TECHNICAL MEMORANDUM

663

Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts

Renate Hagedorn[1], Roberto Buizza, Thomas M. Hamill[2], Martin Leutbecher and T.N. Palmer

Research Department

January 2012

[1] Current affiliation: Deutscher Wetterdienst, Offenbach, Germany

[2] NOAA Earth System Research Laboratory, Boulder, Colorado

Accepted for publication in the

Quarterly Journal of the Royal Meteorological Society

European Centre for Medium-Range Weather Forecasts

Europäisches Zentrum für mittelfristige Wettervorhersage

Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

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

Abstract

Forecasts provided by the THORPEX Interactive Grand Global Ensemble (TIGGE) project were compared with reforecast-calibrated ensemble predictions from the European Centre for Medium-Range Weather Forecasts (ECMWF) in extra-tropical regions. Considering the statistical performance of global probabilistic forecasts of 850-hPa and 2-m temperatures, a multimodel ensemble containing nine ensemble prediction systems (EPS) from the TIGGE archive did not improve on the performance of the best single-model, the ECMWF EPS. However, a reduced multimodel system, consisting of only the four best ensemble systems, provided by Canada, the US, the UK and ECMWF, showed an improved performance. The multimodel ensemble provides a benchmark for the single-model systems contributing to the multimodel. However, reforecast-calibrated ECMWF EPS forecasts were of comparable or superior quality to the multimodel predictions, when verified against two different reanalyses or observations. This improved performance was achieved by using the ECMWF reforecast dataset to correct for systematic errors and spread deficiencies. The ECMWF EPS was the main contributor for the improved performance of the multimodel ensemble; that is, if the multimodel system did not include the ECMWF contribution, it was not able to improve on the performance of the ECMWF EPS alone. These results were shown to be only marginally sensitive to the choice of verification data set.

1 Introduction

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 contributions to the general objective, and therefore none of the single efforts can be judged as more important than another. The largest improvement is expected when post-processing and model development complement each other.

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 in order to provide improved predictions. However, here we will concentrate on two specific methods: (i) combining single-model forecasts into a multimodel forecast, and (ii) calibrating single-model forecasts with the help of specific training datasets. Note that the term single-model or multimodel does not refer only to the forecast model itself but encompasses the whole prediction system including the data assimilation system. Both of these approaches have been proven in the past to be successful in improving forecast quality. For example, the concept of multimodel forecasting has been extensively studied in the DEMETER project (Palmer et al., 2004). Results concerning the potential superiority of multimodel predictions on the seasonal timescale are presented for instance in the special issue on the DEMETER project in Tellus-A 57(3). Studying the rationale behind the success of multimodel ensembles, Hagedorn et al. (2005) concluded that “the key to the success of the multimodel concept lies in combining independent and skilful models, each with its own strengths and weaknesses.” In particular the fact that the performance of the single-model ensembles varies, and thus in an operational environment the “best” model cannot be easily identified, makes the multimodel ensemble overall the most reliable choice. However, based on systematic toy model simulations, Weigel at al. (2008) and Weigel and Bowler (2009) demonstrated that even under the assumption that there is a clearly identifiable best single-model system, a multimodel ensemble can still improve the

performance of this best model. This result is particularly relevant in the context of applying the multimodel concept to medium-range weather forecasts, which has been at the heart of the THORPEX Interactive Grand Global Ensemble (TIGGE) project (Bougeault et al., 2010). First results from comparisons of the performance of individual TIGGE models indicated that in contrast to the seasonal timescale, where it can be difficult to define a "best" single-model which outperforms all other models on virtually all aspects, on the medium-range timescale it is much easier to identify a single-model which is clearly superior to all other models (Park et al., 2008). Therefore, the initial research question posed in the context of seasonal forecasting: "Does the combination of single-model ensembles with overall similar levels of skill lead to a more skilful multimodel ensemble?" changes in the context of medium-range forecasting to: "Does adding information from less skilful models to the best model lead to a more skilful multimodel ensemble?" As Weigel and Bowler (2009) pointed out in their theoretical study "it is possible to construct and combine reliable forecasts such that the multimodel has indeed higher skill than the best component forecast alone", and early diagnosis of the TIGGE dataset confirms this theoretical result with real forecast data (Park et al. 2008, Matsueda and Tanaka, 2008; Johnson and Swinbank, 2009).

The second post-processing method explored here is the calibration of single forecast systems with the help of specific training datasets. A number of different calibration methods have been proposed for operational and research applications, and a recent comparison of several methods can be found in Wilks and Hamill (2007). As most calibration methods are based on the idea of correcting the current forecast by using past forecast errors, they require some sort of training dataset. With this set of past forecast-observation pairs, correction coefficients for a regression-based calibration scheme can be determined. It has been shown that such calibration techniques are particularly successful when a "reforecast" training dataset is available (Hamill et al. 2004, 2006, 2008; Hamill and Whitaker 2006, 2007; Hagedorn et al. 2008). A reforecast dataset is a collection of forecasts from the past, usually going back for a considerable number of years or decades. In order to ensure consistency between reforecasts and actual forecasts, ideally the reforecasts are produced specifically with the same model and data assimilation system that is used to produce the actual forecasts. The availability of a large number of past forecast-observation pairs consistent with the current forecast model is a major factor of the success of the calibration technique used in this study.

One can expect that both post-processing methods, the multimodel concept and the reforecast calibration, have their own strengths and weaknesses. Hence, it is only natural to compare the potential benefits of both approaches, which is the main aim of this publication. However, it is not our intent to come up with a final judgement on which is the better method, but instead to provide some indication for potential users to decide which approach might be the more appropriate choice for their specific circumstances. In contrast to Weigel et al. (2009), who have investigated a similar question on the seasonal timescale, this study concentrates on the medium-range timescale of forecasts up to 15 days.

A description of the datasets used can be found in section 2. The post-processing methods are presented in section 3. The results are presented in section 4, with a summary and discussion following in section 5.

2 Datasets

2.1 Forecast datasets

In this study forecasts from nine global Ensemble Prediction Systems archived in the TIGGE database at ECMWF are used. The main features of the model systems can be found in Table 1, together with a list of the model centres operationally running the forecast systems and providing the data for the TIGGE archive. Further detailed information on the model systems can be found in Park et al. (2008) or on the TIGGE website at ECMWF (http://tigge.ecmwf.int/models.html). The investigations will focus mainly on the winter season December 2008 to February 2009 (DJF-2008/09), with some additional results also shown for the summer season June to August 2009 (JJA-2009). Results for both upper air fields of 850-hPa temperature (T850) and 500-hPa geopotential height (GH500), as well as the near surface variable 2-m temperature (T2m) will be discussed. Only extra-tropical regions are considered in this study. The main comparisons will be done using forecasts starting at 00 UTC, the start time for which also ECMWF reforecasts are available. The use of the reforecast dataset for calibrating 12 UTC forecasts was tested as well, demonstrating that similar improvements can be achieved (not shown here). We note that when using 00 UTC reforecasts for calibrating 12 UTC forecasts it is especially important to take into account a possibly existing daily cycle of the systematic error, and apply the training dataset accordingly. The comparisons involving all nine single-model forecasts from the TIGGE database are done for 12 UTC forecasts since some of the contributing centres produce forecasts only for 12 UTC and not at 00 UTC.

Table 1: Main features of the nine TIGGE model systems used in this study(in DJF2008/09). BOM: Bureau of Meteorology (Australia), CMA: China Meteorological Administration (China), CMC: Canadian Meteorological Centre (Canada), CPTEC: Centro de Previsão de Tempo e Estudos Climáticos (Brazil), ECMWF: European Centre for Medium-Range Weather Forecasts (International), JMA: Japan Meteorological Agency (Japan), KMA: Korea Meteorological Administration (Korea), NCEP: National Centers for Environmental Prediction (USA), MetOffice: The UK Met Office (United Kingdom)

Centre	Horizontal resolution in archive	No. of vertical levels	No. of perturbed members	Forecast length (days)
BOM	1.5° x 1.5°	19	32	10
CMA	0.56° x 0.56°	31	14	16
CMC	1.0° x 1.0°	28	20	16
CPTEC	N96 (~1.0° x 1.0°)	28	14	15
ECMWF	N200 (~0.5° x 0.5°) N128 (~0.7° x 0.7°)	62	50	15
JMA	1.25° x 1.25°	40	50	9
KMA	1.25° x 1.25°	40	16	10
NCEP	1.0° x 1.0°	28	20	16
MetOffice	1.25° x 0.83°	38	23	15

For the verification against analyses, all forecasts have been interpolated to a common 2.5° × 2.5° grid using the interpolation routines provided by the ECMWF TIGGE data portal (http://tigge.ecmwf.int). Since the resolutions of the models are finer than 2.5° × 2.5° to varying degrees, the values at this lower resolution verification grid can be regarded as representing the average forecast over the 2.5° × 2.5° areas. One might expect that this sort of smoothing of the forecasts could improve the scores. However, sensitivity studies on the impact of the verification resolution have demonstrated that performing the verification on a higher resolution 1.5° × 1.5° grid essentially did not change the result (corresponding figures not shown here). In addition to this 2.5° × 2.5° datasets, forecasts have been also extracted for individual stations in order to be verified against station observations of 2m temperature. For this evaluation, the grid point of the original model grid closest to the respective station has been chosen.

The forecasts are mainly evaluated by calculating the Continuous Ranked Probability Score (CRPS) or its skill score (CRPSS), a diagnostic focussing on the entire permissible range of the evaluated scalar variable (e.g. Hersbach, 2000). The CRPS is very suitable for systems issuing probability forecasts in terms of probability densities. However, here we evaluate all ensemble forecasts in the classical way by using as the cumulative distribution function the empirical distribution function given by the sample of values provided by the ensemble. Area averages of CRPS, mean squared errors and variances are computed by weighting the grid point values with cosine latitude in order to correctly represent spatial integration of additive verification statistics.

2.2 Training datasets

An objective calibration technique needs a set of past forecast-observation pairs, also called training dataset. Usually, it is beneficial to have a training dataset as large as possible to achieve a robust calibration. However, the existence of seasonally varying systematic errors suggests that it can also be beneficial to use only training data from a similar time of year. The ECMWF reforecast dataset (Hagedorn, 2008) has been designed to satisfy both of these requirements. A set of reforecasts are operationally produced at ECMWF once per week, with start dates from the past 18 years. That is, on every Thursday the operational EPS is not only run for the actual date, but also for the same calendar day of the past 18 years. The only difference of these reforecasts to the actual EPS forecasts is the reduced number of perturbed members (4 instead of 50) and that ECMWF reanalyses instead of operational analyses are used in the initialization. Before 12 March 2009, a combination of ERA-40 reanalyses (Uppala et al., 2005) and operational analyses was used for the initialization, and from that date onwards only ERA-Interim analyses (Simmons et al., 2007; Dee et al., 2011) have been used to provide a more consistent initialization dataset. The training dataset used in this study consists of the ECMWF reforecasts produced in the five weeks closest to the target date to be calibrated. In this way, both the seasonally varying aspects are preserved and the number of forecast-observation pairs (18 years x 5 start dates = 90) should be sufficient for a robust estimation of the calibration coefficients, at least for the quasi-Gaussian variables studied here, 500-hPa geopotential, 850-hPa and 2-m temperature.

At present, this type of reforecast dataset is only available for the ECMWF EPS and not for the other models in the TIGGE archive. That is, the reforecast based calibration can only be applied to ECMWF forecasts. However, a simple bias-correction procedure has been applied to all single-model systems and the multimodel ensemble. This calibration is based on a training dataset consisting of the last 30 days before the start date of the forecasts (Hagedorn et al. 2008).

2.3 Verification datasets

A number of considerations have to be taken into account when choosing the verification dataset to assess the performance of different single- and multimodels. On one hand, using verification data like station observations has the advantage of being independent 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 multimodel predictions. The first option is to use each model's own analysis as verification dataset. However, this has the disadvantage that (i) the multimodel ensemble has no genuine own analysis, and (ii) it is difficult to draw conclusions from the resulting scores and skill scores when their calculation is based on different verification data. Another possibility is to use the average of all analyses of the participating models or some weighted average, also called multimodel analysis. However, averaging all analyses, including less accurate ones, might not necessarily lead to an analysis closest to reality. Additionally, such a multimodel 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 with 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 onwards.

There are several candidate reanalysis data sets available, including the ECMWF ERA-Interim (Simmons et al. 2007, Dee et al. 2011), ERA-40 (Uppala et al. 2005), and the NCEP-NCAR reanalysis (Kanamitsu et al. 2002, Kalnay et al. 1996). The ECMWF ERA-Interim re-analysis was chosen as main verification dataset. The two important advantages of this choice are the acknowledged high quality of this analysis product (it used 4D-Var, a recent version of the ECMWF forecast model and a T255L60 resolution) and the availability of this dataset for the entire training and test period (1991 up to near realtime). The obvious drawback of this option is that the ERA-Interim re-analyses are not entirely independent of one of the models in the comparison, the ECMWF model.

The question how much the results and possible conclusions might depend on the chosen verification dataset will be considered after the presentation of the main results obtained with ERA-Interim. At that point, forecasts will be validated against NCEP re-analyses (Kanamitsu et al., 2002) and surface observations to answer the question whether the results obtained with ERA-Interim as verification dataset also hold for other consistent long-term verification datasets.

In tropical areas, the analyses vary more strongly between different model systems than in the extra-tropics (e.g. Park et al., 2008). Consequently, using ERA-Interim as verifying analysis may lead to results that are not valid for other re-analyses or observations. Therefore, results have not been discussed for the tropics in this study.

3 Post-processing methods

3.1 Multimodel combination

The most basic way of constructing a multimodel ensemble is to combine the individual ensemble members from the contributing models with the same weight. This approach is not only an easy and robust way of combining different models, it also has been proven to be quite successful in improving on single-model predictions (Park et al., 2008; Hagedorn et al., 2005; Shin et al., 2003). However, there have been also many attempts to improve on this simple method, with some of these studies claiming to be able to improve on the equal weight method (e.g. Krishnamurti et al., 1999; Robertson et al., 2004), others concluding that it is very difficult to achieve significant improvements (Peng et al., 2002, Doblas-Reyes et al., 2005; Johnson and Swinbank, 2009). The main goal of this study is to compare the general level of improvements possible to achieve by either the multimodel or reforecast-calibration methodology, and not to investigate additional incremental improvements possible through further refinements of the individual methods. Therefore, we will investigate here only the standard multimodel ensemble constructed by giving equal weights to all contributing members. Through the different ensemble sizes, this implies an implicit weighting. That is, model systems with a higher number of ensemble members will have a greater impact in the final multimodel prediction than model systems with fewer members. The performance of TIGGE multimodel ensembles will be compared with single-model forecasts. The first includes all ensemble members from the nine model systems listed in Table 1 (i.e. 248 members, 239 perturbed plus 9 control runs), and the second, called TIGGE-4, consists of the 117 (113 perturbed and 4 unperturbed) members of the CMC, ECMWF, MetOffice and NCEP ensembles. Finally, also the three-model ensembles obtained by excluding one system from TIGGE-4 will be considered. Probabilistic scores are known do be moderately sensitive to the size of the ensemble. In the following comparisons, the scores have not been adjusted for the different ensemble sizes. This aspect will be discussed further in Section 4.4.

3.2 Bias correction

A bias-correction procedure is applied to 2-metre temperature fields. In fact, since the reforecasts are only available for the ECMWF EPS, the calibration procedure applied to the remaining models can only be based on a training dataset consisting of a limited number of previous forecasts. Taking into account this restriction, we apply a bias-correction procedure based on the forecasts that verify in the past 30 days (BC-30). The procedure itself calculates at every grid point x and for every lead time t a correction $c(x,t)$

$$c(x,t) = \frac{1}{N}\sum_{i=1}^{N} v_i(x,t) - e_i(x,t)$$

as the average difference between the verification $v_i(x,t)$ and the ensemble mean $e_i(x,t)$ for all cases in the training dataset (N=30). Verification $v_i(x,t)$ and ensemble mean $e_i(x,t)$ are valid i days prior to the forecast initial time.This correction is added to all ensemble members of the forecast to be corrected, i.e. the ensemble distribution is shifted as a whole by c.

The positive values of the CRPSS shown in Fig. 1 indicate the improvements achieved by applying this bias correction. The calibration corrects the forecasts of all models most effectively at short lead times, with the MetOffice and NCEP achieving the largest skill improvement at a lead time of one day. For the remaining lead times, the MetOffice forecasts continue to gain the most from the bias correction, followed by ECMWF, NCEP, CMC and the multimodel forecast. The multimodel system generally gains the least from this explicit bias correction, because it has overall a bias of smaller magnitude than the contributing models. This implicit bias correction in the multimodel concept arises as different models may have different biases that can compensate each other leading to a lower net bias (Pavan and Doblas-Reyes, 2000).

This type of 30-day bias correction has proven to have a significant impact only on near-surface variables like 2-m temperature. Therefore, the results shown for upper-air variables like 850-hPa temperatures or 500-hPa geopotential are not based on bias-corrected forecasts but compare DMO forecasts with the reforecast-calibrated ECMWF EPS described in the next section.

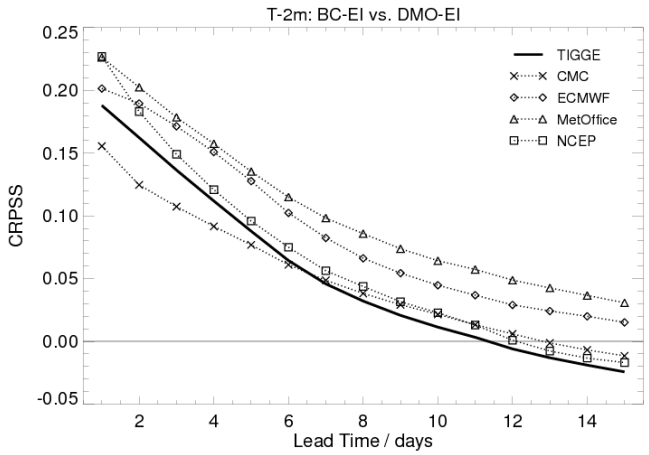


Figure 1 Impact of bias-correction on the relative skill of the predictions of 2-m temperature. The CRPSS is defined as CRPSS = 1 − CRPS(exp) / CRPS(ref) where exp refers to the BC forecast and ref to the DMO forecast. Scores are calculated using ERA-Interim as verification for forecast from the TIGGE-4 multimodel (solid line) and the single-models (CMC, ECMWF, Met Office, NCEP) starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N - 90°N). Here, the TIGGE multimodel consists of the four single models shown.

3.3 ECMWF reforecast calibration

The availability of the ECMWF reforecast dataset enables the application of more sophisticated calibration techniques than the simple bias-correction described above. Here, we use a combination technique “EC-CAL” based on the Non-homogeneous Gaussian Regression (NGR) and results from the 30-day bias-correction (BC-30). The NGR technique itself is described in detail in Gneiting et al. (2005) and has been already previously applied to ECMWF EPS forecasts (Hagedorn et al., 2008). 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 regression-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, in order to be able to compare the performance of the calibrated probabilities with the frequentist probabilities based on 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 the simple bias-correction based on the 30-day training data (BC-30) and the NGR calibration based on reforecasts (NGR-RF) is superior to the pure NGR-RF calibration, in particular for early lead times. As already seen in Fig. 1, the 30-day bias correction can improve the CRPS of the DMO by about 20% for early lead times. The reforecast based NGR calibration is even more effective, with improvements of more than 25% (Fig. 2). However, combining the NGR-RF and BC-30 ensembles can lead to further slight improvements. 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.

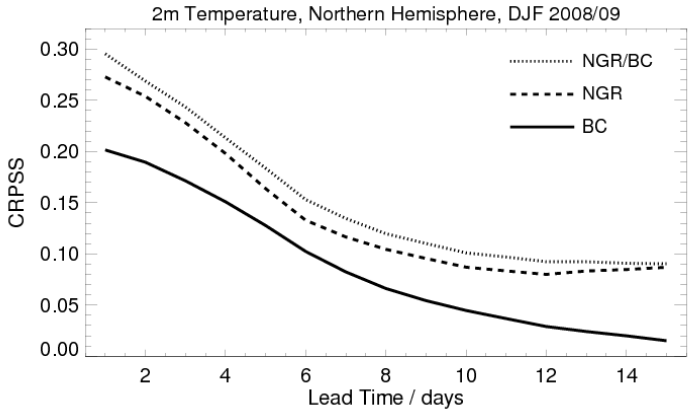


Figure 2 Impact of different calibration methods on the Continuous Ranked Probability Score (CRPS) for ECMWF direct model output of 2-m temperature. The CRPSS is defined as CRPSS = 1 − CRPS(exp) / CRPS(ref), with CRPS(exp) being the CRPS of the bias corrected forecasts (solid), the NGR calibrated forecast (dashed), and the NGR/BC model combination (dotted). CRPS(ref) is in all cases the CRPS of the DMO forecasts. The scores are averaged over start dates in DJF 2008/09 and over the Northern Hemisphere (20°N – 90°N).

4 Results

4.1 TIGGE multimodel ensembles versus single-model systems

A first impression on the level of skill of the single-model systems is given by comparing the CRPSS of the 850-hPa temperature over the Northern Hemisphere (20°N–90°N) for forecasts of the winter season DJF 2008/09 (Fig. 3a). The scores are based on uncalibrated DMO ensembles. The performance of the T850 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, changes considerably from one model to another. The climatological distribution is estimated from ERA-40 reanalyses in the period 1979–2001 (Uppala et al., 2005; Jung and Leutbecher, 2008). Since not all forecasting centres integrate their models to 15 days lead time, the performance of the multimodel 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. Except for the first two forecast days, this multimodel prediction does not significantly improve over the best single-model, i.e. the ECMWF EPS. The significance levels of the difference between the single-model systems and the multimodel ensemble have been assessed using a paired block bootstrap algorithm following Hamill (1999). Similar results can be observed for other variables like e.g. the bias-corrected 2-m temperature (Fig. 3b). Again, the best model is only for the first two to three days significantly worse than the multimodel ensemble, and their performance cannot be distinguished later on.

(a) **(b)**

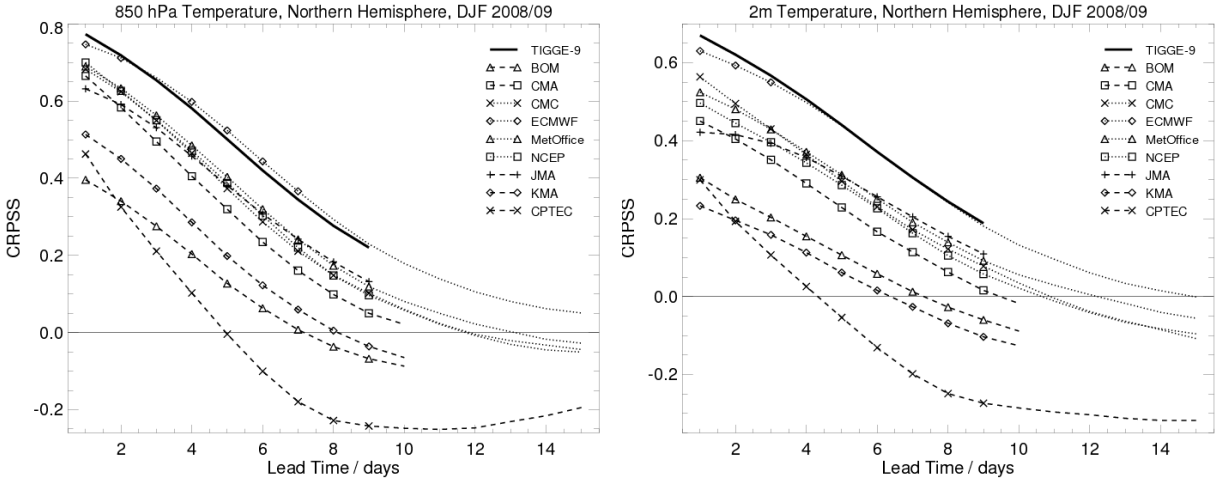


Figure 3: Continuous Ranked Probability Skill Score versus lead time for 850-hPa temperature DMO forecasts (a) and 2-m temperature BC-30 forecasts (b) in DJF 2008/09, averaged over the Northern Hemisphere (20°N – 90°N). Scores are for the TIGGE-9 multimodel and the nine contributing single-models. The climatological distribution is the reference forecast. Symbols are only plotted for cases in which the single-model score significantly differs from the multimodel score at the 1% significance level. All ensembles start from 12 UTC analyses.

The inability of the multimodel ensemble to significantly improve over the best single-model system might be caused by the fact that the nine-model ensemble includes some models with rather poor performance. In order to eliminate these possibly detrimental contributions, a new multimodel (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: CMC, ECMWF, MetOffice, and NCEP (Fig 4a and 4b). In fact, this reduced version of the full multimodel ensemble gives now 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 multimodel predictions. This model selection could also be interpreted as an extreme form of model weighting.

4.2 TIGGE multimodel ensemble versus reforecast-calibrated ECMWF

After having established a new benchmark for the best single-model, the ECMWF EPS, the question is now whether it might be possible to achieve similar improvements by calibrating the ECMWF EPS based on its reforecast dataset. Comparing the CRPSS of the reforecast-calibrated ECMWF EPS (EC-CAL) with the TIGGE-4 multimodel scores reveals that indeed the calibration procedure significantly improves ECMWF's scores (Fig. 4). Overall, the performance of the EC-CAL predictions is as good as the TIGGE-4 multimodel ensemble, and for longer lead times it can be even better. For Northern Hemisphere 850-hPa temperature predictions (Fig. 4a), the CRPSS of EC-CAL lies above the multimodel CRPSS for early lead times, and for longer lead times the skill scores are slightly lower than for the multimodel 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 T850 the reforecast-calibrated ECMWF forecasts are of comparable quality as the TIGGE-4 multimodel forecasts.

This result is confirmed when studying other variables, regions, or seasons (Fig. 4 b–d). In fact, the calibration is even more effective for longer lead times for 2-m temperature forecasts. This indicates that the systematic component of the error is more dominant in the case of 2-m temperature, and thus the calibration procedure is able to further reduce the Root Mean Square Error (RMSE) of the ensemble mean. In addition to the CRPSS, the area under the Relative Operating Characteristic has been compared (not shown). The results are consistent with those for the CRPSS indicating that the calibration also improves the resolution aspect.

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 multimodel forecasts for individual cases at single grid point locations (Fig. 5) gives an indication of how much (or how little) a real forecast product would change. On the one hand, there are locations at which the calibrated or multimodel 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

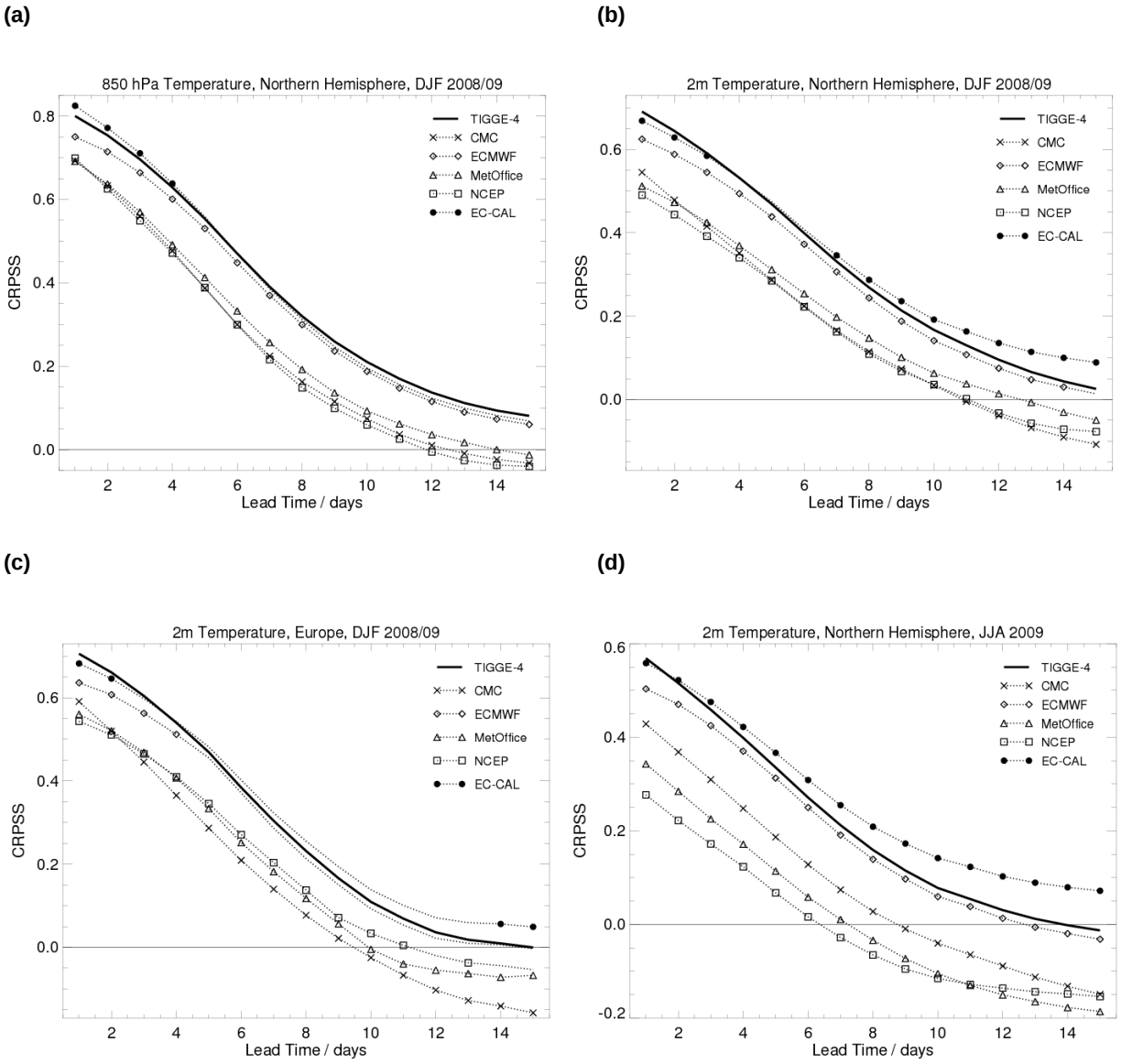
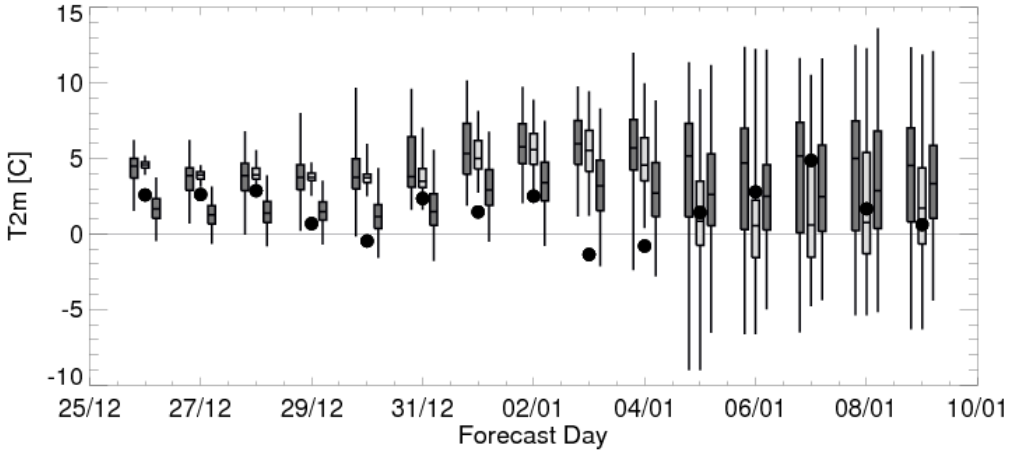


Figure 4 Continuous Ranked Probability Skill Score versus lead time for the TIGGE-4 multimodel (solid line), for the contributing single-models itself (dotted lines with symbols, CMC: cross, ECMWF: diamond, MetOffice: triangle, NCEP: square), and for the reforecast calibrated ECMWF forecasts (dotted lines with bullets). Symbols are only plotted for cases in which the single-model score significantly differs from the multimodel score at the 1% significance level. (a) 850-hPa temperature DMO and EC-CAL forecast scores averaged over the Northern Hemisphere (20°N - 90°N) for DJF 2008/09, (b) as in (a) but for 2-m temperature BC-30 and EC-CAL forecast scores, (c) as in (b) but averaged over Europe, (d) as in (b) but for the summer season JJA 2009. All ensembles start from 00 UTC analyses.

(a)



(b)

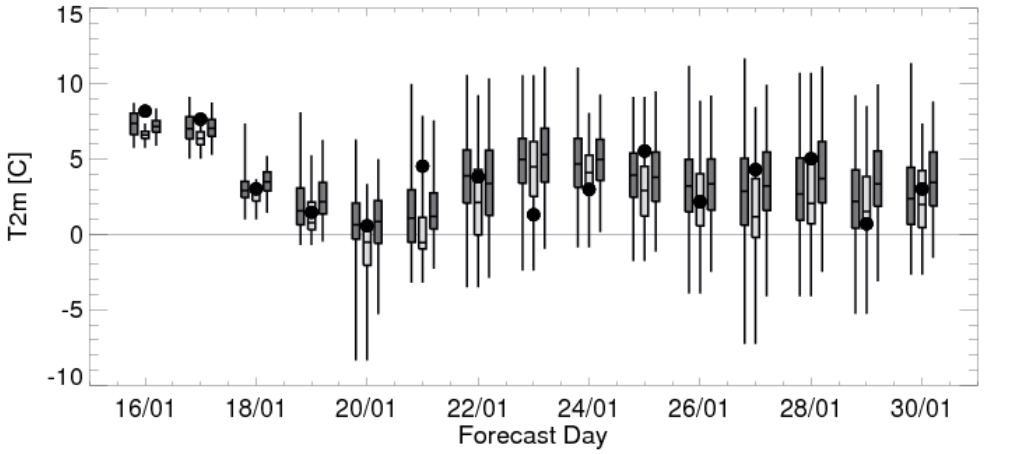


Figure 5 Examples of 2-m temperature forecast distributions at (a) Bologna, IT, (45.0°N, 12.5°E) and (b) London, UK, (52.5°N, 0.0°E) depicted as box-and-whisker plots, also called EPSgrams. (a) forecast started on (a) 25 December 2008 and (b) 15 January 2009. The inner box contains 50% of the ensemble members including the median (horizontal black line inside the box) and the wings represent the remaining ensemble members. For each forecast day three ensembles are shown, with the inner light-grey box-and-whisker depicting the DMO ensemble and the dark-grey box-and-whiskers representing the TIGGE (left) and the reforecast calibrated (right) ensembles. The verifying station observation is shown as black dot.

deficiencies, as can be seen for the early lead times of the EPSgram at the grid point closest to Bologna (Fig. 5a). However, as mentioned above, for longer lead times the predicted distributions tend to be close to the climatological distributions and it is less obvious how the improvements seen in the scores will translate into better decision-making. Additionally, there are also many locations with less pronounced systematic errors or spread deficiencies. At such locations, obviously the calibration has much less impact, as can be seen for example in the EPSgram at the grid point closest to London (Fig. 5b). In general the distributions are much more similar than in the case of the grid point closest to Bologna. Nevertheless, the calibration reduces some of the smaller biases and spread deficiencies. Comparing the multimodel ensemble with the ECMWF EPS suggests that the main improvement for the multimodel ensemble is caused by its larger spread and thus improved reliability rather than a pronounced bias-correction. We note that discussing individual cases as the above is not meant to lead to overall conclusions, but these examples are rather meant as illustration to raise some caution and avoid overinterpretation of potential improvements indicated by overall improved skill scores.

In order to further investigate the mechanisms behind the improvements, Figure 6 shows the RMSE of the ensemble mean and the spread of the different ensembles. Here, the spread is computed as the square root of the averaged ensemble variance and the RMSE as the square root of the averaged mean squared error. Ensemble forecasting aims to construct uncertainty information so that the truth can be considered as indistinguishable from the ensemble members of the forecast. Since both the ensemble mean and the analysis (or observation) have an error, a necessary (but not sufficient) condition for a reliable ensemble is for the sum of the ensemble variance and the variance of the analysis error to be close to the squared difference of the ensemble mean and the analysis (Saetra et al., 2004, Candille et al., 2007). It is difficult to quantitatively estimate the true analysis error variance, however, it is planned to extend the current diagnostic to incorporate this aspect in future work (e.g., using analysis error estimates from the Ensemble of Data Assimilations at ECMWF, Isaksen et al. 2010, multimodel estimates of Langland et al. 2008, or estimates from an Ensemble Kalman filter, e.g., Houtekamer and Mitchell 1998). Until this more quantitative assessment, we state just qualitatively that the spread of the ensemble should be smaller than the RMS error of the ensemble mean, especially at short forecast leads when analysis error is of similar magnitude to ensemble spread and ensemble mean error. For 2-m temperature, all single-model systems have an ensemble standard deviation that is smaller than the RMS error of the ensemble mean (Fig. 6a). CMC starts with the smallest spread deficiency at the beginning of the forecast, but due to a serious mismatch in the growth of spread and error the gap between spread and error grows with lead time. The remaining three models have a similar level of spread, however, the significantly lower RMSE of the ECMWF EPS is one of the main reasons for its significantly better probabilistic scores discussed before. The effect of combining the single-model systems or calibrating the ECMWF EPS can be seen in Figure 6b. The RMSE of the multimodel ensemble is slightly reduced for early lead times, but the most noticeable change is the close match of spread and error in particular up to a forecast range of day-6. According to Langland et al. (2008), the standard deviation of analysis error of temperature in the lower troposphere is expected to be of the order of 1 K. For ensemble standard deviations of 1, 2 and 3 K and an analysis error standard deviation of 1 K, the expected RMS difference of ensemble mean and analysis are 1.4, 2.2 and 3.2 K, respectively. Thus, it is likely that the multimodel ensemble would be diagnosed overdispersive in the early part of the forecast range if analysis uncertainty was accounted for in the verification. In contrast to that, the reforecast-calibrated ECMWF EPS has still a lower spread than error, though it is increased

compared to the original EPS spread. The reason for this is the above discussed methodology of combining the BC-30 and NGR-RF calibration. Applying the pure NGR calibration should lead to a near perfect match of spread and error, but as discussed above, the advantages of possible reductions in the systematic error provided by the 30-day bias-corrected ensemble may outweigh the slight disadvantage of a supposedly poorer 2nd-moment 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.

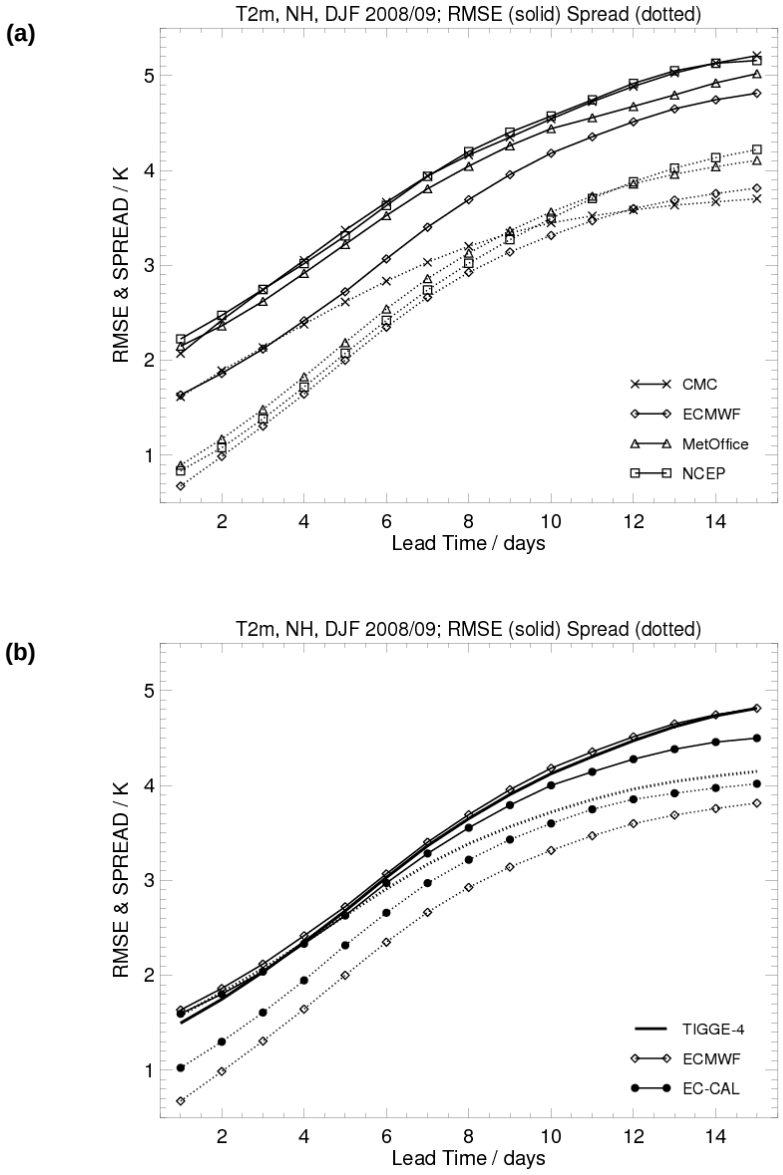


Figure 6 Root Mean Square Error of the ensemble mean (solid lines) and ensemble standard deviation ("spread", dotted lines) versus lead time. (a) results for the single-model BC-30 forecasts from CMC, ECMWF, MetOffice, and NCEP. (b) as in (a) but without the CMC, MetOffice and NCEP, including instead the reforecast calibrated ECMWF (EC-CAL) and TIGGE-4 multimodel. All scores are calculated for 2-m temperature forecasts starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N – 90°N).

It has to be noted that the above described mechanism of how the multimodel and reforecast-calibration techniques can correct for an inadequate representation of uncertainty does not apply to all model variables. As already demonstrated by Johnson and Swinbank (2009) and confirmed in Fig. 7, the multimodel ensemble performs hardly better than the ECMWF EPS when considering more large-scale dynamical variables like, for example, 500-hPa geopotential height forecasts. For early lead times, the CRPSS of the multimodel ensemble is indistinguishable from the CRPSS of the ECMWF EPS, and an apparently slight advantage of the multimodel scores for longer lead times is not statistically significant, except for lead times of 14 and 15 days (Fig. 7). Similarly, also the reforecast-calibrated forecast is indistinguishable from the uncalibrated ECMWF EPS. This result is consistent with the fact that calibration changes the ensemble mean RMS error and the ensemble spread only marginally (not shown). Thus, the bias of 500 hPa geopotential is much smaller than for the other fields. Furthermore, the spread-error relationship in the uncalibrated ECMWF EPS is already close to optimal. Such reliable variance predictions of 500 hPa geopotential were achieved first in the EPS after the model physics had been revised (Bechtold et al., 2008). Only variables which are not reliably predicted can be improved by either the multimodel or reforecast-calibration technique. Those variables that are influenced by surface processes tend to be such variables in present ensembles. However, including perturbations for surface variables like, for example, soil moisture (Sutton et al. 2006), or extending the stochastic physics (Buizza et al, 1999; Palmer et al 2009) to include the surface scheme, might contribute to a better representation of forecast uncertainty for near-surface variables and precipitation.

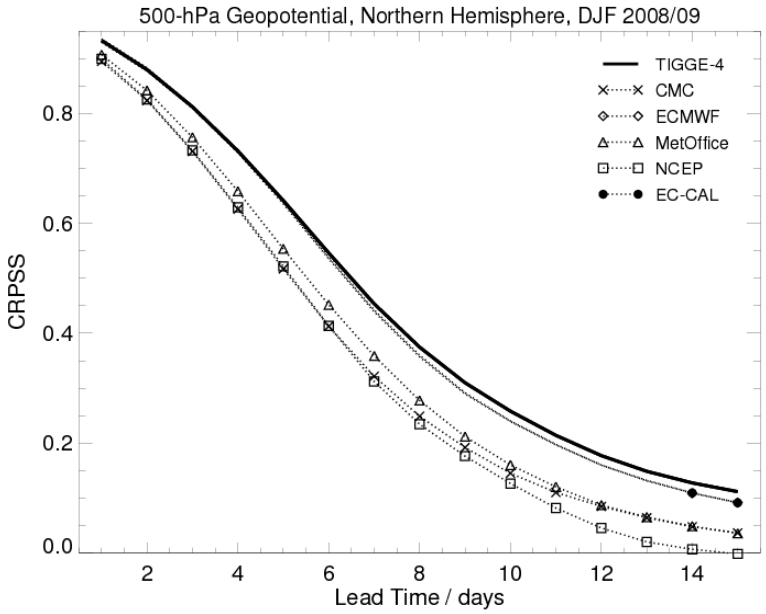


Figure 7 Continuous Ranked Probability Skill Score versus lead time for 500-hPa geopotential averaged over the Northern Hemisphere (20°N – 90°N) for DJF 2008/09. Results are shown for the TIGGE-4 multimodel (solid line), for the contributing DMO single-models itself (CMC, ECMWF, MetOffice, NCEP), and for the reforecast calibrated ECMWF forecasts (EC-CAL). Symbols are only plotted for cases in which the single-model score significantly differs from the multimodel score at the 1% significance level. Since ECMWF and EC-CAL are not significantly different from TIGGE-4, except for lead times of 14 and 15 days, their CRPSS lines are barely distinguishable.

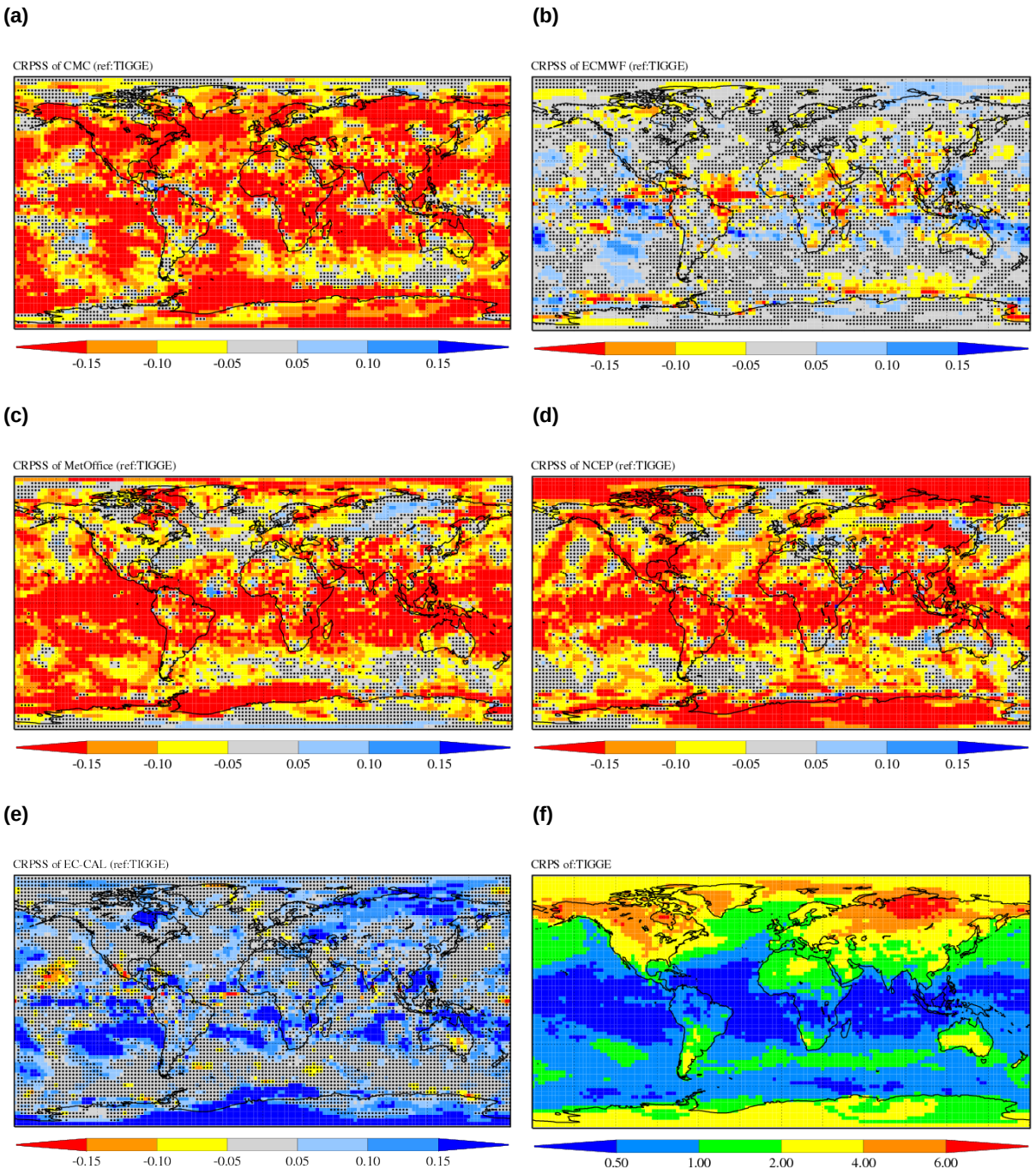


Figure 8: Relative performance of individual single-models with respect to the TIGGE-4 multimodel for 2-m temperature forecasts at 14 days lead time. Panel (a) to (e) display the CRPSS defined as CRPSS = 1 − CRPS(exp) / CRPS(TIGGE-4), with CRPS(exp) the CRPS of the single-models (a: BC-30 CMC, b: BC-30 ECMWF, c: BC-30 MetOffice, d: BC-30 NCEP, e: reforecast calibrated ECMWF). The CRPS of the TIGGE-4 multimodel is shown in panel (f). Scores are calculated at every grid point, in DJF 2008/09. In panel (a) to (e) grid points at which the difference between CRPS(exp) and CRPS(TIGGE-4) is not significant (at the significance level 0.1) are marked with a black bullet.

After considering forecast products for individual cases at grid point locations as well as area-averaged scores, it is interesting to study the character of the global distribution of the differences in the scores, i.e. whether it is possible to identify any geographical pattern where the multimodel ensemble has advantages (or is worse) compared to the single-model systems or the reforecast-calibrated ECMWF ensemble Figure 8 shows the 2-m temperature CRPSS of different models using the CRPS of the TIGGE-4 multimodel ensemble as the reference forecast. The fields are plotted for a lead time of 14 days where the difference in skill between TIGGE-4 and EC-CAL is more pronounced than at earlier lead times (cf. Fig. 4b). Apart from the ECMWF EPS, all single-model systems have a negative CRPSS over large areas, i.e. their performance is worse than the multimodel ensemble. Besides these large areas of negative skill scores, there are a few areas with non-significant differences. These areas with comparable performance between single- and multimodel are most pronounced for the MetOffice ensemble over the Eurasian and North-American continents. The comparison between the ECMWF EPS and the multimodel ensemble (Fig. 8b) reveals that - apart from a few areas mostly in the region of the tropical oceans - largely there are no significant differences. This confirms the results already seen in the area-averaged scores. In contrast to that, there are more distinct areas of significantly positive skill scores for the reforecast-calibrated ECMWF EPS (Fig. 8e). In particular the extended area of positive skill over the Eurasian continent corresponds with the higher CRPS of the multimodel forecasts over that area (Fig. 8f). This points out once again that in particular in areas of low skill (high CRPS values) the calibration can cause significant improvements.

4.3 Sensitivity to the verification dataset

Until now, all results shown have been based on using ERA-Interim reanalysis as verification dataset. In order to illustrate how much the results depend on the choice of verification, Figure 9 compares the CRPS using ERA-Interim as verification with the CRPS using NCEP reanalyses as verification for the 30-day bias corrected forecasts of 2-m temperature. The most obvious difference is that the CRPS generally increases, i.e. the scores of all models deteriorate when verified against NCEP reanalysis (Fig. 9b). The main impact can be found in the early forecast range. While the NCEP forecasts are obviously least affected, MetOffice and CMC forecasts get moderately worse, and ECMWF forecasts are most negatively affected. Thus, when interpreting the early-lead-time results using ERA-Interim as verification, one should bear this sensitivity to the verification data in mind. The slightly surprising effect that ECMWF's CRPS values are higher for the first two forecast days than on day-3 can be explained by the fact that changing the verification dataset leads to an increased RMSE, which is nearly constant for the first three forecast days (not shown here). On the other hand, the spread of the ensemble obviously does not change with the verification dataset, and as such the apparent under-dispersion, which is greatest for the very early forecast range, has an even further negative effect on the CRPS. However, when correcting for this apparent spread deficiency, i.e. applying the reforecast-based calibration technique EC-CAL instead of the simple 30-day bias correction, this effect can be greatly reduced and the scores improve significantly. This increased RMSE suggests that the NCEP reanalysis of 2-m temperature is significantly less accurate than the ERA-Interim reanalysis. Accounting for analysis uncertainty in the verification would reduce the need for inflating the second moment of the predicted distribution — in particular for the least accurate analyses.

Furthermore, comparing the overall performance of all models for lead times beyond four days, the same pattern as in the case of using ERA-Interim as verification (Fig. 9a) emerges. That is, the uncalibrated ECMWF EPS is distinctly better than all the other three single-model systems, the multimodel improves on these ECMWF scores, but the calibrated ECMWF forecasts are at least as good as the multimodel scores, if not better.

Further evidence of the general validity of the above described results and the independence of the findings from the chosen verification dataset is given by calculating the CRPSS at 250 European stations using 2-m temperature observations from WMO's Global Telecommunication System (Fig. 10). The same conclusions can be drawn as from the results obtained with the two different reanalysis datasets: (i) the ECMWF EPS has still overall the best performance compared to the other three single-model systems, (ii) the multimodel ensemble improves on the best single-model system in particular for early lead times, and (iii) the reforecast-calibrated ECMWF EPS is at least as good as the multimodel ensemble if not better. We note though, that there is an interesting difference in terms of the lead time dependency between these results and the corresponding verification against ERA-Interim (Fig. 4c). Using ERA-Interim, the advantage of EC-CAL over TIGGE-4 is largest towards the end of the forecast range while it is largest at the beginning of the forecast range when using station observations as verification. One reason for this apparent inconsistency may be that the calibration against station data also addresses the downscaling from the model scale to the station location.

One might argue that this set of 250 European stations is too limited and thus cannot be taken as supportive evidence for the conclusions drawn from verification against analyses. However, in previous studies we used observations from both the US (Hagedorn et al., 2008, Hamill et al, 2008) and Europe (Hagedorn, 2008). The studies showed a similar impact of calibration on probabilistic skill for the US and Europe. led to similar conclusions, confirming that the conclusions are quasi-independent from the verification area. Thus, one can expect that the general conclusions of this study would be valid for verification against observations in other mid-latitude regions.

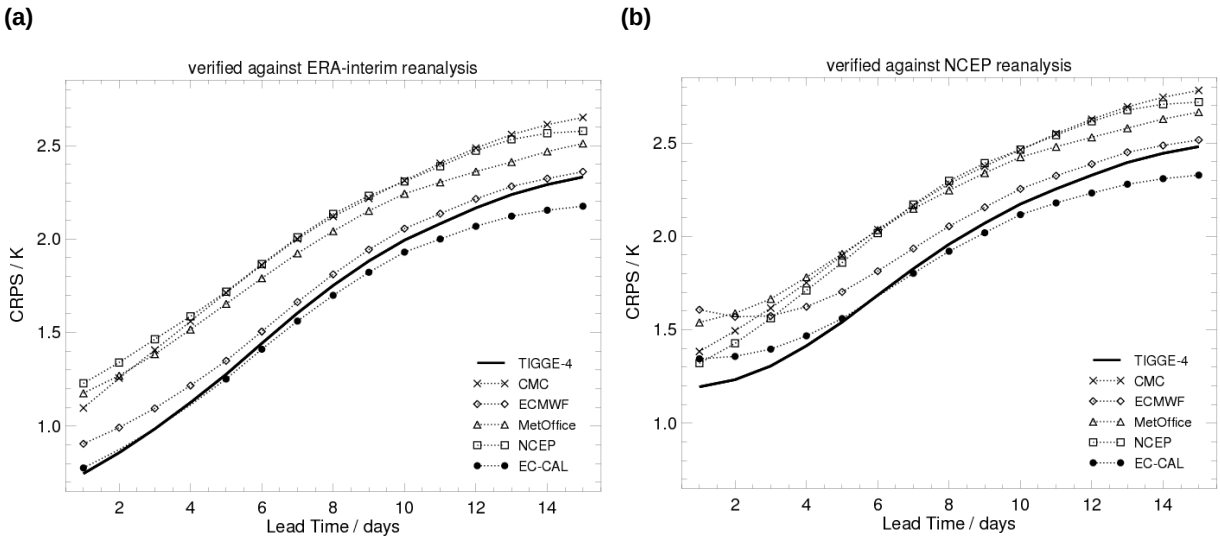


Figure 9: Impact of using (a) ERA-Interim and (b) NCEP reanalysis as verification dataset on the CRPS of 2-m temperature forecasts: TIGGE-4 multimodel, the contributing BC-30 single-models, and the reforecast-calibrated ECMWF forecasts (EC-CAL). CRPS is calculated for forecasts starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N – 90°N). Statistical significance indicated as in Fig. 7.

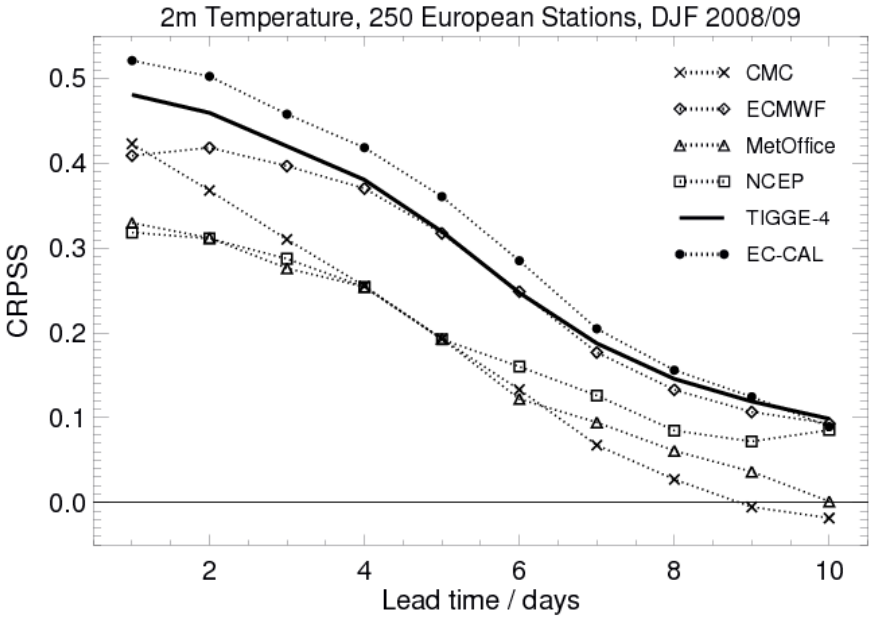


Figure 10: Continuous Ranked Probability Skill Score versus lead time for 2-m temperature forecasts verified against European SYNOP observations. Scores are for forecasts started in DJF 2008/09 and have been computed at 250 European stations. Results are shown for the BC-30 single-models, the TIGGE-4 multimodel and the reforecast calibrated ECMWF model (EC-CAL).

4.4 Individual multimodel contributions

The computational and organizational overhead of generating and collecting all individual model contributions and combining them to a consistent multimodel ensemble grows with the number of contributing models. Therefore, it is worth to investigate the added benefit each individual model can give to the multimodel system. For this purpose, reduced multimodel versions are considered with one individual model component removed from the TIGGE-4 multimodel mix. They are compared against the full multimodel version containing all four models (Fig. 11).

Removing the ECMWF EPS from the multimodel ensemble has the biggest impact, whereas the other models contribute to a lesser extent to the skill of TIGGE-4. It might be argued that one reason for this is the fact that by removing the ECMWF EPS the multimodel ensemble loses 51 members, whereas removing the other models implies only a loss of 21 or 24 members. Since the CRPS 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 the single-model adds to the multimodel 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. First of all, the difference of number of members between the full multimodel ensemble containing 117 members and the most reduced multimodel ensemble containing 66 members would require - according to equation (26) in Ferro et al. (2008) - 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, i.e. only 1% out of an 15% of the increase in the CRPS of the reduced multimodel 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 per se. Secondly, if we want to compare the performance from an operational rather than theoretical point of view, i.e. we are not interested in theoretical questions like “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. Therefore, the CRPS values have not been adjusted 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 which of the single-models contributes most to the multimodel success, a further question is whether the multimodel concept could lead to reduced costs by still keeping the same quality of forecasts. Assuming, for the sake of argument, that ECMWF could not afford anymore to provide its EPS forecasts, could a multimodel consisting of the remaining high-quality ensembles be as good as the ECMWF EPS on its own? Indeed, a TIGGE multimodel 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 2-m temperature forecasts which are as good as the best global ensemble (Fig. 12). However, this is only true for the ECMWF EPS when it has not been reforecast-calibrated. Running the complete ECMWF EPS, including its reforecasts, leads to a performance which cannot be achieved by any current multimodel version not containing ECMWF forecast information. These results are generally confirmed when considering other variables like upper-air temperature or wind components, though small differences in the relative performance, also depending on the region, can be observed (not shown here).

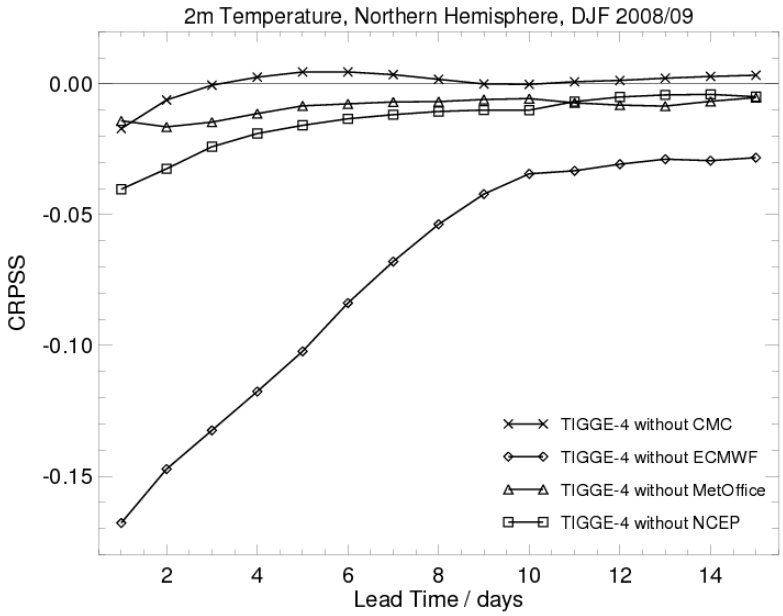


Figure 11: Skill of four different three-model TIGGE ensembles relative to the TIGGE-4 multimodel. The three-model TIGGE ensembles are obtained by removing one of the constituent models from TIGGE-4. All single-models are BC-30 corrected. The CRPSS is defined as CRPSS = 1 − CRPS(exp) / CRPS(TIGGE-4) with CRPS(exp) being the CRPS of one of the three-model TIGGE-(4–1) ensembles, respectively. All scores are calculated for 2-m temperature forecasts starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N – 90°N).

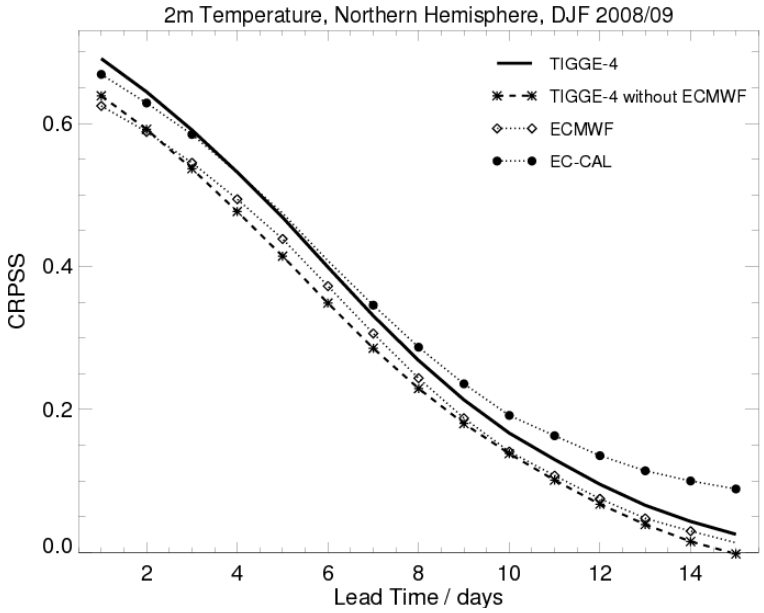


Figure 12 Continuous Ranked Probability Skill Score versus lead time for 2-m temperature forecasts starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N – 90°N). Results are shown for the TIGGE-4 multimodel containing all four BC-30 forecasts from CMC, ECMWF, MetOffice, and NCEP (solid line), the TIGGE-4 multimodel without ECMWF forecasts, i.e. containing only the three BC-30 forecasts from CMC, MetOffice, and NCEP (dashed line with stars), the single BC-30 ECMWF forecasts (dotted line with diamonds), and the re-forecast calibrated ECMWF forecasts (dotted line with bullets). Symbols are omitted for cases in which the score does not significantly differ from the TIGGE-4 multimodel score at the 1% significance level.

5 Summary and discussion

The main aim of this study was to compare the probabilistic skill of TIGGE multimodel forecasts with the reforecast-calibrated ECMWF EPS in the extra-tropics for 850-hPa temperature, 2-m temperature and 500 hPa geopotential. A major issue in any verification study is the choice of verification dataset. Most results were obtained with ERA-Interim. This reanalysis was chosen because: (i) one intent of this study was to evaluate reforecast-calibrated ECMWF products (requiring forecasts and observations or analyses from many years past); (ii) use of analyses rather than observations facilitated a verification over a larger region including oceans, and (iii) the highest-quality analyses were deemed preferable. Since the choice of ERA-Interim as verification data set may favour ECWMF forecasts, we demonstrated the relative impact of using a different reanalyses dataset as verification.

The performance of nine single-model systems from the TIGGE archive was compared with the performance of the full TIGGE multimodel, consisting of all these nine models. This full multimodel version did not improve on the best single-model, the ECMWF EPS. However, when combining only the leading four single-model ensembles (CMC, ECMWF, MetOffice, and NCEP), the multimodel ensemble outperformed the ECMWF-only EPS forecasts. Though, this result does not apply to all model variables and all lead times as demonstrated for 500 hPa geopotential. However, by taking advantage of the reforecast dataset which was available for the ECMWF EPS and using it as training dataset to produce reforecast-calibrated forecasts, the ECMWF EPS scores were improved to such an extent that its overall performance was as good as the TIGGE multimodel system, and often better.

The reforecast calibration procedure was particularly helpful at locations with clearly detectable systematic errors like areas with complex orography or coastal grid points. In such areas, the calibration procedure not only corrects for errors on the spatial scales resolved by the model but also consists of a statistical downscaling of the forecasts. The multimodel approach, in contrast, 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 were improved to a similar extent by both approaches. Experiments assessing the contribution of the individual components of the multimodel system demonstrated that the ECMWF EPS was the single most important source of information for the success of the multimodel ensemble.

Which of the two discussed post-processing methods would be the most appropriate choice for a modelling centre? To answer this, one has to consider also the technical overhead of producing multimodel 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 multimodel system, and if its users or operational forecasters ask for multiple solutions suggested by individual models, then the multimodel concept might be the method of choice. However, for a forecasting centre reluctant to take on the potential risks, the technical overhead, and the potential data unavailability issues inherent in a multimodel system, using the reforecast-calibrated ECMWF EPS forecasts rather than a logistically more complex multimodel system seems to be a more than appropriate choice. On top of the scientific, technical and logistical considerations, decisions on the optimal model and post-processing design might also depend on financial aspects, i.e. the fact that not all single-model forecast systems

discussed here are freely available in real-time for operational applications. As such, every user or operational centre has to decide for themselves on their individual cost-benefit relation and whether it might be worth investing in a system, which initially might require higher investments but potentially in the long run could lead to higher overall benefits. The main driver for providing reforecasts for the ECMWF EPS came from the need to bias-correct longer sub-seasonal forecasts, and once the data were available they proved very valuable for also calibrating the medium-range forecasts. The computational cost of running the reforecasts is about 13% of the cost of running the real-time EPS at ECMWF.

Calibration offers significant prospects for forecast improvement. However, one has to acknowledge that there is probably no universally optimal calibration of probabilistic forecasts. Different users may want to adapt the calibration procedure according to their needs. Someone, who would like to predict values at a station location will calibrate using data from that specific station while someone else who is interested in values on scales of, say O(100 km), would calibrate against analyses or upscaled observations. Some users may need predictions of the joint probability distribution of several variables. For these users, the NGR calibration technique is unsuitable as it would be independent for each variable.

Finally, 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 so distant future. In that case, it would be interesting to study the relative benefits achievable for reforecast-calibrated multimodel or single-model systems, respectively. Furthermore, we suggest exploring the relative merits of multimodel versus reforecast-calibrated predictions for other user-relevant variables like precipitation and wind speed.

Acknowledgements

The authors would like to thank the ECMWF Operational Department, in particular Manuel Fuentes and Baudouin Raoult, for their invaluable technical support related to handling the complex TIGGE datasets. The reviewers' comments on an earlier version of this manuscript are appreciated.

References

Bechtold P, Köhler M, Jung T, Doblas-Reyes F, Leutbecher M, Rodwell MJ, Vitart F, Balsamo G. 2008. Advances in Simulating Atmospheric Variability with the ECMWF model: From Synoptic to Decadal Time-scales. *Quart. J. Roy. Meteor. Soc.* **134:** 1337–1351.

Bougeault P, Toth Z, Bishop C, Brown B, Burridge D, Chen DH, Ebert B, Fuentes M, Hamill TM, Mylne K, Nicolau J, Paccagnella T, Park Y-Y, Parsons D, Raoult B, Schuster D, Silva Dias P, Swinbank R, Takeuchi Y, Tennant W, Wilson L, Worley S. 2010. The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.* **91:** 1059–1072.

Buizza, R, Miller M, Palmer TN. 1999. Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.* **125:** 2887–2908.

Candille, G, Côté C, Houtekamer PL, Pellerin G. 2007. Verification of an Ensemble Prediction System against Observations. *Mon. Wea. Rev.* **135:** 2688–2699.

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kållberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette J-J, Park B-K, Peubey C, de Rosnay P, Tavolato C, Thépaut J-N, Vitart F. 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.* **137:** 553–597. doi: 10.1002/qj.828

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

Ferro CAT, Richardson DS, Weigel AP. 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorol. Appl.* **15:** 19–24.

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

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

Hagedorn R, Doblas-Reyes FJ, Palmer TN. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. *Tellus-A* **57:** 219–233.

Hagedorn R, Hamill TM, Whitaker JS. 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: two-meter temperatures. *Mon. Wea. Rev.* **136:** 2608–2619.

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

Hamill TM, Whitaker JS. 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev*. **134:** 3209–3229.

Hamill TM, Whitaker JS. 2007. Ensemble calibration of 500-hPa geopotential height and 850-hPa and 2-m temperatures using reforecasts. *Mon. Wea. Rev*. **135:** 3273–3280.

Hamill TM, Whitaker JS, Wei X. 2004. Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev*. **132:** 1434–1447.

Hamill TM, Whitaker JS, Mullen SL. 2006. Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.* **87:** 33–46.

Hamill TM, Hagedorn R, Whitaker JS. 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble forecasts. Part II: Precipitation. *Mon. Wea. Rev*. **136:** 2620–2632.

Hersbach H. 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Wea. Forecasting* **15:** 559–570.

Houtekamer PL, Mitchell HL. 1998. Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev*. **126:** 796–811.

Isaksen L, Bonavita M, Buizza R, Fisher M, Haseler J, Leutbecher M, Raynaud L. 2010. Ensemble of data assimilations at ECMWF. *ECMWF Technical Memorandum* **636**.

Johnson C, Swinbank R. 2009. Medium-Range multi-model ensemble combination and calibration. *Quart. J. Roy. Meteor. Soc.* **135:** 777–794.

Jung T, Leutbecher M, 2008. Scale-dependent verification of ensemble forecasts. *Quart. J. Roy. Meteor. Soc*. **134:** 973–984.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D. 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc*.**, 77,** 437–471.

Kanamitsu M, Ebisuzaki W, Woollen J, Yang S-K, Hnilo JJ, Fiorino M, Potter GL. 2002. NCEP-DOE AMIP-II Reanalysis (R-2), *Bull. Amer. Meteor. Soc.* **83:** 1631–1643.

Krishnamurti TN, Kishtawal CM, LaRow TE, Bachiochi DR, Zhang Z, Williford CE, Gadgil S, Surendran S. 1999. Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble. *Science* **285:** 1548–1550.

Langland RH, Maue RN, Bishop CH. 2008. Uncertainty in atmospheric temperature analyses. *Tellus A* **60:** 598–603.

Matsueda M, Tanaka HL. 2008. Can MCGE outperform the ECMWF ensemble? *SOLA* **4:** 77–80.

Palmer TN, Alessandri A, Andersen U, Cantelaube P, Davey M, Délécluse P, Déqué M, Díez E, Doblas-Reyes FJ, Feddersen H, Graham R, Gualdi S, Guérémy J-F, Hagedorn R, Hoshen M, Keenlyside N, Latif M, Lazar A, Maisonnave E, Marletto V, Morse AP, Orfila B, Rogel P, Terres J-M, Thomson MC. 2004. Development of a European Multi-Model Ensemble System for Seasonal to Inter-Annual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.* **85:** 853–872.

Palmer, TN, Buizza R, Doblas-Reyes F, Jung T, Leutbecher M, Shutts GJ, Steinheimer M, Weisheimer A. 2009. Stochastic parametrization and model uncertainty. ECMWF Technical Memorandum **598**, ECMWF, Shinfield Park, Reading RG2-9AX, UK, pp. 42.

Park Y-Y, Buizza R, Leutbecher M. 2008. TIGGE: preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc.* **134:** 2029–2050.

Pavan V, Doblas-Reyes FJ. 2000. Multi-model seasonal hindcasts over the Euro-Atlantic: skill scores and dynamic features. *Climate Dyn.* **16:** 611–625.

Peng P, Kumar A, van den Dool H, Barnston AG. 2002. An analysis of multimodel ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.* **107(D23):** 4710, doi:10.1029/2002JD002712.

Robertson AW, Lall U, Zebiak SE, Goddard L. 2004. Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.* **132:** 2732–2744.

Saetra Ø,Hersbach H, Bidlot J-R, Richardson DS. 2004. Effects of Observation Errors on the Statistics for Ensemble Spread and Reliability. *Mon. Wea. Rev.* **132:** 1487–1501.

Shin DW, Cocke S, Larow TE. 2003. Ensemble Configurations for Typhoon Precipitation Forecasts. *J. Meteor. Soc. Japan.* **81 (4):** 679–696.

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

Sutton C, Hamill TM, Warner TT. 2006. Will perturbing soil moisture improve warm-season ensemble forecasts? A proof of concept. *Mon. Wea. Rev.* **134:** 3174–3189.

Uppala SM, KÅllberg PW, Simmons AJ, Andrae U, Bechtold VDC, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA, Li X, Onogi K, Saarinen S, Sokka N, Allan RP, Andersson E, Arpe K, Balmaseda MA, Beljaars ACM, Berg LVD, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Hólm E, Hoskins BJ, Isaksen L, Janssen PAEM, Jenne R, Mcnally AP, Mahfouf J-F, Morcrette J-J, Rayner NA, Saunders RW, Simon P, Sterl A, Trenberth KE, Untch A, Vasiljevic D, Viterbo, P, Woollen J. 2005. The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.* **131:** 2961–3012.

Weigel AP, Bowler NE. 2009. Comment on 'Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?'. *Quart. J. Roy. Meteor. Soc.* **135:** 535–539.

Weigel AP, Liniger MA, Appenzeller C. 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.* **134:** 241–260.

Weigel AP, Liniger MA, Appenzeller C. 2009. Seasonal Ensemble Forecasts: Are Recalibrated Single Models Better than Multimodels? *Mon. Wea. Rev.* **137:** 1460–1479.

Wilks DS, Hamill TM. 2007. Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.* **135:** 2379–2390.