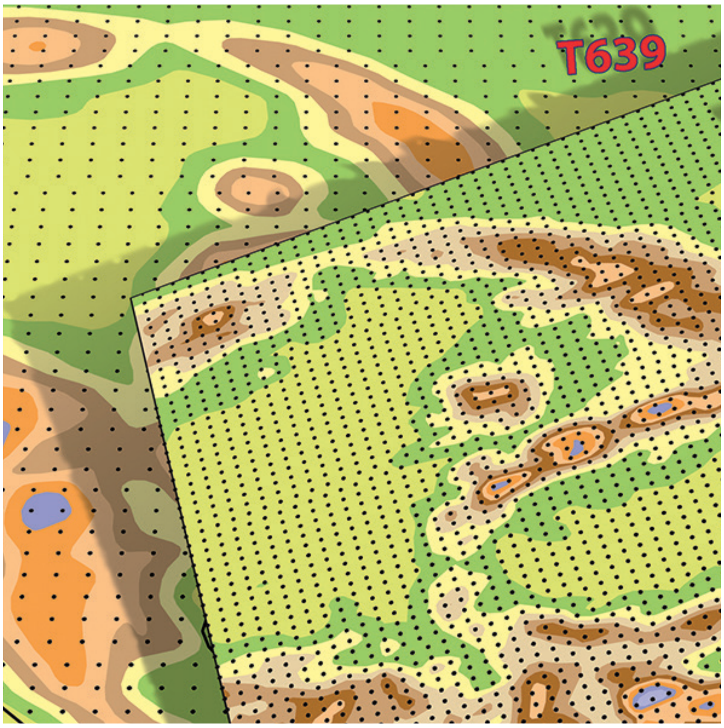
ECMWF Feature article

from Newsletter Number 124 – Summer 2010

METEOROLOGY

On the relative benefits of TIGGE multi-model forecasts and reforecast-calibrated EPS forecasts



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

doi:10.21957/5uxucx8z

This article appeared in the Meteorology section of ECMWF Newsletter No. 124 – Summer 2010, pp. 17–23.

On the relative benefits of TIGGE multi-model forecasts and reforecast-calibrated EPS forecasts

Renate Hagedorn

The main motivation for investing into research activities on Numerical Weather Prediction (NWP) lies in the expectation that improved weather forecasts lead to enhanced socio-economic benefits. As such, the ultimate goal of all research related to NWP is to improve the quality and utility of weather forecasts. There are of course many ways to achieve this goal, ranging from work on the model system per se to research on the provision of user-optimized forecast products. All of these activities are valuable and necessary contributions in their own right, and therefore none of them should be judged as more important than others. On the contrary, only through the complementary diversity of approaches can the overall goal be achieved.

Post-processing of Direct Model Output (DMO) from NWP models is one of the many ways to improve weather forecasts. The term 'post-processing' encompasses any means of manipulating the DMO to provide improved predictions. However, here we will concentrate on two specific methods:

- ***Multi-model.*** Combining single-model forecasts from several models into a multi-model forecast.
- ***Reforecast-calibration.*** Calibrating single-model forecasts with the help of specific training datasets.

Both approaches considered here have proven in the past to be successful in improving forecast quality. For example, the concept of multi-model forecasting has been extensively studied in the context of seasonal forecasting in the DEMETER and ENSEMBLES projects (see *ECMWF Newsletter No. 99 & 103*). It was concluded that overall the multi-model ensemble seems the most reliable approach for seasonal forecasts. However, on the medium-range timescale, it is less well established whether the multi-model concept is as successful as in the case of extended-range forecasting. Thus, one of the main goals of the THORPEX Interactive Grand Global Ensemble (TIGGE) project (see *ECMWF Newsletter No. 116*) is to investigate the applicability and potential benefits of the multi-model concept for medium-range weather forecasts. The method of calibrating the Ensemble Prediction System (EPS) forecasts based on a reforecast dataset has also been studied in the past, and its potential of improving predictions has been documented (see *ECMWF Newsletter No. 117*).

One can expect that both post-processing methods, the multi-model concept and the reforecast calibration, have their own strengths and weaknesses. Hence it is only natural to compare the benefits (and costs) of both approaches, and to investigate the mechanisms behind the potential improvements (the main aim of this article). However, in doing so it is not intended to make a final judgement on which is the better method. Instead the aim is to provide some information that helps users decide which approach might be the more appropriate choice for their specific circumstances.

Performance of TIGGE multi-model versus reforecast-calibrated forecasts

The TIGGE archive at ECMWF contains global ensemble predictions from ten modelling centres. For detailed information on the individual characteristics of the TIGGE models (e.g. resolution and number of ensemble members) refer to the ECMWF TIGGE website: http://tigge.ecmwf.int.

Since the predictions from the Météo-France model are limited to a lead time of 108 hours, here we consider only the remaining nine model contributions: Bureau of Meteorology (BOM, Australia), China Meteorological Administration (CMA), Meteorological Service of Canada (CMC), ECMWF, UK Met Office, National Centers for Environmental Prediction (NCEP, USA), Japan Meteorological Agency (JMA), Korea Meteorological Administration (KMA) and Centro de Previsão de Tempo e Estudos Climáticos (CPTEC, Brazil).

Benchmarking the EPS

A first impression on the level of skill of these nine single-model systems is given by comparing the Continuous Ranked Probability Skill Score (CRPSS) of the 850-hPa temperature over the northern hemisphere for forecasts of the winter season (DJF – December, January, February) of 2008/09 (Figure 1). The performance of these forecasts varies significantly for the different models, with the CRPSS dropping to zero for the worst models at a lead-time of five days and for the best models around day 15. That is, the time range up to which the model predictions are more useful than the reference forecast, which is in this case the climatological distribution, varies considerably from one model to another.

 doi:10.21957/5uxucx8z

Because not all forecasting centres integrate their models out to 15 days, the performance of the multi-model ensemble combining all nine single-model systems can only be assessed up to the maximum forecast range covered by all individual models, which is nine days. The multi-model ensemble is constructed by giving equal weights to all contributing members, noting that through the different number of members in the individual model systems an implicit weighting will be applied. That is, model systems with a higher number of ensemble members will have a greater impact in the final multi-model prediction than model systems with fewer members. Except for the first two forecast days, this multi-model prediction (TIGGE-9) does not significantly improve over the best single model (i.e. the ECMWF EPS). Similar results can be observed for other variables such as the bias-corrected 2-metre temperature.

Note that all results presented in this article are based on using ERA-Interim reanalyses as verification dataset. Further information on the rationale of this choice can be found in Box A.

The inability of the multi-model ensemble to significantly improve over the best single-model system might be caused by the fact that it consists of all nine single models (i.e. it includes the models with rather poor performance). To eliminate these possibly detrimental contributions, a new multi-model (TIGGE-4) containing only the four best single-model systems with lead time up to 15 days was constructed and compared to the four contributing single models from the National Meteorological Services in Canada (CMC), UK (Met Office) and the USA (NCEP), plus ECMWF (Figure 2). In fact, this reduced version of the full multi-model ensemble now gives significantly improved scores over the whole forecast period and for both upper-air and surface variables. This result indicates that a careful selection of the contributing models seems to be important for medium-range multi-model predictions.

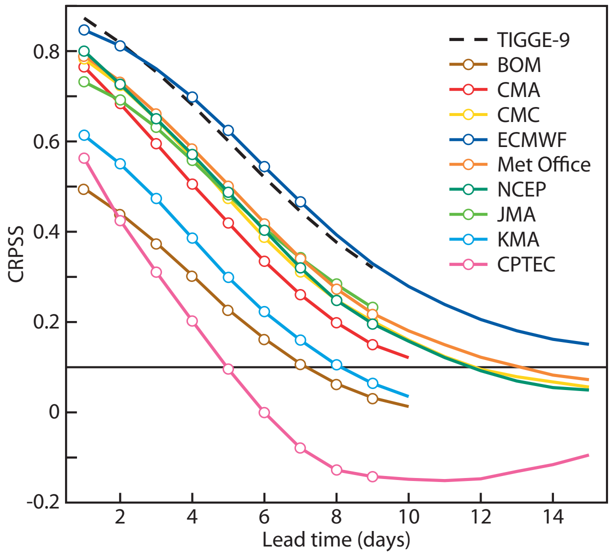


Figure 1 Continuous Ranked Probability Skill Score (CRPSS) versus lead time for 850-hPa temperature forecasts. The TIGGE-9 multi-model composed of nine single models and the scores of all nine contributing single models are shown. Symbols are only plotted for cases in which the single-model score differs significantly from the multi-model score on a 1% significance level. The significance levels have been assessed using a paired block bootstrap algorithm following *Hamill* (1999). All scores are for forecasts starting in DJF (December, January, February) 2008/09 and averaged over the northern hemisphere (20°–90°N).

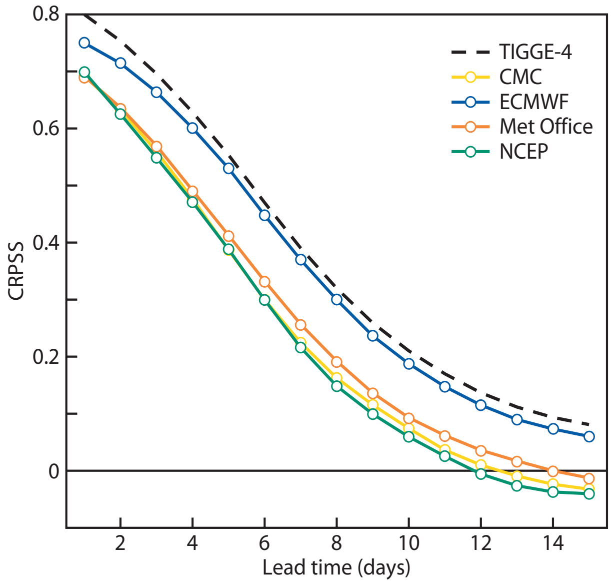


Figure 2 Continuous Ranked Probability Skill Score (CRPSS) versus lead time for 850-hPa temperature forecasts. The TIGGE-4 multi-model composed of the four best single models with lead-time up to 15 days is shown in addition to the CRPSS of the four contributing single models. Symbols are only plotted for cases in which the single-model score significantly differs from the multi-model score on a 1% significance level. All scores are for forecasts starting in DJF 2008/09 and averaged over the northern hemisphere (20°–90°N).

doi:10.21957/5uxucx8z

Choice of verification dataset

A

A number of considerations have to be taken into account when choosing the verification dataset to assess the performance of different single models and multi-models. On the one hand, using model independent verification data, such as station observations, ensures a fair treatment of all models. On the other hand, comparisons of the model performance over larger areas or for variables not directly available in observational datasets require the use of analyses, which commonly exhibit some of the bias of the forecast model used. There are a number of possibilities for the choice of analysis product in the context of comparing single and multi-model predictions.

- Each model's own analysis could be used as the verification dataset. However, there are two issues with this option: (a) the multi-model ensemble has no own analysis, and (b) it would be difficult to draw conclusions from the resulting scores and skill scores when their calculation is based on different reference datasets.
- The average of all analyses of the participating models or some weighted average, also called multi-model analysis, could be used. Such an average analysis would fulfil the condition of being as fair as possible to all models participating in the comparison. On the other hand, averaging all analyses, including less accurate ones, might not necessarily lead to an analysis closest to reality. Additionally, such a multi-model analysis cannot be used as verification dataset in this reforecast-comparison study because it is only available for the TIGGE forecast period (i.e. from 2007 onwards). This is not sufficient because the calibration of ECMWF forecasts based on the reforecast training dataset requires a consistent verification dataset for the entire training and test period (i.e. the verification dataset has to be available from 1991).

A possible compromise between the requirement of being as fair as possible to all models involved and being as accurate as possible is to choose the ECMWF ERA-Interim reanalysis as verification dataset. The two main advantages of this choice are the acknowledged high quality of this analysis product and the availability of this dataset for the entire training and test period (1991 up to near-real time). The obvious drawback of this option is that the ERA-Interim reanalyses are certainly not entirely independent of one of the models in the comparison, the ECMWF model. As such, one might expect that it is more difficult for non-ECMWF models to achieve good scores when verified against the ERA-Interim reanalysis. However, it can be demonstrated that the skill scores of all models are affected by the choice of verification dataset. Using ERA-Interim as verification leads to diagnosing a reduced performance, in particular for early lead times. For longer lead times, the impact tends towards negligible differences.

It is important to note that the performance of all forecasts is similarly affected by the choice of verification dataset, i.e. there is only little impact on the relative performance of individual models with respect to each other. Although model systems that are quite close in their performance (like CMC, NCEP and Met Office) can change their ranking relative to each other, the choice of verification has no impact on the clear superiority of the ECMWF EPS.

For surface variables such as 2-metre temperature the impact of using ERA-Interim is larger, but can be reduced by applying a bias-correction procedure. Overall, we regard using ERA-Interim analyses as general verification dataset to be the best option for this study. Also by keeping in mind the sensitivities towards the choice of verification dataset one can ensure a fair interpretation of the results.

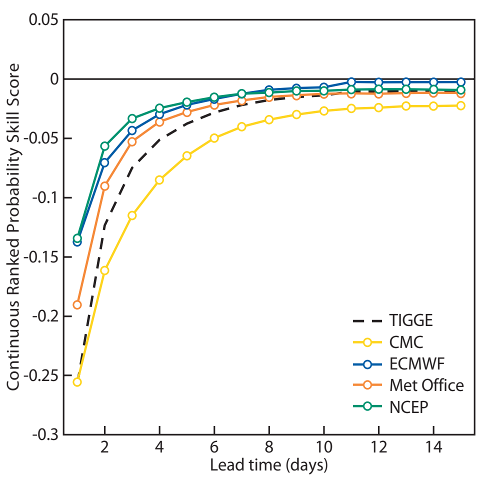


Illustration of the impact of the verification dataset on the relative skill of the predictions. Negative values indicate a worse performance when verified against ERA-Interim reanalyses, a value of zero indicates no impact of the chosen verification dataset. Scores are calculated for forecast of 850-hPa temperature from the TIGGE multi-model and the single models (CMC, ECMWF, Met Office and NCEP) starting in DJF 2008/09 and averaged over the northern hemisphere (20°–90°N).

 doi:10.21957/5uxucx8z

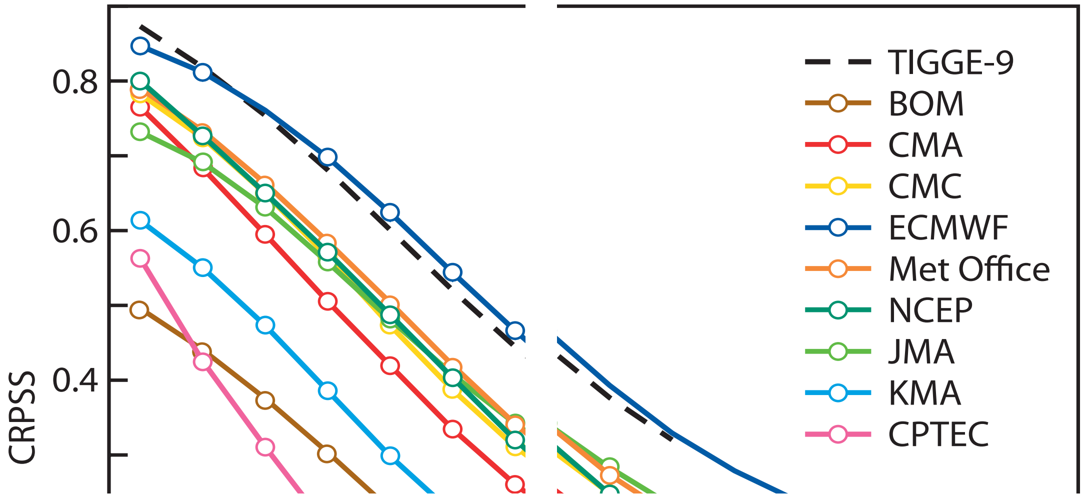


Figure 3 Continuous Ranked Probability Skill Score (CRPSS) versus lead time for (a) 850-hPa temperature forecasts and (b) 2-metre temperature forecasts. The TIGGE-4 multi-model composed of the four best single models with lead-time up to 15 days is shown in addition to the CRPSS of the four contributing single models and the reforecast-calibrated ECMWF EPS (ECMWF-CAL). Symbols are only plotted for cases in which the single-model score differs significantly from the multi-model score on a 1% significance level. All scores are for forecasts starting in DJF 2008/09 and averaged over the northern hemisphere (20°–90°N).

Reforecast-calibration methodology

B

The methodology developed to produce reforecast-calibrated ECMWF EPS forecasts (ECMWF-CAL) is based on combining calibration results from the Non-homogeneous Gaussian Regression technique (NGR) and pure bias-correction (BC).

The NGR technique itself has already been applied to ECMWF EPS forecasts (see Newsletter No. 117). Essentially, NGR is an extension to conventional linear regression by taking into account information contained in the existing spread-skill relationship of the raw forecast. Using the ensemble mean and the spread as predictors, it fits a Gaussian distribution around the bias-corrected ensemble mean. The spread of this Gaussian is on the one hand linearly adjusted according to the errors of the regression model using the training data, and on the other hand depends on the actual spread according to the diagnosed spread-error relationship in the training dataset. Thus, one important feature of this methodology is being able to not only correct the first moment of the ensemble distribution but also correct spread deficiencies.

After applying the NGR calibration, the forecast Probability Density Function (PDF) consists of a continuous Gaussian distribution, not an ensemble of realizations. However, to be able to compare the performance of the calibrated probabilities, retrieved from a full PDF, with the probabilities simply based on counting individual ensemble members, a synthetic ensemble is created from the calibrated Gaussian by drawing 51 equally likely ensemble members from the calibrated PDF. That is, the synthetic ensemble is realized by sampling the members at the 51 equally spaced quantiles of the regressed Cumulative Distribution Function (CDF).

Experimenting with the choice of training dataset and calibration method revealed that combining a simple bias correction using training data from the 30 previous days (BC-30) and the NGR calibration based on reforecasts (NGR-RF) is superior to the pure NGR-RF calibration, particularly for early lead times. The two ensembles are not combined by taking all members from both ensembles to form a new ensemble with twice the number of members, but by first ordering both the bias-corrected and NGR-calibrated ensembles and then averaging the corresponding members. In this way the final combined calibrated system still contains only 51 members. Some experimentation with different weights for the NGR-RF and BC-30 ensembles revealed that applying equal weights at all lead times leads to overall best results.

For the current version, the slightly improved performance might be caused by the fact that the BC-30 calibration contains information on the bias more relevant to the current weather regime than the overall bias diagnosed from the reforecast dataset. However, using a refined version of the NGR-RF calibration by, for example, including soil moisture as an additional predictor might diminish the positive impact the BC-30 contribution can have. A further advantage of adding the BC-30 calibrated ensemble to the Gaussian NGR-RF ensemble is that through this procedure any non-Gaussian characteristics of the original ensemble may be retained to some degree.

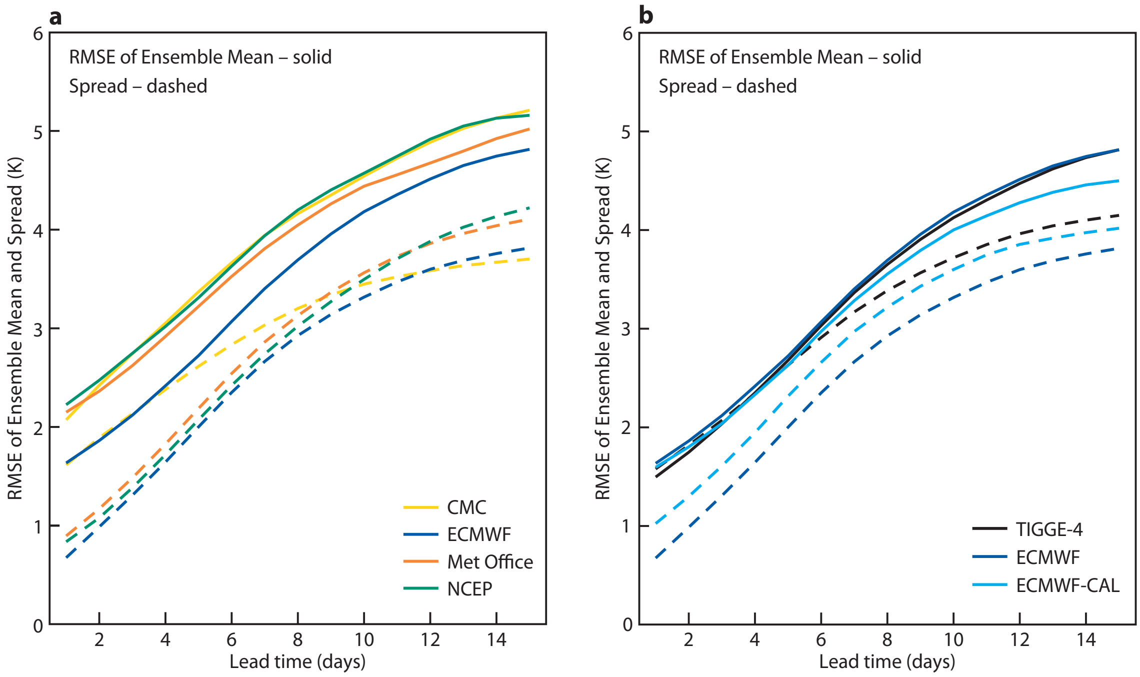


Figure 4 Root-mean-square error (RMSE) of the ensemble mean (solid lines) and ensemble standard deviation ('spread', dotted lines) versus lead-time for 2-metre temperature forecasts. (a) Results for the single-model forecast (CMC, ECMWF, Met Office and NCEP). (b) As (a) but without the CMC, Met Office and NCEP results, including instead the results for the reforecast-calibrated ECMWF (ECMWF-CAL) and TIGGE-4 multi-model results. All scores are for forecasts starting in DJF 2008/09 and averaged over the northern hemisphere (20°–90°N).

Calibrating the EPS

After having established a new benchmark for the best single model, the ECMWF EPS, the next question is whether it might be possible to achieve similar improvements by calibrating the ECMWF EPS based on its reforecast dataset. Detailed information on the methodology applied to create the reforecast-calibrated ECMWF EPS forecasts (ECMWF-CAL) can be found in Box B. Essentially, this calibration methodology corrects both for errors in the mean and spread of the ensemble.

Comparing the CRPSS of the ECWMF-CAL forecasts with the TIGGE-4 multi-model scores reveals that indeed the calibration procedure significantly improves ECMWF's scores (Figure 3). Overall the performance of the ECMWF-CAL predictions is as good as the TIGGE-4 multi-model ensemble, and for longer lead times it can be even better.

For 850-hPa temperature predictions (Figure 3a) the CRPSS of ECMWF-CAL lies above the multi-model value for early lead times, and for longer lead times the skill scores are slightly lower than for the multi-model ensemble, though not statistically significant. Considering the slight advantage in the early lead times for ECMWF forecasts when using ERA-Interim as verification and the lack of statistical significance of the difference in the CRPSS for longer lead times, it can be concluded that for 850-hPa temperature the reforecast-calibrated ECMWF EPS forecasts are of comparable quality as the TIGGE-4 multi-model forecasts.

This result is confirmed when studying other variables, regions or seasons. In fact, for 2-metre temperature forecasts the calibration is even more effective for longer lead times (Figure 3b). This indicates that the systematic component of the error is more dominant for the 2-metre temperature, and thus the calibration procedure is able to further reduce the root-mean-square error (RMSE) of the ensemble mean. However, the general level of skill at those long lead times is very low. Therefore, these improvements – as relevant as they might look in terms of overall scores – might not add very much in terms of improving the usefulness of the predictions in a real forecast situation.

Comparing, for example, the ECMWF EPS with the reforecast-calibrated and TIGGE-4 multi-model forecasts for individual cases at single grid point locations can give an indication of how much (or how little) a real forecast product would change. On the one hand, one can find locations at which the calibrated or multi-model ensemble distributions are significantly different from the ECMWF EPS. These are usually locations with complex orography, where for example different grid resolutions can cause detectable systematic errors. In such cases the NGR calibration is able to correct both such biases and serious spread deficiencies. However, as mentioned above, for longer lead times the predicted

 doi:10.21957/5uxucx8z

distributions are already close to the climatological distributions. Consequently it is not clear whether the improvements seen in the scores can be really translated into practical benefits of better decision-making based on such 'theoretically' improved forecast products. Additionally, there are also many locations with less pronounced systematic errors or spread deficiencies. At such locations, the calibration obviously has much less impact.

Mechanisms behind improvements

To further investigate the mechanisms behind the improvements, Figure 4 focuses on the spread-error relation of the different ensembles. Ensemble forecasting aims to construct uncertainty information so that the observations can be considered as statistically indistinguishable from the ensemble members of the forecast. This requires the spread of the ensemble (ensemble standard deviation) to be close to the root mean square error (RMSE) of the ensemble mean. However, for 2-metre temperature all single-model systems are seriously under-dispersive as shown in Figure 4a. CMC has the lowest spread deficiency at the beginning of the forecast, but due to a serious mismatch in the growth of spread and error it has the worst spread-error relation for longer lead times. The remaining three models have a similar level of spread. However, the significantly lower RMSE of the ECMWF EPS implies not only a slightly better spread-error relation compared to the Met Office and NCEP ensembles, but it is also one of the main reasons for its significantly better probabilistic scores discussed earlier.

The effect of combining the single-model systems or calibrating the ECMWF EPS can be seen in Figure 4b. The RMSE of the multi-model ensemble is slightly reduced for early lead times, but the most noticeable change is the very much improved spread-error relation, particularly up to day 6. In contrast to that, the reforecast-calibrated ECMWF EPS has not such a perfect spread-error relation, though it is improved compared to the original EPS spread. The reason for this is the specific methodology of combining bias-corrected and NGR-calibrated forecasts (see also Box B).

Applying the pure NGR calibration should lead to a near perfect spread-error relation, but the advantages of possible reductions in the systematic error provided by the 30-day bias-corrected ensemble may outweigh the slight disadvantage of a poorer second-moment (i.e. spread) calibration. Since the under-dispersion is not fully corrected in the reforecast-calibrated ensemble, the main improvement of its probabilistic scores comes from the reduction in the RMSE, in particular for longer lead times.

We note that the theoretical disadvantage of the ECMWF-CAL methodology (i.e. the sub-optimal spread correction) in certain situations might even be regarded as a positive aspect. Discussions with operational forecasters revealed that – although theoretically correct – the extent of the full NGR spread is sometimes regarded as counterproductive in real forecast situations. There might be many reasons for this subjective opinion, such as a general reluctance to integrate uncertainty information into operational forecast practice. Although not part of the current investigation, we feel that these aspects are worth considering in further discussions with users on how to achieve our ultimate goal of providing user-optimized forecast products.

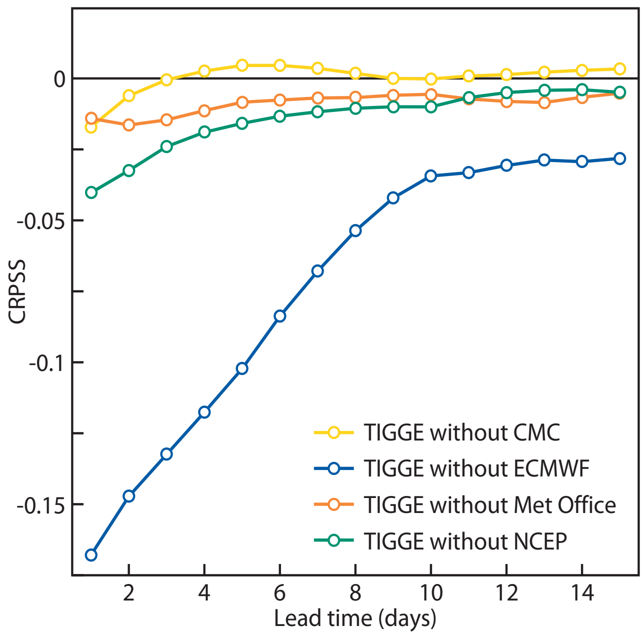


Figure 5 Illustration of gain or loss in skill of 2-metre temperature forecasts versus lead time using the Continuous Ranked Probability Skill Score (CRPSS) depending on which model has been removed from the TIGGE-4 multi-model containing all four single models (CMC, ECMWF, Met Office and NCEP). The CRPSS is defined as *CRPSS* = 1 – *CRPS(exp)/CRPS(ref)*, with *CRPS(ref)* being the CRPS (Continuous Ranked Probability Score) of the TIGGE-4 multi-model and *CRPS(exp)* the CRPS of the reduced multi-model respectively. Negative values indicate a worse performance of the reduced multi-model (i.e. a detrimental effect of removing a particular single model from the multi-model mix). All scores are for forecasts starting in DJF 2008/09 and averaged over the northern hemisphere (20°–90°N).

doi:10.21957/5uxucx8z

Single-model contributions to the TIGGE multi-model

The computational and organizational overhead of collecting all individual model contributions and combining them into a consistent multi-model ensemble grows with the number of contributing models. Consequently it is worth investigating the added benefit each individual model can give to the multi-model system. For this purpose we constructed reduced multi-model versions with individual model components removed from the full multi-model mix and scored them against the full multi-model version containing all four models (Figure 5).

It is obvious that removing the ECMWF EPS from the multi-model ensemble has the biggest impact, whereas the other models contribute to a lesser extent to the multi-model success. It might be argued that one of reasons for this is that by removing the ECMWF EPS the multi-model ensemble loses 51 members, whereas removing the other models produces a loss of only 21 or 24 members. Since the CRPS (Continuous Ranked Probability Score) is expected to go down with increasing number of ensemble members (*Ferro et al.*, 2008), it is not straightforward to distinguish the effect of removing the forecast information that a single model adds to the multi-model from the effect of removing 51 instead of 21 or 24 members. However, there are two reasons why we believe that not explicitly accounting for the difference in the number of members is justified.

- The difference of number of members between the full multi-model ensemble containing 117 members and the most reduced multi-model ensemble containing 66 members would require only a moderate adjustment factor of about 1% CRPS reduction applied to the ensemble with the lower number of members. This is much lower than the difference indicated by a CRPSS between –0.15 and –0.05. Therefore, only 1% out of the 15% increase in the CRPS of the reduced multi-model ensemble is due to the lower number of members and the remaining 14% increase is caused by the withdrawal of the forecast information from that model.
- Suppose we want to compare the performance from an operational rather than theoretical point of view. That is we are not interested in theoretical questions such as "how would these models compare if they had the same number of members?", but we want to answer questions like "how do the operational systems, as they are, compare?" In that case we should not adjust the scores to reflect a potential performance of a model with infinite number of members. Following these considerations, in none of the comparisons of this study are the scores adjusted according to their different numbers of ensemble members.

Apart from the question about which of the single models contributes most to the multi-model success, a further question in the context of the TIGGE project is whether the multi-model concept could lead to reduced costs but still keeping the same quality of forecasts. Assuming, for the sake of argument, that ECMWF could no longer afford to provide its EPS forecasts, could a multi-model consisting of the remaining high-quality ensembles be as good as the ECMWF EPS on its own? Indeed, a TIGGE multi-model ensemble without ECMWF contribution is of comparable quality as the ECMWF EPS alone, i.e. combining the second-, third- and fourth-best global ensembles leads to forecasts which are as good as the best global ensemble (Figure 6). However, this is only true for the ECMWF EPS when it has not been reforecast-calibrated.

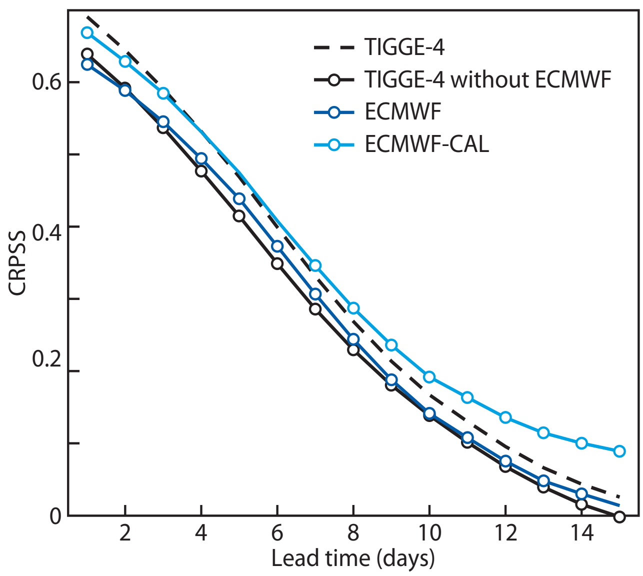


Figure 6 Continuous Ranked Probability Skill Score (CRPSS) versus lead time for 2-metre temperature forecasts. Results are shown for the TIGGE-4 multi-model containing CMC, ECMWF, Met Office, and NCEP forecasts, the TIGGE-4 multi-model without ECMWF forecasts (i.e. containing only CMC, Met Office, and NCEP forecasts), the simple bias-corrected ECMWF forecasts, and the re-forecast calibrated ECMWF forecasts (labelled ECMWF-CAL). Symbols are omitted for cases in which the score does not significantly differ from the TIGGE-4 multi-model score on a 1% significance level. All scores are for forecasts starting in DJF 2008/09 and averaged over the northern hemisphere (20°–90°N).

 doi:10.21957/5uxucx8z

Running the complete ECMWF EPS, including its reforecasts, leads to a performance which cannot be achieved by any current multi-model version not containing ECMWF forecast information. These results are generally confirmed when considering other variables such as upper-air temperature or wind components, though small differences in the relative performance, also depending on the region, can be observed.

Overall costs and benefits

Coming back to the main aim of this article (i.e. comparing the costs and benefits of the multi-model and reforecast-calibration approaches) it is clear that the performance of the reforecast-calibrated ECMWF EPS forecasts is as good as the TIGGE multi-model system, if not better. When considering which post-processing approach leads to better forecast products or can give more useful information in a practical decision-making process, it has to be noted that the calibration procedure is particularly helpful at locations with clearly detectable systematic errors (e.g. areas with complex orography or coastal grid points). In such areas the calibration procedure can correct, for example, for unresolved scales and thus essentially performs a sort of downscaling of the forecasts. This ability is particularly important for all applications needing forecasts at specific locations like, for example, forecasting the wind power production at specific wind farms. The multi-model approach, on the contrary, might be advantageous in situations where it is able to suggest alternative solutions not predicted by the single model of choice.

Further investigations on the mechanisms behind the improvements achieved by the post-processing methods led to the conclusion that both approaches tend to correct similar deficiencies. That is, systematic error and spread deficiencies are improved to a similar extent by both approaches. Experiments assessing the contribution of the individual components of the multi-model system demonstrated that the ECMWF EPS is the single most important source of information for the success of the multi-model ensemble.

For a final assessment which of the two post-processing methods would be the most appropriate choice for a modelling centre, one also has to consider the technical overhead of producing multi-model or reforecast-calibrated single-model forecasts in an operational context. If, for example, a modelling centre has easy and reliable access to all components of the multi-model system, and if its users or operational forecasters ask for multiple solutions suggested by individual models, then the multi-model concept might be the method of choice. However, for a forecasting centre reluctant to take on the potential risks and technical overhead inherent in the increased complexity of a multi-model system, using the reforecast-calibrated ECMWF EPS forecasts rather than a logistically highly complex multi-model system seems to be a more appropriate choice.

Considering the performance improvements made possible by the availability of the ECMWF reforecast dataset, other modelling centres might start providing reforecasts for their model systems in the not too distant future. In that case it would be interesting to study the relative benefits achievable for reforecast-calibrated multi-model or single-model systems. Furthermore, we suggest exploring the relative merits of multi-model versus reforecast-calibrated predictions for other user-relevant variables like precipitation and wind speed, in particular in the context of extreme events.

Further Reading

Ferro, **C.A.T.**, **D.S. Richardson** & **A.P. Weigel**, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorol. Appl.*, **15**, 19–24.

Hamill, **T.M.**, 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167.

doi:10.21957/5uxucx8z