Comparing TIGGE multi-model forecasts with reforecast-calibrated ECMWF ensemble forecasts

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

[1]*European Centre for Medium-Range Weather Forecasts, Reading, UK*

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

Submitted to *Monthly Weather Review*

12 July 2010

Corresponding Author address:

Renate Hagedorn
ECMWF
Shinfield Park
Reading, RG2 9AX
United Kingdom

Tel: +44 (0) 1189499257
Fax: +44 (0) 1189869450
Email: renate.hagedorn@ecmwf.int

ABSTRACT

The relative benefits of multi-model 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). Considering the statistical performance of global probabilistic forecasts of 850-hPa temperature and 2-m temperatures, a multi-model 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 multi-model system, consisting of only the four best ensemble systems, provided by Canada, the US, the UK and ECMWF, showed an improved performance. This multi-model ensemble provided a new benchmark for the single-model systems contributing to the multi-model. However, reforecast-calibrated ECMWF EPS forecasts were of comparable or superior quality to the multi-model predictions, when verified against analyses or observations. This improved performance was achieved by using the ECMWF reforecast dataset to correct for systematic errors and spread deficiencies. Further experimentation revealed that the ECMWF EPS was the main contributor for the improved performance of the multi-model ensemble; that is, if the multi-model 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. On the contrary, only through the diversity of approaches the overall goal can 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 in order to provide improved predictions. However, here we will concentrate on two specific methods: (i) combining single-model forecasts into a multi-model[1] forecast, and (ii) calibrating single-model forecasts with the help of specific training datasets. Both of these approaches have been 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 DEMETER project (Palmer et al., 2004), leading to a number of publications on the potential superiority of multi-model predictions on the seasonal timescale (see e.g. the special issue on the DEMETER project in Tellus-A 57(3)). Studying the rationale behind the success of multi-model ensembles, Hagedorn et al. (2005) concluded that

[1] Note that the term single-model or multi-model does not refer only to the forecast model itself but encompasses the whole model system including the data assimilation system.

"the key to the success of the multi-model 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 multi-model 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 multi-model ensemble can still improve the performance of this best model. This result is particularly relevant in the context of applying the multi-model 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 multi-model 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 multi-model ensemble?" As Weigel and Bowler (2009) pointed out in their theoretical study "it is possible to construct and combine reliable forecasts such that the multi-model 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, Johnson and Swinbank, 2009).

The second post-processing method under discussion 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, 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 model system is a major factor of the success of the calibration technique used in this study.

One can expect that both these 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 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 discussed in section 4 with a summary and conclusions following in section 5.

2. Datasets

a. 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. 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). 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.

All forecasts have been interpolated to a common 2.5° x 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° x 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° x 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 1.5° x 1.5° grid essentially did not change the result (corresponding figures not shown here).

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 a certain variable (Hersbach, 2000). The CRPS is very suitable for systems issuing continuous probability forecasts, however, here we evaluate all forecasts in the classical way of probabilities retrieved from discrete ensemble members.

b. Training datasets

Every calibration 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 also can be beneficial to restrict the use of the available training data to only that data for a similar time of year. Using the ECMWF reforecast dataset (Hagedorn, 2008) enables us 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 were used for the initialization, and from that date onwards only ERA-interim analyses (Simmons et al., 2007) have been used to provide a more consistent initialization dataset. The training dataset used in this study consists of the reforecasts produced for the five calendar days 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, 850-hPa and 2-m temperature.

At present, this type of reforecast dataset is only available for the ECMWF EPS and not for the remaining single-models of 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 multi-model 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).

c. 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 multi-models. On one hand, using model independent verification data like 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. The first option is to use each model's own analysis as verification dataset. However, this has the disadvantage that (i) the multi-model ensemble has no own analysis, and (ii) it would be difficult to draw conclusions from the resulting scores and skill scores when their calculation is based on different reference datasets. Another possibility is to use the average of all analyses of the participating models or some weighted average, also called multi-model analysis. 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 onwards.

There are several candidate reanalysis data sets available, including the ECMWF ERA-interim (Simmons et al. 2007), ERA-40 (Uppala et al. 2005), and the NCEP-NCAR reanalysis (Kanamitsu et al. 2002, Kalnay et al. 1996). Considering the competing requirements of being as fair as possible to all models involved and being as accurate as possible, it was decided to place greater emphasis on the accuracy of

the results and therefore to choose the ECMWF ERA-interim re-analysis (Simmons et al. 2007) 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 T255L40 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 certainly not entirely independent of one of the models in the comparison, the ECMWF model. However, before discussing the relative performance of the forecasts itself, the sensitivity using different analyses as the verification is demonstrated. This will enable the reader to evaluate the potential impact of the choice of verification data set on the subsequent results.

The relative impact of using ERA-interim as verification instead of the multi-model analysis can be estimated by calculating the Continuous Ranked Probability Skill Score (*CRPSS*) from the CRPS that uses ERA-interim as verification ($CRPS_{ERAI}$) and the CRPS that uses the multi-model analysis as verification ($CRPS_{MMA}$).

$$CRPSS = 1 - \frac{CRPS_{ERAI}}{CRPS_{MMA}}$$

If the results were insensitive to the choice of verification dataset, $CRPS_{ERAI}$ and $CRPS_{MMA}$ would be equal, and consequently the CRPSS would be zero. However, the negative values of the CRPSS in Figure 1a and 1b indicate a higher $CRPS_{ERAI}$, i.e. a worse performance of the models when they are verified against ERA-interim instead of the multi-model analysis. Generally, the skill scores worsen for all models, though the level of changes depends strongly on the lead time, variables or areas under investigation. For example, in tropical areas the analyses of the single-models vary

extremely between different model systems (Park et al., 2008), and consequently using ERA-interim as verifying analysis would have a strong impact on the results. Therefore results are not discussed for the tropics in this study. The choice of verification dataset has least impact for longer lead times and extra-tropical upper air fields, with noticeable differences occurring mainly during the first three to four days (Fig. 1a). Assuming that we would like the impact of changing the verification to ERA-interim to be similar for each individual model, the CRPSS curves in Fig. 1a and 1b should be as close as possible. As such, it is important to note that *all models* follow the same general pattern of greater sensitivity to the choice of analysis for early lead times and less sensitivity at longer lead time. However, the performance of the CMC model worsens the most in the given examples, and one should keep in mind this fact when interpreting later results on the relative performance of the forecasts. There are larger differences between the models when looking at near surface variables like 2-m temperature, where each analysis may have its own systematic bias, and therefore the overall sensitivity to the choice of analysis is much more pronounced than for upper-air fields (Fig. 1b, note the different scale to Fig. 1a). Additionally, it can be seen that up to a lead time of three days, the ECMWF T2m forecasts likely benefit from the fact that their initializing analyses are closer to the ERA-interim verification than the initialization of the other models.

In order to reduce this stronger impact close to the surface, a bias correction (BC) is applied to all forecasts using the last 30 days of forecasts and ERA-interim analyses (for details see section 3.b). The positive values of the CRPSS shown in Fig. 1c 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 multi-model forecast. The multi-model 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 multi-model 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).

The final comparison between the bias-corrected forecasts verified against ERA-interim and the uncorrected forecasts verified against the multi-model analysis demonstrates that the bias correction reduces the strong lead-time dependence of the impact of using ERA-interim (Fig. 1d compared to Fig. 1b). Apart from the very early lead times of up to three days, the impact is now relatively similar for all models.

The question how much the results and possible conclusions might depend on the chosen verification dataset will be reconsidered after the presentation of the main results. At that point, forecasts will be validated against NCEP re-analyses (Kanamitsu et al., 2002) and surface observations to demonstrate how robust the findings are when based on using ERA-interim as the verification dataset. All these results, together with the fact that a consistent dataset is indispensable for the reforecast calibration, justify the choice of using ERA-interim as verification dataset. However, for the time being we ask the reader to keep in mind these initial findings on the impact of using ERA-interim as verification dataset, and to relate all further conclusions to this initial discussion.

3. Post-processing methods

a. Multi-model combination

The most basic way of constructing a multi-model ensemble is to simply 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 multi-model 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 multi-model ensemble constructed by giving equal weights to all contributing members, noting that through the different number of members in the individual EPSs 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. The performance of two TIGGE multi-model 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 only the 117 (113 plus 4) members of the CMC, ECMWF, MetOffice and NCEP ensembles.

b. Bias correction

A bias-correction procedure is applied to near-surface fields. As mentioned in section 2.c, this procedure aims to reduce the impact of using ERA-interim as verification dataset (see also Fig. 1c & 1d) and makes the comparison with the reforecast-calibrated ECMWF forecasts fairer. 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 past 30 days of forecasts (BC-30). The procedure itself calculates at every grid point x and for every lead time t a correction $c(x,t)$

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

as the average difference between the ensemble mean $e(x,t)$ and the verification $v(x,t)$ for all N cases in the training dataset. This correction is applied to all ensemble members of the forecast to be corrected, i.e. the ensemble distribution itself is not altered in itself but is shifted as a whole. However, because this type of 30-day bias correction has proven to have a significant impact only on near-surface variables like 2-m temperature, the results shown for upper-air variables like 850-hPa temperatures or 500-hPa geopotential are not based on BC-30 forecasts but compare DMO forecasts with the reforecast-calibrated ECMWF EPS described in the next section.

c. 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 pure bias-correction (BC). 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 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, 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. 1c, 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. 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 the fact that through this procedure any non-Gaussian characteristics of the original ensemble may be retained to some degree.

4. Results

a. TIGGE multi-model 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 for forecasts of the winter season DJF 2008/09 (Fig. 3a). The scores are based on uncalibrated DMO ensembles since (i) applying the bias-correction has no significant impact for upper-air variables like T850 and (ii) we do not need to account for the choice of verification dataset in the case of T850, as demonstrated in section 2.c. 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 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. Except for the first two forecast days, this multi-model prediction does not significantly improve over the best single-model, i.e. the ECMWF EPS. Note that the significance levels of the difference between the single-model systems and the multi-model 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 multi-model ensemble, and their performance cannot be distinguished later on.

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 also the models with rather poor performance. In order 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: CMC, ECMWF, MetOffice, and NCEP (Fig 3c and 3d). In fact, this reduced version of the full multi-model 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 multi-model predictions.

b. TIGGE multi-model 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 multi-model scores previously shown in Figure 3 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 multi-model ensemble, and for longer lead times it can be even better. For 850-hPa temperature predictions the EC-CAL's CRPSS lies above the multi-model CRPSS 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 T850 the reforecast-calibrated ECMWF forecasts are of comparable quality as the TIGGE-4 multi-model forecasts.

This result is confirmed when studying other variables, regions, or seasons (Fig. 4 b-d). In fact, for 2-m temperature forecasts the calibration is even more effective for longer lead times. 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. 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 (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 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, 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 are already close to the climatological distributions, i.e., 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, obviously the calibration has much less impact, as can be seen for the example of 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 multi-model ensemble with the ECMWF EPS suggests that the main improvement for the multi-model 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 indeed further investigate the mechanisms behind the improvements, Figure 6 shows the RMSE of the ensemble mean and the spread of the different ensembles. Ensemble forecasting aims to construct uncertainty information so that the observations can be considered as indistinguishable from the ensemble members of the forecast. Since both the ensemble mean and the analysis have an error, a necessary (but not sufficient) condition for a reliable ensemble is for the sum of the squared spread of the ensemble 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., multi-model estimates of Langland et al. 2008, or direct estimates via the ensemble Kalman filter, e.g., Houtekamer and Mitchell 1998). Until this more quantitative assessment, we just qualitatively state that the spread of the ensemble should be somewhat smaller than the standard deviation of the ensemble mean, especially at short forecast leads when analysis error is of similar magnitude to spread and error. However, for 2-m temperature all single-model systems are seriously under-dispersive (Fig. 6a), i.e. much more than could be explained by the analysis error variance currently not accounted for. 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 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 relationship compared to the MetOffice and NCEP ensembles, it is also 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 multi-model ensemble is slightly reduced for early lead times, but the most noticeable change is the very much improved spread-error relation in particular up to a forecast range of 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 above discussed methodology of combining the BC-30 and NGR-RF calibration. Applying the pure NGR calibration should lead to a near perfect spread-error relation, 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.

It has to be noted that the above described mechanism of how the multi-model 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 multi-model ensemble performs hardly better than the ECMWF EPS when considering more large-scale dynamical variables like, for example, 500-hPa geopential height forecasts. For early lead times, the CRPSS of the multi-model ensemble is indistinguishable from the CRPSS of the ECMWF EPS, and an apparently slight advantage of the multi-model 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 form the uncalibrated ECMWF EPS. This indicates that the uncalibrated EPS may be suffering from much less bias in this field (Hamill and Whitaker 2007, Fig. 5), and/or it is already optimally tuned in terms of its spread-error relationship, and no other systematic errors can be detected and corrected with the reforecast-calibration or multi-model technique. The inability of both the multi-model and calibration methodology to improve on the ECMWF EPS seems to be caused by the fact that - for 500-hPa geopential height forecasts - uncertainties are already well represented, amongst others by ECMWF's stochastic physics scheme (Buizza et al., 1999; Palmer et al., 2009). Only variables which are more influenced by surface processes can be improved by either the multi-model or reforecast-calibration technique. However, including perturbations for surface variables like, for example,

soil moisture (Sutton et al. 2006), or extending the stochastic physics to include the surface scheme, might contribute to a better representation of forecast uncertainty for near-surface variables and precipitation.

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 multi-model ensemble has advantages (or is worse) compared to the single-model systems or the reforecast-calibrated ECMWF ensemble (Fig. 8). Apart from the ECMWF EPS, all single-model systems have a negative CRPSS over large areas, i.e. their performance is worse than the multi-model 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 multi-model are most pronounced for the MetOffice ensemble over the Eurasian and North-American continents. The comparison between the ECMWF EPS and the multi-model 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 multi-model 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.

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 resulting when using ERA-interim and NCEP reanalyses as verification for the 30-day bias corrected forecasts. 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, as in the previous comparison when the multi-model analysis was used as verification. While the NCEP forecasts are obviously least affected, MetOffice and CMC forecasts get moderately worse, and ECMWF forecasts are most negatively affected. 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. 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 multi-model improves on these ECMWF scores, but the calibrated ECMWF forecasts are at least as good as the multi-model 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). In fact, the general messages of the previous results are still confirmed: (i) the ECMWF EPS has still overall the best performance compared to the other three single-model systems, (ii) the multi-model 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 multi-model ensemble if not better. 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). Both studies led to similar conclusions, confirming that the conclusions are quasi-independent from the verification area. Thus we are confident that the general conclusions also in this case would not change if we used other non-European observations.

c. Individual multi-model contributions

The computational and organizational overhead of collecting all individual model contributions and combine them to a consistent multi-model ensemble grows with the number of contributing models. Insofar it is worth to investigate 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 (Fig. 11). It is very 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 the fact that by removing the ECMWF EPS the multi-model 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 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. First of all, 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 - 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 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 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, then 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 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 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 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 (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 multi-model 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).

5. Summary and conclusions

The main aim of this study was to compare the relative benefits of TIGGE multi-model forecasts versus the reforecast-calibrated ECMWF EPS. A major issue in such a verification study was the choice of verification dataset. ERA-Interim was chosen here 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 different reanalyses as verification. It was shown that the level of advantage the ECMWF EPS gained when using ERA-interim as verification for surface variables was reduced by applying a simple bias-correction to all models using the past 30 days of forecasts and analyses.

The performance of nine single-model systems from the TIGGE archive was compared with the performance of the full TIGGE multi-model, consisting of all these nine models. This full multi-model version did not improve on the best single-model, the ECMWF EPS. However, when combining only the four best single-model ensembles (CMC, ECMWF, MetOffice, and NCEP), the multi-model ensemble outperformed the ECMWF-only EPS forecasts, though we note that this result did not apply to all model variables and all lead times. 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 multi-model 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 essentially applied a statistical downscaling to the forecasts. The multi-model 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 multi-model system demonstrated that the ECMWF EPS was the single most important source of information for the success of the multi-model ensemble.

Which of the two discussed post-processing methods would be the most appropriate choice for a modelling centre? To answer this, 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, the technical overhead, and the potential data unavailability issues inherent in 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 than appropriate choice. On top of the above discussed 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.

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 multi-model or single-model systems, respectively. 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.

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. We acknowledge the comments from the anonymous reviewers and the Editor which stimulated fruitful discussions and helped us strengthening the arguments of the paper.

References

Bougeault, P. and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble (TIGGE). *Bull. Amer. Meteor. Soc.*, in press.

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

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

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

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.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 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., F. J. Doblas-Reyes and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. *Tellus-A*, **57**, 219-233.

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

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

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev*., **134**, 3209-3229.

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

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434-1447.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33-46.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 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, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev*., **126**, 796-811.

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

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

Kalnay, E., and coauthors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc*., **77,** 437-471.

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

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, S. Surendran, 1999: Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble. *Science*, **285**, 1548-1550.

Langland, R. H., Maue., R. N., and C. H. Bishop, 2008: Uncertainty in atmospheric temperature analyses. *Tellus A*, **60**, 598-603.

Palmer, T.N. and Coauthors, 2004: Development of a European Multi-Model Ensemble System for Seasonal to Inter-Annual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Research Department Technical Memorandum No. 598, ECMWF, Shinfield Park, Reading RG2-9AX, UK, pp. 42.

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

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

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

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

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

Shin, D. W., S. Cocke, and T. E Larow, 2003: Ensemble Configurations for Typhoon Precipitation Forecasts. *J. Meteor. Soc. Japan.*, **81 (4)**, 679-696.

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

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

Uppala, S. M. and Coauthors, 2005: The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961-3012.

Weigel, A. P., and N. E. Bowler, 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, A. P., M. A. Liniger, C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.*, **134**, 241-260.

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

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

LIST OF FIGURE CAPTIONS

Figure 1: Illustration of the impact of the verification dataset and bias-correction on the relative skill of the predictions. The CRPSS is defined as CRPSS = 1 - CRPS(exp) / CRPS(ref). Scores are calculated for forecast from the TIGGE multi-model (solid line) and the single-models (dotted lines with symbols; CMC: crosses, ECMWF: diamonds, Met Office: triangles, NCEP: squares) starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N - 90°N). (a): 850-hPa temperature DMO forecast using ERA-interim as verification (exp) and the multi-model analysis as verification (ref); (b): 2-m temperature DMO forecast using ERA-interim as verification (exp) and the multi-model analysis as verification (ref); (c): 2-m temperature BC forecast using ERA-interim as verification (exp) and DMO forecast using ERA-interim as verification (ref); d: 2-m temperature BC forecast using ERA-interim as verification (exp) and DMO forecast using the multi-model analysis as verification (ref).

Figure 2: Illustration of gain in skill depending on the calibration method applied to ECMWF direct model output. 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. 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 3: Continuous Ranked Probability Skill Score versus lead time for 850-hPa temperature DMO forecasts (left column, subpanel a and c) and 2-m temperature BC-30 forecasts (right column, subpanel b and d) in DJF 2008/09, averaged over the

Northern Hemisphere (20°N - 90°N). Figures (a) and (b) in the upper row show results for the TIGGE-9 multi-model (solid line) composed of nine single-models and the scores of all nine contributing single-models (dashed and dotted lines with symbols). Figures (c) and (d) in the lower row show results for the TIGGE-4 multi-model (solid line) composed of the four best single-models with lead time up to 15 days and the scores of the four contributing single-models (dotted lines with symbols). Symbols are only plotted for cases in which the single-model score significantly differs from the multi-model score on a 1% significance level.

Figure 4: Continuous Ranked Probability Skill Score versus lead time for the TIGGE-4 multi-model (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 multi-model score on a 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.

Figure 5: Visualization of 2-m temperature forecast distributions at individual grid point locations, depicted as box-and-whisker plots, also called EPSgrams. 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 verification is shown as black bullet. (a) EPSgram at the gridpoint closest to Bologna, IT, (45.0°N, 12.5°E) for the forecast started on 25 December 2008, (b) EPSgram at the gridpoint closest to London, UK, (52.5°N, 0.0°E) for the forecast started on 15 January 2009.

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 depicted using lines with symbols (CMC: cross, ECMWF: diamond, MetOffice: triangle, NCEP: square). b: as in (a) but without the CMC, MetOffice and NCEP results, including instead the results for the reforecast calibrated ECMWF (lines with bullets as symbol) and TIGGE-4 multi-model results (curves without symbols). 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 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 multi-model (solid line), for the contributing DMO 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 multi-model score on a 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 hardly distinguishable.

Figure 8: Illustration of the relative performance of individual single-models with respect to the TIGGE-4 multi-model. Panel (a) to (e) display the CRPSS defined as CRPSS = 1 - CRPS(exp) / CRPS(ref), 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) and CRPS(ref) the CRPS of the TIGGE-4 multi-model, shown in panel (f). Scores are calculated at every grid point for 2-m temperature forecasts at 14 days lead time, starting in DJF 2008/09. In panel (a) to (f) grid points at which the difference between CRPS(exp) and CRPS(ref) is not significant (on a significance level 0.1) are marked with a black bullet.

Figure 9: Illustration of the impact of using (a) ERA-interim and (b) NCEP reanalysis as verification dataset. The CRPS is calculated for 2-m temperature forecasts from the TIGGE multi-model (solid line), the contributing BC-30 single-models (dotted lines with symbols; CMC: crosses, ECMWF: diamonds, Met Office: triangles, NCEP: squares), and the reforecast-calibrated ECMWF forecasts (dotted line with bullets) starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N - 90°N).

Figure 10: Continuous Ranked Probability Skill Score versus lead time for 2-m temperature forecasts in DJF 2008/09 at 250 European stations. Results are shown for the BC-30 single-models (dotted lines with symbols, CMC: cross, ECMWF: diamond, MetOffice: triangle, NCEP: square) contributing to the TIGGE-4 multi-model (solid line without symbols) and the reforecast calibrated ECMWF model (dotted line with bullet symbols).

Figure 11: Illustration of gain or loss in skill depending on which model has been removed from the TIGGE-4 multi-model containing all four BC-30 single-models (CMC, ECMWF, MetOffice, NCEP). The CRPSS is defined as CRPSS = 1 -

CRPS(exp) / CRPS(ref), with CRPS(ref) being the CRPS of the TIGGE-4 multi-model and CRPS(exp) the CRPS of the reduced multi-model 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 multi-model containing all four BC-30 forecasts from CMC, ECMWF, MetOffice, and NCEP (solid line), the TIGGE-4 multi-model 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 multi-model score on a 1% significance level.

Tables

Table 1: Main features of the nine TIGGE model systems used in this study. 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 Centres 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

Figures

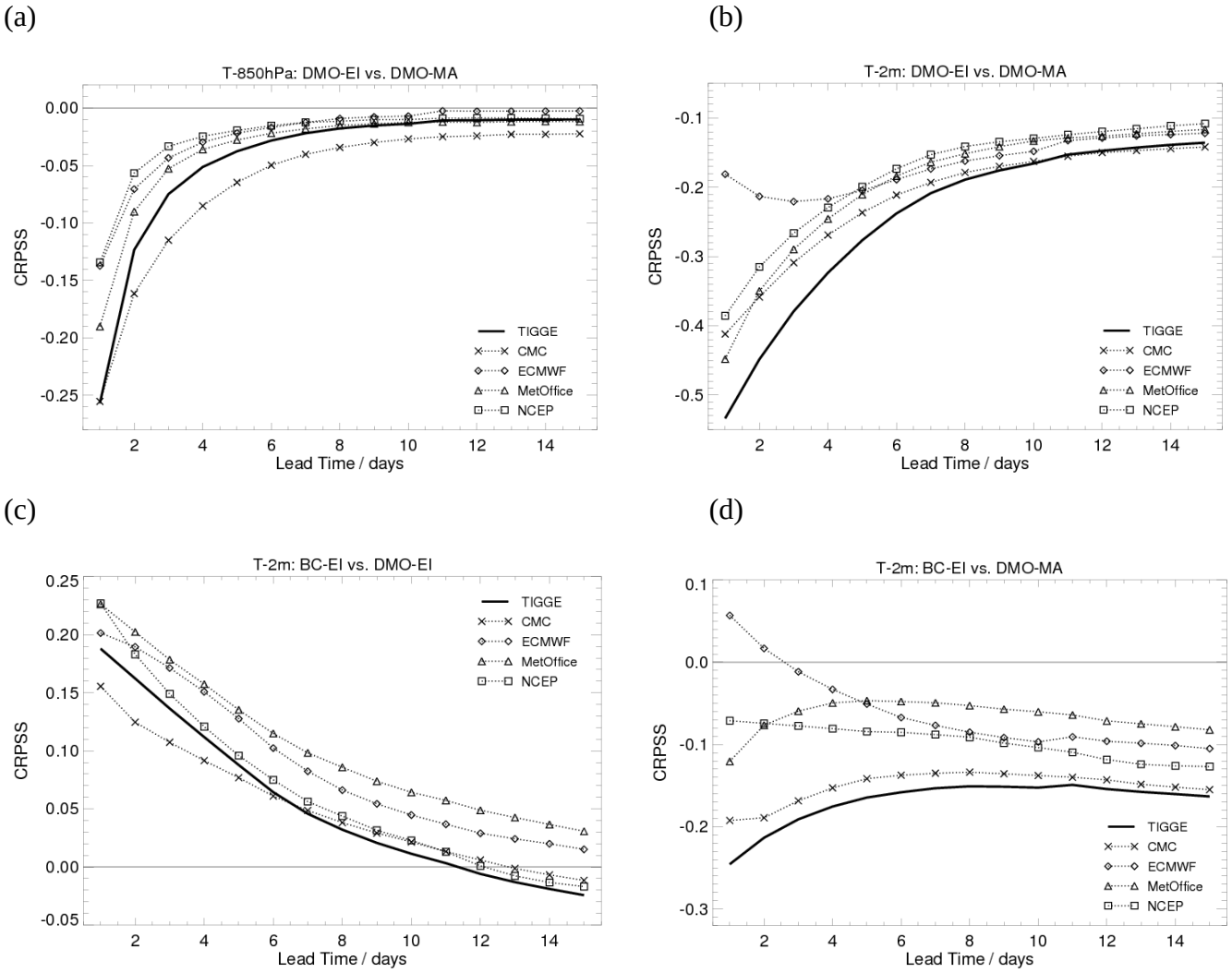


Figure 1: Illustration of the impact of the verification dataset and bias-correction on the relative skill of the predictions. The CRPSS is defined as CRPSS = 1 - CRPS(exp) / CRPS(ref). Scores are calculated for forecast from the TIGGE multi-model (solid line) and the single-models (dotted lines with symbols; CMC: crosses, ECMWF: diamonds, Met Office: triangles, NCEP: squares) starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N - 90°N). (a): 850-hPa temperature DMO forecast using ERA-interim as verification (exp) and the multi-model analysis as verification (ref); (b): 2-m temperature DMO forecast using ERA-interim as verification (exp) and the multi-model analysis as verification (ref); (c): 2-m temperature BC forecast using ERA-interim as verification (exp) and DMO forecast using ERA-interim as verification (ref); d: 2-m temperature BC forecast using ERA-interim as verification (exp) and DMO forecast using the multi-model analysis as verification (ref).

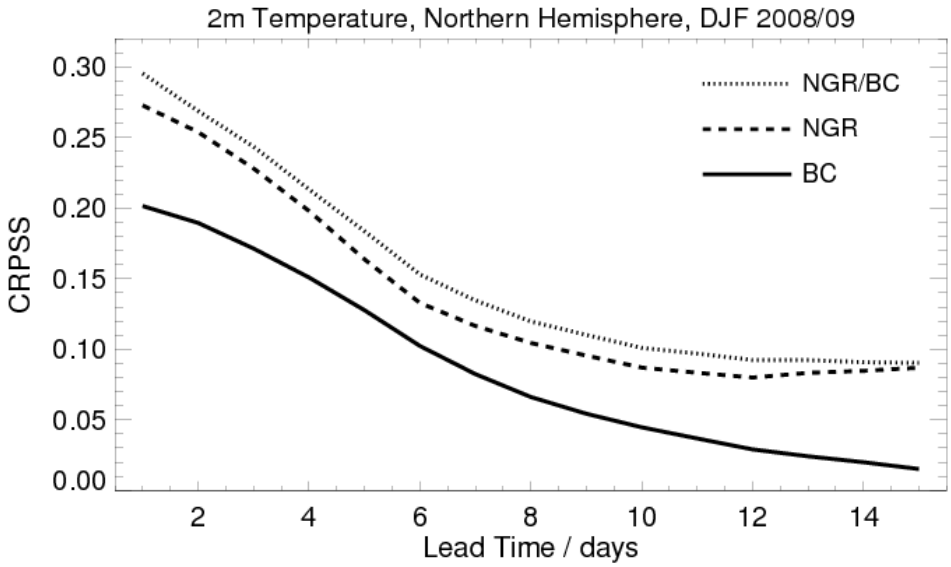


Figure 2: Illustration of gain in skill depending on the calibration method applied to ECMWF direct model output. The CRPSS is defined as CRPSS = 1 - CRPS(exp) / CRPS(ref), with CRPS(exp) being the CRPS of the bias corrected forecasts (solid), the NGR calibrated forecast (dashed), and the NGR/BC model combination (dotted). CRPS(ref) is in all cases the CRPS of the DMO forecasts. All scores are calculated for 2-m temperature forecasts starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N - 90°N).

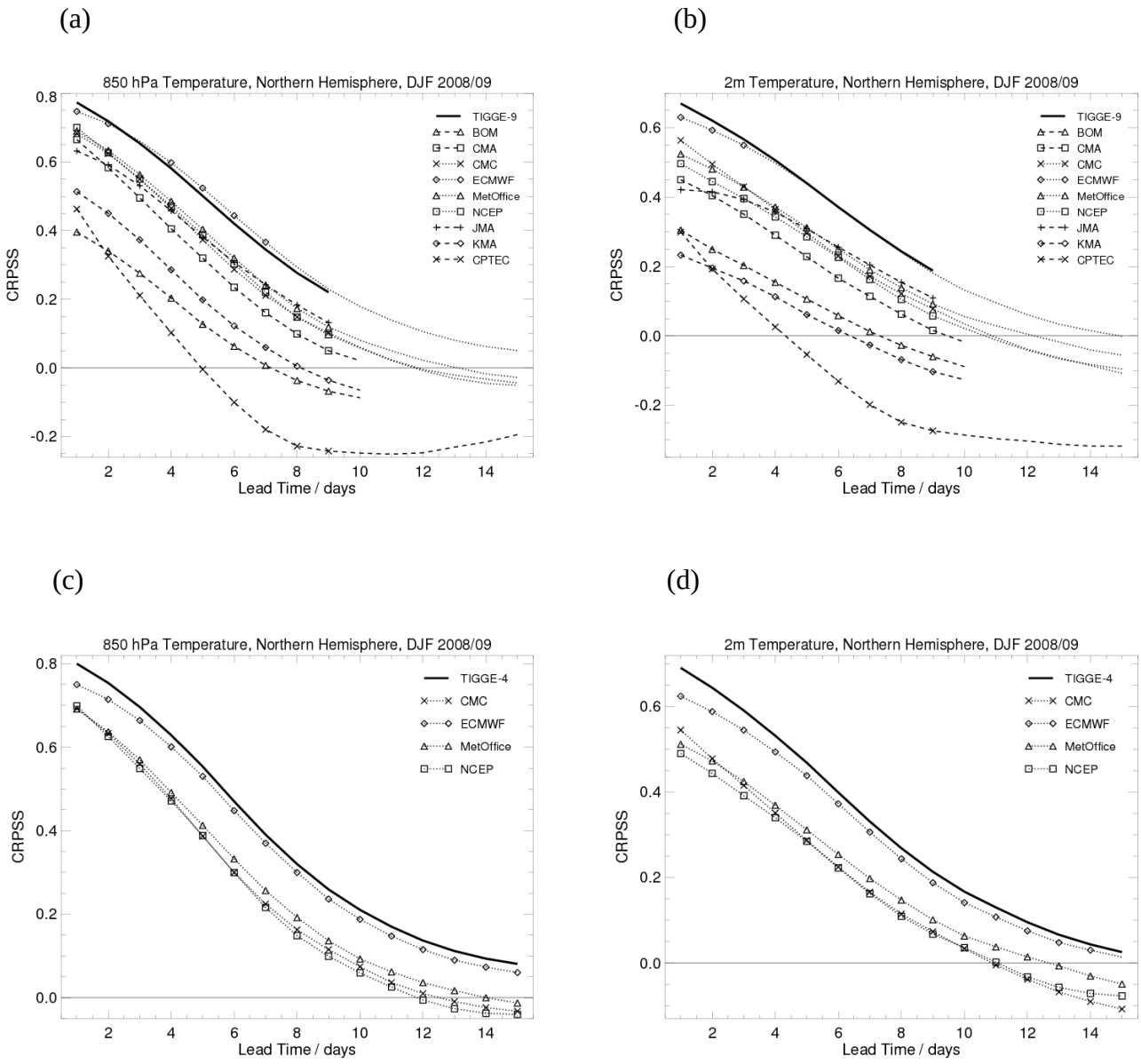


Figure 3: Continuous Ranked Probability Skill Score versus lead time for 850-hPa temperature DMO forecasts (left column, subpanel a and c) and 2-m temperature BC-30 forecasts (right column, subpanel b and d) in DJF 2008/09, averaged over the Northern Hemisphere (20°N - 90°N). Figures (a) and (b) in the upper row show results for the TIGGE-9 multi-model (solid line) composed of nine single-models and the scores of all nine contributing single-models (dashed and dotted lines with symbols). Figures (c) and (d) in the lower row show results for the TIGGE-4 multi-model (solid line) composed of the four best single-models with lead time up to 15 days and the scores of the four contributing single-models (dotted lines with symbols). Symbols are only plotted for cases in which the single-model score significantly differs from the multi-model score on a 1% significance level.

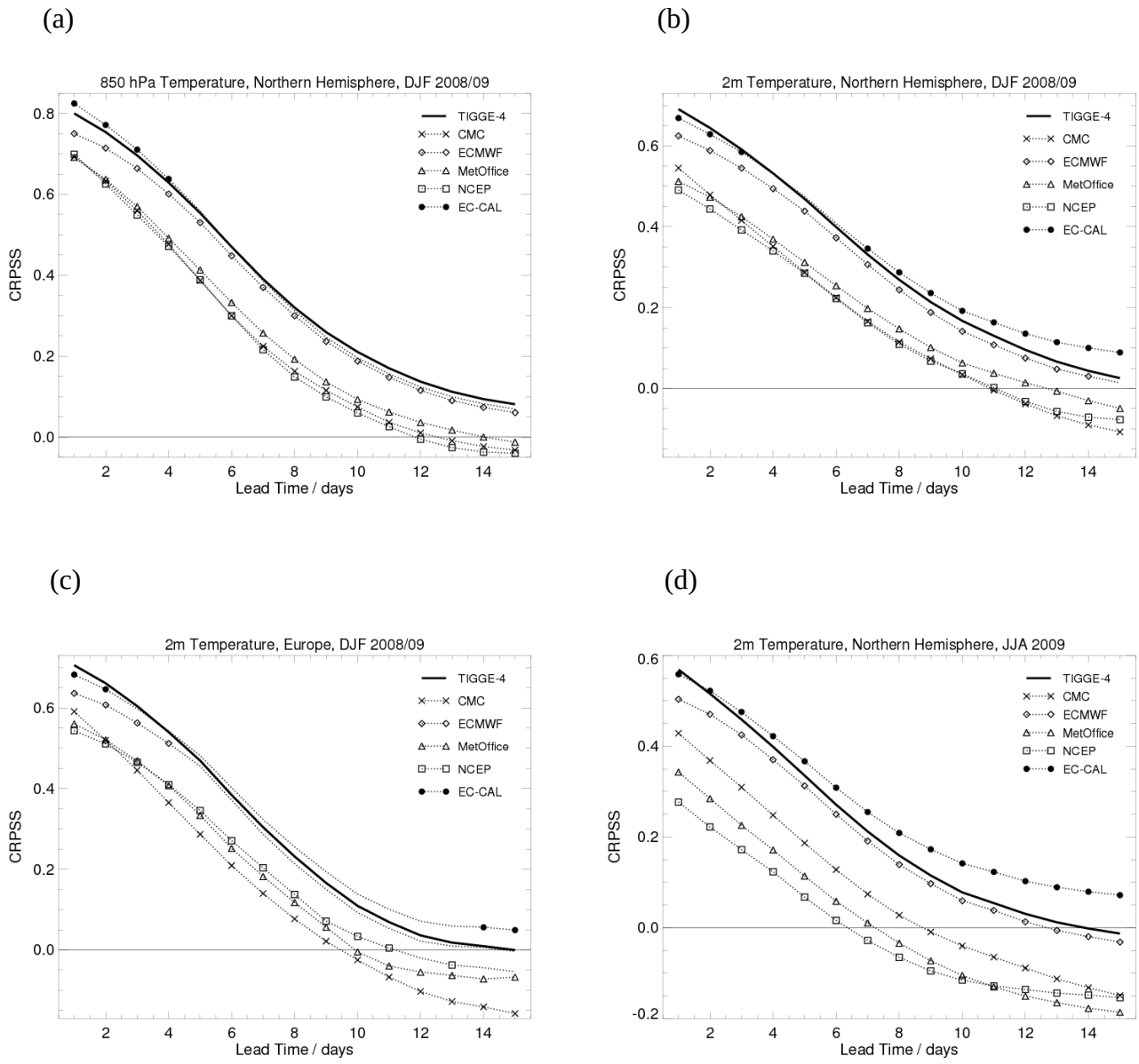
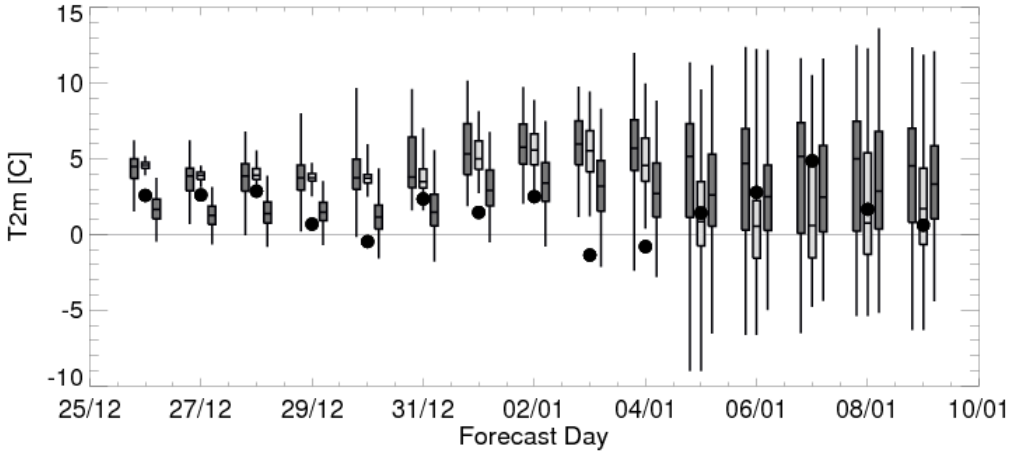


Figure 4: Continuous Ranked Probability Skill Score versus lead time for the TIGGE-4 multi-model (solid line), for the contributing single-models itself (dotted lines with symbols, CMC: cross, ECMWF: diamond, MetOffice: triangle, NCEP: square), and for the reforecast calibrated ECMWF forecasts (dotted lines with bullets). Symbols are only plotted for cases in which the single-model score significantly differs from the multi-model score on a 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.

(a)



(b)

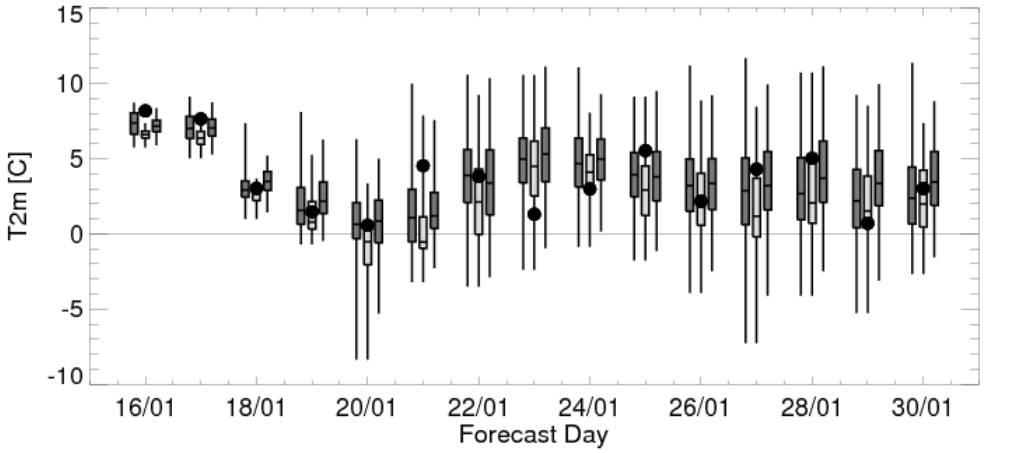
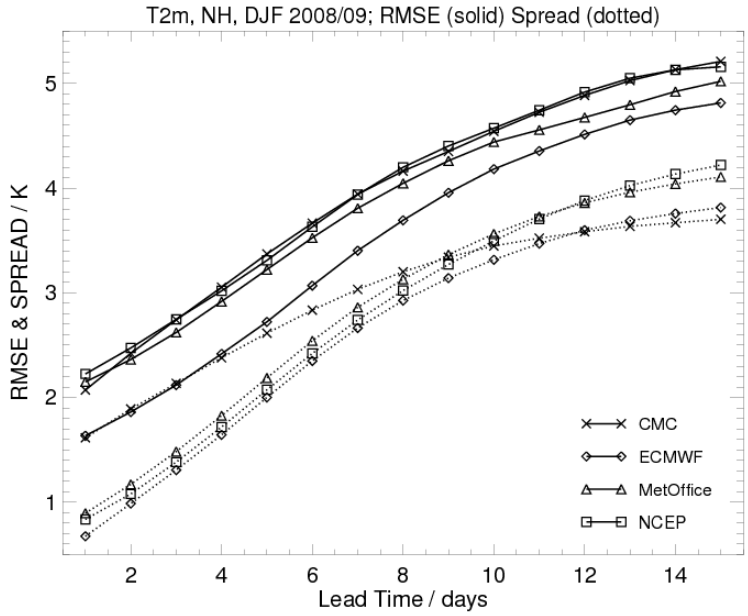


Figure 5: Visualization of 2-m temperature forecast distributions at individual grid point locations, depicted as box-and-whisker plots, also called EPSgrams. The inner box contains 50% of the ensemble members including the median (horizontal black line inside the box) and the wings represent the remaining 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 verification is shown as black bullet. (a) EPSgram at the gridpoint closest to Bologna, IT, (45.0°N, 12.5°E) for the forecast started on 25 December 2008, (b) EPSgram at the gridpoint closest to London, UK, (52.5°N, 0.0°E) for the forecast started on 15 January 2009.

(a)



(b)

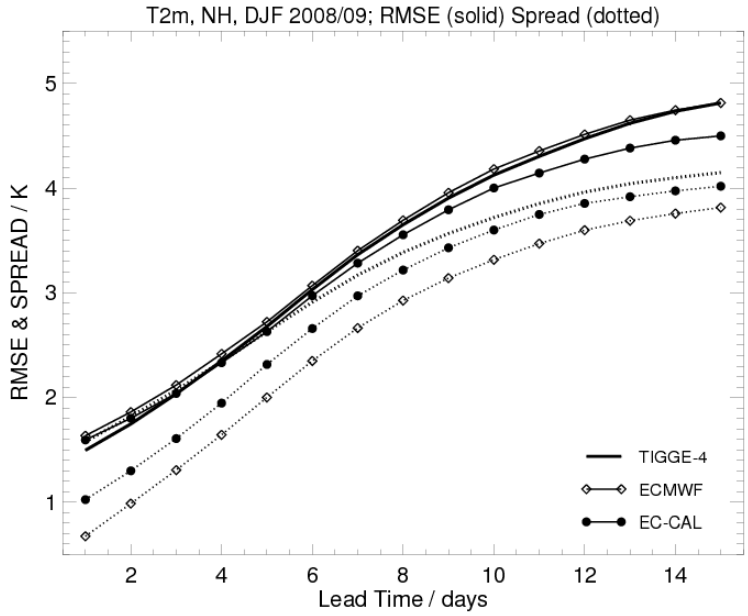


Figure 6: Root Mean Square Error of the ensemble mean (solid lines) and ensemble standard deviation ("spread", dotted lines) versus lead time. a: results for the single-model BC-30 forecasts depicted using lines with symbols (CMC: cross, ECMWF: diamond, MetOffice: triangle, NCEP: square). b: as in (a) but without the CMC, MetOffice and NCEP results, including instead the results for the reforecast calibrated ECMWF (lines with bullets as symbol) and TIGGE-4 multi-model results (curves without symbols). All scores are calculated for 2-m temperature forecasts starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N - 90°N).

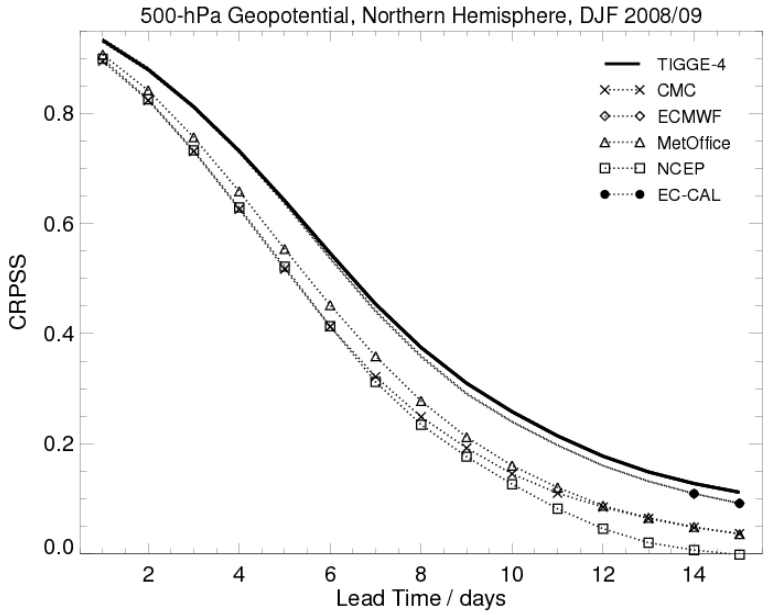


Figure 7: Continuous Ranked Probability Skill Score versus lead time for 500-hPa geopotential averaged over the Northern Hemisphere (20°N - 90°N) for DJF 2008/09. Results are shown for the TIGGE-4 multi-model (solid line), for the contributing DMO single-models itself (dotted lines with symbols, CMC: 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 multi-model score on a 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 hardly distinguishable.

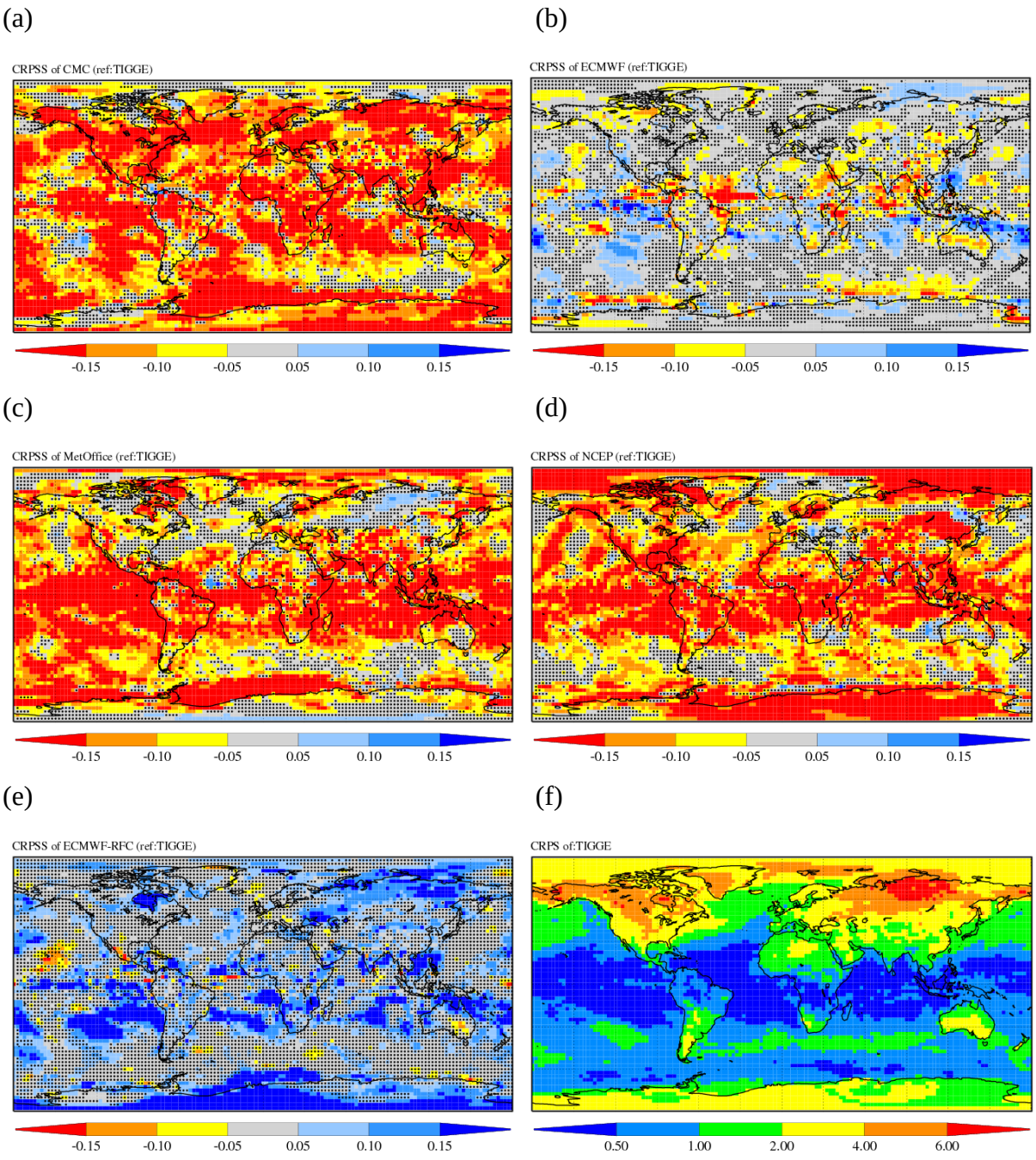


Figure 8: Illustration of the relative performance of individual single-models with respect to the TIGGE-4 multi-model. Panel (a) to (e) display the CRPSS defined as CRPSS = 1 - CRPS(exp) / CRPS(ref), with CRPS(exp) the CRPS of the single-models (a: BC-30 CMC, b: BC-30 ECMWF, c: BC-30 MetOffice, d: BC-30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

(a) (b)

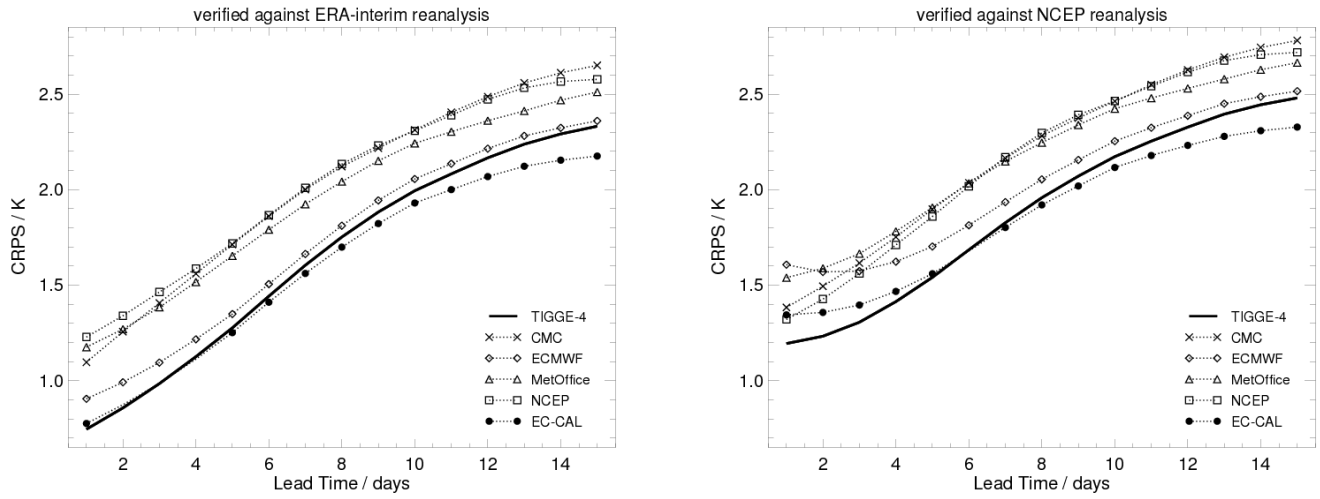


Figure 9: Illustration of the impact of using (a) ERA-interim and (b) NCEP reanalysis as verification dataset. The CRPS is calculated for 2-m temperature forecasts from the TIGGE multi-model (solid line), the contributing BC-30 single-models (dotted lines with symbols; CMC: crosses, ECMWF: diamonds, Met Office: triangles, NCEP: squares), and the reforecast-calibrated ECMWF forecasts (dotted line with bullets) starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N - 90°N).

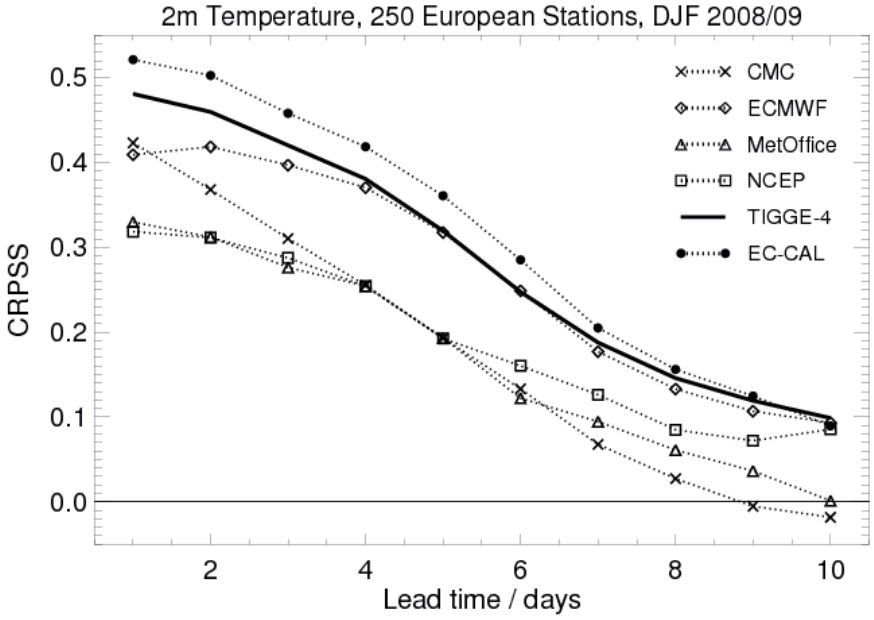


Figure 10: Continuous Ranked Probability Skill Score versus lead time for 2-m temperature forecasts in DJF 2008/09 at 250 European stations. Results are shown for the BC-30 single-models (dotted lines with symbols, CMC: cross, ECMWF: diamond, MetOffice: triangle, NCEP: square) contributing to the TIGGE-4 multi-model (solid line without symbols) and the reforecast calibrated ECMWF model (dotted line with bullet symbols).

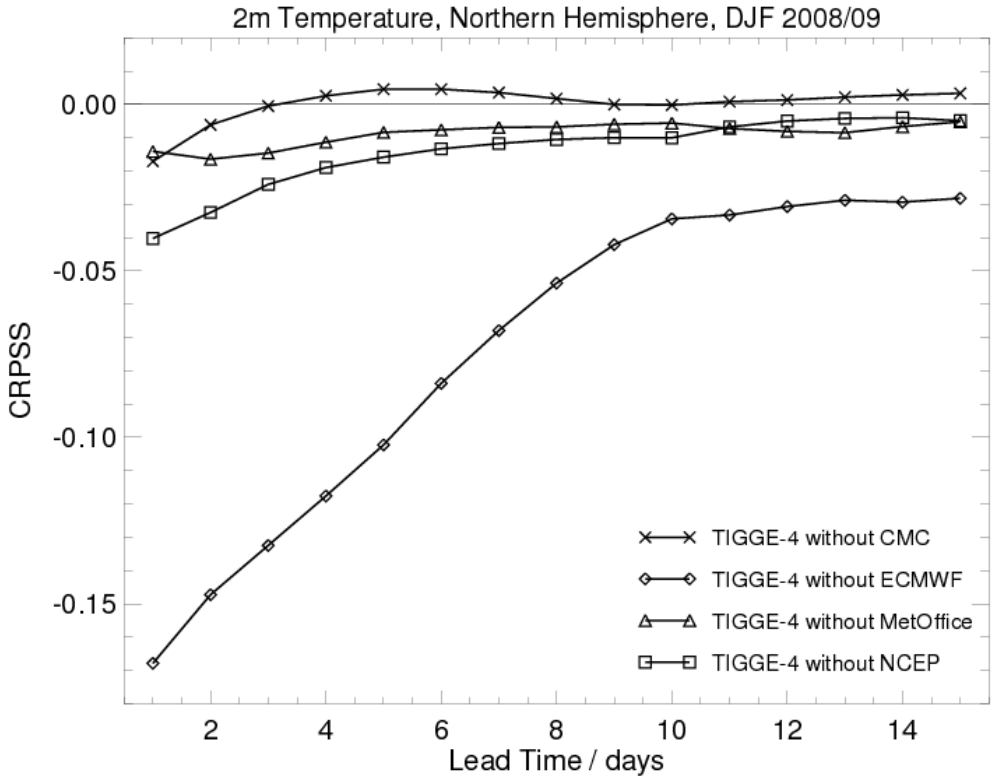


Figure 11: Illustration of gain or loss in skill depending on which model has been removed from the TIGGE-4 multi-model containing all four BC-30 single-models (CMC, ECMWF, MetOffice, NCEP). The CRPSS is defined as CRPSS = 1 - CRPS(exp) / CRPS(ref), with CRPS(ref) being the CRPS of the TIGGE-4 multi-model and CRPS(exp) the CRPS of the reduced multi-model 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).

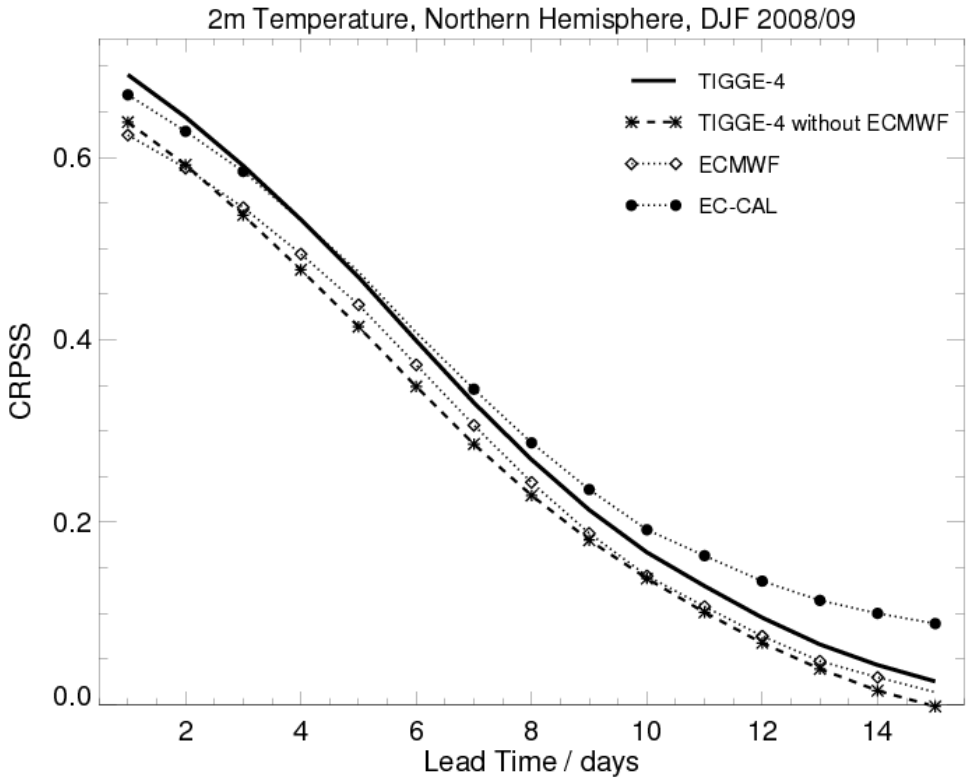


Figure 12: Continuous Ranked Probability Skill Score versus lead time for 2-m temperature forecasts starting in DJF 2008/09 and averaged over the Northern Hemisphere (20°N - 90°N). Results are shown for the TIGGE-4 multi-model containing all four BC-30 forecasts from CMC, ECMWF, MetOffice, and NCEP (solid line), the TIGGE-4 multi-model 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 multi-model score on a 1% significance level.