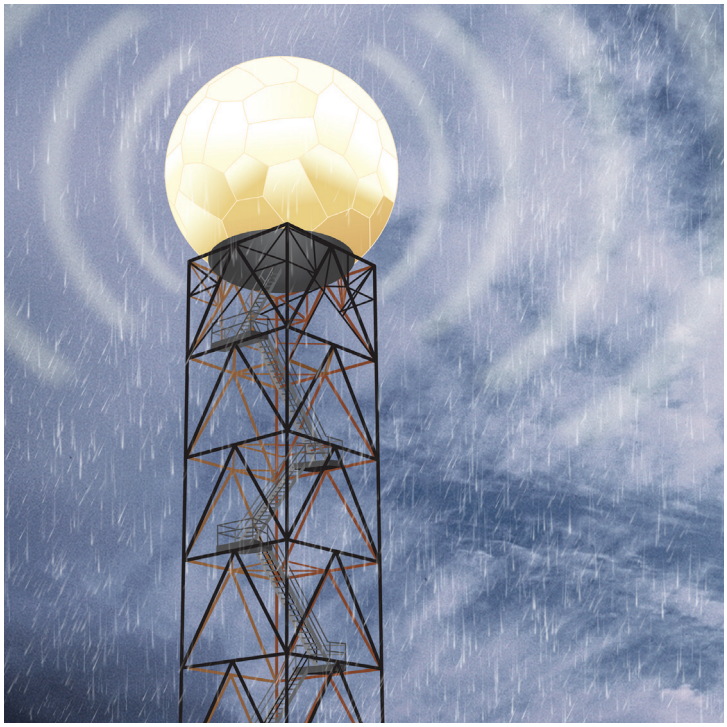
ECMWF Feature article

from Newsletter Number 117 – Autumn 2008

METEOROLOGY

Using the ECMWF reforecast dataset to calibrate EPS forecasts



www.ecmwf.int/en/about/news-centre/media-resources

doi:10.21957/tlk9s1tzzr

This article appeared in the Meteorology section of ECMWF Newsletter No. 117 – Autumn 2008, pp. 8-13.

Using the ECMWF reforecast dataset to calibrate EPS forecasts

Renate Hagedorn

With the unification of the ECMWF medium-range Ensemble Prediction System (EPS) and the Monthly Forecasting System on 11 March 2008 (see *ECMWF Newsletter No. 115*) a new reforecast dataset has become available for a variety of applications. A reforecast dataset is a collection of forecasts with start and prediction dates from the past, usually going back for a considerable number of years. In order to ensure consistency between reforecasts and actual forecasts, reforecasts are produced specifically with the same model system that is used to produce the actual forecasts. Before the unification of the medium-range and monthly forecast systems, reforecasts were only produced – and thus applicable – for the monthly forecast system. However, through the unification of both systems, it is now possible to use the reforecasts produced with the unified system for both the EPS and the monthly forecasts.

Originally, the reforecasts of the monthly forecast system were mainly used to determine the model climate and forecast anomalies with respect to this model climate. Now, with the reforecasts also being applicable to the medium-range EPS forecasts, new applications are possible. One of these new applications is the calibration of the medium-range EPS forecasts. Testing various calibration methods has shown that the forecasts can be significantly improved through calibration, in particular for near-surface weather parameters.

In this article we are going to discuss various questions related to calibration methods, their impact on the performance of the EPS, the added benefit of using reforecasts for calibration, and the design of the new operational reforecast dataset. Last but not least, we will make the case for ECMWF users to consider taking advantage of this new dataset, which we believe can be of enormous value for a variety of applications.

How do we apply calibration using reforecasts?

Calibration or more generally post-processing of uncalibrated Direct Model Output (DMO) is a well established technique. Many National Meteorological Services of ECMWF Member States apply this technique, also known as Model Output Statistics (MOS) or statistical adaptation, to ECMWF's DMO. A number of different calibration methods have been proposed for operational and research applications and a recent comparison of the main methods can be found in *Wilks & Hamill* (2007). Most calibration methods are based on the idea of correcting the current forecast by using past forecast errors. As such, they all require a so-called training dataset (a number of past forecast-observation pairs) to determine the optimal correction.

Until now, such post-processing activities have been mainly based on operationally available training datasets, which are either relatively short datasets or – if they cover longer times – are inconsistent datasets containing data from different model cycles or even different model resolutions. More recently it has been suggested that calibration can lead to even greater improvements if large datasets of consistent reforecasts are available and large operational weather forecast centres have been urged to provide such reforecasts (*Hamill et al.,* 2006). However, before embarking on such a reforecast programme it had to be examined whether the level of improvements, which had been demonstrated only for forecasts with relatively low quality, could also be achieved for the higher-quality ECMWF forecasts.

The reforecast dataset produced to investigate this question covers the period 1 September to 1 December, with one reforecast per week, i.e. 14 cases or start dates are available (01/09, 08/09,...,01/12). For each of these start dates, 20 reforecasts covering the years 1982–2001 are available. The reforecasts were produced with the model cycle and setup which was operational during September–December 2005 (Cy29r2, T255), except that the initial conditions were taken from ERA-40 reanalysis. Furthermore, the reforecast ensemble consists of only 15 members (1 control + 14 perturbed) instead of the operational set of 51 members. Ideally, the reforecast dataset should contain the same number of members as the real-time ensemble. However, since the production of such a full set of reforecasts seems not to be affordable in an operational setting, this option was not considered in this study – only the maximum affordable number of members were produced for this test reforecast dataset.

 doi:10.21957/tlk9s1tzzr

The first step in the calibration process is creating the training dataset. Two aspects have to be considered here: on the one hand it is desirable to have the largest possible number of training data available whilst on the other hand the training data should be as close as possible to the climate of the forecast date to be calibrated. Thus, the training dataset should be composed of reforecasts from a window centred around the date of the forecast to be calibrated. Figure 1 is a schematic showing how to compile the training dataset from the available reforecasts. The size of the window is determined by the minimal number of reforecasts needed for a reliable calibration. Window sizes of three, five, and seven weeks were tested, with five weeks turning out to be a reasonable size.

After creating the training dataset it needs to be decided which calibration method is most suitable for the specific purpose at hand. In this article we compare the results of two calibration methods:

- ***Linear Bias Correction (BC)*** – a very simple and computationally inexpensive method.
- ***Non-homogeneous Gaussian Regression (NGR)*** – a more advanced and computationally expensive method.

Whereas the BC method attempts to only correct a possible systematic shift of the ensemble mean, NGR also accounts for spread deficiencies. Further information on the calibration methods can be found in the Box A or in *Hagedorn et al.* (2008) and references therein.

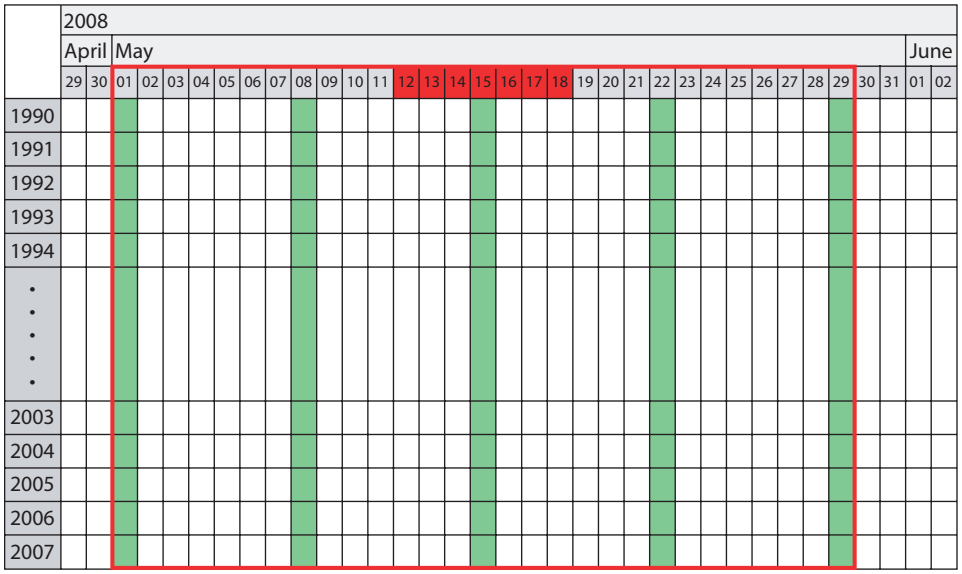


Figure 1 Schematic of available reforecast dataset and five-week window of reforecasts used as training dataset. The red frame indicates the time window used to compile the training dataset used for calibrating the forecasts started on the dates in the centre of the time window, also marked red. That is, the training dataset for calibrating the forecasts started between 12 and 18 May 2008 is composed of the reforecasts started on 1, 8, 15, 22 and 29 May, each date comprising the reforecasts from 1990–2007. The time window moves with the dates of the forecasts to be calibrated.

doi:10.21957/tlk9s1tzzr

Calibration methods

A

The In order to assess the different levels of improvements achievable with different calibration methods, two calibration methods have been tested.

Bias Correction

In this simplest calibration scheme, the long-term systematic error of the ensemble mean $b(x,t,l)$ is determined from the mean difference between the ensemble mean forecast $f(x,t,l)$ and the observations $o(x,t)$ in a training dataset:

$$b(x,t,l) = \frac{1}{N}\sum_{n=1}^{N} f_n(x,t,l) - o_n(x,t)$$

with: x the location, t the date of forecast, l the lead time and n the number of training cases (n = 1,..,N).

This long-term systematic error is then subtracted from each ensemble member of the forecast to be calibrated. Thus only the ensemble mean, but not the ensemble spread, is affected by this procedure.

Non-homogeneous Gaussian Regression

Non-homogeneous Gaussian Regression (NGR) is an extension to conventional linear regression. The basic idea is to construct a Probability Density Function (PDF) in the shape of a Gaussian, with mean and variance determined by a regression equation. The method is called "non-homogeneous" because the variance is allowed to be non-homogeneous, i.e. not the same for all values of the predictor. In this implementation of NGR, the mean forecast temperature and sample variance interpolated to the station location were predictors, and observed 2-metre temperature at station locations were the predictands. We assumed that stations had particular regional forecast biases sometimes distinct from those at nearby stations. Hence, the training did not composite the data. For example, the fitted parameters for London were determined only from London forecasts and not from a broader sample of locations around and including London.

To describe NGR more formally, let ~N (α, β) denote that a random variable has a Gaussian distribution with mean α and variance β. Let $\overline{x}_{\mathrm{ens}}$ denote the interpolated ensemble mean and s^2_{ens} denote the ensemble sample variance. Then NGR estimated regression coefficients a, b, c and d so as to fit:

$$N\left(a + b\,\overline{x}_{\mathrm{ens}}, c + d\,s^2_{ens}\right)$$

When d = 0, no spread-error relationship is in the ensemble, and the resulting distribution resembles the form of linear regression with its constant-variance assumption. The four coefficients are fitted iteratively by minimizing the Continuous Ranked Probability Score.

What is the impact of calibrating the EPS?

The first issue to be addressed is the level of improvement that can be achieved when applying the different calibration methods. In other words, what is the impact of calibrating the EPS?

It is well known that in general the greatest impact of calibration can be seen in near-surface weather parameters since model deficiencies are most important for these (*Hamill & Whitaker,* 2007). Therefore our evaluation focuses on comparing the performance of the 2-metre temperature forecast of the uncalibrated DMO with calibrated forecasts at 250 European stations (see Figure 4 for the locations of the stations).

In order to evaluate the gridded model forecasts at irregularly spaced station location, the model forecasts were interpolated onto these stations. The main performance measure is the Continuous Ranked Probability Skill Score (CRPSS), since the CRPSS gives a good general assessment of the probabilistic forecast performance by taking into account the whole range of possible events to be forecast. A perfect forecast is assigned a skill score of 1, and a CRPSS below 0 characterizes a forecast system with less skill than the reference forecasts which here is chosen to be climatology.

Figure 2 compares the CRPSS, calculated over all 250 stations and all forecasts from 1 September to 30 November 2005, for the Direct Model Output, the Bias Corrected forecasts and the NGR calibrated forecasts. It is evident that both calibration methods significantly improve the performance of the uncalibrated model. For example, the performance of the Direct Model Output at 1-day lead time is at the same level as the performance of a 4–5 day calibrated forecast, i.e. through calibration a gain in lead time of 3–4 days can be achieved. For longer forecast lead times this gain is still around two days. When comparing the performance of the two different calibration methods it becomes clear that, particularly for early lead times, the NGR calibrated forecasts are better than purely Bias Corrected forecasts. In general, NGR can improve on Bias Corrected forecasts by two days early in the forecast range and about half a day later in the forecast range.

doi:10.21957/tlk9s1tzzr

What is the reason for the improvements in the calibrated EPS?

It is of interest to analyse the reasons for the improvements achieved by the calibration procedures. Analysing the root mean square (rms) errors and spread of the different forecasts (Figure 3) gives insight into what is happening during the calibration process. First of all, both calibration methods, BC and NGR, reduce the rms error significantly. The reduction is virtually the same for both methods, with the red and blue lines hardly being distinguishable. However, by considering additionally the changes in the spread of the forecasts it becomes clear why the NGR calibrated forecasts are improved even more compared to the Bias Corrected forecasts. It is evident that the spread of the uncalibrated DMO is much too low. Since the BC procedure does not affect the spread of the DMO, the blue and black lines are identical. In contrast, the spread of the NGR calibrated forecasts is much improved, with the spread now matching the rms error more closely. As the spread deficiency is particularly evident in the early forecast range, the NGR calibration can significantly improve the DMO over and above the BC calibration, especially at these lead times.

Figure 2 and 3 gave an overall assessment of the performance improvements for all 250 stations. However, it is also interesting to investigate the impact of the calibration at individual stations. Figure 4 gives this information by showing the CRPSS of the two-day forecasts of the Direct Model Output at individual stations (Figure 4a) and the difference in the CRPSS between NGR calibrated and uncalibrated forecasts (Figure 4b). In general, the CRPSS of the uncalibrated forecasts ranges between 0.3 and 0.7; however, there are some stations with quite low and even negative CRPSS. These stations are located mainly in areas of inhomogeneous terrain such as coastal or mountainous areas, where simple interpolation methods from gridded model forecasts to station locations are not sufficient, even when taking into account different land-sea masks etc. Obviously, at such locations calibration can be of particular value, and in fact it is the case that the differences between NGR and DMO forecasts are especially positive at these stations (Figure 4b). That is, at locations with already a fairly good performance in the uncalibrated forecasts, only moderate improvements around 0.1 can be achieved. However, in cases with particularly bad performance in the uncalibrated forecasts, calibration can achieve improvements of more than 0.2 in the CRPSS.

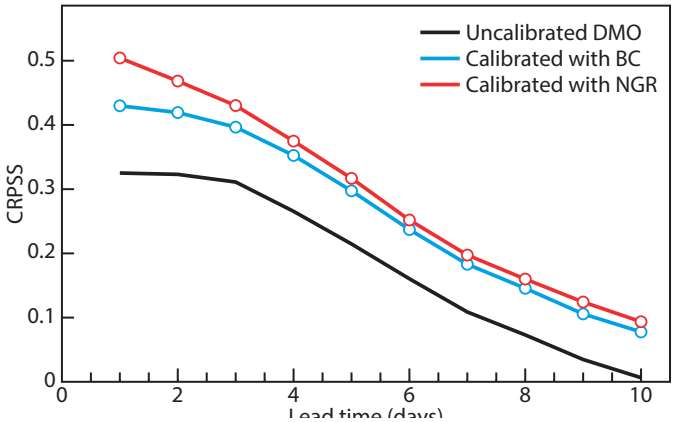


Figure 2 Continuous Ranked Probability Skill Scores of 2-metre temperature predictions at 250 European stations and for 91 cases (1 September to 30 November 2005) versus lead time. *Black line:* uncalibrated Direct Model Output. *Blue line:* calibrated predictions using the BC method. *Red line:* calibrated predictions using the NGR method.

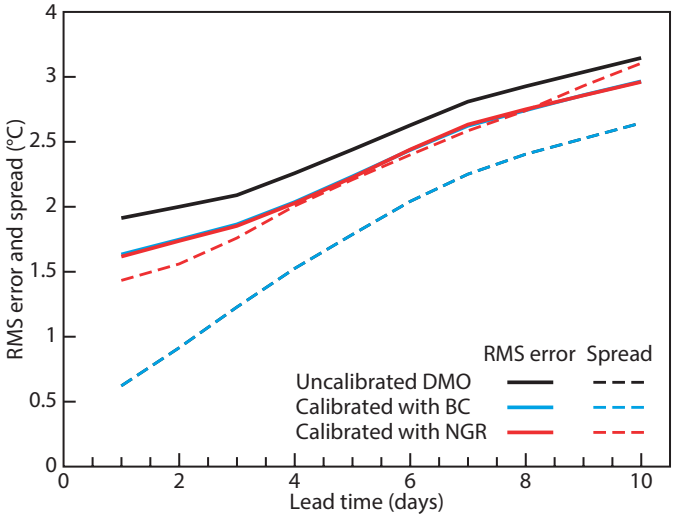


Figure 3 RMS error (solid lines) and spread (dashed lines) of 2-metre temperature predictions at 250 European stations and for 91 cases (1 September to 30 November 2005) versus lead time. *Black lines:* uncalibrated Direct Model Output. *Blue lines:* calibrated predictions using the BC method. *Red lines:* calibrated predictions using the NGR method.

doi:10.21957/tlk9s1tzzr

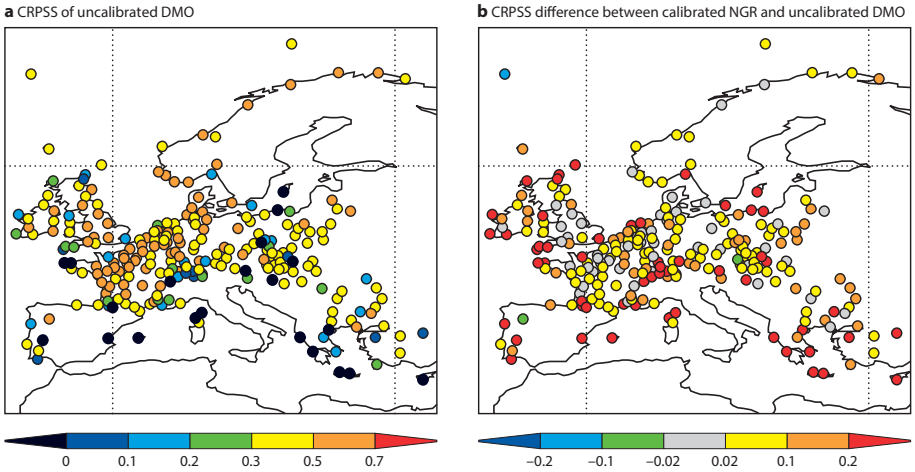


Figure 4 Continuous Ranked Probability Skill Scores of 2-metre temperature predictions at single locations, averaged over 91 cases (1 September to 30 November 2005). (a) CRPSS of uncalibrated Direct Model Output. (b) Differences between the CRPSS of the calibrated NGR and uncalibrated DMO forecasts; positive values indicate improvements by the calibration.

What is the added benefit of using the reforecast dataset?

All the calibration results shown so far were based on using reforecasts as the training dataset. However, one could certainly ask the question whether these improvements also could have been achieved by using operational forecasts from say the previous 30 days. In other words, is there really an added benefit of using a reforecast dataset?

The comparison of the performance achieved by NGR calibration using reforecasts versus the last 30 days of operational forecasts as training dataset demonstrates the level of added improvement using the reforecast dataset (Figure 5). In the early forecast range the calibration using operational forecasts as training dataset can improve the DMO nearly as much as the calibration using reforecasts. For the later forecast range, however, its performance is much worse and the calibration is no longer able to improve significantly on the uncalibrated forecasts.

So why is using the reforecast dataset particularly helpful for longer lead times? It is suggested that there are at least three contributing factors. First, the prior 30-day training data set was 9 days older for a 10-day forecast (training days –39 to –10) than for a 1-day forecast (training days –30 to –1). If errors were synoptically dependent and a regime change took place in the intervening 9 days, the training set at 1-day lead will include samples from the new regime while the training set at 10-days lead will not. The second reason might be due to the fact that at long leads, the proportion of the error attributable to bias shrinks due to the rapid increase of errors due to chaotic error growth. Consequently, as the overall error grows and a larger proportion is attributable to random errors, determining the bias requires a bigger sample. The third reason could be related to the fact that for the operational training data the short-lead forecasts tend to have more independent errors than the longer-lead forecasts. By contrast, the reforecast dataset, being produced only once a week, should be comprised of truly independent samples.

How is the new operational reforecast dataset designed?

Another question which had to be answered when setting up the operational production of this new reforecast dataset concerned the optimal design of this dataset, i.e. what is the best compromise in terms of costs and benefits? Decisions to be made included: “How many ensemble members are necessary and can we afford?” and “How many years should be included?” In order to answer such questions, some experiments were carried out comparing the performance of the calibration using reduced/increased reforecast datasets (Figure 6).

Increasing the number of ensemble members from 5 to 15 only adds significant benefits at longer lead times (Figure 6a). By contrast, reducing the number of available reforecast years in the training dataset from 20 to 12 reduces the performance of the calibration both in the later and earlier forecast ranges (Figure 6b). Taking into account these results, the new operational reforecast dataset comprises 5 ensemble members (1 control + 4 perturbed) and produces reforecasts for the past 18 years (currently 1990 to 2007).

First results using these operational reforecasts to calibrate most recent EPS forecasts for April to June 2008 confirm that the level of improvements actually achieved is similar to the results of the experimental calibration of the September to November 2005 forecasts. Figure 7 shows the CRPSS for the uncalibrated DMO, Bias Corrected and NGR calibrated forecasts, i.e. displays the results corresponding to Figure 1, but here for the most recent period and using the operational reforecasts as training dataset.

doi:10.21957/tlk9s1tzzr

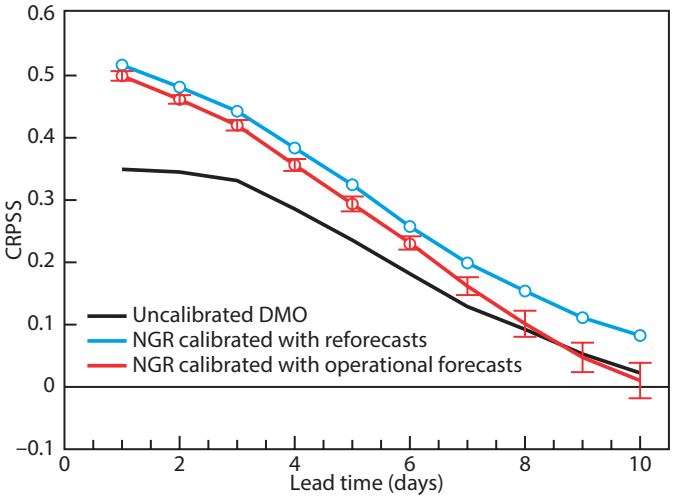


Figure 5 Continuous Ranked Probability Skill Scores of 2-metre temperature predictions at 250 European stations and for 91 cases (1 September to 30 November 2005) versus lead time. *Black line:* uncalibrated Direct Model Output. *Blue line:* NGR calibrated predictions using reforecasts as training dataset. *Red line:* NGR calibrated predictions using the last 30 days of operational forecasts as training dataset. Significance levels (0.05) of the calibration results using operational forecasts as training dataset, with respect to the calibration results using reforecasts, are denoted by the red vertical bars.

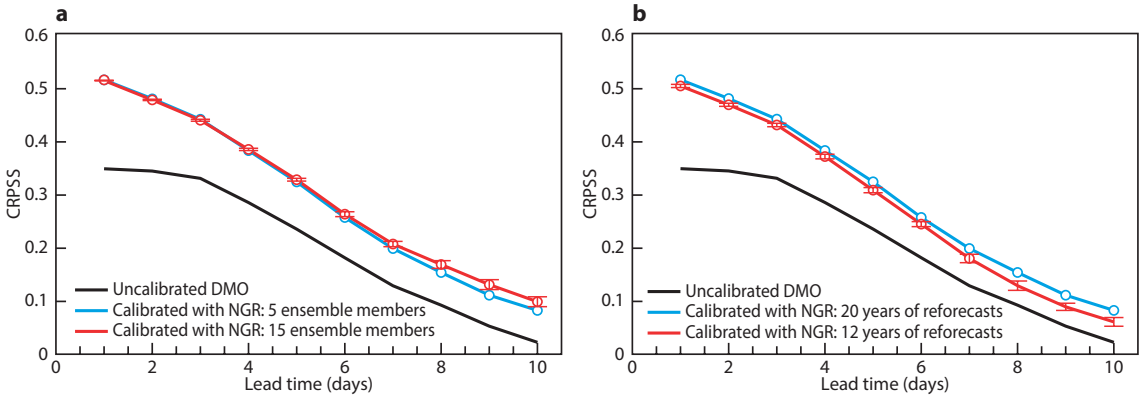


Figure 6 Continuous Ranked Probability Skill Scores of 2-metre temperature predictions at 250 European stations and for 91 cases (1 September to 30 November 2005) versus lead time. (a) Impact of increased number of ensemble members in reforecast dataset. *Black line:* uncalibrated Direct Model Output. *Blue line:* NGR calibrated predictions using only 5 enemble members of the reforecasts as training dataset. *Red line:* NGR calibrated predictions using all 15 members of the reforecasts as training dataset. Significance levels (0.05) of the calibration results using 15 members as training dataset, with respect to the calibration results using 5 members, are denoted by the red vertical bars. (b) Impact of reduced number of years. *Black line:* uncalibrated Direct Model Output. *Blue line:* NGR calibrated predictions using all 20 years of the reforecasts as training dataset. *Red line:* NGR calibrated predictions using only 12 years of the reforecasts as training dataset. Significance levels (0.05) of the calibration results using 12 years as training dataset, with respect to the calibration results using 20 years, are denoted by the red vertical bars.

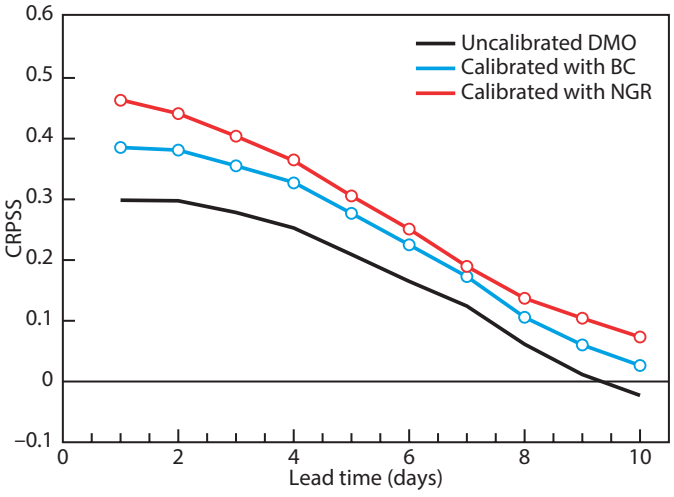


Figure 7 Continuous Ranked Probability Skill Scores of 2-metre temperature predictions at 250 European stations and for 91 cases (1 April to 30 June 2008) versus lead time. *Black line:* uncalibrated Direct Model Output. *Blue line:* calibrated predictions using the BC method. *Red line:* calibrated predictions using the NGR method. Both calibration methods use the new operational reforecasts as training dataset.

doi:10.21957/tlk9s1tzzr

Who should use ECMWF's operational reforecasts?

The new set of operationally available reforecasts is opening the way to a number of applications such as site specific calibration of weather parameters, global calibration of near-surface and upper-air fields, regime dependent calibration and calibration of parameters important for specific customers. The variety of applications in itself would probably demand a variety of calibration methods, which are best developed by ECMWF Member States. Individual users will have their own requirements and observational datasets, and we encourage them to take full advantage of the new dataset for their specific purposes. However, a common set of calibrated products for the more standard applications could also be made available by ECMWF, should the users require so.

Apart from using the reforecasts for calibration purposes, there are also a number of other possible applications. For example, the reforecast dataset can be used for diagnostic studies including monitoring model performance and consistent assessment of variations in spread from year to year.

Another application is using the reforecast dataset in the context of ECMWF's activities on the Extreme Forecast Index (EFI). To determine the EFI, a reliable assessment of the model climate is necessary. Before the introduction of the operational reforecast dataset, the EFI climate was determined by running a 2-day forecast of the EPS-control every day for the last 30 years. Now the EFI is based on the model climate determined from the reforecast dataset. This has the advantage that the model climate can now be determined with a lead-time dependence for the whole forecast range and not only for the first two days of the forecast. Furthermore, the information added by having available five ensemble members instead of only the control also seems to be beneficial. These two advantages outweigh the slight disadvantage that the new operational reforecasts are produced only for the last 18 years and only once a week.

In summary, we hope that the new operational reforecast dataset will be useful not only directly for calibrating the ECMWF EPS forecasts, but also for a whole range of other possible applications.

Further Reading

Hagedorn, R., T.M. Hamill & **J.S. Whitaker,** 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble forecasts. Part I: 2-metre temperature. *Mon. Wea. Rev.,* **136,** 2608–2619.

Hamill, T.M., J.S. Whitaker & **S.L. Mullen,** 2006: Reforecasts – An important dataset for improving weather predictions. *Bull. Am. Meteorol. Soc.,* **87,** 1–33.

Hamill, T.M. & **J.S. Whitaker,** 2007: Ensemble calibration of 500 hPa geopotential height and 850 hPa and 2-metre temperatures using reforecasts. *Mon. Wea. Rev.,* **135,** 3273–3280.

Wilks, D.S. & **T.M. Hamill,** 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.,* **135,** 2379–2390.

 doi:10.21957/tlk9s1tzzr