ECMWF Feature article

from Newsletter Number 142 – Winter 2014/15

METEOROLOGY

Calibration of ECMWF forecasts

Based on an image from mrgao/iStock/Thinkstock

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

doi:10.21957/45t3o8fj

This article appeared in the Meteorology section of ECMWF Newsletter No. 142 – Winter 2014/15, pp. 12–16.

Calibration of ECMWF forecasts

David Richardson, Stephan Hemri, Konrad Bogner, Tilmann Gneiting,
Thomas Haiden, Florian Pappenberger, Michael Scheuerer

ECMWF has studied the benefits of calibrating the ECMWF medium-range forecasts, based on statistical post-processing, to improve probabilistic predictions of four near-surface weather parameters. The motivation was the expert review of calibration methods carried out for ECMWF by Prof Tilmann Gneiting, who has recently been appointed as one of the inaugural ECMWF Fellows. The study was carried out in collaboration with Prof Gneiting and members of his Group on Computational Statistics at the Heidelberg Institute for Theoretical Studies (HITS).

The aim of the study was to demonstrate the benefits of using state-of-the-art calibration for the ECMWF forecasts, including an objective approach to combine the various components of ECMWF's forecast. It was found that calibration can provide substantial additional skill compared to the raw ECMWF forecasts.

Data and method

Calibration was carried out for four surface parameters.

- *T2M*: 2-metre temperature
- *PPT24*: 24-hour accumulated precipitation
- *V10*: near-surface wind speed
- *TCC*: total cloud cover

Synoptic observations (SYNOP) from a large number of stations across the globe were used for verification. SYNOP stations with suspicious data or significant missing data were excluded from the study. With these stations removed, around 4,000 stations for T2M and V10 and 3,000 stations for PPT24 and TCC were used in the study. Observations were used for 12 UTC only.

The ECMWF forecast was considered as a 52-member ensemble comprising the high-resolution forecast (HRES), the ensemble control (CTRL), and the ensemble forecast (ENS) consisting of 50 perturbed members. Operational forecasts were used from 12 UTC for the period 1 January 2002 to 20 March 2014.

The performance of the forecasts was measured using the continuous ranked probability score (CRPS). The CRPS is negatively oriented – lower scores indicate better forecasts, with a lower limit of zero for perfect forecasts. CRPS is a widely used measure of performance for probabilistic forecasts, and the ECMWF headline scores for the ensemble probabilistic forecasts of 850 hPa temperature and precipitation use the CRPS.

The aim of the calibration is to generate a probabilistic forecast with lower CRPS than the raw forecasts. A reduction in the CRPS indicates that the calibrated forecasts provide more skill value than the raw ensemble for the individual stations. During preliminary work a number of calibration methods were tested. For each parameter it was found that the best results were obtained using the method known as ensemble model output statistics (EMOS). This is a technique that converts a raw ensemble of discrete forecasts into a continuous probability distribution – see Box A.

AFFILIATIONS
David Richardson, Konrad Bogner, Thomas Haiden, Florian Pappenberger: ECMWF, Reading, UK
Stephan Hemri, Tilmann Gneiting: Heidelberg Institute for Theoretical Studies, Germany
Michael Scheuerer: University of Heidelberg, Germany; now at NOAA/ESRL, Boulder, USA

doi:10.21957/45t3o8fj

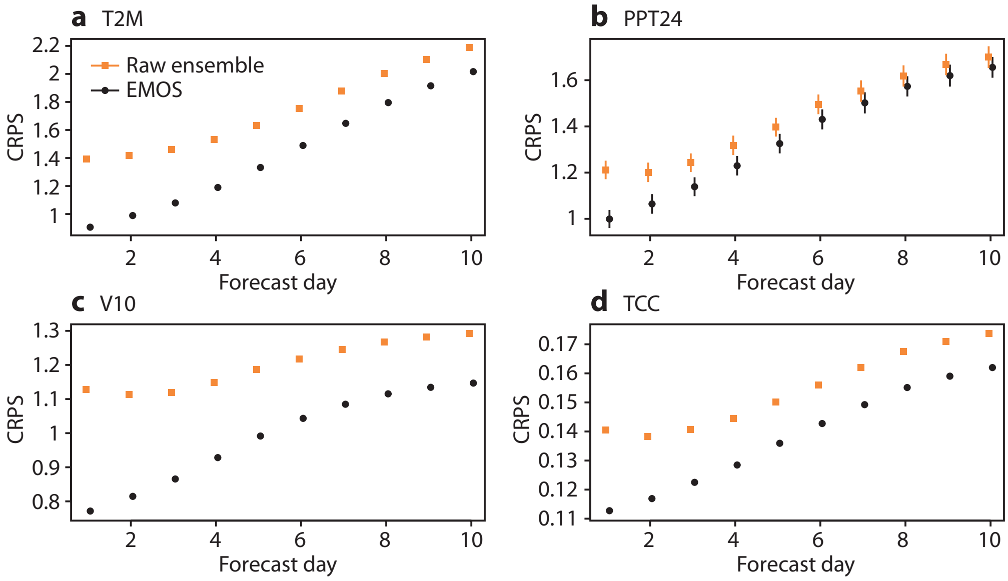

Figure 1 Mean CRPS for raw ensemble and calibrated ensemble for forecast lead times of one to ten days over whole verification period for European stations for (a) T2M, (b) PPT24, (c) V10 and (d) TCC. The vertical bars correspond to 90% confidence intervals for the expected average CRPS over all stations in the European subset (these bars are only large enough to show in panel b).

Overall impact of calibration

We first compare the mean CRPS values over the entire verification period for each of the calibrated parameters. Figure 1 shows the results averaged over European stations: the benefit of the EMOS calibration can be seen throughout the 10-day forecast range. For T2M the calibration brings a lead-time gain of around two days; for example the CRPS of the calibrated T2M forecast at day 6 is approximately the same as that of the 4-day raw forecast. The same lead-time gain is obtained for the TCC forecasts while the improvement for V10 is even larger. PPT24 shows the smallest benefit from the calibration, although there is still a one day gain or more in CRPS at all forecast lead times.

To put these results into some context, the overall increase in performance of the ECMWF forecasting system due to (a) model developments and (b) improved availability and use of data is typically one day per decade. In other words, the calibration brings similar gains in skill for forecasts at specific locations as is achieved for the basic atmospheric fields with 10–20 years of development of the Integrated Forecasting System (IFS). As we show in a later section, as the IFS has improved so has the skill of the calibrated forecasts. This shows that the modelling improvements and the calibration are complementary, both contributing to the overall skill of the final point forecasts.

Geographical variation of results

We now investigate how the effect of the calibration varies between stations. It should be noted that the selection of the best calibration method (i.e. EMOS) and training period was made based on results from the European stations, and may not be optimal for other regions.

Figure 2 shows the percentage change in CRPS at all evaluated stations for forecast days 5 and 10 for T2M. CRPS is improved significantly for almost all stations at lead times up to five days. Beyond day 5, there is an increasing number of stations for which CRPS cannot be improved significantly by calibration. Nevertheless, even for the 10-day forecasts the majority of stations show a performance improvement. There are only four out of over 4,000 stations at which CRPS deteriorates.

As for temperature, calibration significantly improves the CRPS of PPT24 for the vast majority of stations. With increasing forecast lead time, there is a growing number of stations, especially in North Africa, on the Arabian Peninsula and in central Asia, where there is no significant difference in CRPS between the raw ensemble and the calibrated forecast. However, there are no stations at which calibration deteriorates the CRPS.

For V10, calibration improves the skill in terms of CRPS compared to the raw ensemble at almost all stations for all lead times. Even for the later steps, including day 10, there are very few stations at which CRPS is not significantly reduced by calibration, and there are none where this increases (i.e. worsens) the CRPS. This confirms the European results on the global scale – that the largest and most consistent impact of the calibration is achieved for the 10 m wind speed.

For TCC, calibration leads to better skill in terms of CRPS compared to the raw ensemble for the vast majority of stations. However, there are a few stations for which there is a deterioration in the forecast skill; further analysis has shown that this is probably due to problems in the numerical optimization procedure used in the calibration process. This problem should be resolvable. Generally the relative improvement in skill by calibration decreases with increasing lead time, but it remains significant even at a forecast range of 10 days.

Calibration using ensemble model output statistics (EMOS)

A

Calibration using EMOS converts a raw ensemble of discrete forecasts into a continuous probability distribution. The most appropriate distribution will be different for the different forecast parameters.

Temperature (T2M)

For T2M we use a normal density distribution with mean *m* and variance σ^2. In the original EMOS the mean of the forecast distribution is given by

$$m = a_1 f_{\mathrm{HRES}} + a_2 f_{\mathrm{CTRL}} + a_3 f_{\overline{\mathrm{ENS}}}$$

where the parameters a_1, a_2 and a_3 can be interpreted as the relative weights given to the HRES, CTRL and the set of ENS members. In the present study a variant of this approach is used to account for the seasonal cycle of T2M: the departures of the observed temperatures from the climatological mean are related to those of the forecasts. A regression model using a combination of sine and cosine functions is applied to both observations and forecasts over the training period.

The variance of the forecast distribution is

$$\sigma^2 = b_0 + b_1 s^2$$

where s^2 is computed as the standard deviation across all 52 members of the ECMWF forecast.

The five parameters a_1, a_2, a_3, b_0 and b_1 are estimated from a set of training data, separately for each observation station.

Precipitation (PPT24), wind speed (V10) and total cloud cover (TCC)

Different distributions are appropriate for the other surface variables used in the study. For PPT24 we used a left-censored (cut-off at zero) generalised extreme value (GEV) distribution, while for V10, the most appropriate choice was found to be a left-truncated (at zero) normal distribution applied to the square-root transformed variables. For more details see *Hemri et al.* (2014). A mixed approach was found to be best for total cloud cover: the model needs to be able to allocate probabilities for zero cloud or totally cloudy as well as a continuous range in between.

Model fitting

For each of the forecast variables the parameters of the relevant forecast distribution are estimated by minimising the CRPS over a training period *T*. The training period for each verification day consists of the *n* days preceding the initialisation date of the forecast. A number of different lengths of the training period were considered, using data for a subset of European stations. The best results were obtained for a training period of 720 days (2 years) for T2M, 365 days (1 year) for V10, and 1816 days (5 years) for PPT24 and TCC. In principle, longer training periods should give the most robust parameter estimates. However, the long training periods will almost all include model upgrades and sometimes changes to the ensemble configuration. Such changes may have an adverse effect on the parameter estimates.

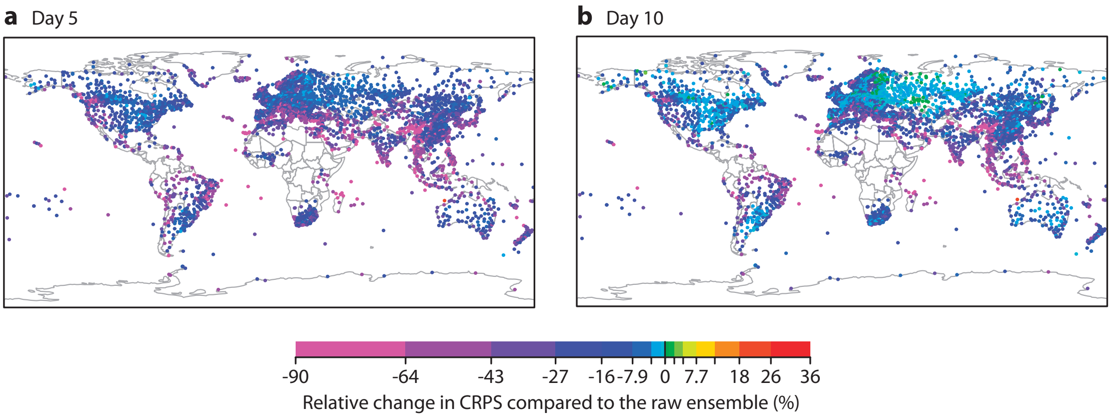

Figure 2 Relative change (%) in CRPS by EMOS compared to the raw ensemble at all stations for T2M for (a) day 5 and (b) day 10.

Trend in CRPS over time

The performance of the raw ensemble has changed significantly between the beginning and end of the 10-year verification period used in the study. The skill of the calibrated ensemble will also have changed as a result. In this section, we investigate whether the benefits of calibration decrease as the skill of the raw ensemble improves. Figure 3 shows how the percentage change in CRPS between the calibrated forecasts and the raw ensemble (over all European stations) has changed over time for T2M for the 5-day forecast. The plot shows selected quantiles of these differences: the median change is shown together with the 5%, 25%, 75% and 95% values; a temporal smoothing is applied to reduce the sampling variability. The distribution is not symmetrical about the median value – there are occasions where the calibration can result in very large improvements compared to the average change. However, there is no clear trend in these results: the benefits of calibration in terms of the percentage reduction in CRPS are about the same in 2014 as they were in 2004. This also applies for the 10-day forecast.

The results for PPT24 also show no strong overall trend. Both V10 and TCC show larger variations over time than T2M and PPT24, particularly for the lower quantiles. For example, calibration of V10 resulted in reduction of CRPS by up to 60% in 2008–2010, while maximum benefits are now closer to 35%. However, the median improvement has remained more constant over the years at around 10–15%. For TCC, there has been some increase over the years in the maximum benefit that the calibration can achieve.

For PPT24 and TCC there are some periods that show some increase in the number of cases where the calibration degrades the forecasts. This could be related to changes in the model. Some operational upgrades have introduced substantial changes to the model physics. It could be that the calibration using previous operational forecasts is no longer sufficient for the new model cycle, at least for some aspects of the forecast. However, further investigation (and a longer period of verification) would be needed to confirm this. Nevertheless, it is worth noting that the operational reforecasts (which always use the current model cycle) are designed explicitly to account for such model changes.

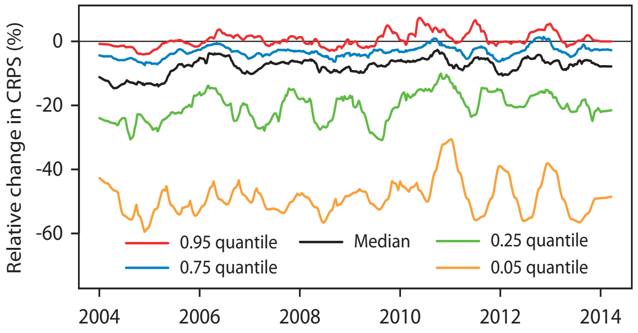

Figure 3 Change in CRPS by calibration compared to the raw ensemble against date for 5-day T2M forecasts. The lines correspond to a continuous smoothed box-plot showing the 0.05, 0.25, 0.5, 0.75, and 0.95 quantiles of the CRPS difference between the calibrated and the raw forecast among the European stations.

doi:10.21957/45t3o8fj

Weighting of HRES and CTRL

The calibration treats the HRES, CTRL and ENS members together as a 52-member ensemble. All 50 perturbed members are considered equally (all have the same weight), but the HRES and CTRL are allowed different weights. The preliminary tests, using a sample of European stations, assessed different basic configurations, for example excluding either the HRES or CTRL, or even excluding the ENS altogether and just using the HRES. Including the HRES together with the ENS was shown to give the best results, significantly improving the CRPS.

Overall, the HRES has a very high weight for the first few days. This decreases with increasing lead time, but even at day 10 the HRES is weighted significantly more than an individual ENS member. The CTRL has a much lower weight than the HRES, especially at shorter lead times. Although there is some variation between stations and parameters, the weight of the CTRL generally increases with forecast lead time and the CTRL has higher weight than an ENS member (greater than 1/51) for most forecast steps. If the HRES is not included in the calibration then the weight increases for the CTRL. This behaviour is illustrated by Figure 4, which shows the weights for T2M for two of the European stations: Vienna (representative of central Europe with modest terrain effects) and Skopje (in south-east Europe with more complex terrain).

The results for TCC are somewhat different from the other parameters. The HRES has lower weight and in particular the control forecast has decreasing weight as the forecast range increases (becoming less than 1/51 towards the end of the forecast). This would be consistent with TCC being the least predictable of the parameters being considered, and therefore having the most need for the full ensemble distribution.

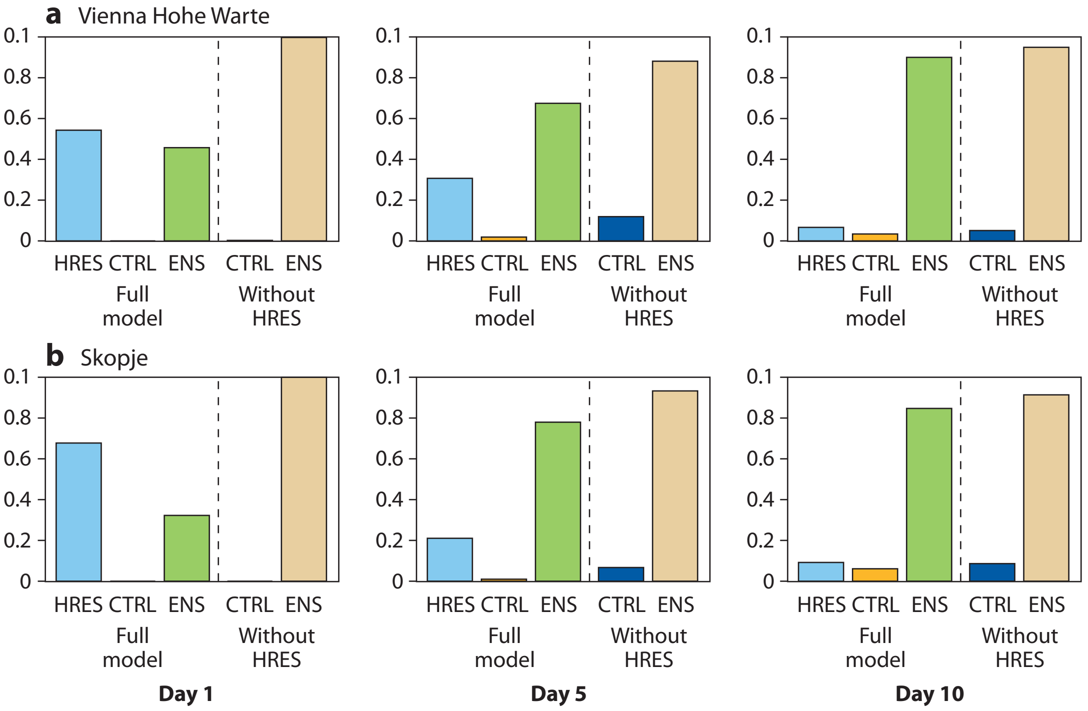

Figure 4 The left-hand side of each panel (labelled ‘Full model’) shows the weights assigned to HRES, CTRL, and ENS, respectively by EMOS for T2M at (a) Vienna Hohe Warte and (b) Skopje. The right-hand side of each panel (labelled ‘Without HRES’) shows the weights for CTRL and ENS when the HRES is not included in the calibration.

doi:10.21957/45t3o8fj

Use of reforecasts

The results in the previous sections used the traditional approach of training the calibration on a sliding window of previous operational forecasts. This has the drawback that it does not account for changes to the IFS: a calibration applied to a new model cycle based on results from a previous cycle may be inconsistent and could degrade the performance. Although the results show that overall the benefits outweigh these disadvantages, some potential adverse effects were noted.

ECMWF runs a set of ensemble reforecasts as part of the operational suite of products. Once a week, 5-member ensembles are re-run for the equivalent date in each of the last 20 years using the current version of the IFS. These 'reforecasts' are used to calibrate the monthly forecast products as well as to generate the Extreme Forecast Index (EFI).We can use these reforecasts to calibrate the medium-range ensemble and compare the results with those using past data shown in the previous sections, which we will refer to as the sliding window approach. However, since there is no reforecast data set for the HRES, we exclude the HRES from the sliding window results in this comparison.

Figure 5 shows initial results for T2M forecasts during winter 2013/14 at the European stations. The evolution of CRPS with forecast lead time from one to ten days is shown for the raw ensemble and the different calibration methods. The vertical lines show the 90% confidence intervals. Both the sliding window approach and the reforecasts give very similar results, with no significant difference between the two methods at any forecast step. The benefit of combining the sliding window and reforecast methods was also investigated. The results are shown for two slightly different combination methods: both show some potential, but no overall significant extra benefit.

One major difference between the reforecast data and the operational forecasts is the ensemble size. An important aspect of the EMOS calibration method is the need to estimate the ensemble spread and only 5 members is not sufficient to give a good estimate of this. A new reforecast configuration using 11-member ensembles (and running twice a week) will be introduced soon. This has the potential to substantially improve the reforecast approach to the calibration.

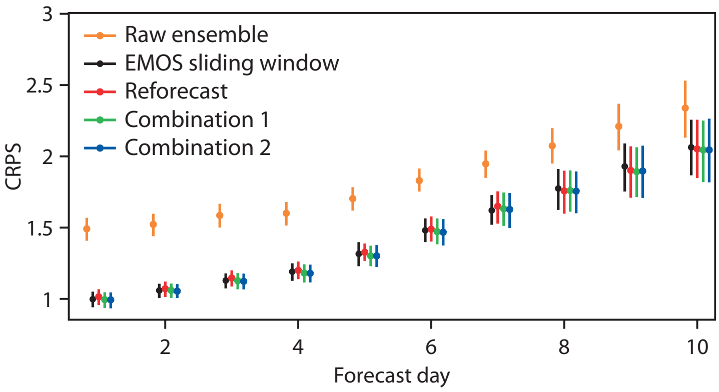

Figure 5 Mean CRPS over all considered stations for winter 2013/2014 for T2M for the raw ensemble, the EMOS sliding window approach, the reforecast approach and two versions of a combination of reforecast and sliding window forecasts. The vertical bars correspond to 90% confidence intervals for the expected average CRPS over all stations in the European subset.

Summary and outlook

A study was carried out to assess the benefits of calibrating the ECMWF medium-range forecasts to improve probabilistic predictions of four near-surface weather parameters. The main conclusions from the study are summarised below.

- Overall, state-of-the art methods of calibration provided substantial additional skill compared to the raw ECMWF forecasts. The reduction in CRPS for point forecasts is typically equivalent, and complementary, to 10–20 years of model system development.
- The skill of the calibrated forecasts has increased over time at a similar rate to the raw ensembles: in relative terms, the benefit of calibration is the same now as it was 10 years ago, suggesting that model development and calibration improve different aspects of the forecast error. It is expected that similar relative benefits will be obtained by calibration for the foreseeable future.
- Treating the complete set of ECMWF forecasts (HRES, CTRL and ENS members) as one forecasting system, with appropriate weight to each component, provides the greatest benefit.
- Although it was not primarily designed for such calibration, the current reforecast data gives equivalent results to the alternative and more traditional approach of using a sliding window training period using previous operational forecasts.

A number of relevant aspects were not addressed in the present study, which focused on individual locations and on overall performance as measured by the CRPS. Important areas for further study include the spatial and temporal structure of calibrated products and the impact of calibration on the forecasting of extreme events. The enhanced reforecast dataset to be introduced in 2015 will allow ECMWF to begin investigating these topics. ECMWF will explore the potential for calibration of gridded fields (against analyses). This work will allow the development of 'seamless' forecast products that cover all time-ranges from the medium-range to seasonal.

Resources for this work were made available through the externally-funded EFAS and GEOWOW projects.

Further reading

Gneiting, **T**, 2014: Calibration of medium-range weather forecasts. *ECMWF Tech. Memo. No. 719*. http://old.ecmwf.int/publications/library/do/references/show?id=91014

Hemri, **S.**, **M. Scheuerer**, **F. Pappenberger**, **K. Bogner** & **T. Haiden**, 2014: Trends in natural calibration of raw ensemble weather forecasts. Submitted to *Geophys. Res. Lett*.

European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, England

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error or omission or for loss or damage arising from its use.

doi:10.21957/45t3o8fj