TECHNICAL MEMORANDUM

548

# TIGGE: preliminary results on comparing and combining ensembles

Young-Youn Park[1], Roberto Buizza and Martin Leutbecher

Research Department

[1] Korea Metrological Administration, Seoul, Republic of Korea (www.kma.go.kr)

Submitted to the Q. J. Roy. Meteorol. Soc.

January 2008

This paper has not been published and should be regarded as an Internal Report from ECMWF.
Permission to quote from it should be obtained from the ECMWF.

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

**© Copyright 2008**

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

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. This publication is not to be reprinted or translated in whole or in part without the written permission of the Director. Appropriate non-commercial use will normally be granted under the condition that reference is made to ECMWF.

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

**Abstract**

TIGGE, the THORPEX Interactive Grand Global Ensemble, is a World Weather Research Programme to accelerate the improvements in the accuracy of 1-day to 2 week high-impact weather forecasts. This report reviews the status of the TIGGE archive, and discusses some preliminary results from predictability studies that would not have been possible without TIGGE. First, the key characteristics of the eight ensemble systems available in the TIGGE database at the time of writing (December 2007) are compared, and the strengths and weaknesses of each system are highlighted. Then, issues related to the generation of multi-model/multi-analysis ensemble products are discussed, and the potential value of combining different ensembles to generate medium-range products with a grand multi-model/multi-analysis global ensemble is investigated. Results should help developers of the different ensemble systems to better understand the characteristics of their ensemble, and should provide valuable information on how to improve their performance. This work proves the value of the TIGGE initiative, and illustrates some of the issues that could be addressed with the TIGGE data.

# 1. Introduction

TIGGE, the THORPEX Interactive Grand Global Ensemble (see THORPEX in the list of references), is a key component of THORPEX: a World Weather Research Programme to accelerate the improvements in the accuracy of 1-day to 2 week high-impact weather forecasts. TIGGE has among its objectives to develop a deeper understanding of the contribution of observation, initial and model uncertainties to forecast error, and to investigate new methods of combining ensembles from different sources and of correcting systematic errors (biases, spread over-/under-estimation). This report briefly reviews the status of the TIGGE archive (three centres are acting as data collection centres, ECMWF, CMA and NCAR), and discusses results from predictability studies designed to achieve the two objectives mentioned above:

- Compare the performance of single ensemble prediction systems, and identify their strengths and weaknesses

- Assess the potential value of combining different ensembles to generate multi-model/multi-analysis grand global ensemble products

Medium-range ensemble prediction started in December 1992, when the National Centres for Environmental Prediction (NCEP, *Toth & Kalnay* 1993 and 1997) and the European Centre for Medium-Range Weather Forecast (ECMWF, *Palmer et al* 1993, *Buizza & Palmer* 1995, *Molteni et al* 1996, *Buizza et al* 2007) started producing global ensemble predictions as part of their operational products. In 1995, the Meteorological Service of Canada (MSC, *Houtekamer et al* 1996) implemented its ensemble prediction system. Following these examples, six other centres started running global ensemble prediction systems daily. At the time of writing (December 2007), ten meteorological centres are running a medium-range global ensemble prediction system[1]: ECMWF, NCEP, MSC, the Australian Bureau of Meteorology (BMRC, *Bourke et al* 1995 & 2004), the Chinese Meteorological Administration (CMA), the Brazilian Centro de Previsao de Tempo e Estudos Climatico (CPTEC), the United States Fleet Numeric Meteorological Operational Center (FNMOC), the Japanese Meteorological Administration (JMA), the Korean Meteorological Administration (KMA, *Goo & Moon* 2003), and the UK Met Office (UKMO). Other centres (e.g. Meteo France, Met

Norway, the COSMO Consortium established by the German, Greece, Italian, Polish and Swiss National Meteorological Services, and the Spanish Instituto National de Meteorologia) are running or testing a short-range regional system.

All these ensemble prediction systems have been designed to simulate, explicitly or implicitly, the effect on weather forecasts of observation uncertainties, model uncertainties (e.g. due to a lack of resolution, simplified parameterization of physical processes, effect of unresolved processes), imperfect boundary conditions and data assimilation assumptions (e.g. due to the assumed statistics). Although it may be difficult to clearly distinguish between these sources of forecast error, for simplicity and to facilitate the comparison between the different approaches used by the ten centers, they are often grouped into two broad classes, namely initial and model uncertainties. (The fact that it is difficult to separate clearly between initial and model uncertainties is linked to the fact that initial conditions are constructed using a model-based data-assimilation procedure; thus, errors that are sometimes defined as 'initial condition errors' may actually be due to model errors.)

All these ensemble prediction systems are based on *N* time integrations of a numerical weather prediction model, with one (the control forecast) starting from a 'central' analysis, usually the unperturbed analysis generated by a data-assimilation procedure, and the others (the perturbed forecasts) starting from perturbed initial conditions defined to simulate the effect of initial uncertainties. Following the examples of ECMWF and NCEP, nine of the ten centres listed above (i.e. all but MSC) simulate the effect of initial uncertainties by adding perturbations to the 'central' analysis (see Table A for a summary of the key characteristics of the ten ensembles). Considering the simulation of initial uncertainties, three centres (ECMWF, Meteo France and BMRC) use singular vectors (*Buizza & Palmer* 1995, *Bourke et al* 2004), three centres (CMA, JMA and KMA) use bred-vectors (*Toth & Kalnay* 1997), two centres (NCEP and UKMO) use an Ensemble Transform or an Ensemble Transform Kalman Filter approach (ET or ETKF, *Wei et al* 2006, *Bishop et al* 2001; see *Bowler et al* 2007 for a description of the UKMO system), and one centre (CPTEC) uses an EOF-based method (*Zhang & Krishnamurti* 1999). MSC is the only centre that adds random perturbations to the observations, and generates the perturbed analyses with an ensemble Kalman filter. Considering the simulation of model uncertainties, only three of the ten operational systems, the ECMWF, MSC and UKMO ones, also simulate the effect of model uncertainties: ECMWF simulates random model errors due to physical processes by adding a stochastic perturbation to the model tendencies due to the physical processes (*Buizza et al* 1999), UKMO simulates the effect of model errors due to energy dissipation with a stochastic backscatter scheme (*Shutts* 2005) and MSC uses two schemes similar to the Buizza et al (1999) and the *Shutts* et al (2005) schemes and, in addition, it uses several different physical parameterisation schemes (*Houtekamer & Lefaivre* 1997).

The key characteristics of the three global ensemble systems implemented at ECMWF, MSC and NCEP, and their performance for a 3-month period (spring 2002), were discussed in *Buizza et al* (2005), who concluded that *"… for all three global systems, the spread of ensemble forecasts are insufficient to systematically capture reality, suggesting that none of them is able to simulate all sources of forecast uncertainty."*.

[1] In this report, a medium-range global ensemble system is an ensemble system designed to provide probabilistic forecasts for up to 7 days and for the whole globe. By contrast, a short-range regional ensemble prediction system is a system designed to provide probabilistic forecasts for up to 3 days and for a limited geographical region.

*Table A: Characteristics of the 10 TIGGE ensembles*

| Centre | Initial pert method (area) | Model error simul | Horizon resolution | # vert lev | Fcst length (days) | # pert mem | # runs per day (UTC) | # mem per day | Initial date of TIGGE operational mode |
|---|---|---|---|---|---|---|---|---|---|
| **BMRC (Australia)** | SVs (NH,SH) | NO | TL119 | 19 | 10 | 32 | 2(00/12) | 66 | 3 Sep 07 |
| **CMA (China)** | BVs (globe) | NO | T213 | 31 | 10 | 14 | 2(00/12) | 30 | 15 May 07 |
| **CPTEC (Brazil)** | EOF-based (40S:30N) | NO | T126 | 28 | 15 | 14 | 2(00/12) | 30 | On test |
| **ECMWF (Europe)** | SVs (globe) | YES | TL399 | 62 | 0-10 | 50 | 2(00/12) | 102 | 1 Oct 06 |
| | | | TL255 | 62 | 10-15 | | | | |
| **JMA* (Japan)** | BVs (NH+TR) | NO | TL159 | 40 | 9 | 50 | 1(12) | 51 | 1 Oct 06 |
| **KMA (Korea)** | BVs (NH) | NO | T213 | 40 | 10 | 16 | 2(00/12) | 34 | On test |
| **MeteoFrance (France)** | SVs (targeted) | NO | TL358 | 41 | 2.5 | 10 | 1(18) | 11 | 26 Oct 07 |
| **MSC (Canada)** | EnKF (globe) | YES | TL149 | 28 | 16 | 20 | 2(00/12) | 42 | 3 Oct 07 |
| **NCEP (USA)** | BVs (globe) | NO | T126 | 28 | 16 | 20** | 4(00/06/12/18) | 84 | 5 Mar 07 |
| **UKMO (UK)** | ETKF (globe) | YES | 1.25x0.83deg | 38 | 15 | 23 | 2(00/12) | 48 | 1 Oct 06 |

Table legend:

*) Changed to SVs(NH+TR), TL319 with 60 vertical levels from 21 Nov 2007.

**) Start dates 18 UTC 27 Mar 2007. 15 members from 12 UTC 14 Dec 2006-12 UTC 27 Mar 2007. 11 members from 00 UTC 1 Nov 2006 - 06 UTC 14 Dec 2006.

To address the sub-optimal simulation of model uncertainties and of the limited ensemble size, MSC and NCEP have decided to combine their operational ensemble systems in the North American Ensemble Forecasting System (see NAEFS web page, http://www.emc.ncep.noaa.gov/gmb/ens/NAEFS.html) and started disseminating joint products, generated using both ensemble systems. The other centres have decided to investigate the potential benefit of combining ensemble forecasts generated by different centres by establishing TIGGE. The first TIGGE workshop was held at ECMWF on 1-3 March 2005: sixty scientists, from international organizations, national and regional meteorological and hydrological services, universities and private companies, attended the workshop. The workshop discussed the scientific aims, user requirements and infrastructure for TIGGE data bases and Centres (readers can access the workshop report from WMO: see WMO series WMO/TD-No. 1273 WWRP/THORPEX No. 5). Since then, three centres (CMA, ECMWF and NCAR) have been developing capabilities to become TIGGE Data Centres, and have started collecting the TIGGE data. The collaboration between the three archiving centres has proved to be excellent, with each center acting as back-up in case of missing data in the others.

The status of the TIGGE archive is briefly reviewed in section 2. Section 3 discusses how the TIGGE data can be used to compare the characteristics of the different ensemble systems. Section 4 discusses some preliminary results obtained by combining different ensemble systems. Section 5 draws some conclusions, and discusses possible future developments of TIGGE.

## 2. The TIGGE data-base, variables and verification measures

At the time of writing (December 2007), eleven organizations expressed interest to participate in the TIGGE project, with ten of them (all but NCAR) providing ensemble forecasts (for more information see the ECMWF TIGGE web page http://tigge.ecmwf.int/tigge/d/tigge/):

- BMRC, the Bureau of Meteorology, Melbourne, Australia (http://www.bom.gov.au/)
- CMA, the China Meteorological Administration, Beijing, China (http://www.cma.gov.cn/cma_new/)
- MSC, the Meteorological Service of Canada, Montreal, Canada (http://www.weatheroffice.gc.ca/canada_e.html)
- CPTEC, the Centro de Previsao Tempo e Estudos Climaticos, Cachoeira Paulista, Brazil (http://www.cptec.inpe.br/)
- EC, the European Centre for Medium-Range Weather Forecasts, Reading, Europe (http://www.ecmwf.int/)
- JMA, the Japan Meteorological Agency, Tokyo, Japan (http://www.jma.go.jp/jma/indexe.html)
- KMA, the Korea Meteorological Administration, Seoul, Korea (http://www.kma.go.kr)
- Météo France, the French Meteorological Service, Toulouse, France (http://www.meteofrance.com/FR/index.jsp)
- UKMO, the UK MetOffice, Exeter, United Kingdom (http://www.metoffice.gov.uk/)
- NCAR, the National Center for Atmospheric Research, Boulder, CO, USA (http://www.ncar.ucar.edu/)
- NCEP, the National Centres for Environmental Prediction, Washington, DC, USA (http://www.ncep.noaa.gov/)

Table A lists the key characteristics of the ten ensemble prediction systems. Note that nine of the ten organizations produce daily medium-range global ensemble forecasts, while one (MeteoFrance) produces daily one short-range ensemble forecast. In October 2006, ECMWF, JMA and UKMO started delivering data to TIGGE. In March 2007, NCEP started delivering data, and in May 2007 CMA joined the other production centres. In December 2007, eight of the ten organizations (BMRC, CMA, ECMWF, JMA, KMA, MSC, NCEP and UKMO) are delivering medium-range global forecasts and one centre (Météo France) is delivery short-range forecasts optimized for Europe to the TIGGE archive. A ninth organization (CPTEC) has started sending test data, and is expected to start sending data routinely very soon.

Unfortunately, data from all eight centres are available only since October 2007 (see Table B for a list of the available data), while data from fewer centres are available for the earlier periods (note that due to transmission problems, there are still few missing data in the archive which are currently being recovered). Thus, this work will discuss results from four different periods (which include a total of 281 cases), chosen so that for each period data from at least three medium-range forecasts (and the corresponding verification analyses) were available:

- October-November 2007 (ON07, 45 cases), eight centres (BMRC, CMA, EC, JMA, KMA, NCEP, MSC and UKMO)
- Winter 2006/07 (DJF07, 90 cases), three centres (EC, JMA and UKMO)
- Spring 2007 (AM07, 62 cases), four centres (EC, JMA, NCEP and UKMO)
- Summer 2007 (JJA07, 84 cases), five centres (BMRC, CMA, EC, JMA and UKMO)

In section 4, to test the effect of combination on a large data set, results based on two periods will also be analyzed based on an extended period:

- February to August 2007 (FMAMJJA07, 204 cases), two centres (EC and UKMO)
- Summer 2007 (JJA07, 86 cases) with May 2007 (M07, 30 cases) for training, four centres (CMA, EC, JMA and UKMO)

*Table B: Schematic representation of the status of the TIGGE archive: grey shading identifies available data, x symbols represent missing dates. The four top black bars denotes the four main periods used in the report.*

| | | | | | DJF07 | | | AM07 | | | JJA07 | | | | ON07 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Centers | IT (UTC) | 2006 | | | 2007 | | | | | | | | | | | |
| | | 10 | 11 | 12 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| BMRC | 00 | | | | | | | | | xx | | | | | |
| | 12 | | | | | | | | | | | | | | |
| CMA | 00 | | | | | | | | | | | | | | |
| | 12 | | | | | | | | | x | | | | | |
| ECMWF | 00 | | | | | | | | | | | | | | |
| | 12 | | | | | | | | | | | | | | |
| JMA | 12 | | | | | | | | | | | | | | |
| KMA | 00 | | | | | | | | | | | | | | |
| | 12 | | | | | | | | | | | | | | |
| MetFr | 18 | | | | | | | | | | | | | | |
| MSC | 00 | | | | | | | | | | | | | | |
| | 12 | | | | | | | | | | | | | | |
| NCEP | 00 | | | | | | | | x | x | | | | | |
| | 06 | | | | | | | | | xxx | | x | x | xx | |
| | 12 | | | | | | | | | x | x | x | | x | |
| | 18 | | | | | | | | | xxx | x | | | x | |
| UKMO | 00 | | | x | | | | | | | | | | | |
| | 12 | | | | | | | | | | | | | | |

Note that since this work focuses on medium-range ensemble forecasts, only data from the available medium-range ensemble systems have been used (i.e. Meteo France forecasts, with a forecast length of only 60 hours, were not used).

For ease of comparison, forecasts have all been verified on a latitude/longitude 1.25 degree grid, using only the forecasts starting at 12 UTC. Forecasts are verified at 00 and 12 UTC for lead times from 12 to the last available forecast step. To limit the number of diagrams to a reasonable amount and for reason of space, attention will focus on two variables, the 500 hPa geopotential height (Z500), and the 850 hPa temperature (T850), and two regions, the Northern Hemisphere (NH, latitudes from 20°N to 90°N) and the tropics (TR, latitudes from 20°S to 20°N). In Section 3, the performance of each ensemble system is assessed against its corresponding analysis (which has been defined as the control forecast at step zero), while in Section 4 all systems are verified against the EC analysis (see section 4 for details). Again for reason of space, only four measures of ensemble accuracy will be used (the reader is referred to *Wilks* 1995 for a definition of these scores): the root-mean-square-error (RMSE) of the control (i.e. the ensemble member starting from the unperturbed initial conditions) forecast, the RMSE of the ensemble-mean forecast, the ensemble standard deviation as a measure of ensemble spread, and the ranked probability skill score (RPSS) for probabilistic predictions. The RPSS is based on 10 climatologically equally likely categories, and it uses a climatological probability density function as reference “forecast”. The climatological probability density function is estimated from ERA-40 analyses in the so called satellite era (1979–2001). Further details are given in Palmer *et al.* (2007).

To give a first impression of the similarities and differences between the different ensembles, Fig. 1 shows the t+120h Z500 ensemble mean forecast from 12 UTC of 18 October 2007 given by the eight available centres. As a reference, Fig. 1 also shows the EC verifying analysis. Figure 1 shows that both the ensemble-mean forecasts and the ensemble spread (measured by the ensemble standard deviation) can be rather different, as it is the case over North America and Europe, two regions characterized by a cut-off low development.

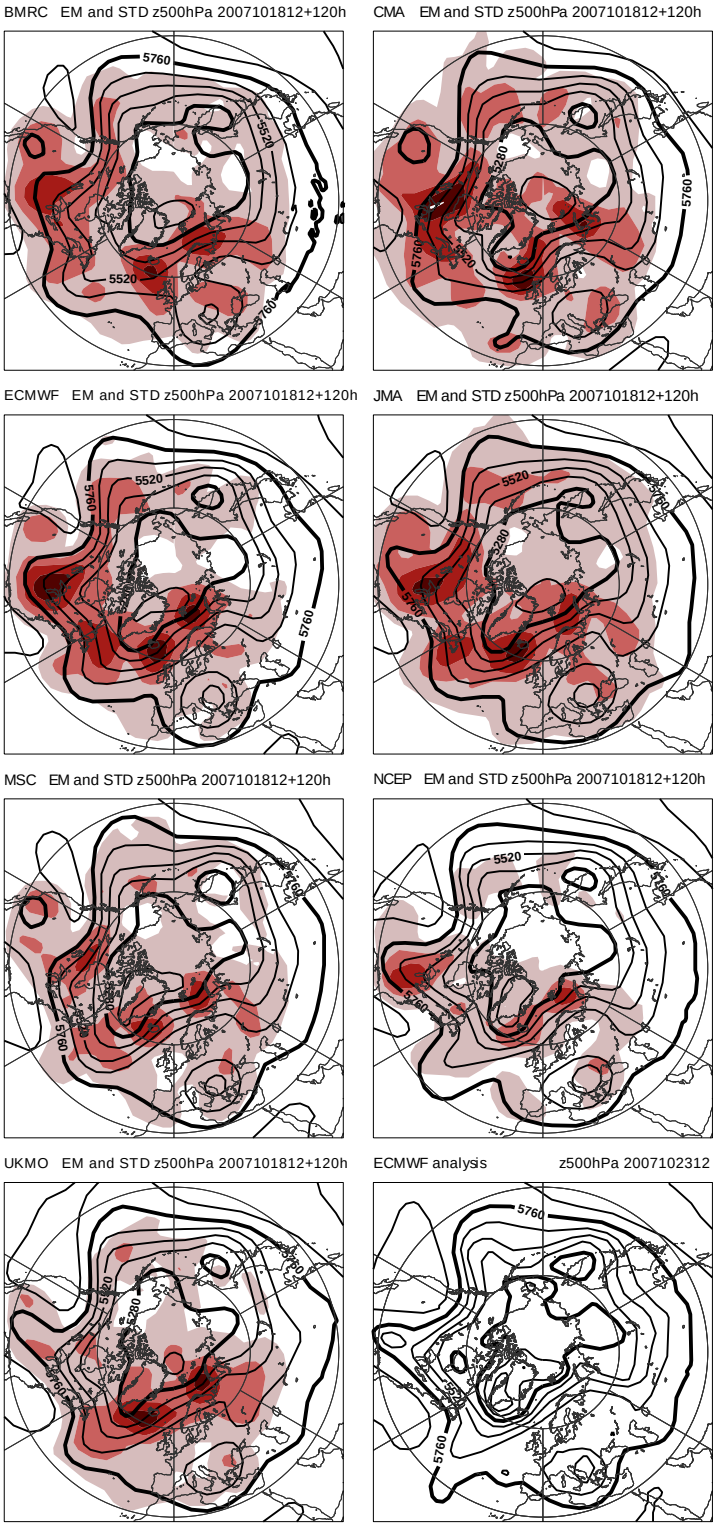


*Figure 1: TIGGE ensemble forecasts: t+120h ensemble-mean (full lines, with a contour interval of 120m) and standard deviation (shading, with a contour interval of 40m) for Z500 over NH for ensembles starting at 12 UTC of 18 October 2007. The last panel shows the EC analysis.*

# 3. Comparison of the characteristics of different ensembles in the TIGGE archive

In the following three sub-sections, the performance of single and probabilistic ensemble forecasts, and the ensemble spread, is analyzed for four different periods. Each ensemble (including the control forecast) is verified against its own analysis.

## 3.1. RMSE of single forecasts: control and ensemble-mean

Figure 2 shows the RMSE of the control forecasts of Z500 over NH. Results averaged for ON07 (45 cases, Fig. 2a) indicate that the EC control has the lowest RMSE up to t+8d, while the NCEP control has the lowest RMSE from t+8d. Up to t+8d, four control forecasts (UKMO, NCEP, MSC and JMA) have comparable second-best performance, while CMA, KMA and BMRC controls have a larger RMSE. Results for the other three periods confirm the very good performance of the EC control forecast, with the UKMO and JMA controls performing in the group of the second-best forecasts. Overall, results indicate that up to t+8d, the EC control forecast has the equivalent of 6-12 hours gain in predictability (i.e. the RMSE of the t+132h EC forecasts is comparable to the RMSE of the t+120h forecast of the second best ensemble).

Figure 3 shows the RMSE of the ensemble-mean forecasts of Z500 over NH. The first conclusion that can be drawn from this figure is that, for all ensembles, the RMSE of the ensemble-mean forecast is lower than the RMSE of the corresponding control forecast. The second one is that the EC ensemble-mean forecast has the lowest RMSE for all four periods and all forecast steps. Note also that the difference between the RMSE of the EC ensemble-mean and the RMSE of the second-best ensemble is larger than the difference between the corresponding control forecasts shown in Fig. 2, especially in AM07 and JJA07 (Figs. 3c,d). Note that these increased differences is smaller for the UKMO and the MSC ensembles: as it will be discussed later, this is most likely linked to the fact that the EC, MSC and UKMO ensembles are the ones with the best tuned ensemble spread. In other words, the fact that the other ensembles have an inferior spread-error relationship contributes to enlarge the difference in their performance.

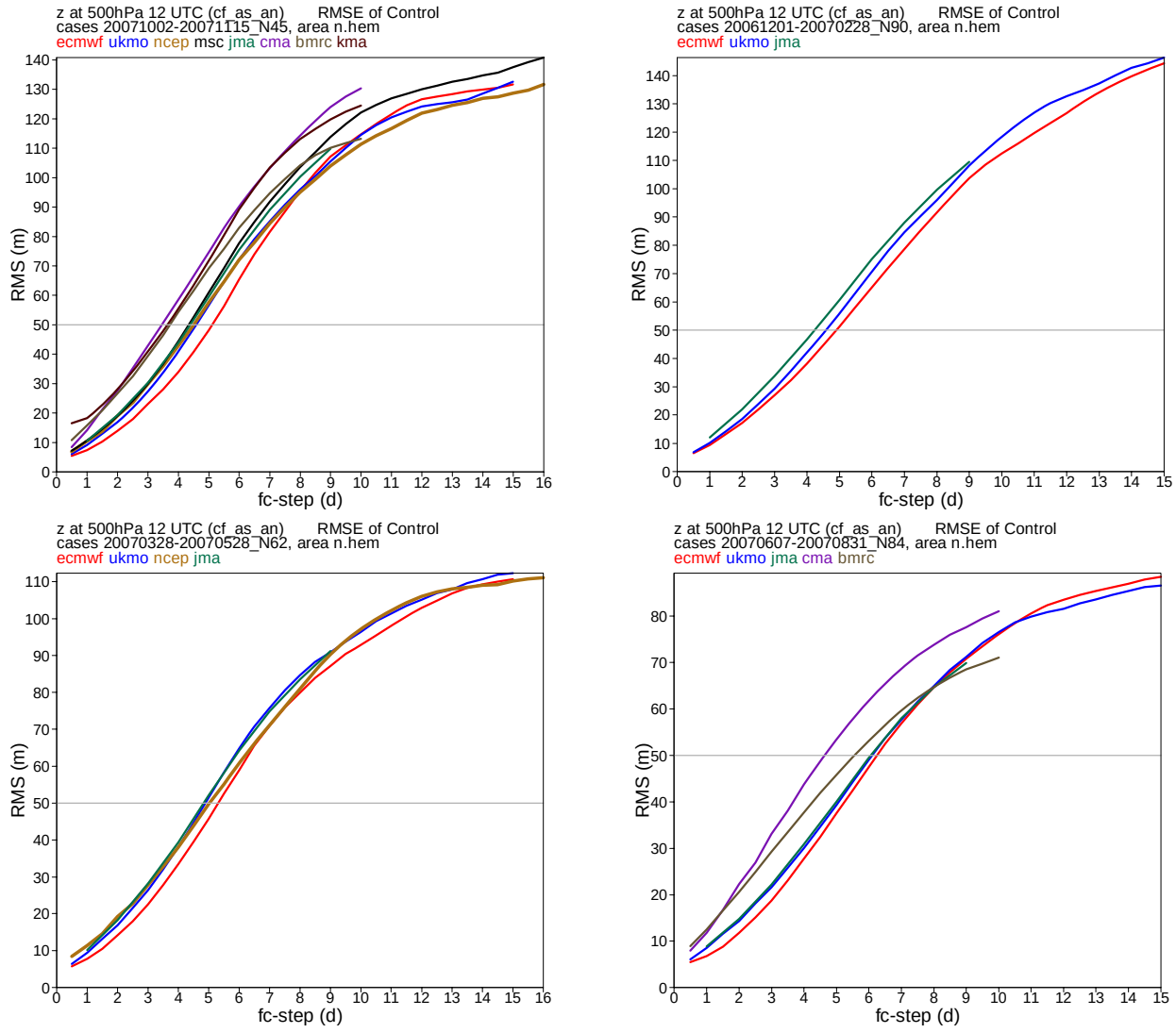


*Figure 2: Average root-mean-square-error of the control forecasts of Z500 over NH of the EC (red line), UKMO (blue line), NCEP (yellow line), MSC (black line), JMA (green line), CMA (violet line), BMRC (purple line) and KMA (black line) ensembles, each verified against its own analysis, for four periods (due to data availability, not all forecasts were available for all periods):*

*a: ON07 (45 cases), EC, UKMO, NCEP, MSC, JMA, CMA, BMRC and KMA;*

*b: DJF07 (90 cases), EC, UKMO and JMA;*

*c: AM07 (62 cases), EC, UKMO, NCEP and JMA;*

*d: JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC*

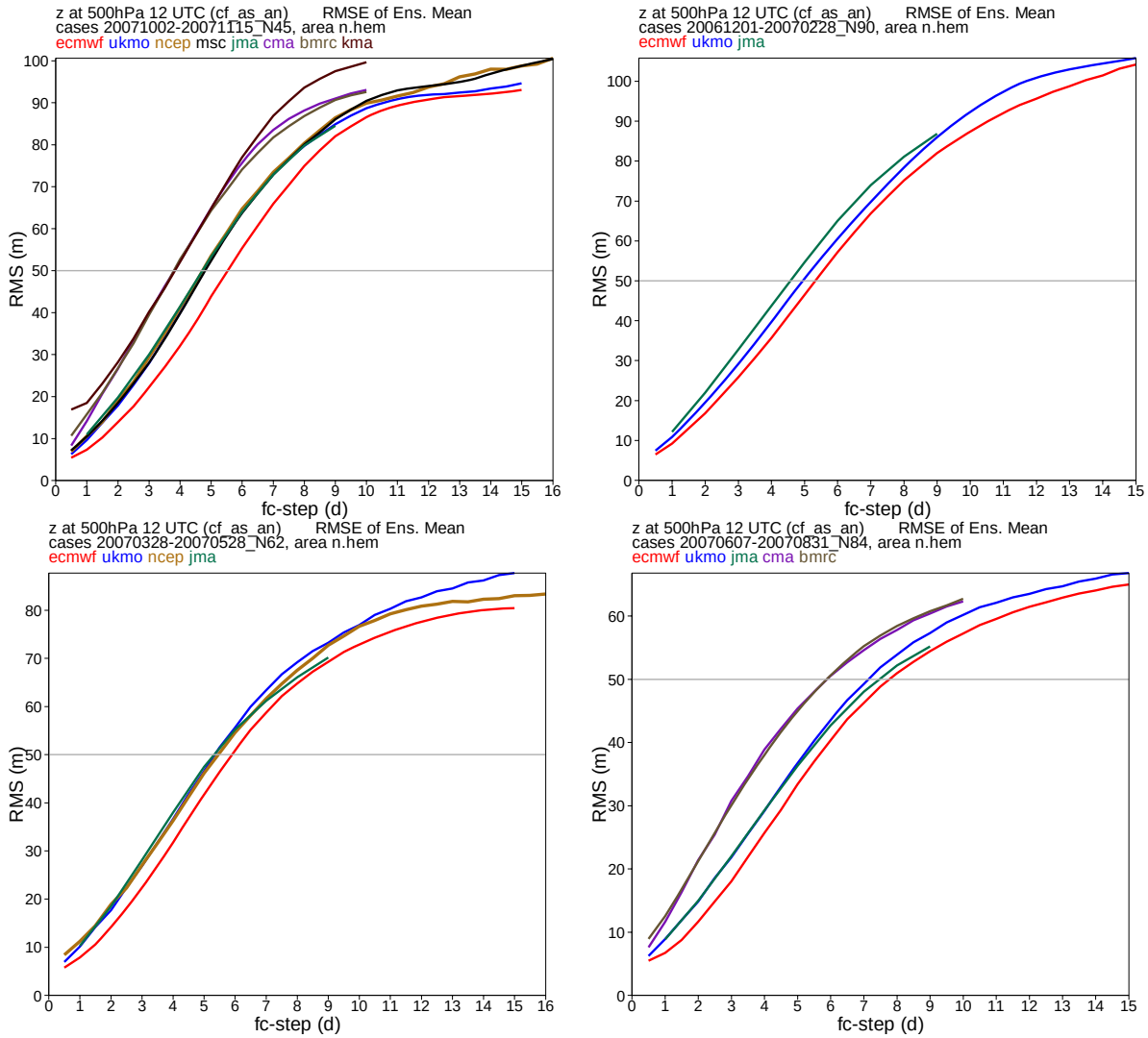


*Figure 3: Average root-mean-square-error of the ensemble-mean forecasts of Z500 over NH of the EC (red line), UKMO (blue line), NCEP (yellow line), MSC (black line), JMA (green line), CMA (violet line), BMRC (purple line) and KMA (black line) ensembles, each verified against its own analysis, for four periods (due to data availability, not all forecasts were available for all periods):*

*a: ON07 (45 cases), EC, UKMO, NCEP, MSC, JMA, CMA, BMRC and KMA;*

*b: DJF07 (90 cases), EC, UKMO and JMA;*

*c: AM07 (62 cases), EC, UKMO, NCEP and JMA;*

*d: JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC*

### 3.2. Ensemble spread measured by the ensemble standard deviation

Figure 4 shows the ensemble standard deviation in terms of Z500 over NH. Differences in ensemble spread are larger than the differences in the RMSE of the ensemble-mean (Fig.3) or the control (Fig. 2) forecasts. Figure 4 shows that in the medium-range (say after forecast day 3), the spread difference between the EC and UKMO ensembles is small for two periods (DJF07 and AM07, Figs. 4b,c), while it is large for the other periods (ON07 and JJA07, Figs. 4a,d). Figure 4 also shows that in the short range, the EC spread is always the lowest and grows the fastest. The fact that the EC spread grows fastest is linked to the use of singular vectors to simulate the effect of initial uncertainties (a similar growth difference between the SV-based EC ensemble and the MSC and NCEP ensembles was detected and documented in *Buizza et al* 2005). Note that in ON07 and JJA07 the BMRC spread grows as fast as, or even faster than, the EC one in the early forecast range: this is due to the fact that the BMRC ensemble also uses initial singular vectors to define the initial perturbations. The two ensembles differ in the medium range. This could be due to several effects (see Table A): different resolution ($T_L$119L19 versus $T_L$399L62) and different model activity, use of a different number of initial singular vectors (16 versus 50), use of evolved singular vectors in the EC ensemble, and stochastic physics in the EC system.

Considering the other ensembles, Fig. 4 shows that the NCEP and the BMRC ensembles have the smallest spread, while the CMA and the JMA ensembles have the largest spread. Unfortunately the MSC ensemble is available only for the most recent period (ON07, Fig. 4a), thus considerations can be made only for this limited period: for this period, the MSC spread is almost identical to the EC spread. Overall, Fig. 4 shows that the spread of the different ensembles varies up to a factor of 2 for Z500 over NH.

It is interesting to compare the spread of the ensembles for another region, one where the EC system is known to underestimate the ensemble spread, for example considering T850 over the tropics (Fig. 5). Differences over this area are larger than for Z500 over NH, up to a factor of 6 for some forecast times. The MSC and the JMA ensembles have the largest spread, the UKMO and the EC ensembles have similar values, and the NCEP, KMA and BMRC ensembles show again the lowest values.

In a perfect ensemble, i.e. an ensemble in which the members are drawn from the same distribution as the true state, the ensemble standard deviation is equal to the RMSE of the ensemble mean when a sufficiently large sample is considered. Figures 6 and 7 show the difference between the RMSE of the ensemble-mean and the ensemble standard deviation for Z500 over NH and T850 over the tropics. In terms of Z500 over NH (Fig. 6), the EC, MSC and UKMO ensembles have the best tuned spread. The EC ensemble tends to overestimate the ensemble spread up to forecast day 5, and the UKMO does the same for a shorter forecast time. The JMA ensemble severely overestimates the ensemble spread for the whole forecast period, while the NCEP, KMA and BMRC ensemble severely underestimate the ensemble spread. Results are different in terms of T850 over the tropics: the MSC and the JMA ensembles are slightly over dispersive, while the other ensembles are under dispersive. For the EC system, this is a known problem due to the limited coverage of the tropical initial perturbations to the areas where a tropical cyclone has been detected in the analysis (*Barkmeijer et al* 2001): work is in progress to address this weakness (e.g. by using an ensemble of data assimilation together with singular vectors, *Buizza & Palmer* 1998, *Leutbecher et al* 2007). Care must be taken when interpreting these T850 results in the tropics: preliminary results (*Martin Leutbecher and Edit Hagel*, personal communication) indicate that if analysis uncertainties are taken into consideration (*Saetra et al* 2004), the level of under dispersion is reduced.

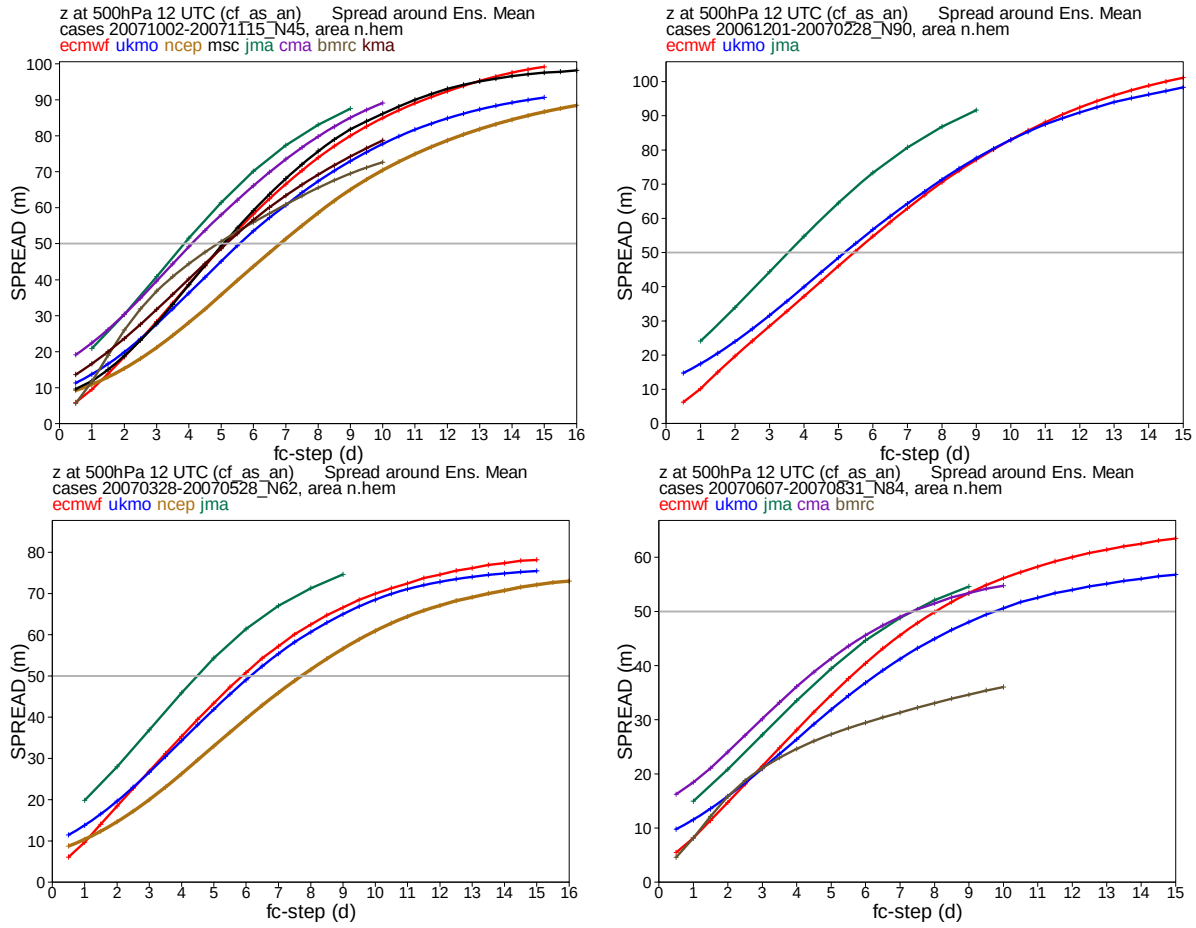


*Figure 4: Average ensemble standard deviation for Z500 over NH of the EC (red line), UKMO (blue line), NCEP (yellow line), MSC (black line), JMA (green line), CMA (violet line), BMRC (purple line) and KMA (black line) ensembles for four periods (due to data availability, not all forecasts were available for all periods):*

*a: ON07 (45 cases), EC, UKMO, NCEP, MSC, JMA, CMA, BMRC and KMA*

*b: DJF07 (90 cases), EC, UKMO and JMA*

*c: AM07 (62 cases), EC, UKMO, NCEP and JMA*

*d: JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC*

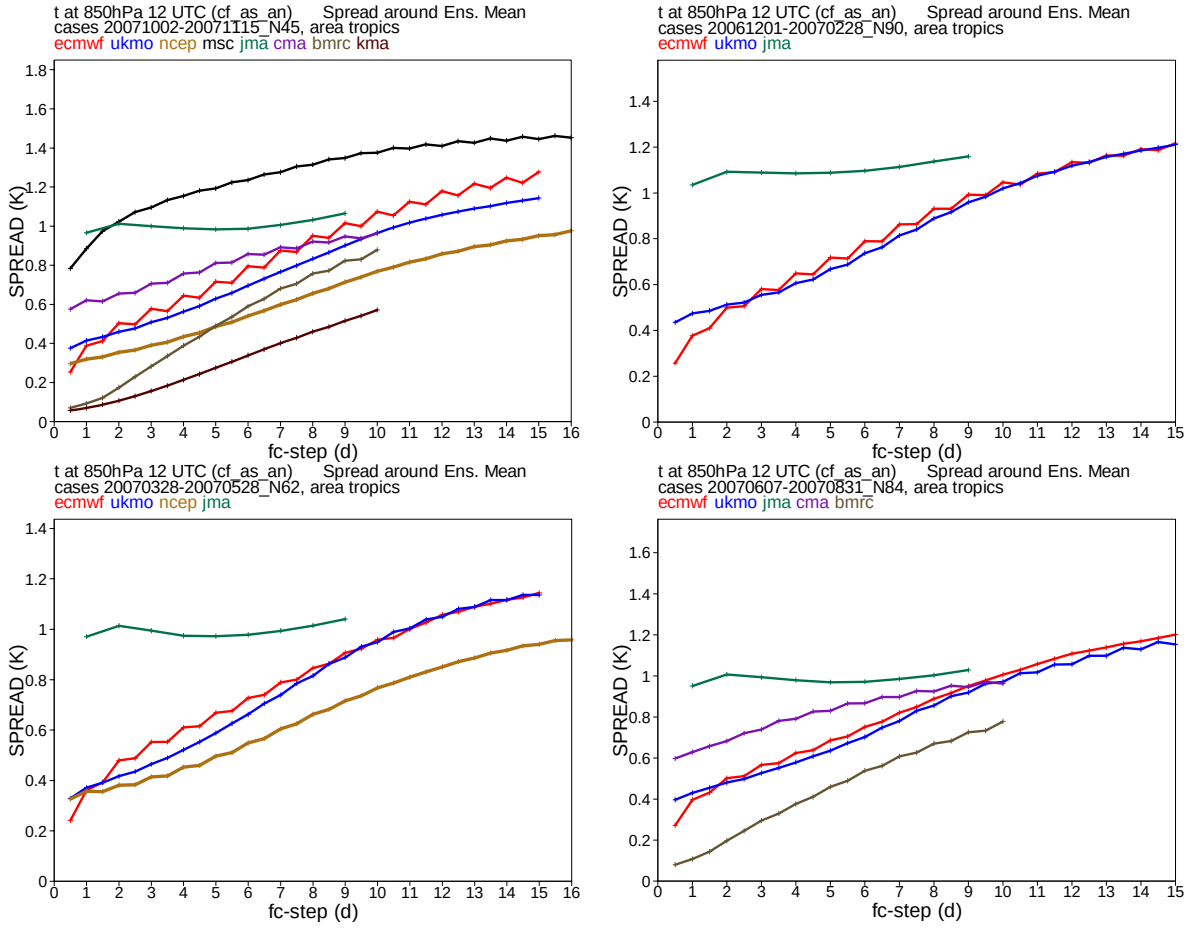


*Figure 5: Average ensemble standard deviation for T850 over the tropics of the EC (red line), UKMO (blue line), NCEP (yellow line), MSC (black line), JMA (green line), CMA (violet line), BMRC (purple line) and KMA (black line) ensembles for four periods (due to data availability, not all forecasts were available for all periods):*

*a: ON07 (45 cases), EC, UKMO, NCEP, MSC, JMA, CMA, BMRC and KMA*

*b: DJF07 (90 cases), EC, UKMO and JMA*

*c: AM07 (62 cases), EC, UKMO, NCEP and JMA*

*d: JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC*

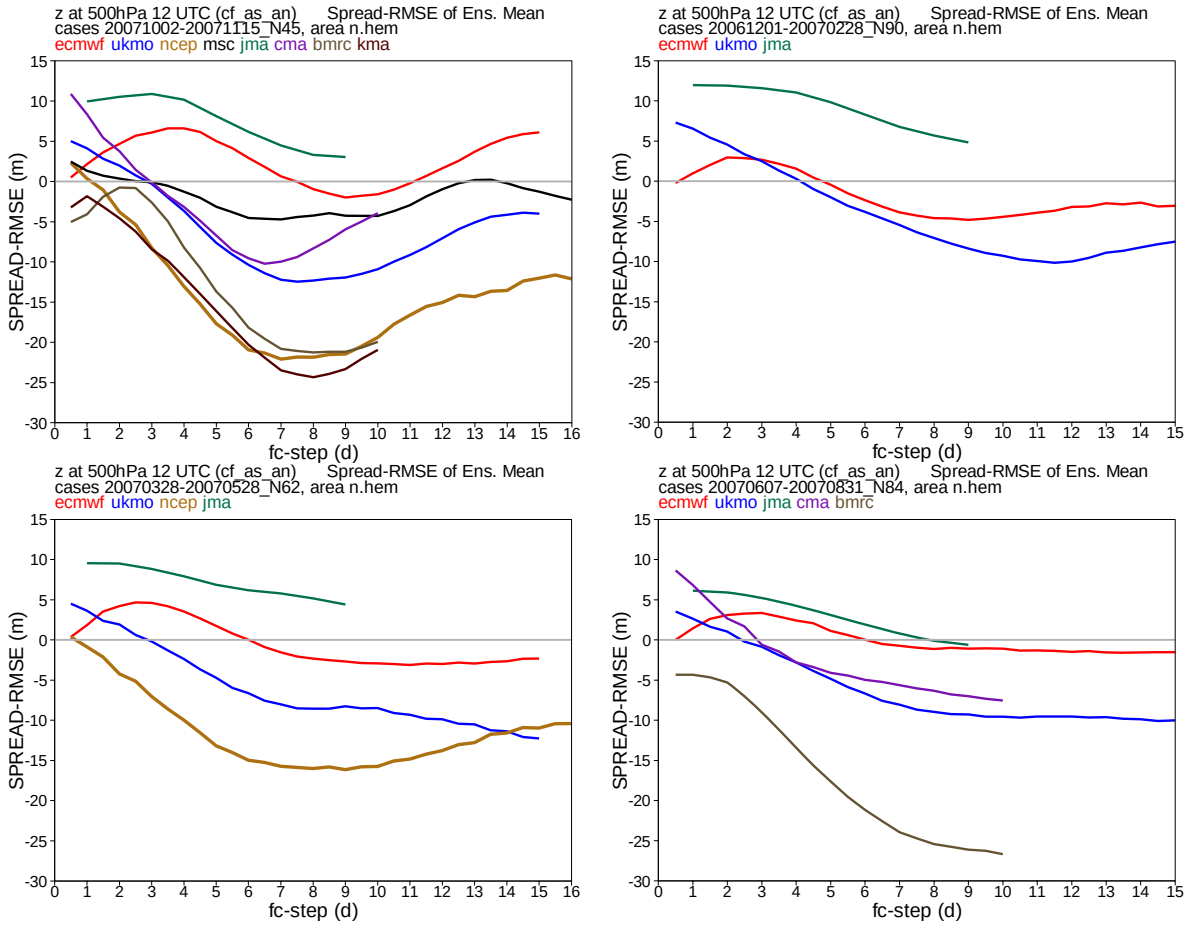


*Figure 6: Difference between the average root-mean-square-error of the ensemble-mean and the average ensemble standard deviation [STD-RMSE(EM)] for Z500 over NH of the EC (red line), UKMO (blue line), NCEP (yellow line), MSC (black line), JMA (green line), CMA (violet line), BMRC (purple line) and KMA (black line) ensembles for four periods (due to data availability, not all forecasts were available for all periods):*

*a: ON07 (45 cases), EC, UKMO, NCEP, MSC, JMA, CMA, BMRC and KMA*

*b: DJF07 (90 cases), EC, UKMO and JMA*

*c: AM07 (62 cases), EC, UKMO, NCEP and JMA*

*d: JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC*

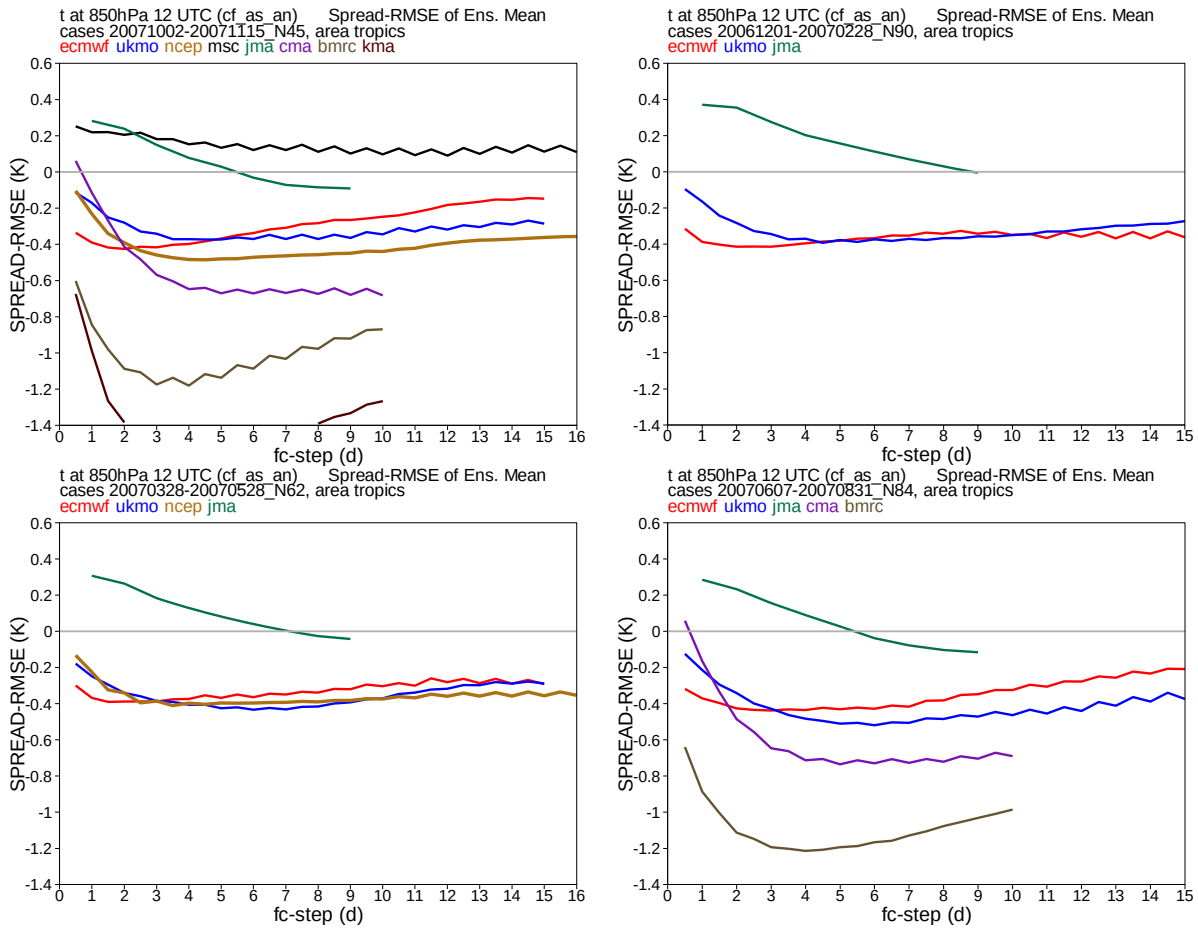


*Figure 7: Difference between the average root-mean-square-error of the ensemble-mean and the average ensemble standard deviation [STD-RMSE(EM)] for T850 over the tropics of the EC (red line), UKMO (blue line), NCEP (yellow line), MSC (black line), JMA (green line), CMA (violet line), BMRC (purple line) and KMA (black line) ensembles for four periods (due to data availability, not all forecasts were available for all periods):*

*a: ON07 (45 cases), EC, UKMO, NCEP, MSC, JMA, CMA, BMRC and KMA*

*b: DJF07 (90 cases), EC, UKMO and JMA*

*c: AM07 (62 cases), EC, UKMO, NCEP and JMA*

*d: JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC*

### 3.3. Skill of probabilistic predictions

Figure 8 shows the RPSS for the probabilistic prediction of Z500 over NH. For all four periods, the EC ensemble shows the highest values for all forecast times. In ON07 (Fig. 8a), four centres (MSC, UKMO, JMA and NCEP) show RPSS differences from EC in the medium-range equivalent to about 1 day of predictability. In the other periods, the differences in the medium-range between the EC ensemble and the group of second best ensembles is smaller, equivalent to about 12 hours in DJF07 (Fig. 8b), and equivalent to about 18 hours in the other two periods. Note that the difference between the RPSS of the different ensembles is larger than the differences between the RMSE of the ensemble-mean forecasts. These results confirm the comment made earlier (sections 3.1 and 3.2), that the level of skill of the ensemble probabilistic forecasts is linked not only to the quality of the data-assimilation and forecasting model, but also to the correct simulation of the ensemble spread. In other words, key ingredients of a skilful ensemble system are a good analysis, a skilful model, and a well tuned ensemble spread obtained with a good simulation of initial and model uncertainties.

Figure 9 shows the RPSS for the probabilistic prediction of T850 over the tropics. Note that over this region the EC ensemble performs worse than many other systems, with the UKMO system performing best in three periods (DJF07, AM07 and JJA07, Figs. 9b-d) for the whole forecast period. However, care must be taken when interpreting the results of Fig. 9 due to the significant biases in the analyses and forecasts for T850 in the tropics. Consider for example the average RPSS scores for DJF07 (Fig. 9b) and for JJA07 (Fig. 9d). By definition, the RPSS of each centre shown in Fig. 9 has been computed using its own analysis as reference, i.e.:

$$RPSS_j = 1 - \frac{RPS_j}{RPS_{cli-j}},$$

where j=EC, UKMO, JMA, CMA and BMRC, respectively. Figure 10 shows the ranked probability score of each forecast ($RPS_j$) and of the reference ($RPS_{cli-j}$, i.e. the RPS of the forecast based on the ERA-40 climatological probability density function verified against analysis j). Consider for example the EC and the UKMO values in DJF07 (Fig. 10a,b): the difference between the RPS of the different ensembles (Fig. 10a) is significantly smaller than the difference between the RPSS (Fig. 9b). The larger difference in RPSS is due to the large difference between the $RPS_{cli-j}$ of the climatological reference forecast (Fig. 10b), with $RPS_{cli-UKMO} > RPS_{cli-EC}$. The larger RPS of the climatological forecast verified against UKMO analysis is caused by the large systematic difference between the UKMO analyses and the ECMWF analyses, which implies that the ERA-40 climatological probability density function is a relative poor prediction of the UKMO analysis. Due to a lack of a re-analysis datasets for the other operational analyses, two approaches are feasible to address this issue. Firstly, one could use a bias corrected climatological distribution when verifying ensembles against their own analyses. Secondly, one could verify all ensembles after bias correction against one reference analysis. The second approach is followed in section 4.2. It is interesting to note that with this second approach the ECMWF ensemble performs better than the UKMO ensemble for T850 in the tropics. This investigation suggests that the interpretation of differences in forecast performance for variables and regions that are affected by significant biases in the initial conditions need to take into account the consistency of the climatological probability density function and the verifying analysis.

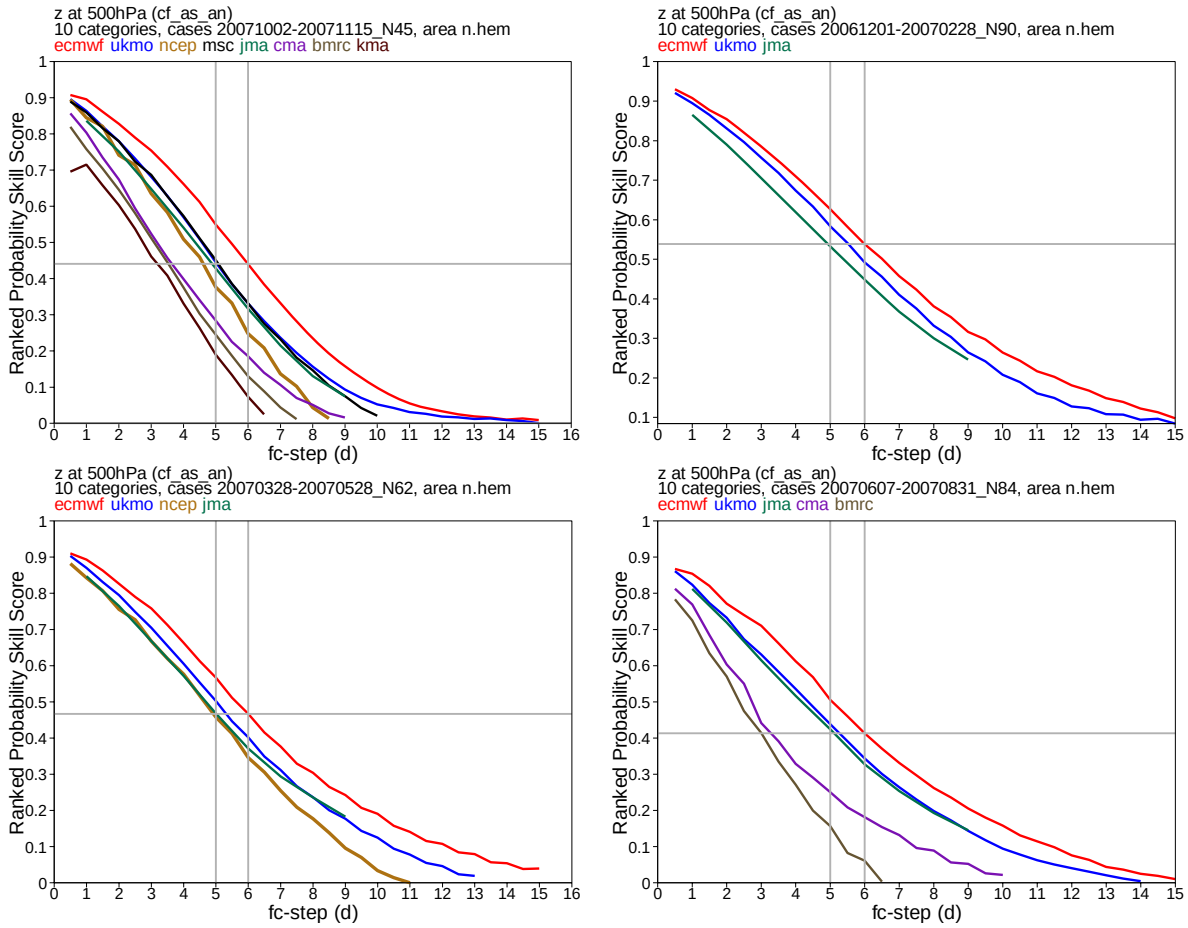


*Figure 8: Average ranked probability skill score for the probabilistic prediction of Z500 over NH of the EC (red line), UKMO (blue line), NCEP (yellow line), MSC (black line), JMA (green line), CMA (violet line), BMRC (purple line) and KMA (black line) ensembles, each verified against its own analysis, for four periods (due to data availability, not all forecasts were available for all periods):*

*a: ON07 (45 cases), EC, UKMO, NCEP, MSC, JMA, CMA, BMRC and KMA*

*b: DJF07 (90 cases), EC, UKMO and JMA*

*c: AM07 (62 cases), EC, UKMO, NCEP and JMA*

*d: JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC*

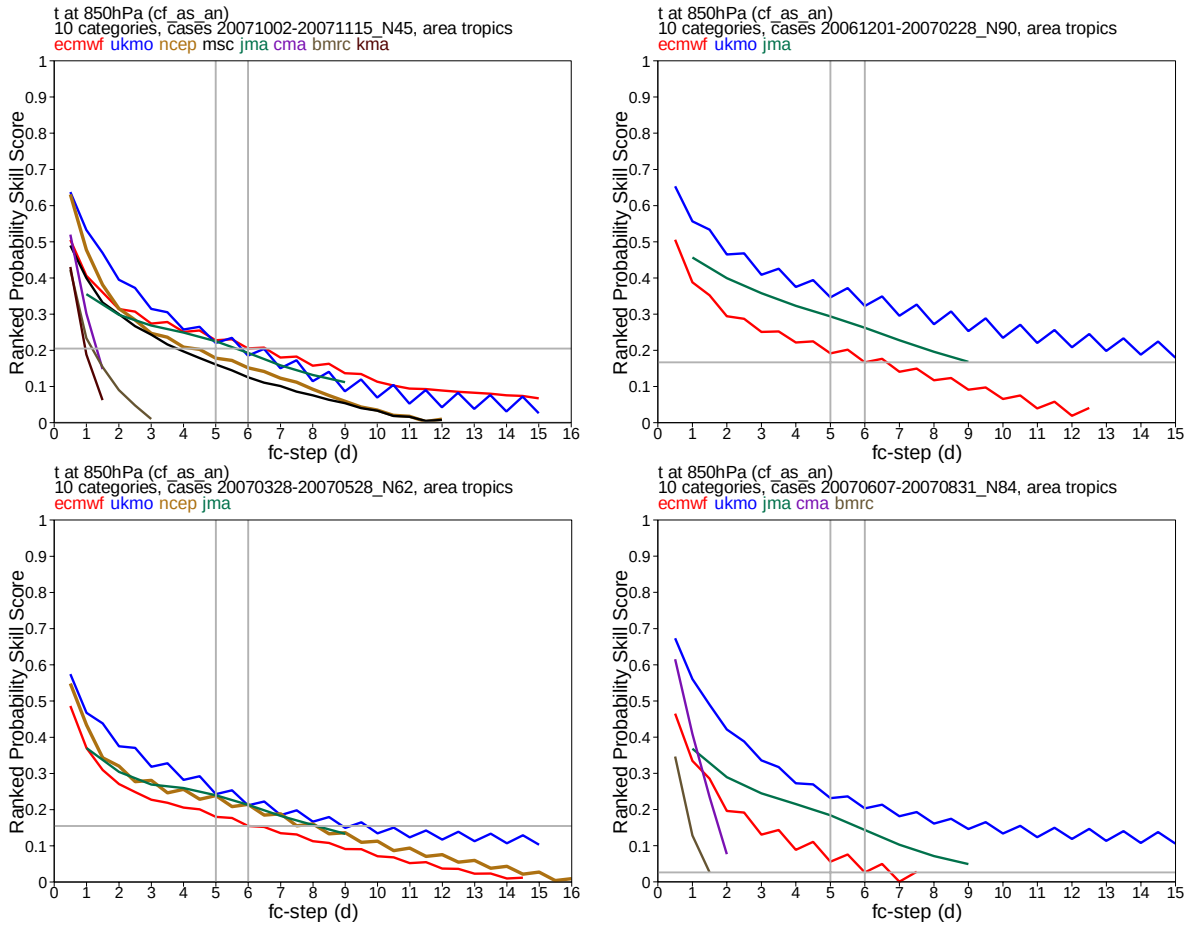


*Figure 9: Average ranked probability skill score for the probabilistic prediction of T850 over the tropics of the EC (red line), UKMO (blue line), NCEP (yellow line), MSC (black line), JMA (green line), CMA (violet line), BMRC (purple line) and KMA (black line) ensembles, each verified against its own analysis, for four periods (due to data availability, not all forecasts were available for all periods):*

*a: ON07 (45 cases), EC, UKMO, NCEP, MSC, JMA, CMA, BMRC and KMA*

*b: DJF07 (90 cases), EC, UKMO and JMA*

*c: AM07 (62 cases), EC, UKMO, NCEP and JMA*

*d: JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC*

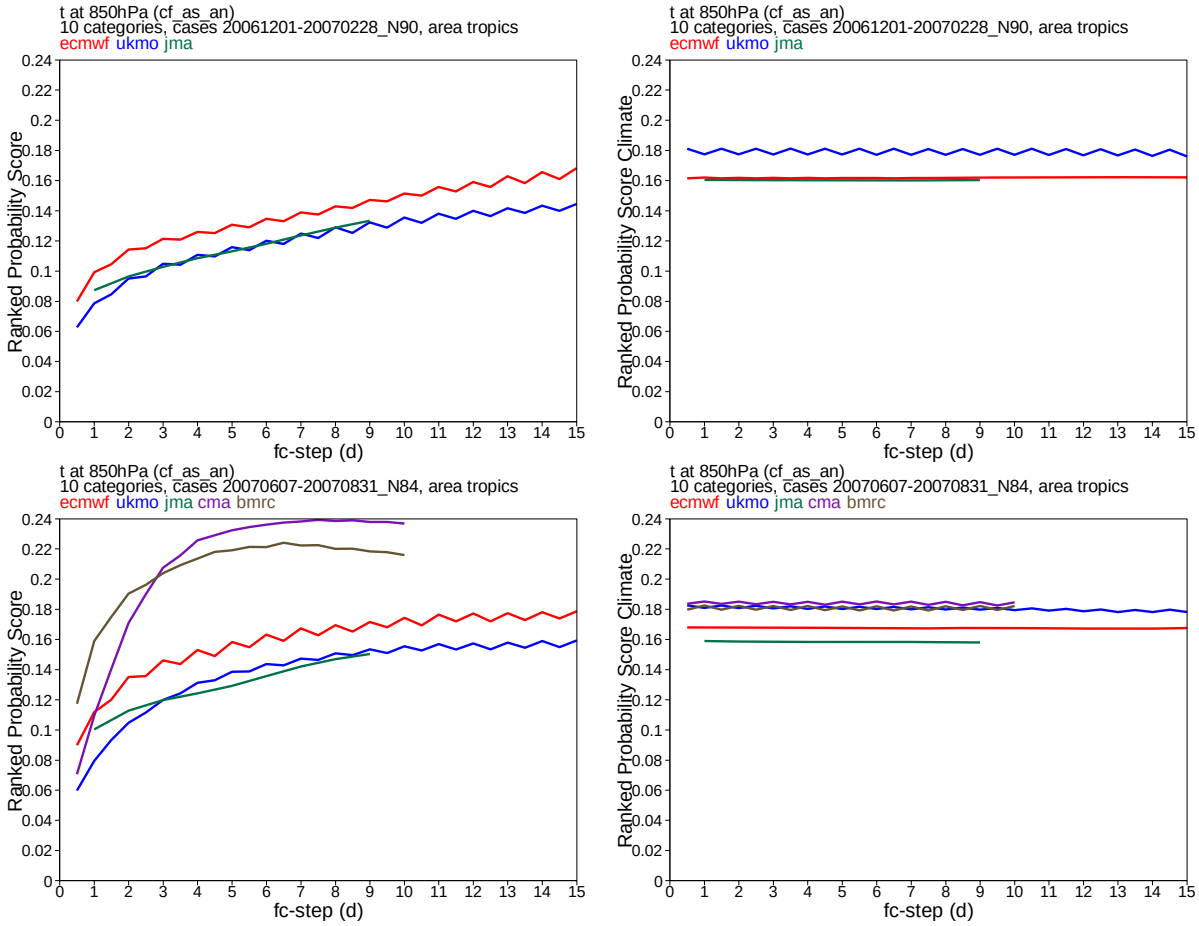


*Figure 10: Average ranked probability score for the ensembles and climatological distribution of T850 over the tropics of the EC (red line), UKMO (blue line), JMA (green line), CMA (violet line), BMRC (purple line) and KMA (black line) ensembles, each verified against its own analysis, for four periods (due to data availability, not all forecasts were available for all periods):*

*a: RPS, DJF07 (90 cases), EC, UKMO and JMA*

*b: RPS climate, DJF07 (90 cases), EC, UKMO and JMA*

*c: RPS, JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC*

*d: RPS climate, JJA07 (84 cases), EC, UKMO, JMA, CMA and BMRC*

# 4. Preliminary results on the potential value of multi-model/multi-analysis ensemble products

Two different, very simple combination methods have been tested to assess the impact on forecast skill of combining different ensemble systems. The first *'equal-weight'* method is based on constructing a multi-model/multi-analysis ensemble by giving the same weight to each single member: ensembles constructed using this method are named 'combined ensembles'. Note that, since the same weight has been given to each single ensemble member, the ensembles with the largest size contributed more than the ones with the smallest size (see Table A for membership information). The second *'equal-weight bias-correction'* method, is based on constructing a multi-model/multi-analysis ensemble by giving the same weight to each single member after applying a bias correction: ensembles constructed in this way are named 'bias-corrected combined' ensembles. This second method relies on three choices: first, the definition of a reference analysis to be used to compute the model biases of all centres, second the definition of the training period used to compute the model biases, and third on the way the bias is computed. On this latter point, biases have been computed by giving the same weight to all the forecasts of the training period (i.e. it has not been given a higher weight to the most recent period). Issues linked to the first two choices are discussed in the following sub-sections.

## 4.1. Reference analysis

The definition of the verification field when an ensemble of analyses is available is a non trivial problem, especially since few of the available analyses have equivalent quality, and each analysis may perform best over a different area. Although it is beyond the scope of this work to identify which is the 'best' analysis, it is worth discussing this issue, to understand how results are sensitive to the choice of the verifying analysis.

Figure 11 shows the area-average value of six analyses in terms of Z500 over the NH and T850 over the tropics. Note that for both variables and regions the differences are rather large, up to 10m for Z500 over NH and almost 2 degrees for T850 over the tropics. Over both regions, the EC and the NCEP analyses rank in the middle, CMA and UKMO show the highest values and JMA and BMRC show the lowest values. The fact

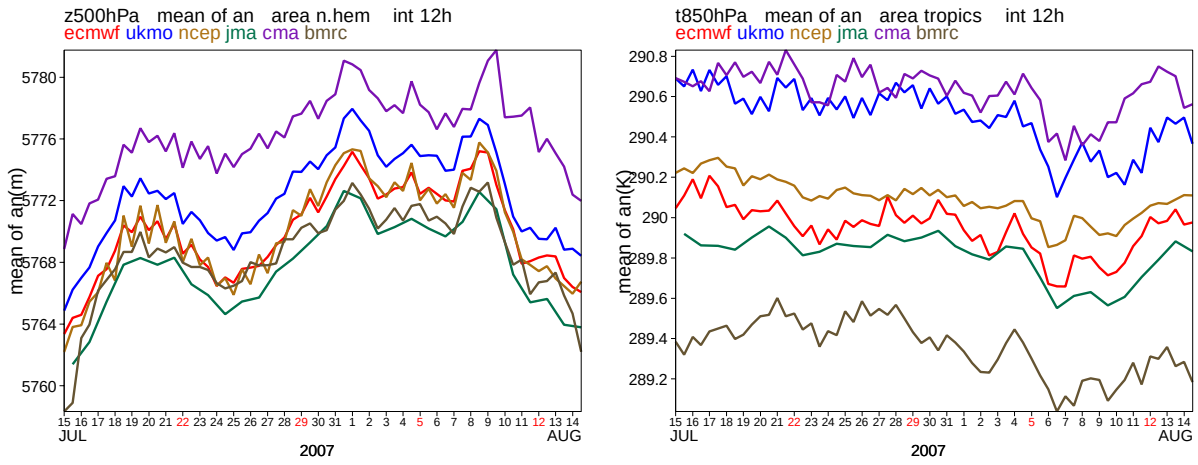


*Figure 11: Time series of the geographical average value of the EC (red line), UKMO (blue line), NCEP (yellow line), JMA (green line), CMA (violet line) and BMRC (purple line) analysis, for (a) Z500 over NH and (b) T850 over the tropics.*

that the analyses have these rather large differences raises the issues of which analysis should be used to compute model biases. (The same problem would have to be addressed in more sophisticates combination/calibration schemes that computes optimal weights to be given to the different ensemble systems.) To provide some more evidence on the relevance of this issue, Figs. 12-14 show the impact on ensemble scores of using a different verification field.

Figure 12.a shows the difference between the root-mean-square-error of the EC ensemble-mean forecast when verified against the EC analysis, the UKMO analysis, the NCEP analysis, and the root-mean-square-error of the EC ensemble-mean forecast when verified against the mean of the EC-UKMO-NCEP analyses. More precisely, the red, blue and yellow line shows the following differences:

$$
\begin{aligned}
d_1(EC) &= RMSE(<EC>, AN_{EC}) - RMSE(<EC>, <AN>) \\
d_2(EC) &= RMSE(<EC>, AN_{UKMO}) - RMSE(<EC>, <AN>) \\
d_3(EC) &= RMSE(<EC>, AN_{NCEP}) - RMSE(<EC>, <AN>)
\end{aligned}
,
$$

where <*AN*> denotes the mean of the three analyses. Similarly, Fig. 12.b and 12.c show the corresponding differences for the UK and NCEP ensemble-mean forecasts. The first conclusion that can be drawn from Fig. 12 is that differences are larger in the short forecast range, say up to forecast day 2, but then they are small compared to the RMSE. Furthermore, results indicate that $d_1$ is smaller than $d_2$ and $d_3$, suggesting that for Z500 over NH the difference between the scores computed using the EC analysis and the scores computed using the mean of the EC-UKMO-NCEP analyses is rather small. Figure 13 shows the RPSS of the three ensembles when verified against the EC analysis, the UKMO analysis, the NCEP analysis, and the mean of the EC-UKMO-NCEP analyses for Z500 over NH. Results indicate that the differences are rather small, detectable mainly in the early forecast range. They also show that, for Z500 over NH, the difference between verifications computed using the EC analysis, or the mean of the EC-UKMO-NCEP analyses is very small. Figure 14 is the same as Fig. 13 but for T850 over the tropics. Over this region the difference between the scores computed using different analyses is very large, with each ensemble achieving the highest score when verified against its own analysis.

These results indicate that for the NH and for variables that represent the large scale synoptic flow (e.g. Z500) the sensitivity to the choice of the verification field is small, and detectable mainly in the early forecast range. By contrast, the sensitivity is large for variables such as T850 over the tropical region. As mentioned above, it is beyond the scope of this work to investigate which is the best analysis. Figures 12-14 can be used to estimate the potential impact that choosing a different analysis might have on the results discussed in section 4.2. But since a choice has to be made, in view of the fact that the EC analysis is the highest resolution one, possibly one of the best since single forecasts started from it have the lowest RMSE (Fig. 2), and has values that lies in the middle of the distribution of the available analysis (Fig. 11), the EC analysis is used in the final part of the study.

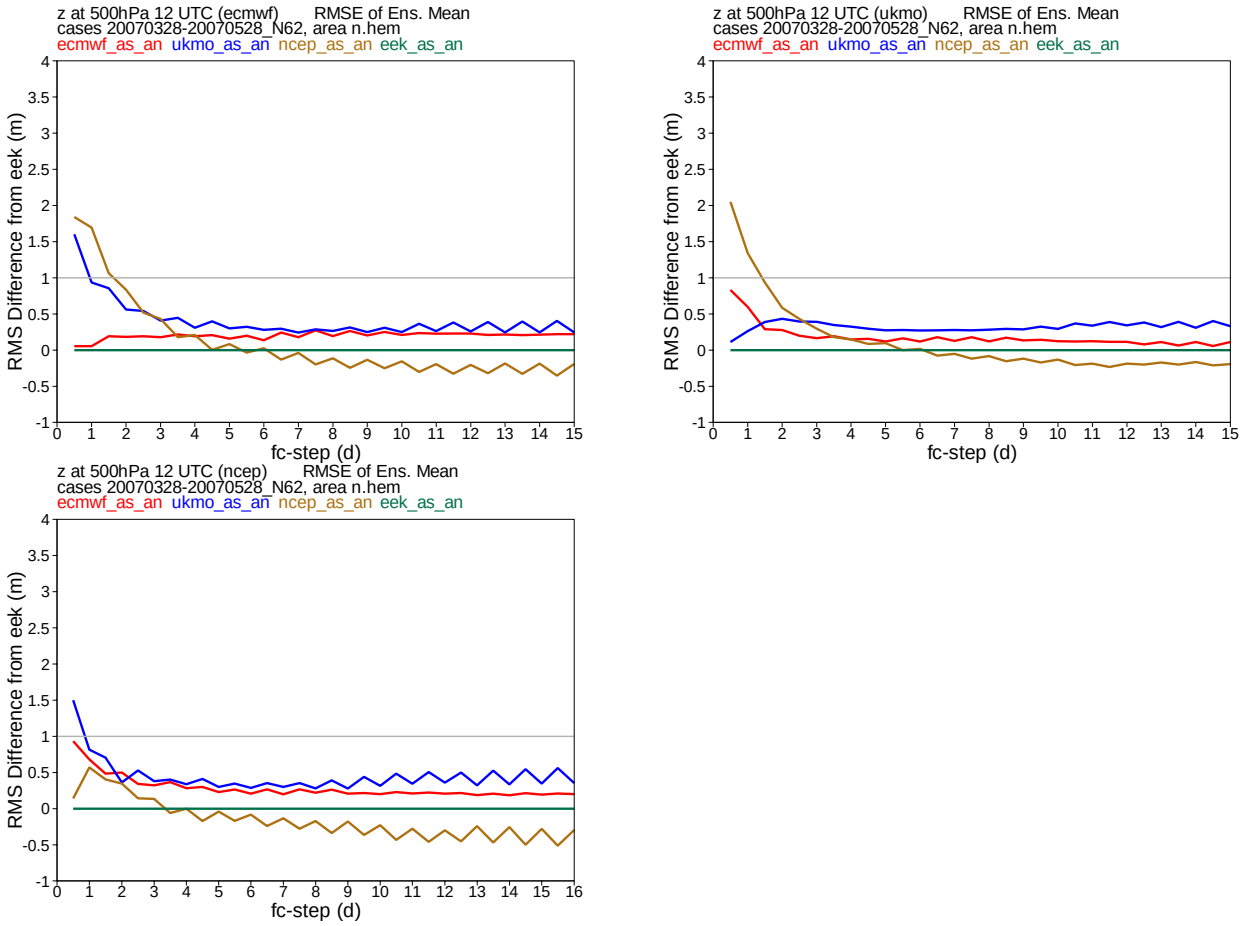


*Figure 12: Sensitivity to verification analysis. (a): difference between the root-mean-square-error of the EC ensemble-mean forecasts given by verified against the EC analysis (red line), the NCEP analysis (yellow line) and the UKMO analysis (blue line) and the root-mean-square-error of the EC ensemble-mean forecast verified against the mean of the three analysis (green line). (b): as (a) but for the UK ensemble-mean forecast. (c): as (a) but for the NCEP ensemble-mean forecast. Results refer to the AM07 (62 cases) average for Z500 over NH.*

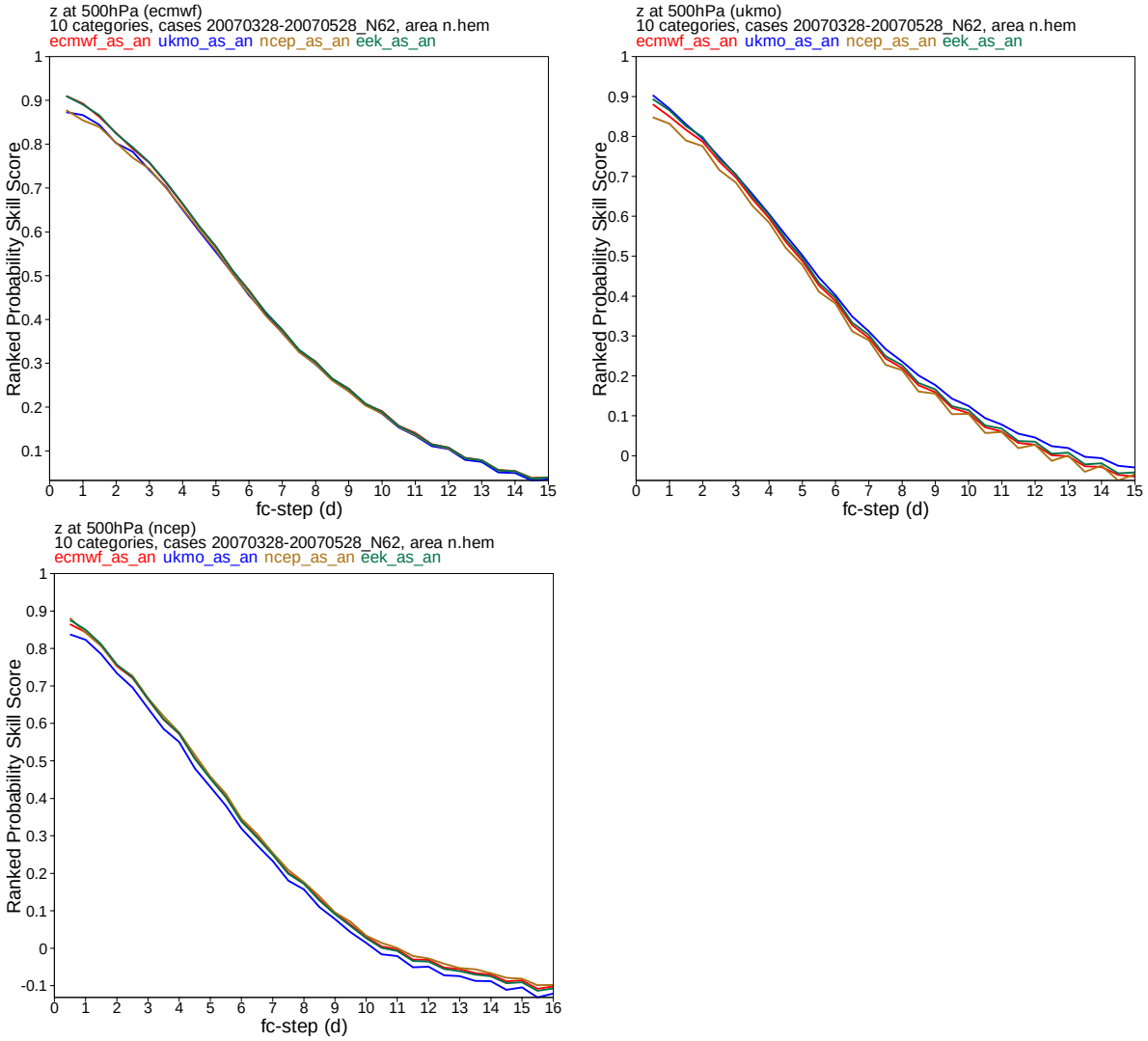


*Figure 13: Sensitivity to verification analysis: AM07 (62 cases) average rank probability skill score of the probabilistic forecasts of Z500 over NH given by (a) the EC ensemble, (b) the NCEP ensemble and (c) the UKMO ensemble verified against the EC analysis (red line), the NCEP analysis (yellow line), the UKMO analysis (blue line) and the mean of the three analysis (green line).*

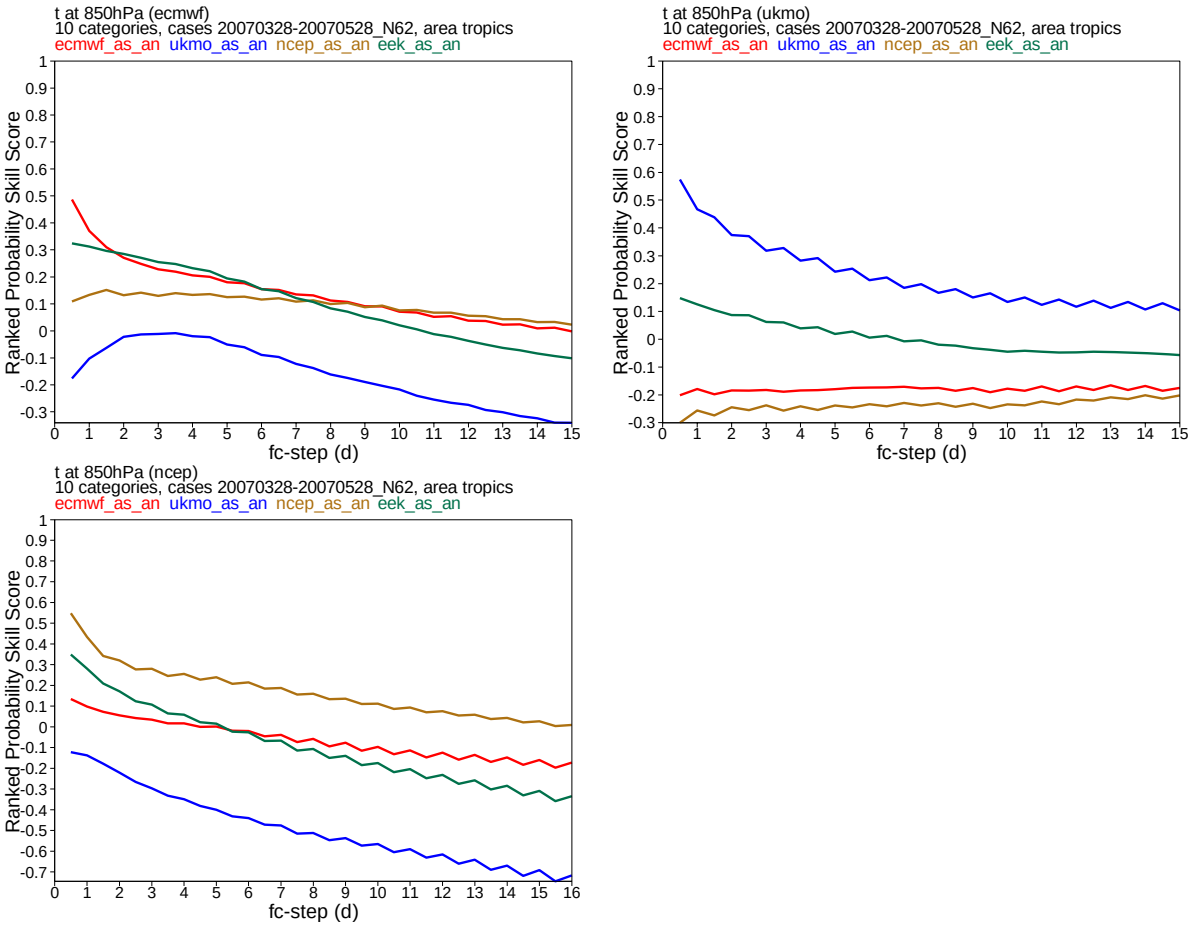


*Figure 14: Sensitivity to verification analysis: AM07 (62 cases) average rank probability skill score of the probabilistic forecasts of T850 over the tropics given by (a) the EC ensemble, (b) the NCEP ensemble and (c) the UKMO ensemble verified against the EC analysis (red line), the NCEP analysis (yellow line), the UKMO analysis (blue line) and the mean of the three analysis (green line).*

## 4.2. Training period for bias-correction

The sensitivity of the bias-correction method to the length of the training data set has been tested using two of the best ensemble systems that were available for the longest period, the EC and the UKMO ensembles. As discussed above, the reference analysis that has been used to compute the model biases is the EC analysis. Furthermore, due to the limited data set available, the training period used to compute the model biases has been limited 60 days. It is worth pointing out that choosing a longer training period is not possible: given the fact that each centre changes the ensemble system at least one time a year (e.g. due to changes in the data assimilation or the model, or in the ensemble resolution, size, frequency), and the fact that centres do not usually perform re-forecasts with the most recent ensemble system, it is practically impossible to have long periods of data available for all centres. Furthermore, using a short training period has the advantage of resolving seasonal variations in the biases.

Figure 15 shows the impact of using a 15-, 30- or 60-day training period on the skill of combined (bias-corrected and non) EC-UKMO ensemble forecasts obtained for the FMAMJJA07 period (204 cases). For Z500 over the NH (Fig. 15a), bias-correction does not have any positive effect: the RMSE of the bias-corrected combined ensemble-mean (Fig. 15a) increases significantly if a 15-day training period is used, while if a 30 or a 60 day period is used the RMSE gets closer to the RMSE of the non-bias-corrected combined ensemble-mean. Similarly, the probabilistic prediction of the bias-corrected combined ensemble with a 15-day training period is worse than the one of the non-bias-corrected combined ensemble (Fig. 15b): increasing the training period from 15 to 30 or 60 days brings the performance closer to the one of a simply combined ensemble. For T850 over the tropics (Fig. 15c,d), bias-correction has a positive effect, with scores obtained using a 30 or 60 days being very close. These results, although based on a limited sample, suggest that care should be taken when combining different ensemble systems: these preliminary results indicate that in the case of the EC and the UKMO ensembles, for some variables and periods a simple combination method might be superior to a bias-correction method. A possible explanation of this finding is that the training period is too short. If this is the reason, due to the lack of longer dataset of ensemble forecasts in TIGGE, it might be better to use equal-weights when combined different ensemble systems. It is interesting to point out that this is the approach currently followed by NAEFS.

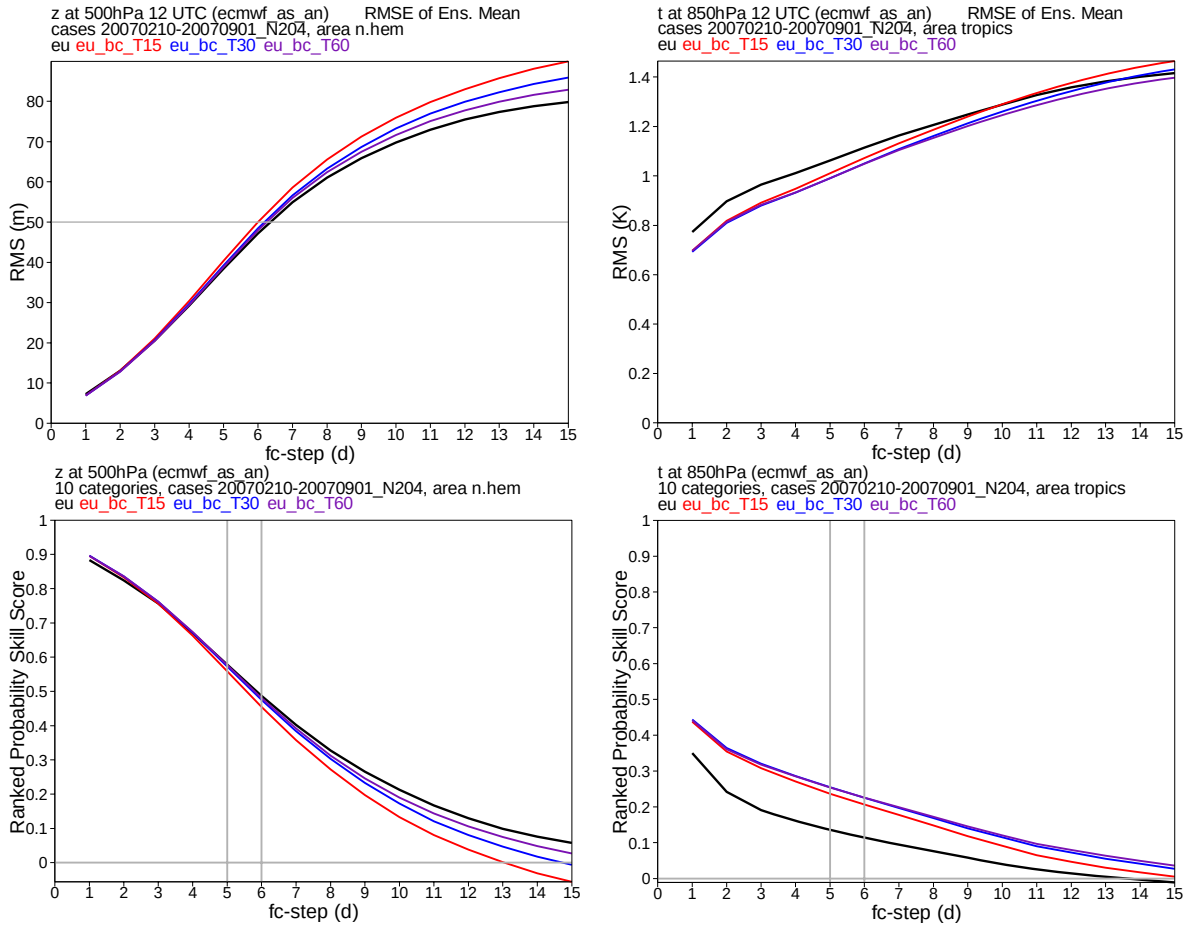


*Figure 15: Sensitivity to the length of the training period: FMAMJJA07 (204 cases) average scores of combined EC and UKMO (eu, black line) and of combined bias-corrected ensembles with biases estimated with a 15 (eu_bc_T15, red line), 30 (eu_bc_T30, blue line) and 60 (eu_bc_T60, violet line) day training period:*

*a: root-mean-square-error of the ensemble mean forecast of Z500 over NH*

*b: root-mean-square-error of the ensemble mean forecast of T850 over the tropics*

*c: the rank probability skill score for Z500 over NH*

*d: the rank probability skill score for T850 over the tropics*

## 4.3. Combination of ensemble systems to generate grand global ensemble products

Before discussing the impact on the ensemble scores of combining ensembles from different centres, it is interesting to compare the impact of bias-correction on the skill of two ensembles, EC and CMA, for another period, JJA07 (86 cases). Note that since the forecast length of the CMA ensemble is 10 days, the ensemble has been scored up to forecast day 10 only. Figure 16 confirms the results discussed above (shown in Fig. 15) that for Z500 over NH (Fig. 16a,c) the impact of bias correction is negative, while the impact is positive over the tropics. Figure 16 also compares the scores of single ensembles (without or with bias correction) with the score of bias-corrected combined EC, UKMO, JMA and CMA ensembles (since the forecast length of the JMA ensemble is 9 days, the combined ensemble has been scored up to forecast day 9 only). The comparison of the four centre combined ensemble-mean with the (single centre) EC ensemble-mean indicates that for Z500 over NH the impact of adding three ensembles to the EC one has a negligible impact, but it has a positive impact for T850 over the tropics.

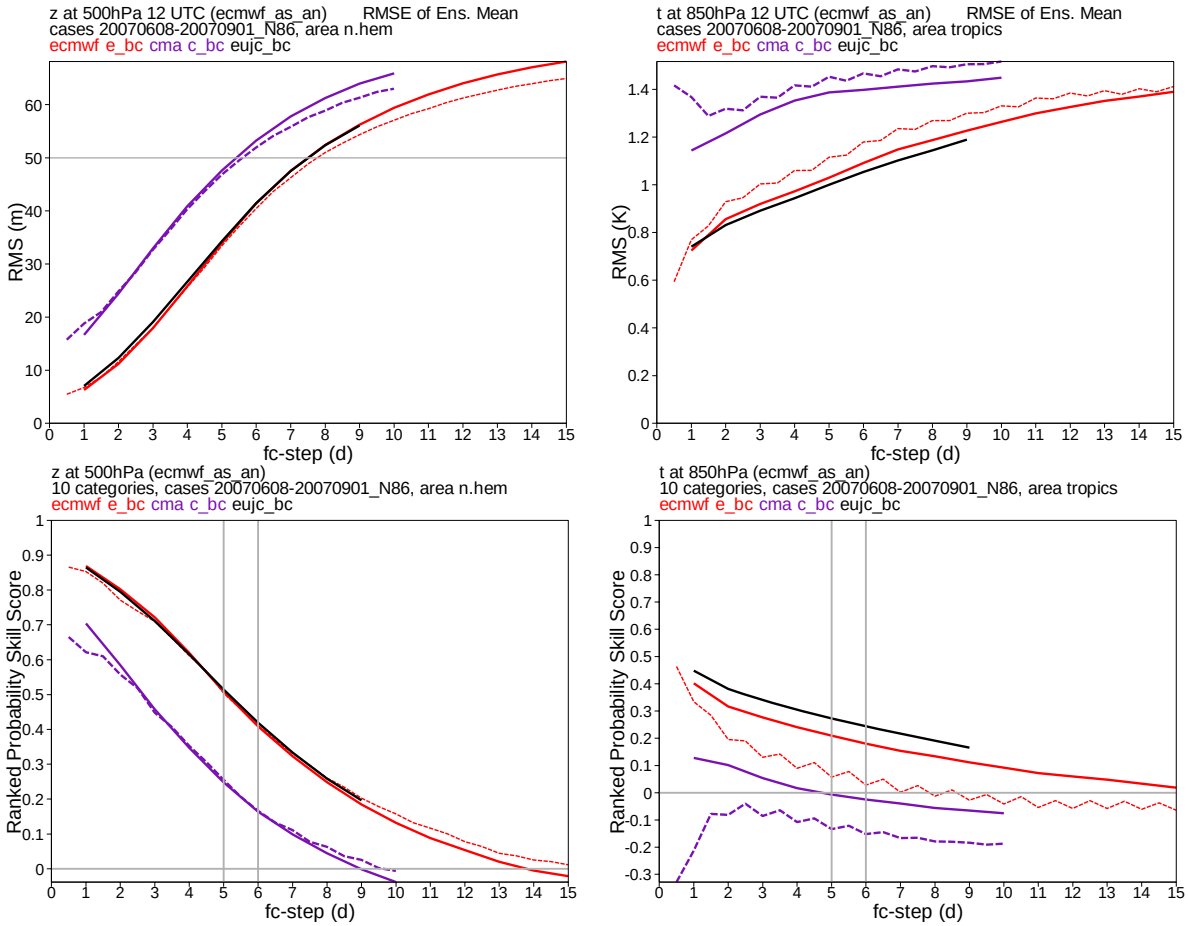


*Figure 16: Combination results: JJA07 (86 cases) average scores of EC (ecmwf, dashed red line), bias-corrected EC (e_bc, solid red line), CMA (cma, dashed violet line) and bias-corrected CMA (cma_bs, solid violet line) and bias-corrected combined EC, UKMO, JMA and CMA (eujc_bc, solid black line), with biases estimated using a 30-day training period:*

*a: root-mean-square-error of the ensemble mean forecast of Z500 over NH*

*b: root-mean-square-error of the ensemble mean forecast of T850 over the tropics*

*c: the rank probability skill score for Z500 over NH*

*d: the rank probability skill score for T850 over the tropics*

Figure 17 shows more results on the effect of combination of ensemble systems for the same period, JJA07. More precisely, Fig. 17 compares the scores of the single EC ensemble, with or without bias correction, with the scores of bias-corrected combined EC-UKMO, EC-UKMO-JMA and EC-UKMO-JMA-CMA ensembles. Results indicate that for Z500 over NH the scores of all these ensembles are very close, all slightly worse than the score of the single EC ensemble without bias correction. By contrast, for T850 over the tropics, adding the UKMO and the JMA ensembles to the EC improves the scores, but further adding the CMA ensemble does not bring any extra improvement. It is interesting to point out that most of the improvements are brought by the combination of the EC and UKMO ensembles, two systems characterized by a similar, high-resolution and a large membership (Table A).

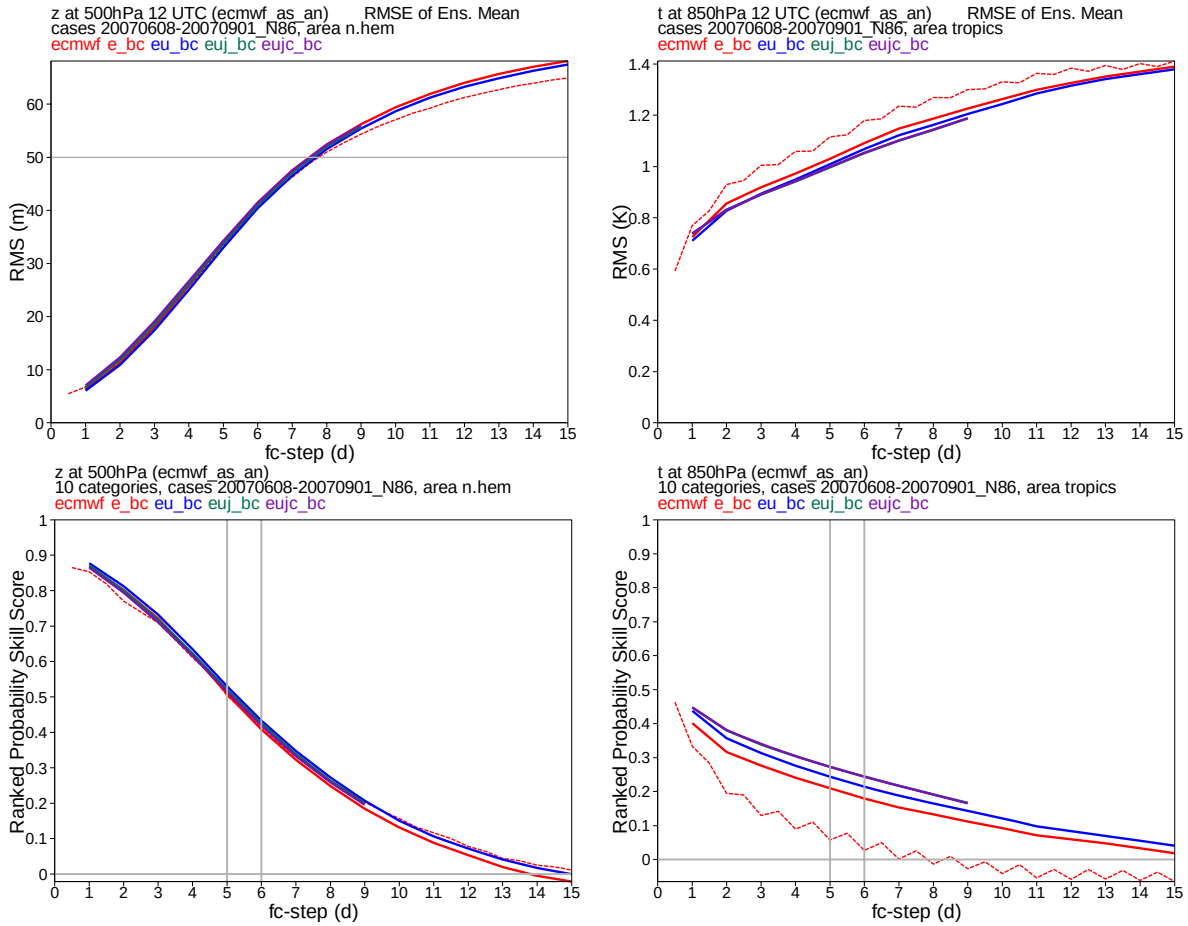


*Figure 17: Combination results: JJA07 (86 cases) average scores of EC (ecmwf, dashed red line), bias-corrected EC (e_bc, red line), bias-corrected combined EC and UKMO (eu_bc, blue line), bias-corrected combined EC, UKMO and JMA (euj_bc, green line) and bias-corrected combined EC, UKMO, JMA and CMA (eujc_bc, violet line), with biases estimated using a 30-day training period:*

*a: root-mean-square-error of the ensemble mean forecast of Z500 over NH*

*b: root-mean-square-error of the ensemble mean forecast of T850 over the tropics*

*c: the rank probability skill score for Z500 over NH*

*d: the rank probability skill score for T850 over the tropics*

To further investigate the impact of merging two ensembles with rather similar characteristics, either with or without bias-correction, Fig. 18 compares the performance of the EC and UKMO ensembles (without bias correction), with the performance of the combined EC-UKMO ensembles with and without bias correction for JJA07. For Z500 over the NH, results (Fig. 18a,c) show that the single EC ensemble performs better than the single UKMO ensemble, and that in the short-range the combined bias-corrected EC-UKMO ensemble performs best, but in the long range the combined non-bias-corrected EC-UKMO ensemble performs best. By contrast, for T850 over the tropics it is the combined EC-UKMO with bias-correction that outperforms the other systems.

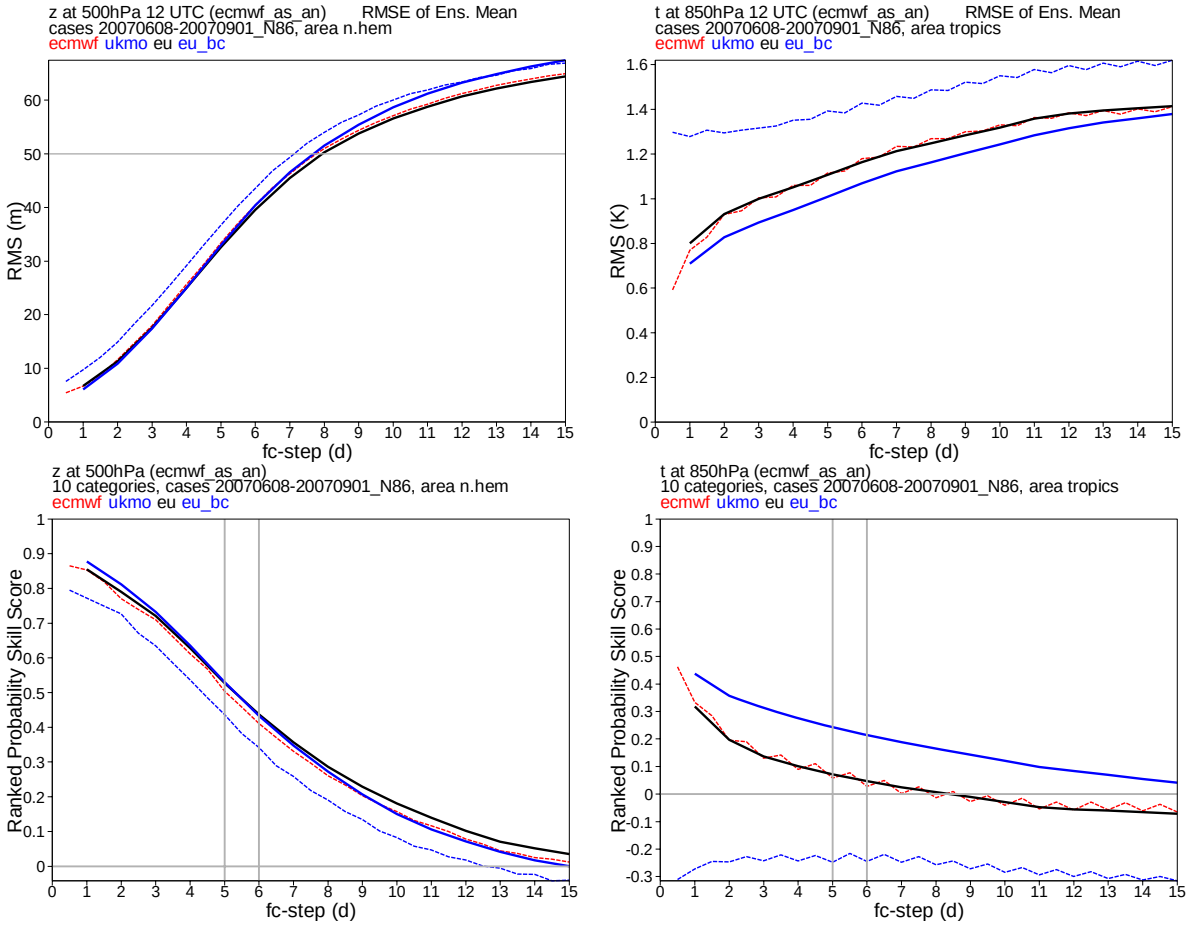


*Figure 18: Combination results: JJA07 (86 cases) average scores of EC (ecmwf, dashed red line), UKMO (ukmo, dashed blue line), combined EC and UKMO (eu, solid black line) and bias-corrected combined EC and UKMO (eu_bc, solid blue line), with biases estimated using a 30-day training period:*
*a: root-mean-square-error of the ensemble mean forecast of Z500 over NH*
*b: root-mean-square-error of the ensemble mean forecast of T850 over the tropics*
*c: the rank probability skill score for Z500 over NH*
*d: the rank probability skill score for T850 over the tropics*

These results indicate that care must be taken when combining ensemble systems: for Z500 over NH, results indicate that a simple equal-weight combination without bias correction of only few of the best ensembles brings a better performance than the combinations of bias-corrected ensembles. Results for T850 over NH are different (not shown). But for T850 over the tropics a positive impact of bias correction was detected. Note that this is the area where most of the single ensembles have a rather poorly tuned ensemble spread (see Fig. 7), which suggests that, at least for the variables considered in this study, the improvements that very

simple combination methods can bring to a good ensemble system (e.g. the EC one) can be rather limited. This result might be due to the fact that the training period used to compute the model biases was too short: but practical reasons might make it impossible to extend the training set to include a much longer period (it was impossible for us to do so with the current available data set). More work on this issue, which takes into account the existing practical constraints, is required to understand whether different combination methods can lead to larger improvements.

# 5. Discussion and conclusions

This report briefly reviewed the status of the TIGGE archive, and illustrated the value of the TIGGE database by presenting some preliminary results obtained using all the available data at the time of writing (December 2007). In the first part of this report, the performance of eight ensembles (BMRC, CMA, EC, JMA, KMA, MSC, NCEP and UKMO) has been compared, and their strengths and weaknesses have been analyzed. In the second part of the report, issues linked to the combination of ensemble systems to generate a grand multi-model/multi-analysis ensemble have been discussed, and the potential value of combining different ensembles has been investigated. The preliminary results presented in this work should provide valuable information to the scientists responsible for the future development of each single ensemble system. A more complete analysis of the TIGGE ensembles based on more seasons will be performed as soon as ensemble forecasts from all TIGGE contributors are available for longer periods.

One of the key results of this investigation has been the quantification of the difference in performance of the different ensembles. Results have indicated that there is a large difference between the performance of the single ensembles: for Z500 over NH, in the medium-range (say around forecast day 5), the difference between the worst and the best control or ensemble-mean forecasts is about 2 days of predictability (Figs. 2-3), while the difference between the worst and the best probabilistic predictions can be larger, about 3 days of predictability (Fig. 8).

Another key result has been the quantification of the difference between the skill of the EC ensemble and a combined ensemble generated considering up to four different ensemble systems. Results have indicated that the difference is very small in areas where the EC ensemble system has a well tuned ensemble spread, equivalent to less than 6 hours of predictability in the medium range (Figs. 17 and 18).

Although these results are based on a limited sample of cases and variables, it is thought that they provide some useful indications of the status of ensemble prediction, and on the issues that need to be addressed to combine different ensemble systems in an effective way. Scientists are encouraged to access the TIGGE database, and try to answer some of the key questions that were raised by this investigation, such as the following ones:

- Which is the best combination method that should be used to generate GRAnd multi-Model/Multi-Analysis (GRAMMA) ensemble products? Should ensembles be bias-corrected first? Can the TIGGE centres generate larger data-sets that could be used to design the combination methods? How long should the training data-set be?

- Should the ensembles be given an equal or a different weight? Which weight should be given to the different ensembles?

- Is it better to use only few well-tuned ensembles in the GRAMMA ensemble, or should all available ensembles be used?

- What is the sensitivity of GRAMMA's value to the forecast variable/area?
- Should different combination methods be used for different variables?
- Can a single (calibrated) ensemble outperform GRAMMA?
- Is GRAMMA the best approach to simulate the effect of model uncertainty on forecast quality?
- If GRAMMA is the way forward in ensemble prediction, which is the best way to use the resources available at the different centres? Should all centres try to run forecasts with a similar resolution? In other words, which is the best GRAMMA?
- Why are the analyses from the different centres so different? Which of them is closest to the true state of the atmosphere?

## Acknowledgements

Young-Youn Park is very thankful to KMA and ECMWF for arranging and supporting her 1-year visit to ECMWF. The ten TIGGE contribution centres and the three data centres are thanked for providing data. Many ECMWF staff and consultants are acknowledged and thanked for their support: this study would not have been possible without all their work. Thanks go in particular to Philippe Bougeault, for his very valuable comments during the whole project, to Claude Gilbert for his help in developing the EPS verification software, and Baudouin Raoult for his role in the design and realization of the TIGGE archive. Tim Palmer is acknowledged for comments to an earlier version of this paper.

## References

Barkmeijer, J., Buizza, R., Palmer, T. N., Puri, K., & Mahfouf, J.-F., 2001: Tropical singular vectors computed with linearized diabatic physics. *Q. J. R. Meteorol. Soc*., **127**, 685-708.

Bishop, C. H., Etherton, B. J. & Majumdar, S. J., 2001: Adaptive sampling with the ensemble transform kalman filter. Part I: theoretical aspects. *Mon. Wea. Rev.*, **129**, pp. 420-436.

Bowler, N. E., Arribas, A., Mylne, K. R., & Robertson, K. B., 2007 : The MOGREPS short-range ensemble prediction system. Part I: system description. *MetOffice NWP Technical Report N. 497*, available from The MetOffice, FitzRoy Rd, Exeter, EX1 3PB, UK, pp. 18 (see also UKMO web page).

Bourke, W., T. Hart, P. Steinle, R. Seaman, G. Embery, M. Naughton, & L. Rikus, 1995: Evolution of the Bureau of Meteorology's Global Assimilation and Prediction system. Part 2: resolution enhancements and case studies. *Aust. Met. Mag.***, 44,** 19-40**.**

Bourke, W., Buizza, R., & Naughton, M., 2004: Performance of the ECMWF and the BoM Ensemble Systems in the Southern Hemisphere. *Mon. Wea. Rev.* **132**, 2338-2357.

Buizza R., & T. N. Palmer. 1995: The singular vector structure of the atmospheric general circulation. *J. Atmos. Sci.*, **52**, 1434-1456.

Buizza R., & Palmer, T. N., 1998: Impact of ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **124**, 1935-1960.

Buizza R., M. Miller, & T. N. Palmer. 1999: Stochastic representation of model uncertainties in the ECMWF EPS. *Q. J. Roy. Meteor. Soc.*, **125**, 2887-2908.

Buizza R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, & Y. Zhu. 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076-1097.

Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G., & Vitart, F., 2007: The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). *Q. J. Roy. Meteorol. Soc.*, 133, 681-695.

Goo, T.-Y., S.-O. Moon, J.-Y. Cho, H.-B. Cheong, & W.-J. Lee, 2003: Preliminary results of medium-range ensemble prediction at KMA: Implementation and performance evaluation as of 2001. *Korean J. Atmos. Sci.*, **6**, 27-36.

Houtekamer, P L & L. Lefaivre, 1997: Using ensemble forecasts for model validation. *Mon. Wea. Rev*, **125**, 2416-2426.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie & H. L. Mitchell, 1996. A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225-1242.

Leutbecher, M, Buizza, R, & Isaksen, L, 2007: Ensemble forecasting and flow-dependent estimates of initial uncertainty. *Proceedings of the ECMWF workshop on Flow-dependent aspects of data assimilation*, ECMWF, 11-13 June 2007, pp. 185-201. Available from ECMWF, Shinfield Park, Reading RG2 9AX, UK.

Molteni F., R. Buizza, T. N. Palmer, & T. Petroliagis. 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Q. J. Roy. Meteor. Soc.*, **122**, 73-119.

NAEFS, the North American Ensemble Forecasting System, see NAEFS web page hosted by NCEP: http://www.emc.ncep.noaa.gov/gmb/ens/NAEFS.html.

Palmer, T. N., Molteni, F., Mureau, R., Buizza, R., Chapelet, P., & Tribbia, J., 1993: Ensemble prediction. Proceedings of the *ECMWF Seminar on Validation of models over Europe: vol. 1*, ECMWF, Shinfield Park, Reading RG2-9AX, UK, 285 pp.

Palmer, T N, Buizza, R., Leutbecher, M., Hagedorn, R., Jung, T., Rodwell, M, Virat, F., Berner, J., Hagel, E., Lawrence, A., Pappenberger, F., Park, Y.-Y., van Bremen, L., Gilmour, I., Smith, L., 2007: The ECMWF Ensemble Prediction System: recent and on-going developments. A paper presented at the 36th Session of the ECMWF Scientific Advisory Committee. *ECMWF Research Department Technical Memorandum n. 540*, ECMWF, Shinfield Park, Reading RG2-9AX, UK.

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

Shutts, G. 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems, *Q. J. Roy. Meteor. Soc.*, **131**, 3079-3100.

THORPEX, THe Observing system Research and Predictability Experiment. See THORPEX web page hosted by WMO: http://www.wmo.ch/thorpex/.

Toth, Z. & E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330.

Toth, Z., & Kalnay, E., 1997: Ensemble Forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Zhang, Z., & Krishnamurti, T.N., 1999: A perturbation method for hurricane ensemble predictions. *Mon. Wea. Rev.*, **127**, 447-469.

Wei, M., Toth, Z., Wobus, R., Zhu, Y., Bishop, C., & Wang, X., 2006: Ensemble Transform Kalman Filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus A*, **58**, 28-44.

Wilks, D. S., 1995: Statistical Methods in the Atmospheric Sciences-An Introduction. International Geophysics Series, Vol.59, Academic Press, 467 pp.