts are also imperfect as they suffer from initial-condition errors and model deficiencies. Therefore, in addition to the normally-distributed error between true value and the observation $\varepsilon_O = X_N - X_O$, we include an error between the true value and the value predicted by model variables $\varepsilon_C = X_N - X_C$. The issue concerning the choice of error variances remains to be settled, however, it is natural to assume ε_O and ε_C to have a variance $\sigma_O^2 = \langle X_O^2 - \langle X_O \rangle^2 \rangle$ and $\sigma_C^2 = \langle X_C^2 - \langle X_C \rangle^2 \rangle$, respectively. We can then define the cost function $\mathcal{J}_{EV}$, the minimum of which yields the most probable estimate for X:

$$\mathcal{J}_{EV} = \frac{\langle (X_C - X_N)^2 \rangle}{\sigma_C^2} + \frac{\langle (X_O - X_N)^2 \rangle}{\sigma_O^2}.$$

A minimization with respect to X_N, α and β_p $(p = 1, \ldots, P)$ has a solution:

$$\boldsymbol{\beta}_{EV} = \frac{\sigma_O \left(\overline{\boldsymbol{V}}^T \overline{\boldsymbol{V}}\right)^{-1} \left(\overline{\boldsymbol{V}}^T \overline{\boldsymbol{X}}_O\right)}{\sqrt{\left(\overline{\boldsymbol{V}}^T \overline{\boldsymbol{X}}_O\right)^T \left(\overline{\boldsymbol{V}}^T \overline{\boldsymbol{V}}\right)^{-1} \left(\overline{\boldsymbol{V}}^T \overline{\boldsymbol{X}}_O\right)}}, \tag{3a}$$

$$\alpha_{EV} = \langle X_O \rangle - \sum_{p=1}^{P} \beta_{EV,p} \langle V_p \rangle. \tag{3b}$$

Comparing the OLS and EVMOS regression coefficients, we conclude:

$$\boldsymbol{\beta}_{EV} = \frac{\sigma_O \boldsymbol{\beta}_{OLS}}{\sqrt{\boldsymbol{\beta}_{OLS}^T \left(\overline{\boldsymbol{V}}^T \overline{\boldsymbol{X}}_O\right)}}. \tag{4}$$

The EVMOS-corrected forecast as characterized by these coefficients, satisfies both conditions on equality of mean $\langle X_C \rangle = \langle X_O \rangle$ and equality of variability $\sigma_C = \sigma_O$.

Note that several linear regression methods were investigated in Van Schaeybroeck and Vannitsem (2011) where it was concluded that for ensemble forecasting, EVMOS was the most appropriate. For example the total-least-squares (TLS) methods incorporates errors in the model variables by assuming a random noise term for each of the predictors instead of introducing a noise term additional to X_N as is done by EVMOS. TLS, however, becomes unstable when a predictor is uncorrelated to the observation.

How well are the assumptions of random noise errors underlying the EVMOS formalism satisfied? For two-meter temperature the Gaussian error approximation is a good assumption while this is not really the case for the wind variables. On the other hand, the randomness of errors will be well approximated when considering the hindcast data variables since the different hindcasts are initialized with delays of at least one week.

Lastly, in what follows we also apply a third post-processing method called bias correction. The bias-corrected forecast of a model variable V_{j1} is simply:

$$X_{C,j} = \langle X_O \rangle - \langle V_1 \rangle + V_{j1}.$$

This correction could be seen as a one-variable regression with $\alpha = \langle X_O \rangle - \langle V_1 \rangle$ and $\beta_1 = 1$.

4 The predictor-selection procedure

In the last section, we have provided the details of the method of training considering the predictors V_p $(p = 1, \ldots, P)$ as given. As specified before, we have more than eighty candidate predictors at our disposal but performing the training using the entire set would lead to overfitting or unstable behavior of the regression. Overfitting refers to the situation wherein the regression relation performs well on the training set but, when applied to an independent data set yields wild results and bad scores. Moreover, overfitting may occur as a consequence of using too many regression parameters, or because of multicollinearity which involves the presence of two or more highly-correlated predictors. Therefore a predictor-selection procedure is necessary to restrict the size of the predictor set. We use a forward-selection method procedure (Wilks, 1995) which we explain now.

Consider a data set of observations and the corresponding hindcasts. We start by subdividing this set into a training and a verification set and performing an EVMOS training with two predictors on the training set. The first predictor is the one corresponding to the observation variable while the second one is taken from the set of $P-1$ remaining predictors (in our case $P = 81$). We next construct a corrected forecast using the verification data set and the EVMOS regression relations and determine a corresponding mean square error (MSE). By taking several different resamplings separating the data set into training and verification sets we determine an average MSE. We continue by determining a MSE for each $P-1$ "second" predictors. The candidate predictor corresponding with the lowest MSE is then retained in order to perform EVMOS using three predictors. We then proceed by iterating over the $P-2$ residual candidate predictors using the same resampling procedures. Again we select the predictors corresponding to the lowest MSE. We stop this predictor-selection procedure when a predetermined number of predictors is found. Finally, using that predictor list, the corresponding regression parameters are calculated once more, now using the entire training data set.

The separation into disjoint training and verification sets was required to avoid overfitting. Typically the verification data set was seventeen times smaller than the one of the training; increasing the resampling number did not amount to substantial changes of the selected predictors. It was also checked that the selection of each new predictor resulted in a reduction of MSE, and this was almost always the case for up to five predictors.

5 Validation

To assess the EVMOS training and the predictor-selection procedure we perform the validation using the hindcast data set and the observations at several stations in Belgium for ten-meter wind speed and two-meter temperature. Our control period consists of fourteen weeks of hindcasts in 2011 and extends from June 9 until September 8 with forecast lead times up to one week. The data window available for training extends three more weeks before and after this validation window, thus from May 19, one day after the model change, until September 29. No model change was experienced during the entire verification period. All predictor values are linearly interpolated to the positions of the nine synoptic stations, all of which have available observations for the past eighteen years. Our inland stations include Ukkel, Elsenborn, St-Hubert, Kleine-Brogel, Florennes, Bierset and Chièvres while the coastal stations are Middelkerke and Koksijde, located in the North-West of Belgium. We perform validation on the hindcast data set only, using all ensemble members on equal footing, both for training and validation. Details of the double cross-valiation procedure used to obtain our results are given in Appendix A.

As already mentioned, the availability of a sufficiently large set of data is indispensable in order to avoid overfitting of any kind. Yet, assembling data from different seasons to build a training data set may turn out detrimental for the quality of the error statistics. Indeed, as a consequence of the dissimilar seasonal atmospheric conditions, the important predictors as well as the systematic errors may differ. Therefore, one would prefer the training set to be initialized using the same seasonal conditions, say, hindcasts ranging from a few weeks preceding the forecast date until a few weeks after. However, ECMWF guarantees only the availability of hindcast data of the last Thursday[1]. Therefore, it is

[1]Note, however, that hindcasts for the upcoming three weeks are available via MARS at ECMWF.

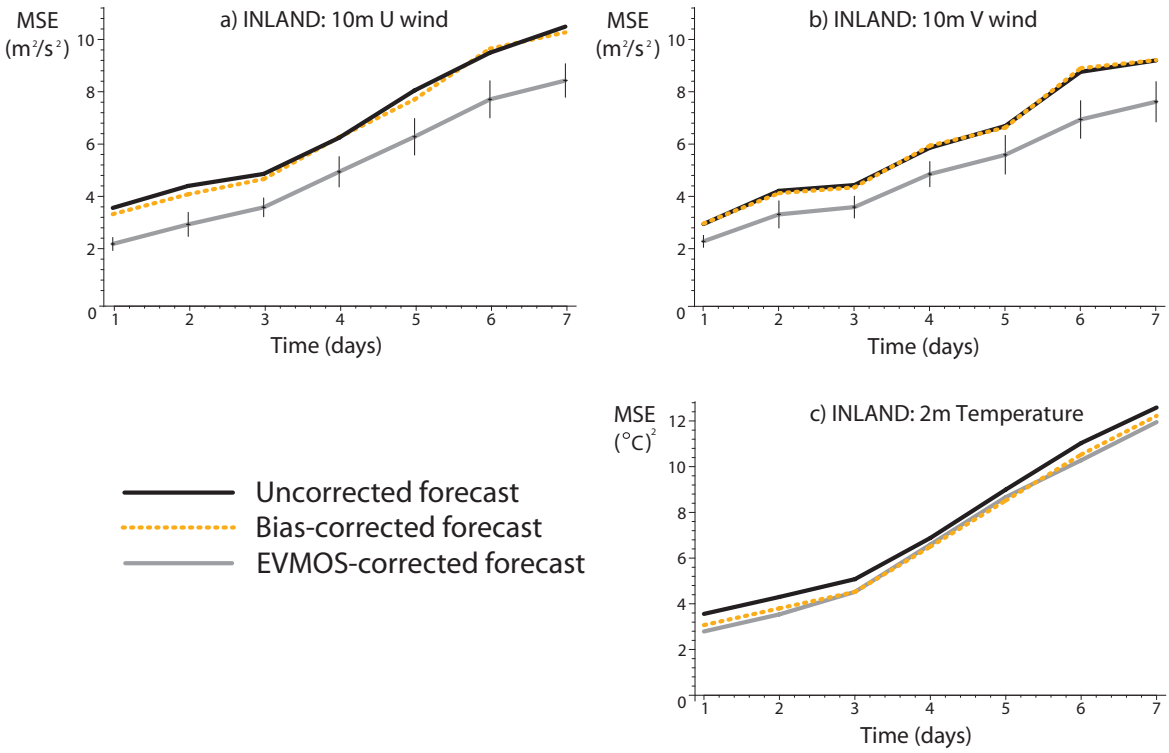

Figure 1: Mean square error (MSE) against lead time (in days) for 10m zonal wind (a), 10m meridional wind (b) and 2m temperature (c), averaged over seven inland stations. The three lines indicate different forecasts: the raw forecast (full black line), the bias-corrected forecast (orange dotted line) and the EVMOS-corrected forecast (full grey line). For both corrected forecasts we use a training window of seven weeks of hindcast data, centered around the week of validation (centered training). Also we use daily clustering and three predictors are implemented in the EVMOS framework. The intervals around the EVMOS-lines for wind quantify the consistency of the EVMOS-correction over the fourteen verification weeks. The interval widths are twice the standard deviation of the MSE reduction which is defined as the difference in MSE of the EVMOS-corrected and uncorrected forecasts.

essential to find the optimal balance between having sufficient data (a long enough training set) and the reduction of seasonal effects (a short enough training set). A key goal of this study is to optimize the parameters which attribute to the aforementioned balance so as to obtain the best, as well as the most reliable results. The parameters concerned include the number of forecast steps to cluster, the number of training weeks and the number of predictors used. In the following sections, we evaluate these parameters and estimate their optimal values.

5.1 Validation of EVMOS against the bias-corrected forecast

Figures 1a-c show the MSE averaged over seven inland stations for the ten-meter U-wind, the ten-meter V-wind and the two-meter temperature, henceforth denoted as 10m U wind, 10m V wind and 2m T, respectively. The MSE of the EVMOS-corrected forecast against the MSE of the bias-corrected forecast and the raw forecast are depicted. Three predictors for EVMOS and a training window of seven weeks centered around the verification week

are used. Note that a week of training represents eighteen hindcasts initialized on the same day but over all past eighteen years. Also, the regression coefficients for lead times within the same day (0h and 12h) are the same (this issue will be further discussed in Sect. 5.3). This is done to enlarge the training sets. Note that the MSE of the bias-corrected forecasts are also determined using cross-validation.

The results for the 10m U and the 10m V wind (Figs. 1a and b) are very much alike. The score of the bias-corrected wind forecast is very close to the one of the uncorrected forecast indicating unbiased model representation. EVMOS involves a substantial improvement in forecast accuracy as is obvious by its reduction of MSE, which, for both variables seems independent of lead time and is around 1.5 m^2/s^2.

We estimate the consistency of the EVMOS improvement over the fourteen verification weeks for the wind variables by showing intervals around the EVMOS-line. The amplitudes of the intervals are twice the standard deviations of the difference in MSE of the EVMOS-corrected and the uncorrected forecasts. It is clear that the intervals are below the MSE of the uncorrected forecast and therefore the EVMOS forecast improvement can be considered consistent.

A more modest, though still significant improvement, is provided by EVMOS for the 2m T. The main contribution of the corrections may be identified as stemming from systematic errors as seen in Fig. 1c by the proximity of the EVMOS and the bias-corrected line. Again a consistent improvement arises which is of the order of 0.5 $(°C)^2$. A more skilful correction method for temperature will be presented in Sect. 5.3.

5.2 Results for varying training periods

In the following we consider different choices of training periods. We select two distinct approaches and call them "centered training" and "forward training". For the former, the training window is centered around the week for which we desire to correct the forecast. Note that the number of training weeks must be uneven. Forward training indicates the fact that the training window contains only weeks prior or equal to the week for which we want to correct the forecast. Note that the results of the last section were obtained using centered training. Even though the best results may be expected for centered training, to date the lack of (guarantee for) availability of forthcoming weeks of hindcast data restricts the operational implementation to the use of forward training.

In Figs. 2a-c we use the centered training scheme and show the MSE scores against lead time for different numbers of training weeks (3, 7, 13 and 17 weeks) for the 10m wind and the 2m T, yielding considerably different results. EVMOS forecasts with training periods of five (not shown) or more weeks are consistently more accurate than the uncorrected forecasts while training periods of three weeks only deliver a modest improvement. One can state that thirteen weeks, or about one season, of training is optimal for both wind and temperature. By extending the training period to seventeen weeks, forecasts become slightly worse again and this may be attributed to the fact that seasonal effects are no longer optimally taken into account.

In Fig. 3a-c we show the same results obtained using the forward training scheme with 7,9 and 13 weeks of training. For comparison we also show the MSE of centered 13 weeks training. The situation is less clear-cut: although worse than centered training and consistently better than the raw forecast, none of the forward training periods may be termed the best. While thirteen weeks of training is superior for some lead times, seven

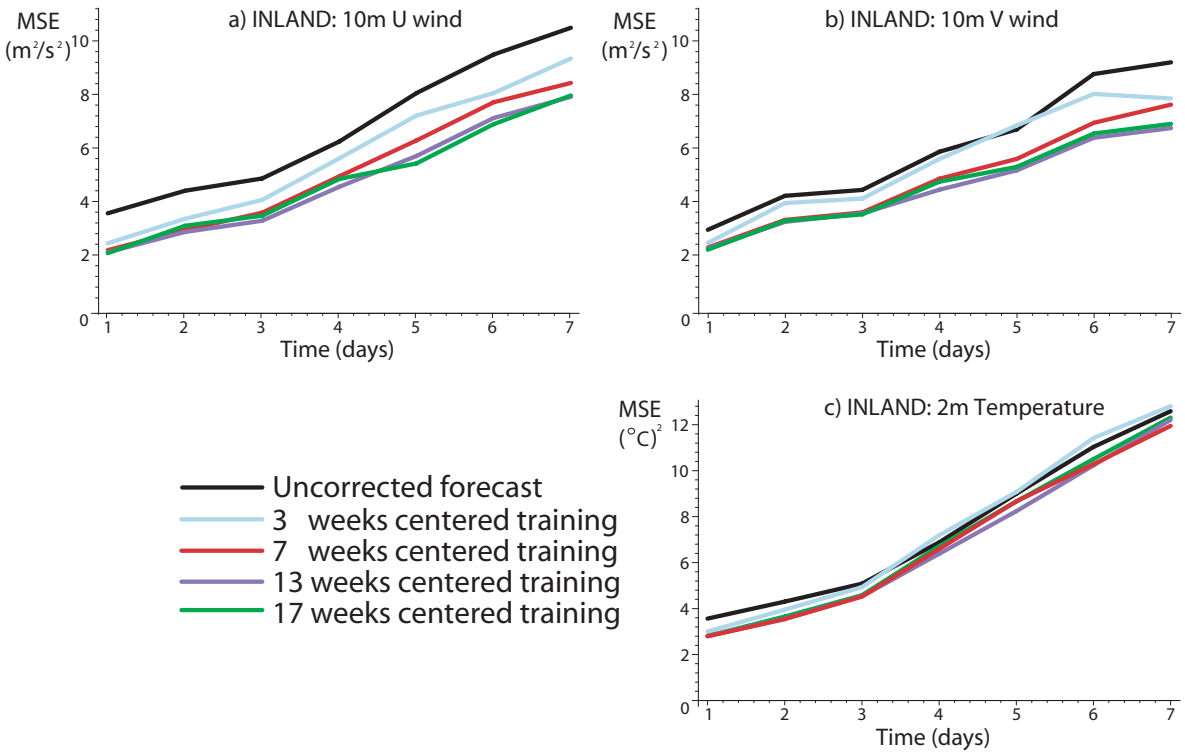

Figure 2: Mean square error (MSE) against lead time (in days) for 10m zonal wind (a), 10m meridional wind (b) and 2m temperature (c), averaged over seven inland stations. The full black line indicates the MSE of the raw forecast. All other lines involve EVMOS-corrections using centered training, that is, using a training window centered around the week of validation but with a variable number of training weeks. Also we use daily clustering and three predictors are implemented in the EVMOS framework.

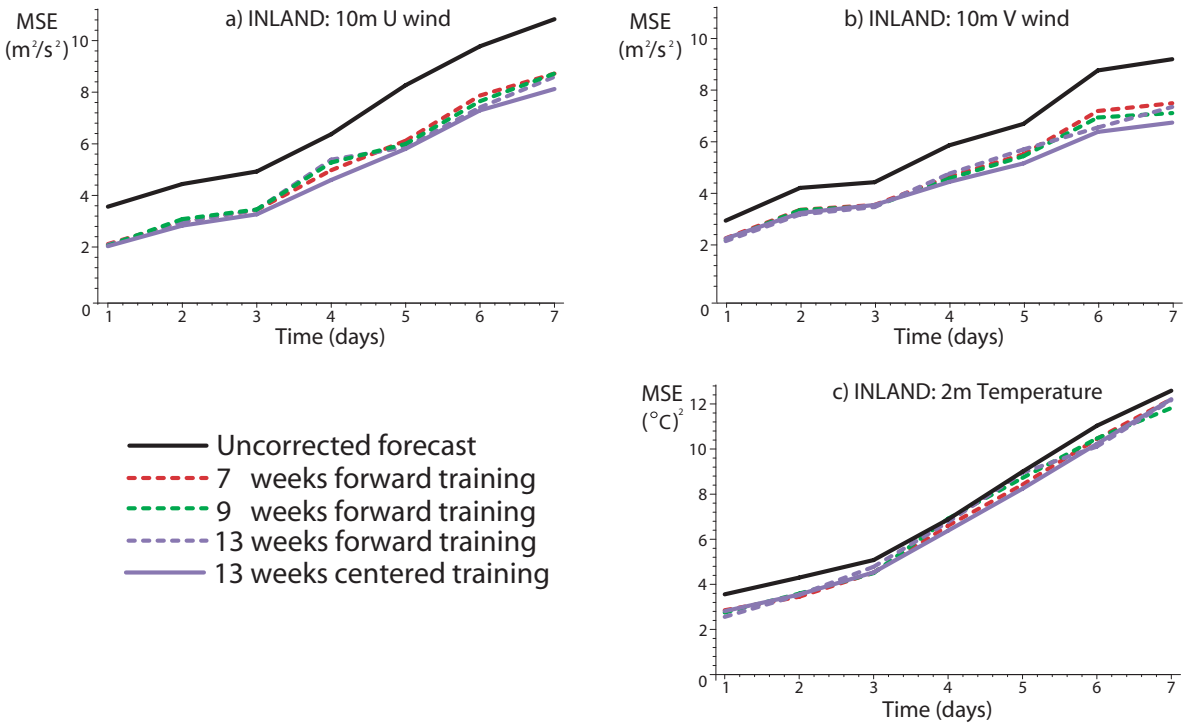

Figure 3: Mean square error (MSE) against lead time (in days) for 10m zonal wind (a), 10m meridional wind (b) and 2m temperature (c), averaged over seven inland stations. The full black line indicates the MSE of the raw forecast. All other lines indicate EVMOS-corrected forecasts using forward training, that is, the training set includes only weeks equal to or preceding the week of validation. This is shown for a variable number of forward training weeks and as a reference we show the result obtained using 13 weeks of centered training. Also we use daily clustering and three predictors are implemented in the EVMOS framework.

weeks gives the most consistent results. Again, the removal of seasonal effects seems indispensable for obtaining optimal corrections.

5.3 Results with different data clusterings

In order to enlarge the set of training points, and for all results shown so far, we use an approach which we call "daily clustering". We thereby take the regression coefficients to be the same for all forecast times within each forecast day. Given the fact that in practice only forward training may be possible, it seemed the best option to increase the training set. This implies that for the training we assemble the data for lead times within the same day: we cluster lead times 0h and 12h for day one, 24h and 36h for day two and so forth for all other days. We have also tested training without daily clustering and it was revealed that, although the training data set per coefficient was reduced, the regression was nevertheless stable. The most important conclusion was the fact that systematic biases were much better corrected without clustering. In fact, for the 2m T, the bias-corrected forecast turned out to beat EVMOS. The fact that correcting without clustering improves the skill, is not surprising since it is expected that the daytime temperature depends on parameters present in the radiation scheme of the model whereas the night-time temperature is more sensitive to wind speed and cloud cover.

The results with and without daily clustering are shown in Figs. 4a-c using seven weeks of centered training. For the 10m U wind and the 2m T, it is clearly seen that the bias-corrected forecasts without clustering perform better than the forecasts with clustering. For the 10m V wind, the elimination of clustering deteriorates the forecast. Compared to all previously-shown results for the 2m T, the bias-corrected forecast is the best at days three and four and provides a correction which is invariant in time. Surprisingly, the EVMOS approach without clustering, although perfectly capable of correcting biases, is worse than the bias-corrected forecast without clustering. The MSE of the latter is the same at days three and four as the EVMOS result shown in Fig. 2c using thirteen weeks of centered training.

In order to disentangle the different contributions when correcting the 2m T, we show in Fig. 5 the corrections at all lead times. It is clear that, as opposed to the bias correction improvement which is more or less invariant in time, the quality of the EVMOS-corrected forecast is good only for lead times which are multiples of 24 hours, that is, for the night-time (0h) forecast. Note that, since all forecasts are initialized at 0h, lead times which are integer multiples of 24 hours indicate the night-time forecasts. Very similar results are obtained for the bias correction using forward training. This leads us to conclude that an improved post-processing procedure would involve a combination of an EVMOS approach for the night-time forecast and a bias-correction for the day-time forecast. The result of such approach is shown in Fig. 4c by the red line. A consistent improvement of the order of 0.8 $(^{\circ}\mathrm{C})^2$ arises. Since the intervals which associated standard deviation of the improvement are below the MSE of the uncorrected forecast, the forecast improvement is consistent.

Figure 6 shows the training data set of seven weeks of hindcasts and resulting linear regressions for 2m T at Ukkel. The observations are put against the forecast for lead times of 48 and 60 hours. EVMOS regressions with one predictor is applied to three data sets: the 48h data set (green line), the 60 hours data set (black line) and the entire data set (red line). The red line therefore represents the result of the daily-clustering method.

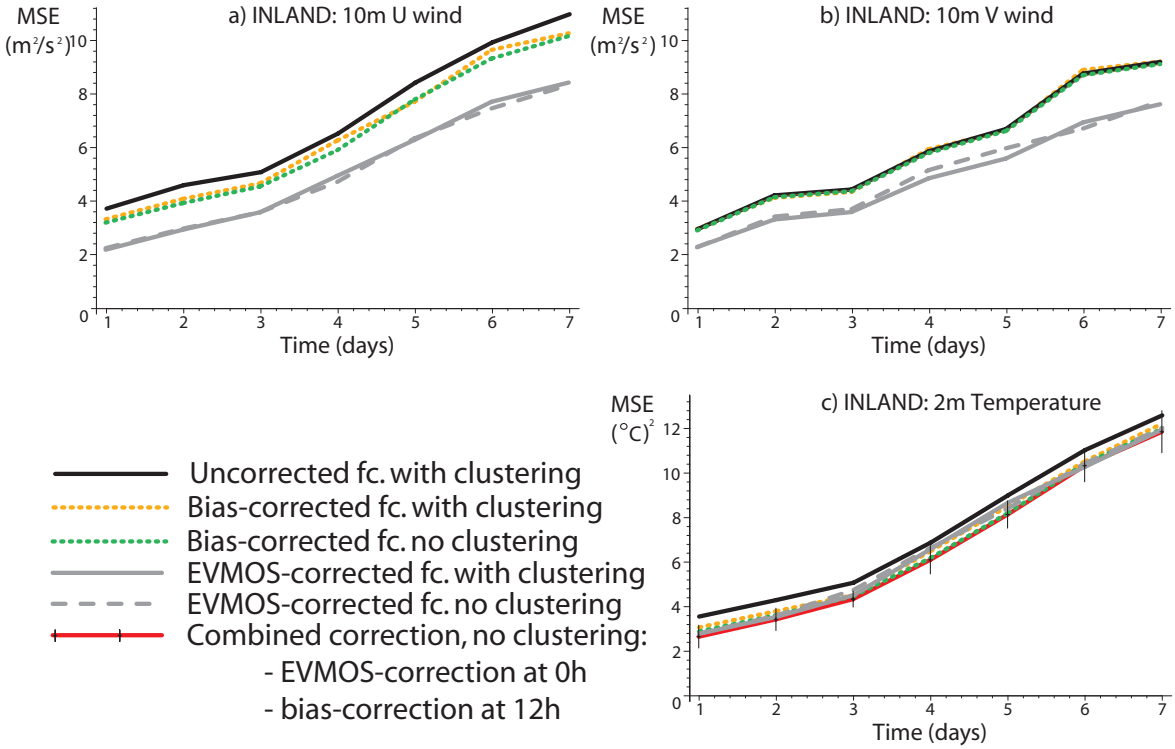

Figure 4: Mean square error (MSE) against lead time (in days) for 10m zonal wind (a), 10m meridional wind (b) and 2m temperature (c), averaged over seven inland stations. The full black line indicates the MSE of the raw forecast. All other lines indicate corrected forecasts obtained using seven weeks of centered training with and without clustering of the lead times per day. Three predictors are implemented in the EVMOS regression. The intervals around the red line in figure (c) quantify the consistency of the combined correction over the fourteen verification weeks. The interval widths are twice the standard deviation of the MSE reduction which is defined as the difference in MSE of the combined-corrected and uncorrected forecasts.

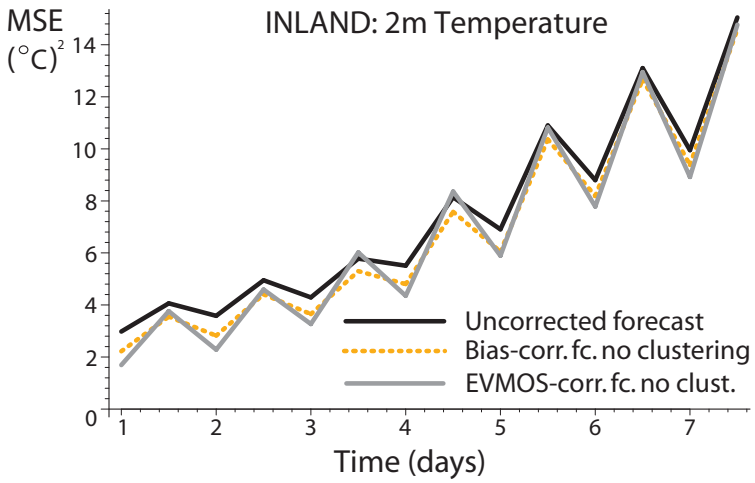

Figure 5: Mean square error (MSE) against lead time (in days) for 2m temperature, averaged over seven inland stations. The full black line indicates the MSE of the raw forecast. The two other lines are obtained using seven weeks of centered training *without* clustering of the lead times per day. The dotted line indicates the bias-corrected forecast while the grey line represents EVMOS-corrected forecasts.

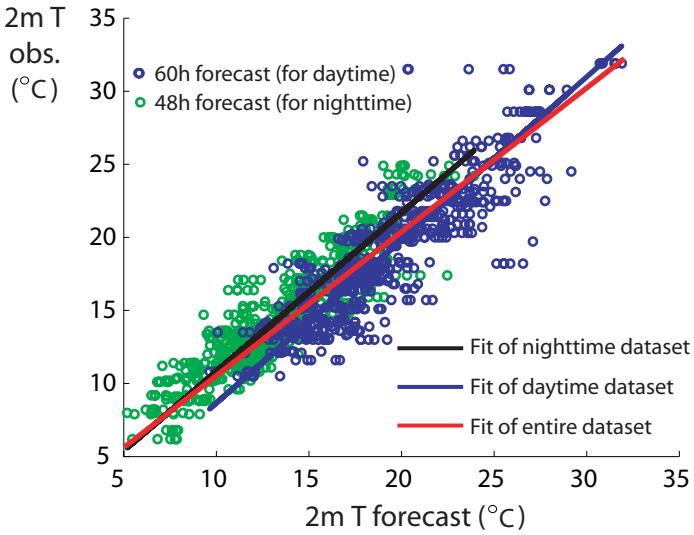

Figure 6: Observed 2m temperature against forecast 2m temperature for lead times 48h (daytime) and 60h (nighttime) using seven weeks of training data at Ukkel. The lines indicates EVMOS fits with one predictor: the blue line fits the blue data points, the black line fits the green points and the red line is a fit to all data points. The red line is the result obtained with daily clustering.

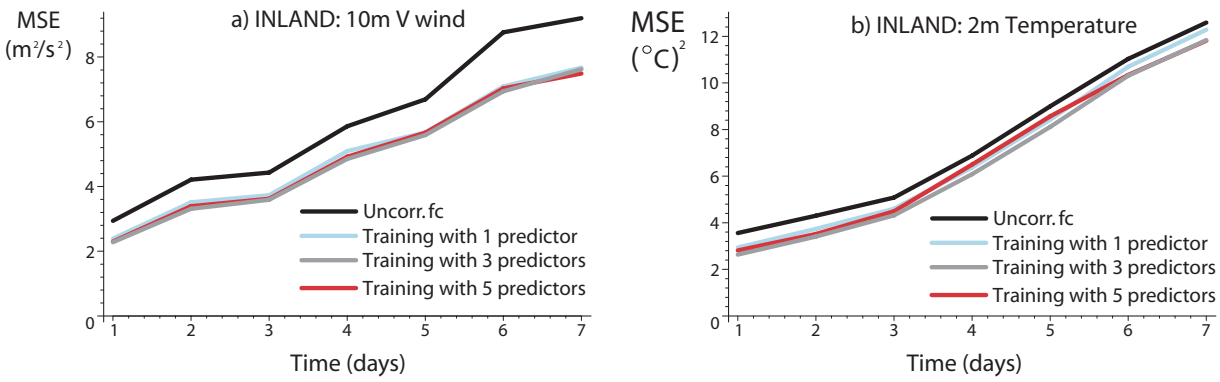

Figure 7: Mean square error (MSE) against lead time (in days) for ten-meter meriodional wind (a) and two-meter temperature (b), averaged over seven inland stations. The full black line involve the MSE of the raw forecast. All other lines indicate corrected forecasts using seven weeks of centered training but with a variable number of predictors. Wind corrections were performed using EVMOS only whereas temperature calibration was done using EVMOS for the night-time (0h) forecast and a bias-correction for the daytime (0h) forecast.

Due to a different daytime and nighttime forecast bias (black and blue line) and a larger average daytime temperature, the fit of the entire data set has a reduced slope, implying that, for 2m T, a corrected forecast trained using no clustering will in general be better than the one with clustering.

5.4 Dependence on the number of predictors

Increasing the number of regression predictors should improve the regression, yet, when using too many predictors, the corrected forecast becomes prone to overfitting. Concerning the ten-meter wind, as opposed to the number of training weeks, the influence of the predictor number is weak insofar more than one predictor is implemented. This is shown in Fig. 7a where we show the MSE of meridional wind using centered training of seven weeks and for one, three and five predictors. Apparently regression with three predictors is slightly better than with one predictor and the nominal difference with the five-predictor case suggests the sufficiency of three predictors. Figure 7b shows the influence of the amount of predictors on temperature used for EVMOS-correcting for our best approach: while the night-time (0h) forecast uses EVMOS, the daytime (12h) forecast is bias corrected. Clearly the skill of the forecast calibration with three predictors is larger than the one corrected with one predictor and five predictors.

5.5 Results for stations at sea

Due to the different weather regimes and model physics close to sea, we have performed a separate post-processing of two stations near the coast, namely Koksijde and Middelkerke. As compared to the inland stations the variability of coastal wind is larger while the temperature variability is reduced. A study similar to the one already given for inland stations led to the results presented in Figs. 8a-c. Corrections for the inland and coastal stations are similar in the sense that systematic biases for wind are negligible while the

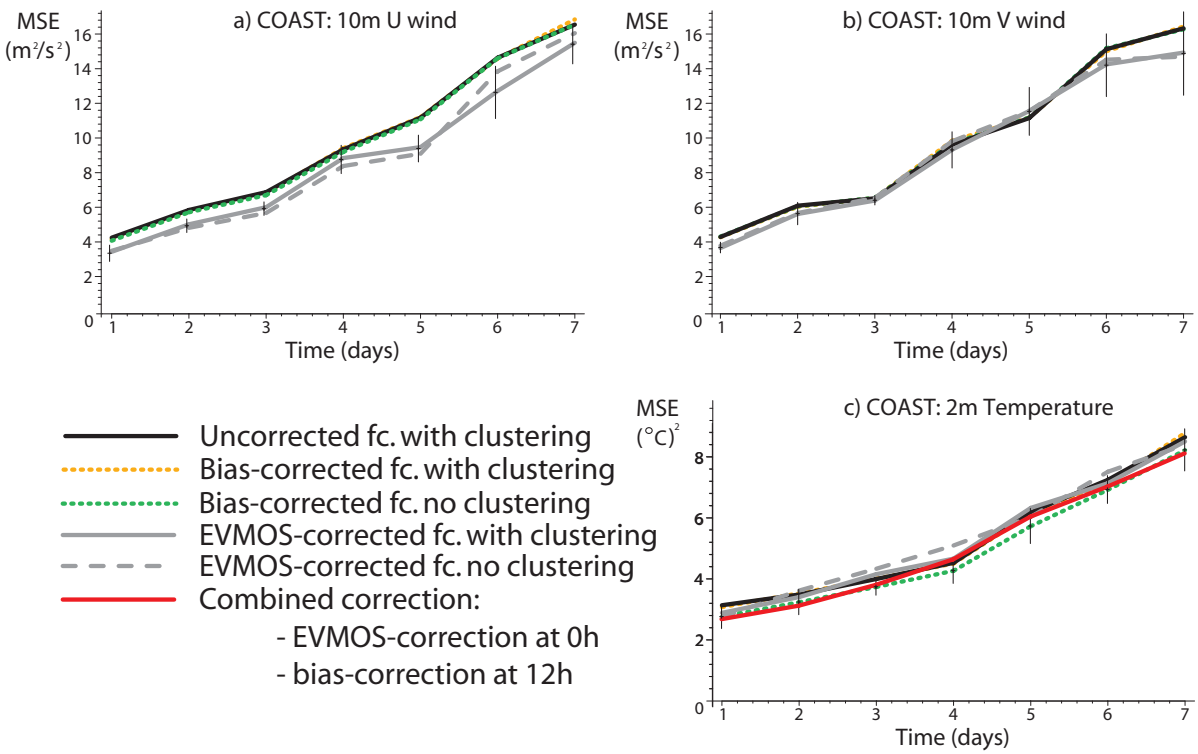

Figure 8: As Fig. 4 but averaged over two stations at the coast (Middelkerke and Koksijde).

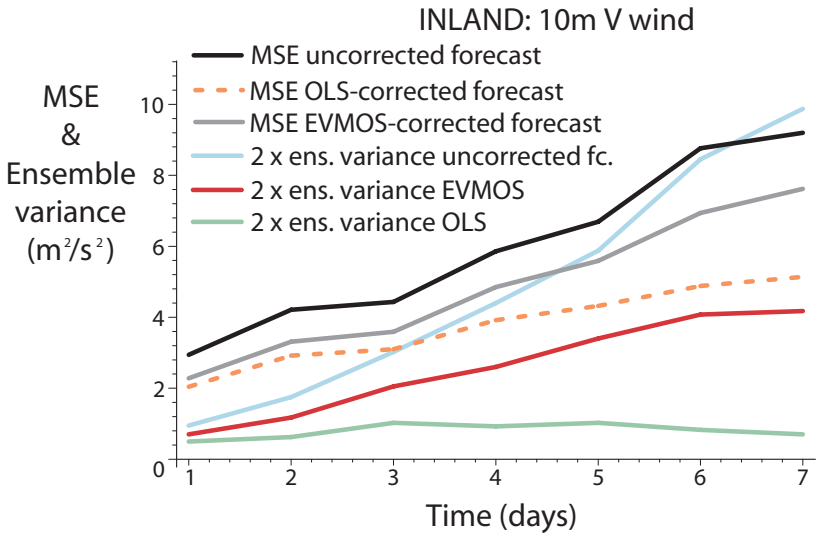

Figure 9: Mean square error (MSE) and two times the ensemble variance against lead time (in days) for 10m meridional wind. The black and blue lines indicates the MSE and twice the ensemble variance of the raw forecast, respectively. All other lines involve corrected forecasts obtained using seven weeks of centered training using daily clustering. Three predictors are implemented in the EVMOS and OLS regression.

ones for 2m T dominate the corrections. Wind corrections based on EVMOS are still considerable although substantially smaller than for the inland stations and best when using thirteen weeks of centered training (not shown). The bias-corrected forecast for 2m T obtained without clustering of the daily data is far better than any other approach. The intervals around the EVMOS indicate that the gain in skill for the 10m U wind is consistent while this is not the case for the 10m V wind where the gain in skill is small anyway. For the 2m T, bias-correcting gives rise to consistent improvements and is better than the combined approach of bias-correcting the day-time forecasts and EVMOS-correcting the night-time forecasts.

6 Ensemble features of the corrected forecasts

Let us now briefly discuss the ensemble features of the corrected forecasts. As mentioned earlier, for long lead times an ensemble corrected by OLS will have a vanishing ensemble spread. This is evidenced in Fig. 9 where we depict the mean ensemble variance of the uncorrected forecast for meridional wind against the mean ensemble variance of forecasts obtained with OLS and EVMOS. The mean ensemble variance is defined as $\langle (X_C - \ll X_C \gg_e)^2 \rangle$ with the double brackets denoting the ensemble mean and X_C the corrected variable. We also show the MSE of the raw and OLS-corrected forecasts. The conspicuous strong reduction of MSE of the OLS-corrected forecast as compared to the one of EVMOS can be almost entirely ascribed to a gradual decay of the predictand towards climatology. Therefore, although the ensemble mean of the OLS-corrected forecast may be a good estimate of the forecast, the statistical information of the corrected ensemble is unreliable.

For a perfect ensemble, two times the ensemble variance should equal the MSE as proved in Appendix B. From Fig. 9, it must be concluded that the EPS is underdispersive. Although this is a well-known fact, other factors contribute here to the discrepancy between MSE and ensemble variance. These factors include the representativity and observational errors. The former factor can be estimated using analysis data as done in Vannitsem and Hagedorn (2010). The ensemble variance converges to the observation variance σ_O^2 at long lead times since then the ensemble members are dispersed over the attractor while we know that $\sigma_C = \sigma_O$ (as shown in section 3.2).

The aforementioned underdispersiveness of the EPS can be corrected by a post-processing approach which simply adds Gaussian noise to the ensemble members in order to account for the representativity and observational error at the level of the station. For reasons outlined in Van Schaeybroeck and Vannitsem (2011) and Vannitsem and Hagedorn we prefer not to adapt such an approach.

7 Selected model-based predictors

Based on the predictors selected when performing the validation, we have extracted the most frequently-used predictors. Except for the corresponding variables, by far the most often used predictors are boundary layer height (for 10m U), East-west surface stress (for 10m U), 10 U wind (for 10m V), North-South surface stress (for 10m V), temperature at 925hPa (for 2m T) and maximal 2m T in the last 6h (for 2m T). Other frequently-generated predictors are tabulated in Table 1. Note that the variable top thermal radiation is by far the most important predictor for the day-time (12h) two-meter temperature

Table 1: List of most frequently-selected predictors for stations in Belgium for three variables (2m T, 10m U-wind and 10m V-wind). Note that only the predictors used for correcting the night-time (0h) forecast of the 2m T are listed.

	10m U wind	10m V wind	2m T
10m U wind	X	X	X
Boundary Layer Height	X		
100 U wind	X	X	
East-West surface stress	X		
V Velocity (925hPa)	X		
(10m U wind)^3	X		
10m V wind		X	
100 V wind		X	
North-South surface stress	X	X	
U Velocity (500 hPa, 850 hPa, 925hPa)		X	
(10m U wind)^2		X	
2 meter temperature			X
Max. 2m T in last 6hs			X
Temperature (500 hPa, 850 hPa, 925hPa)			X
Soil temp. level 1,2			X

forecast but since this does not result in forecast improvement (see Fig. 5), we have excluded all concerned predictors from Table 1. Although most predictors are common to all locations, some are not. For example, U velocity at 925hPa is only regularly chosen as a good predictor for the 10m U wind at Ukkel, and finally (10m U wind)^2 for 10m V as well as (10m U wind)^3 for 10m U at St-Hubert. Note that for different seasons we expect to find a slightly different predictor set.

8 Conclusions

We have validated possible procedures to implement an operational post-processing of 10-meter wind and two-meter temperature for ensemble forecasts. The training method, EVMOS, is a deterministic regression method suited for ensemble forecasts and is combined with a predictor-selection procedure to avoid overfitting. From the validation we conclude that a strong MSE reduction is obtainable for two-meter temperature as well as for the ten-meter wind. The best results emerge when applying centered training, that is, for a training period centered around the date of forecast. The ideal training period in that

case seems to be thirteen week; training with fewer weeks give rise to less stable regression while adding weeks is also counterproductive due to the loss of essential seasonal information in the error statistics. For both wind components, good results are obtained by applying the three-predictor EVMOS regression with clustering of lead time data per day. On the other hand, a simple bias-correction forecast was optimal for predicting the daily 12h two-meter temperature. Although mainly correcting systematic errors, additional improvements are obtained with EVMOS when forecasting the 2m T at night-time (0h). The number of predictors integrated in the EVMOS regression seems not to be of crucial importance as long as it is more than one and smaller than five. All these conclusions are valid both for the inland as well as for the coastal stations. However, post-processing at coastal stations only yields appreciable and consistent gain for temperature and zonal wind but not for the meridional wind. The decrease of corrections at the coastal stations may be caused by either initial condition errors which are of non-systematic nature, or to model errors uncorrelated with the predictors. The latter express the failure of the model to accurately represent physical processes at the coastal zone such as sea breeze. Note that our validation was fully based on the hindcast data set and validation using the EPS data would be useful although similar conclusions are expected.

We have also post-processed surface pressure forecasts. Due to its strong sensitivity with respect to height differences, a simple bias correction for stations inland yields a relative MSE improvement of more than 95%. Corrections additional to the bias corrections were only noteworthy in the coastal zone and were of the order of 10% of the original MSE.

In addition to the hindcast data, the YOTC database[2] present at ECMWF was utilized for post-processing during the years 2008 and 2009. The YOTC database was chosen because it includes dynamical tendencies of common meteorological variables and we wanted to test the hypothesis that these are particularly good model predictors. However, this was found not to be the case. Also available in the YOTC data set was the instantaneous surface stress and friction velocity, the fraction of which turned out to be a very powerful predictor for wind speed. Unfortunately both predictors are unavailable in the hindcast data set.

Based on the validation results, we propose as an operational implementation at RMI the use of EVMOS post-processing for the wind with three predictors, at least seven weeks of training and clustering of the daily data. For the temperature, on the other hand, we suggest a combination of bias-correction for the 12h forecasts, and, for the 0h forecasts, a three-predictor EVMOS method, both using at least five training weeks and no clustering of lead times per day. It is also concluded that the EPS provides a good representation of the mean surface wind and the surface temperature variability, at least at the scale of the model.

In order to render the post-processing operational at the RMI we propose a weekly training each Thursday upon the arrival of the ECMWF hindcast data. The regression coefficients along with the concomitant predictors thus obtained must be stored and, for the entire following week, reloaded twice daily for post-processing the disseminated 51-member EPS. Although we have confined our validation up to a lead time of seven days we are planning to post-process the EPS with a forecast horizon of fifteen days, as is allowed by the hindcast data set. In addition to the 2m T and the 10m wind,

[2]The Year of Tropical Convection (YOTC) refers to the project conducting a year of coordinated observing, modeling and forecasting of organized tropical convection. See also http://www.ucar.edu/yotc.

we plan the correction of minimal and maximal daily temperatures and, in the long run, a (preferentially statistical) post-processing procedure for precipitation (Roulin and Vannitsem, 2011).

Note that at the present stage there is no analog to the EPS hindcast for the deterministic IFS forecast at ECMWF even though it would be useful for post-processing purposes.

9 Acknowledgement

We acknowledge the kind assistance of Hans Van Hauteghem, Liliane Frappez and Dominique Lucas for obtaining the required data. Discussions with Pascal Mailier, Emmanuel Roulin and Alex Deckmyn and their comments on the manuscript are highly appreciated.

A Appendix: Details of verification

To illustrate our verification method, we represent in Fig. 10 eighteen years of hindcast data sets by squares, all at one fixed station and for one fixed lead time. In case clustering of the daily data is used, each square represents two ensembles of five members each, while it represents only one ensemble in case of no clustering. In the example shown, we employ a training period of seven weeks and final verification is done using central training which means that the data set used for verification includes the eighteen squares of week 0. Note that in this appendix we denote by a "point" all data associated with one square; different members of one ensemble are never used for training and verification separately due to their possible correlation and hence they are grouped.

We use cross-validation which involves a resampling of the hindcast set by a subdivision into a verification set and a trainings set. In practice we do this by selecting one out of the eighteen data points from the data set used for final verification; we indicate this verification point by a yellow color in Fig. 10. All but this data point will be used for training when performing the final verification; however, prior to performing the final verification, we need to determine the most ideal predictor set. As specified in Sect. 4 our predictor-selection method involves another resampling procedure. In our example we select one year (row) of hindcast data indicated by the green color in Fig. 10 which will be used for verification of the predictor-selection method. Given a certain predictor set, a training is then performed with all blue data points and verification is done by calculating an average MSE of the green points. Moreover, this procedure repeated by iterating the green data set over all remaining seventeen years (the row of the yellow point is excluded). Each predictor set is then given a total MSE score which is then an average over the seventeen times seven data points. The set associated with the lowest score is used to validate the yellow point; this is done by a new training using all blue plus green data points and determining the MSE of the yellow point. The above procedure is then repeated by iterating over all eighteen possible yellow points, the average score of which determines our MSE at a certain lead time and station and for one validation week. Finally the training as specified above is repeated for fourteen different validation weeks and the scores presented in our plots are averages over these weeks and over different stations.

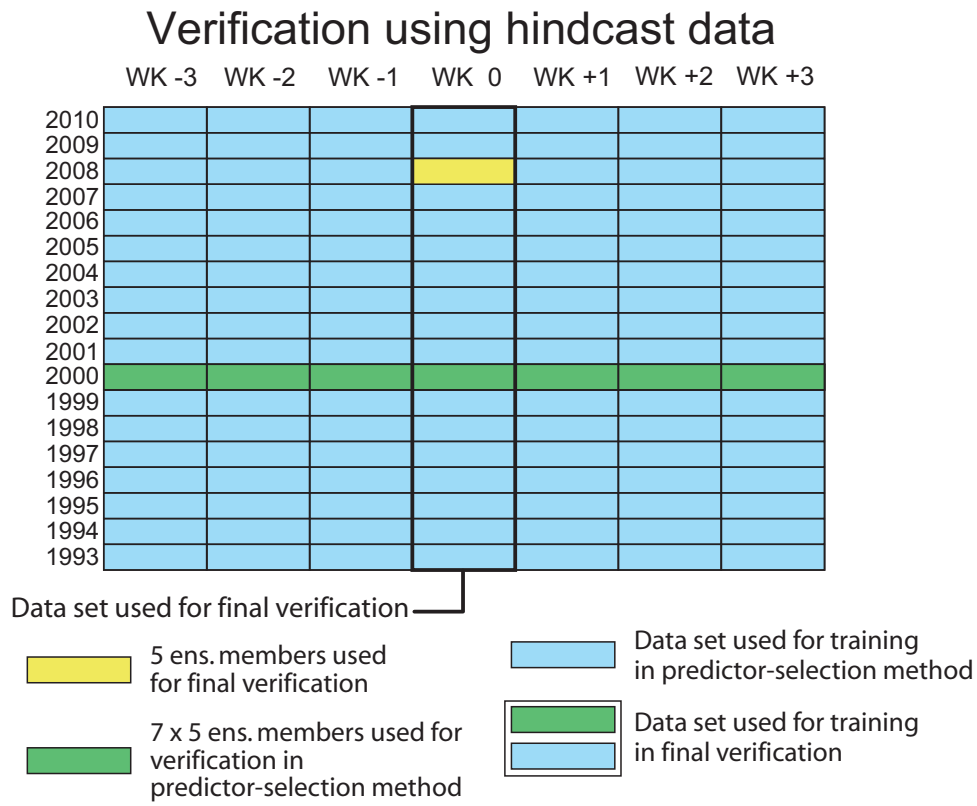

Figure 10: Representation of seven weeks (columns) and the eighteen years (rows) of hindcast data for a fixed station and a fixed lead time. Each square represents one or two ensembles of five members, depending on whether or not clustering of the daily forecast data is used. The different colors indicate different subsets which are used for our validation approach. The dataset for final verification is the one of week zero (WK 0) since central training is used. In case of forward training, this set would be the of week three. Note that for our final verification we use fourteen such sets and we average over different stations.

For determining the score of one predictand using three predictors at a fixed lead time and for all stations inland the associated resampling number is therefore roughly 14 (weeks) x 7 (locations) x 18 (verification resamplings) x 80 x 79 (predictors) x 17 (predictor-selection resamplings) ($\approx 2 \times 10^8$) resamplings. A final MSE score for stations inland at fixed lead time ℓ (and in case of clustering) is therefore determined by the following average:

$$\mathrm{MSE}_{\mathrm{Inland},\ell} \propto \sum_{\text{Week } w=1}^{14} \sum_{\text{Station } s=1}^{7} \sum_{\text{Year } y=1}^{18} \sum_{\text{Ens. } e=1}^{2} \sum_{\text{Ens. mbr. } m=1}^{5} \mathrm{MSE}_{w,s,y,e,m,\ell}.$$

Here each $\mathrm{MSE}_{w,s,y,e,m,\ell}$ is determined after the predictor-selection method. The determination of the bias-corrected forecast, on the other hand, involves only 14 (weeks) x 7 (locations) x 18 (verification resamplings) (≈ 1800) resamplings.

B Appendix: Perfect ensemble

Here we prove that for a perfect ensemble the mean square error should be equal to two times the ensemble variance. The MSE is defined as $\langle (X - X_{_O})^2 \rangle$ with X the forecast and $X_{_O}$ the observed value. By adding and subtracting the ensemble mean $\ll X \gg_e$ to $X - X_{_O}$, we arrive at:

$$\mathrm{MSE} = \langle (X - \ll X \gg_e)^2 \rangle + \langle (X_{_O} - \ll X \gg_e)^2 \rangle - 2 \langle (X - \ll X \gg_e)(X_{_O} - \ll X \gg_e) \rangle. \tag{5}$$

The first two terms can be identified as the ensemble variance and the mean-square error of the ensemble mean. The last term is zero which can be obtained by recognizing that the average $\langle \cdot \rangle$ is twofold; an average over all ensembles $\ll \cdot \gg_a$ and an average over all ensemble members $\ll \cdot \gg_e$. Since the observation $X_{_O}$ is the same for all ensemble members of the same ensemble the last terms of Eq. (5) can be rewritten as:

$$-2 \ll (\ll X - \ll X \gg_e \gg_e)(X_{_O} - \ll X \gg_e) \gg_a .$$

Clearly the first term in brackets is zero. In a perfect ensemble the observation $X_{_O}$ is sampled from the same distribution as the ensemble members X and therefore the first and second term of Eq. (5) will be equal. We have therefore proven that the MSE of a perfect ensemble should be twice the ensemble variance or also twice the mean-square error of the ensemble mean.

References

[1] Casella, G., and R.L. Berger, Statistical Inference. Brooks Cole Publishing, 1989.

[2] Glahn, H.R., and D.A. Lowry: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. J. Appl. Meteor., 11, 1203-1211, 1972.

[3] Glahn, B., Peroutka, M., Wiedenfeld, J., Wagner, J., Zylstra, G., Schuknecht, B., and Jackson, B.: MOS Uncertainty Estimates in an Ensemble Framework. Mon. Wea. Rev., 137, 246-268, 2009.

[4] Gneiting T, Raftery AE, Westveld AH III, Goldman T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review 133: 10981118,2005.

[5] Roulin, E., Vannitsem, S.: Post-processing of ensemble precipitation predictions with extended logistic regression based on hindcasts, Mon. Wea. Rev., To Appear, 2011.

[6] Unger, D. A., van den Dool, H., O'Lenic, E., and Collins, D.: Ensemble Regression, Mon. Wea. Rev., 137, 2365-2379, 2009.

[7] Van Huffel, S., and Vandewalle, J.: The total least-square problem: Computational aspects and analysis, SIAM, Philadelphia, 1991.

[8] Van den Dool, H.M.: Empirical Methods in Short-Term Climate Prediction, Oxford University Press, 2006.

[9] Vannitsem, S., and Nicolis, C.: Dynamical Properties of Model Output Statistics Forecasts. Mon. Wea. Rev., 136, 405-419, 2008.

[10] Vannitsem, S.: Dynamical Properties of MOS Forecasts: Analysis of the ECMWF Operational Forecasting System. Wea. Forecasting, 23, 1032-1043, 2008.

[11] Vannitsem S.: A unified linear Model Output Statistics scheme for both deterministic and ensemble forecasts, Quart. J. Roy. Meteorol. Soc., 135, 1801-1815, 2009.

[12] Vannitsem, S., and Hagedorn, R.: Ensemble forecast post-processing over Belgium: Comparison of deterministic-like and ensemble regression methods. In press in Meteorological Applications, 2010.

[13] Van Schaeybroeck, B., and Vannitsem, S.: Post-processing through linear regression, Nonlin. Processes Geophys., 18, 147-160, 2011.

[14] Wilks D.S.: Comparison of ensemble-MOS methods in the Lorenz 96 setting. Meteorol. Appl., 13, 243-256, 2006.

[15] Wilks, D.S.: Statistical methods in the atmospheric sciences. Academic Press, London, 1995.

ISSN nr. D2012/0224/061