RMetS
Royal Meteorological Society

Verification of global numerical weather forecasting systems in polar regions using TIGGE data

Thomas Jung[a]* and Mio Matsueda[b,c]
[a]*Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany*
[b]*Center for Computational Sciences, University of Tsukuba, Japan*
[c]*Department of Physics, University of Oxford, UK*

*Correspondence to: T. Jung, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bussestrasse 24, D-27570 Bremerhaven, Germany. E-mail: Thomas.Jung@awi.de

High-latitude climate change is expected to increase the demand for reliable weather and environmental forecasts in polar regions. In this study, a quantitative assessment of the skill of state-of-the-art global weather prediction systems in polar regions is given using data from the THORPEX Interactive Grand Global Ensemble (TIGGE) for the period 2006/2007–2012/2013. Forecast skill in the Arctic is comparable to that found in the Northern Hemisphere midlatitudes. However, relative differences in the quality between different forecasting systems appear to be amplified in the Arctic. Furthermore, analysis uncertainty in the Arctic is more of an issue than it is in the midlatitudes, especially when it comes to near-surface parameters over snow- and ice-covered surfaces. Using NOAA's reforecast dataset, it is shown that the changes in forecast skill during the 7-year period considered here can largely be explained by flow-dependent error growth, especially for the more skilful forecasting systems. Finally, a direct comparison between the Arctic and Antarctic suggests that predictions of mid-tropospheric flow in the former region are more skilful.

Key Words: polar prediction; deterministic predictability; probabilistic predictability; flow dependence, analysis uncertainty

Received 26 February 2014; Revised 22 May 2014; Accepted 19 August 2014; Published online in Wiley Online Library 13 October 2014

1. Introduction

Concerns about the amplification of anthropogenic climate change has led to a growing interest in the polar regions in recent years (Emmerson and Lahn, 2012). Furthermore, increased economic and transportation activities in the polar regions are leading to more demands for reliable weather predictions (Jung *et al.*, 2013). Given that *global* forecasting systems are used by most numerical weather prediction centres, forecasts for the polar regions are readily available. However, partly as a result of a strong emphasis of previous work on lower and middle latitudes, relatively little is known about the performance of weather forecasts in polar regions.

To our knowledge, the only study in which the performance of a global weather forecasting system in the Arctic has been investigated more thoroughly has been published by Jung and Leutbecher (2007). Their comparison between different analysis products suggests that synoptic-scale features in the Arctic are relatively well represented by state-of-the-art analysis systems. Furthermore, they show that the improvement in deterministic forecast error for the European Centre for Medium-range Weather Forecasts (ECMWF) forecasting system in the Arctic from the early 1980s to the mid-2000s follows closely that reported in previous studies for the Northern Hemisphere (NH) as a whole. The analysis of the ECMWF Ensemble Prediction System (EPS) reveals substantial medium-range probabilistic forecast skill down to synoptic scales for 500 hPa geopotential height ($Z500$) fields in the polar regions. While providing some insight into the quality of weather forecasts in polar regions, their study was somewhat limited due to a strong focus on the the free atmosphere, the Arctic, the boreal winter season and one particular forecasting system.

More progress has been made in the development and verification of regional prediction systems such as the Antarctic Mesoscale Prediction System (AMPS; Bromwich *et al.*, 2005; Adams, 2004; Powers *et al.*, 2012; Bromwich *et al.*, 2013).

The purpose of this study is to expand on the results by Jung and Leutbecher (2007) in order to provide a more comprehensive assessment of the quality of state-of-the-art *global* weather prediction systems in the polar regions. More specifically, this study aims to address the following questions:

- How much predictive skill do state-of-the-art forecasting systems have in the polar regions?

- How has predictive skill changed over time?
- How does the predictive skill in polar regions compare to that in the midlatitudes?

Here the focus will be on the Arctic and boreal winter. However, results for the other seasons and Antarctica will be mentioned where deemed important. In order to assess the forecast quality of different global forecasting centres in the polar regions, data from the THORPEX* Interactive Grand Global Ensemble (TIGGE; Park *et al.*, 2008; Bougeault *et al.*, 2010; Matsueda and Tanaka, 2008) will be used. Flow-dependence of interannual changes in skill is explored using NOAA's second generation Global Ensemble Forecast System (GEFS) reforecast dataset (Hamill *et al.*, 2013).

The structure of this article is as follows. The data and methods used will be described in the next two sections. This is followed by the results section, which comprises a comprehensive discussion of deterministic and probabilistic forecast skill along with some consideration of analysis uncertainty. Finally, conclusions will be given and discussed.

2. Methods

2.1. Data

The main dataset used in this study comes from TIGGE (Bougeault *et al.*, 2010), which has been developed as part of THORPEX. TIGGE provides operational medium-range ensemble forecast data for non-commercial research purposes through its data portals (http://tigge.ecmwf.int; accessed 29 August 2014).

The operational ensemble prediction systems used here include the Australian Bureau of Meteorology (BoM), the China Meteorological Administration (CMA), the Canadian Meteorological Center (CMC), the Brazil Centro de Previsão de Tempo e Estudos Climáticos (CPTEC), the Japan Meteorological Agency (JMA), the Korea Meteorological Administration (KMA), the US National Centers for Environmental Prediction (NCEP), the United Kingdom Meteorological Office (UKMO) and the ECMWF, as of July 2013.

The operational configurations of the EPSs including the data assimilation system differ among the different systems (Bougeault *et al.*, 2010, give details). ECMWF, JMA and UKMO, for example, use 4D-Var as a data assimilation method for their analysis, whereas the other centres use different data assimilation methods. The horizontal resolutions of forecast models vary from T_L119L19 (160 km horizontal resolution) for BoM to T_L639L62 (30 km horizontal resolution) for ECMWF. CMC, KMA, NCEP, and UKMO adopt an ensemble Kalman filter (EnKF) technique to represent initial uncertainty, whereas the other centres use either the singular vector method (ECMWF and JMA), the breed vector method (CMA), or a method based on Empirical Orthogonal Functions (CPTEC). Model uncertainties are also considered by introducing stochastic perturbations of model physics tendencies (CMC, ECMWF, JMA, NCEP, and UKMO) or by means of multi-parametrizations (CMC). The maximum (minimum) forecast length available is 384 h (216 h) for NCEP (JMA). The smallest (largest) ensemble includes 21 (51) members for NCEP (ECMWF and JMA). JMA currently conducts their medium-range ensemble forecast at 1200 UTC every day, whereas the other NWP centres run their forecasts two to four times daily. In order to ensure comparability, therefore, only the ensemble forecasts initialised at 1200 UTC are used here.

The operational NWP systems included in TIGGE have undergone frequent changes during the period considered in this study. Therefore, year-to-year changes in prediction skill might have their origin in forecast system changes. In order to distinguish flow-dependent predictability from the influence of forecast system changes, NOAA's second generation GEFS reforecast dataset is used as well (Hamill *et al.*, 2013). For the period from December 1985 to 2013, the GEFS reforecast uses the same NWP model, the same initial perturbation method (the ensemble size of 11), and the same data assimilation system (but up to 22 May 2012) as the operational EPS used at NCEP in 2012. NCEP's climate forecast system reanalysis at 0000 UTC is used as initial conditions for the GEFS reforecast. Only the GEFS reforecasts during the TIGGE period are used here.

This study focuses on $Z500$ and 2 m temperature ($T2$m) forecasts during the period from October 2006 to May 2013. The TIGGE forecasts have been verified against either the ERA-Interim reanalysis (Dee *et al.*, 2011) or their own analyses for four different regions: the Arctic (65–90°N), the NH midlatitudes (20–60°N), Southern Hemisphere (SH) midlatitudes (20–60°S), and Antarctica (65–90°S). The operational analysis for each NWP centre was taken from the respective control forecast at initial time. Prior to the verification, the forecast data and ERA-Interim were interpolated onto a common horizontal grid with a spacing of 2.5°. For the GEFS reforecast, its own analysis (that is defined as the control forecast at the initial time of the forecast) was used as verifying analysis, and verification has been carried out on a 1° grid. The climatologies for the TIGGE data and the GEFS reforecast are estimated using the ERA-Interim and own analysis, respectively. ERA-Interim was chosen to compute climatologies since the relative shortness of the TIGGE period (7 years) poses problems when computing statistics.

2.2. Verification scores

Forecast skill is quantified using Anomaly Correlation Coefficients (ACC) and Ranked Probability Skill Scores (RPSS), which are widely used in deterministic and probabilistic forecast verification, respectively (Wilks, 2011).

The ACC is defined by:

$$\mathrm{ACC} = \frac{\sum_i (f_i - c_i)(a_i - c_i)}{\sqrt{\sum_i (f_i - c_i)^2}\sqrt{\sum_i (a_i - c_i)^2}}, \tag{1}$$

where f_i, a_i, and c_i indicate forecast, analysis, and climatology, respectively, and the summation is taken over each verification area. The ACC indicates a pattern correlation between forecast anomaly and analysis anomaly. The ACC has a maximum value of one for a perfect forecast. For upper-air fields such as Z500, a score below 0.6 indicates that a forecast has little useful skill.

The RPSS is a skill score of the Ranked Probability Score (RPS) and defined by:

$$\mathrm{RPSS} = \frac{\mathrm{RPS}_{\mathrm{ref}} - \mathrm{RPS}_{\mathrm{fcst}}}{\mathrm{RPS}_{\mathrm{ref}}} \tag{2}$$

and

$$\mathrm{RPS} = \frac{1}{N}\sum_{i}^{N}\left\{\frac{1}{J-1}\sum_{m=1}^{J}\left(\sum_{j=1}^{m} p_j^i - \sum_{j=1}^{m} o_j^i\right)^2\right\}, \tag{3}$$

where $\mathrm{RPS}_{\mathrm{ref}}$ is for a reference forecast (in this study the climatological forecast), and p_j^i and o_j^i are forecast and observed probabilities, respectively, in the jth ($j = 1, \ldots, J$) climatologically equal category ($J = 10$ in this study) on a grid point i. N denotes the the total number of grid points over the verification area. o_j^i is 1 if an event occurs and 0 if it does not. The RPS is a squared measure that compares the cumulative density function (CDF) of a probabilistic forecast with the CDF of the corresponding observations over given probabilistic categories. The RPS is zero for a perfect forecast and is positive otherwise (Weigel *et al.*, 2007). The RPSS has a maximum value of unity for a perfect forecast, and a value of 0 for a probabilistic forecast comparable in skill to a climatological forecast.

*THe Observing system Research and Predictability EXperiment

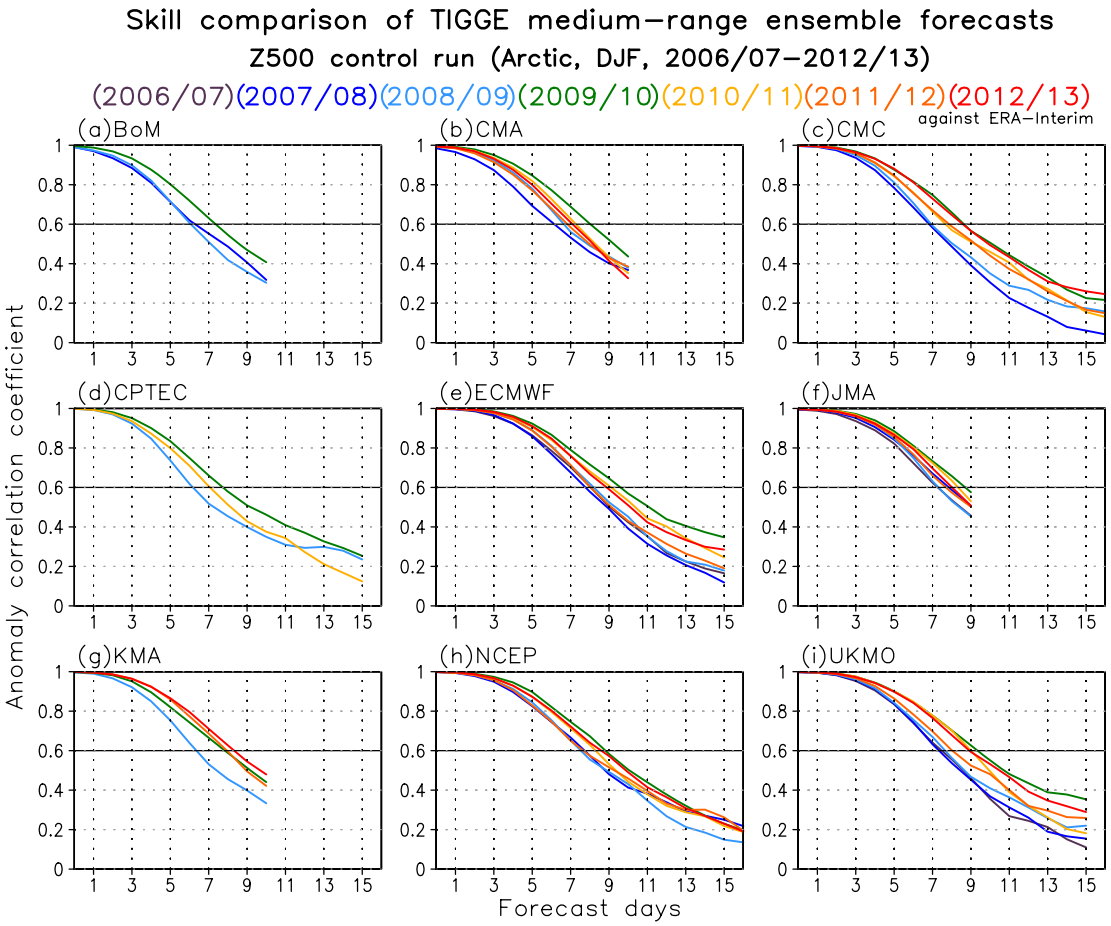

Figure 1. Anomaly correlation coefficient of 500 hPa geopotential height control forecasts for the Arctic (north of 65°N) and winters (December–February) of the years 2006/2007–2012/2013 (different colours) for nine different global forecasting systems: (a) BoM, (b) CMA, (c) CMC, (d) CPTEC, (e) ECMWF, (f) JMA, (g) KMA, (h) NCEP, and (i) UKMO. Forecasts were verified against ERA-Interim reanalysis data.

3. Results

3.1. *Verification of 500 hPa geopotential height*

3.1.1. *Deterministic scores*

The anomaly correlation coefficient of *Z*500 control forecasts in the Arctic verified against ERA-Interim data is shown in Figure 1 for the different TIGGE models and the winters from 2006/2007 to 2012/2013. The average deterministic predictive skill–as measured by the forecast lead time at which ACC drops below a value of 0.6 –ranges between 6 and 9.5 days. The ECMWF model appears to be performing best with UKMO, NCEP, JMA and CMC following closely. It cannot be excluded that ECMWF performs best simply because ERA-Interim data is used for verification. However, the fact that all models perform similarly well in the early short range (with ACC close to one) suggests that *Z*500 fields in the Arctic are relatively well constrained by the observations and therefore the verifying analysis actually used plays a secondary role.

Each of the forecasting systems shows sizeable year-to-year variability in deterministic forecast skill for *Z*500 during winter in the Arctic (Figure 1). This raises the question as to whether the differences in forecast skill are due to forecasting system improvements (e.g. Simmons and Hollingsworth, 2002; Jung, 2005) or flow-dependent forecast error growth (e.g. Ferranti *et al.*, 2002; Jung and Leutbecher, 2007). In order to address this question, NOAA GEFS reforecasts with a frozen forecasting system were evaluated in the same way as the TIGGE data (Figure 2). The winter of 2009/2010 turns out to be the most predictable one for all lead times during the period considered. This is consistent with the very persistent negative phase of the Arctic Oscillation during this winter (Jung *et al.*, 2011). The winters of 2007/2008 and 2008/2009 turn out to be the least predictable (two days less than in 2009/2010). A very similar result is found for all TIGGE models, which suggests that the year-to-year differences in deterministic forecast skill of *Z*500 in the Arctic shown in Figure 1 is primarily due to flow-dependent perturbation growth rather than forecasting system development. This notion is also consistent with the fact that all different TIGGE models show the same winters to be more (or less) predictable.

It is worth putting the results for the Arctic into perspective by comparing with the deterministic forecast skill for *Z*500 in the much better studied NH midlatitudes (Figure 3). The first thing to notice is that the year-to-year variability of deterministic skill is much lower in the midlatitudes than in the Arctic. The standard deviation of interannual ACC variability of 9-day ECMWF forecasts for the period 2006/2007–2012/2013, for example, is twice as large in the Arctic as it is in the NH midlatitudes. One possible explanation for this finding is that the midlatitude belt (20–60°N) represents a larger area where different flow regimes may occur simultaneously. Furthermore it turns out that the skill in the most predictable winters is comparable in the midlatitudes and the Arctic; the major difference lies in fewer midlatitude winters with relatively poor forecast skill. Overall, this translates into a slightly higher average level of predictive skill for *Z*500 in the midlatitudes than in the Arctic (see also Jung and Leutbecher, 2007). Interestingly, winters in the Arctic which tend to be more (less) predictable show also a higher (lower) level of predictive skill in the midlatitudes. This suggest that the skill is influenced by the same atmospheric circulation regimes. This notion is consistent with the fact that the relatively predictable winter of 2009/2010 was charaterised by the predomindance of the negative phase of the AO (Jung *et al.*, 2011), which influences both the Arctic and the midlatitudes.

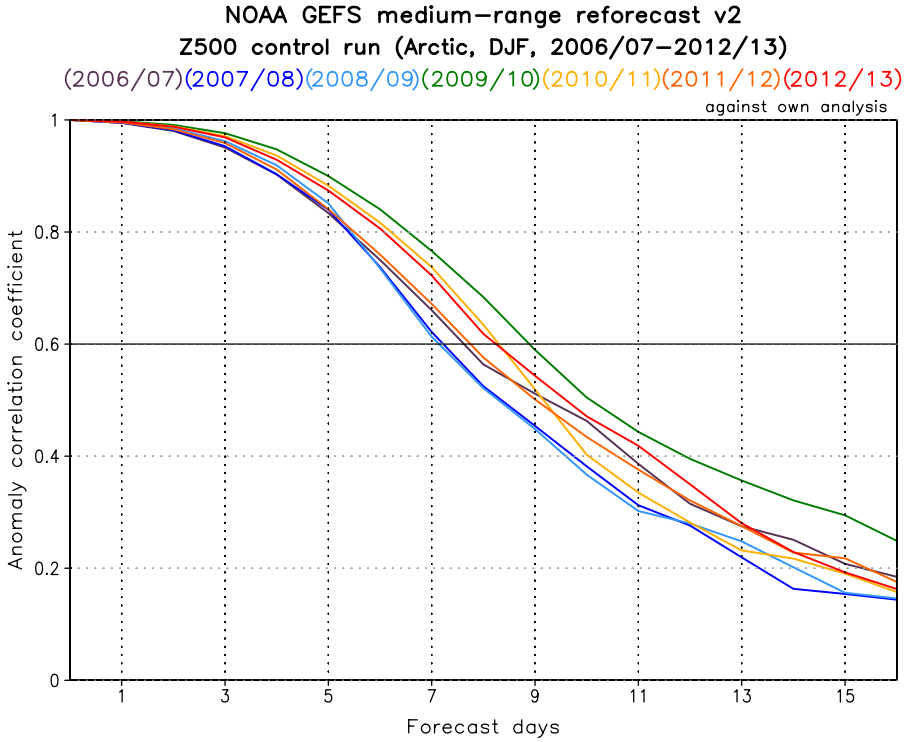

Figure 2. Anomaly correlation coefficients of 500 hPa geopotential height control forecasts from NOAA's GEFS reforecast system (frozen forecasting system) over the Arctic (north of 65°N) for winters (December–February) of the years 2006/2007–2012/2013. Reforecast data were verified against their own reanalysis.

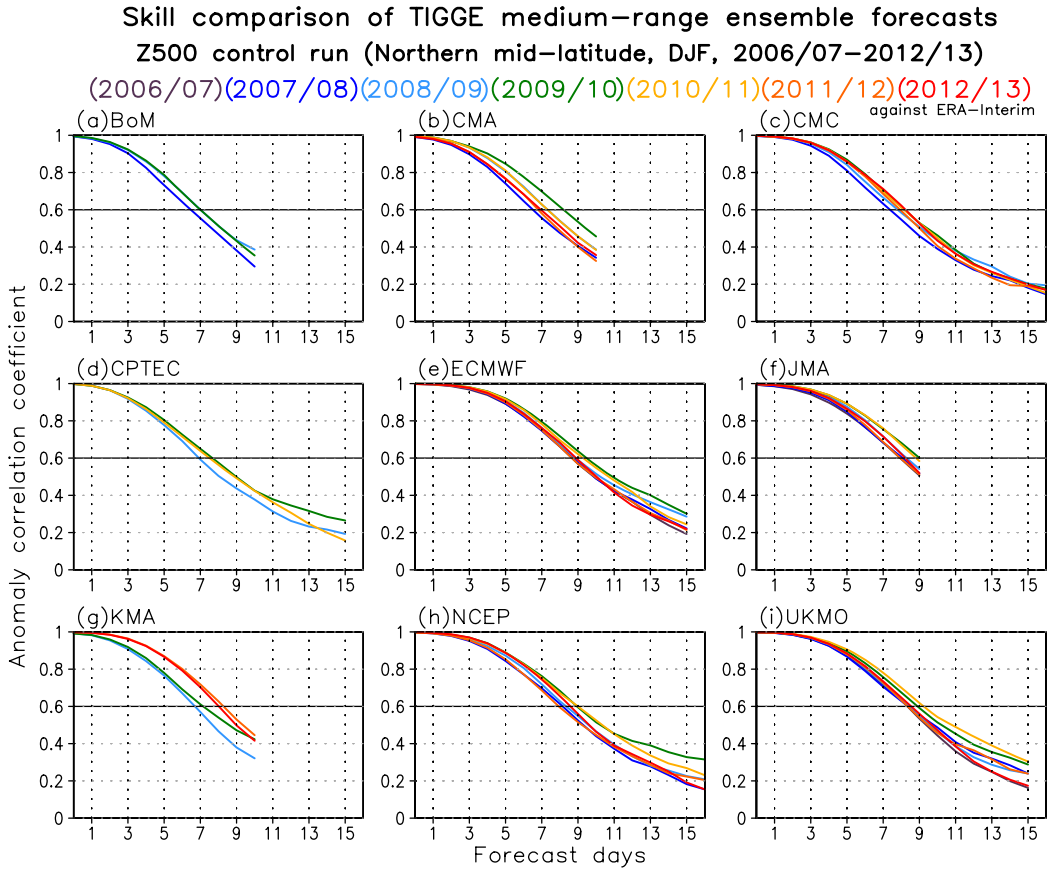

Figure 3. As Figure 1, but for the Northern Hemisphere midlatitudes (20–60°N).

3.1.2. *Probabilistic scores*

After having discussed the deterministic skill of *Z*500 forecasts, in the following the performance of ensemble forecasts will be assessed in a probabilistic framework.

Ranked Probability Skill Scores for *Z*500 ensemble forecasts over the Arctic are shown in Figure 4 for winters of the period 2006/2007–20012/2013 and different TIGGE systems. The first thing to notice is the large difference in performance between the different ensemble prediction systems; whereas the best systems show RPSS of about 0.6 at day 5, the worst systems lie consistently below 0.4 for the same lead time. In fact, differences in the performance of the various systems appear to be larger for probabilistic than for deterministic forecasts. These differences

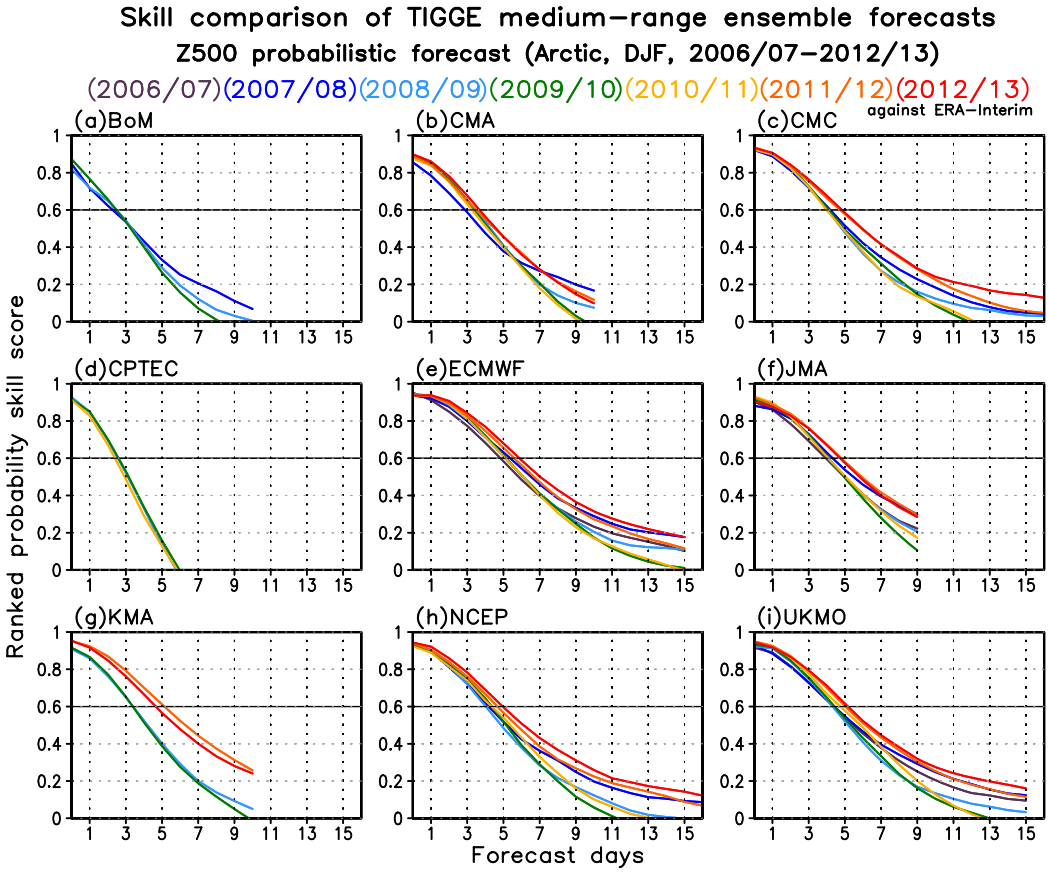

Figure 4. As Figure 1, but showing Ranked Probability Skill Score of 500 hPa geopotential height ensemble forecasts.

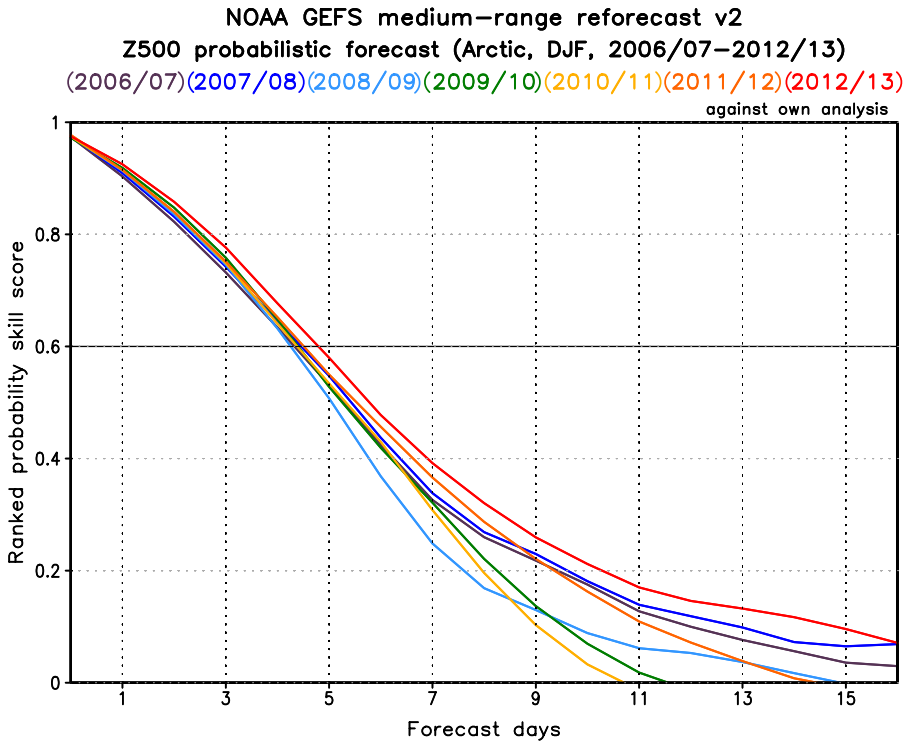

Figure 5. As Figure 2, but showing Ranked Probability Skill Score of ensemble reforecasts (frozen forecasting system) from NOAA's GEFS.

are presumably due to difference in the quality of the methods used for representing initial and model uncertainties. It is also possible that the climatological standard deviation from ERA-Interim is more representative for some systems than it may be for others.

Like for the deterministic scores, large year-to-year variability in predictive skill is found. The fact that temporal (year-to-year) evolution of the skill of probabilistic $Z500$ forecasts in the Arctic is similar for different forecasting systems suggest that flow-dependent perturbation growth is the main cause of interannual changes in probabilistic skill. This notion is confirmed by Figure 5, which shows a very similar time dependency for the frozen NOAA GEFS system.

Interestingly, the flow-dependence of probabilistic skill is different from that of deterministic skill (cf. Figures 5 and 2). From a deterministic perspective, the winter 2009/2010, for example, looks relatively predictable; however, from a probabilistic

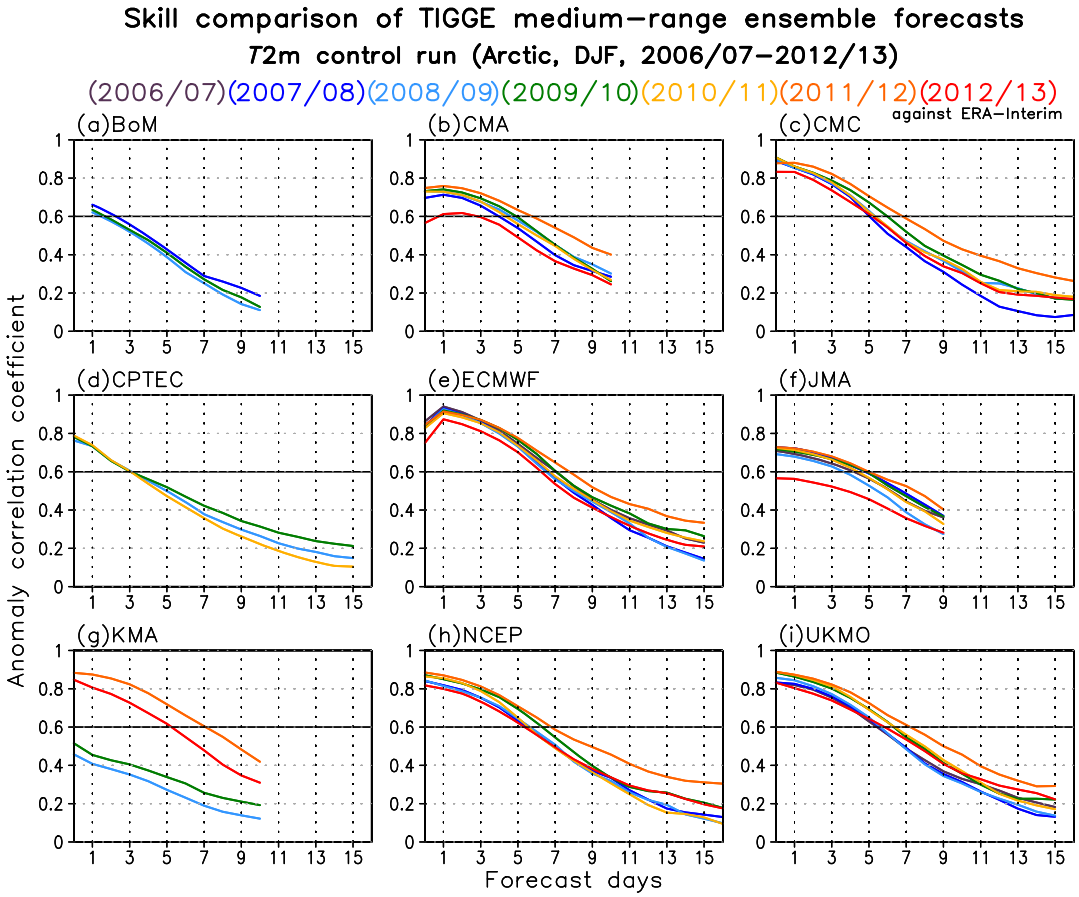

Figure 6. As Figure 1, but for 2 m temperature control forecasts.

perspective it is not. One possible explanation is that for strongly negative AO anomalies, ACC provides quite different answers from scores that take the magnitude of the error into account (Langland and Maue, 2012) such as RMS and RPPS. Therefore, difference between ACC and RPSS suggest that models had problems in predicting the correct amplitude of the flow anomaly, especially during the winter of 2009/2010.

The largest changes in RPSS (and also ACC; Figure 1) are found for KMA during the period considered here. Given the relatively large size of this change compared to what can be expected from flow-dependence, it seems plausible that probabilistic forecasts with the KMA system have undergone genuine improvements in recent years. This is consistent with the fact that KMA introduced the UKMO system into operations in 2010. (Differences in skill between KMA and UKMO after 2010 can be explored in more detail at the TIGGE Museum: http://gpvjma.ccs.hpcc.jp/TIGGE/index.html; accessed 29 August 2014.)

In general, the RPPS for probabilistic $Z500$ forecasts does not seem to be overly sensitive to the analysis used. This interpretation is supported be the fact that the EPS from CMC, ECMWF, NCEP and UKMO show very similar RPSS despite the fact that all were verified against non-native ERA-Interim data. In this context, note that even the ECMWF system, which is closer to ERA-Interim than any of the other systems, shows RPSS values very similar to CMC, NCEP and UKMO during the first 24 h of the forecast.

3.2. *Verification of 2 m temperatures*

In the following, results from the verification of $T2m$ forecasts will be described. This parameter is much more relevant from a user perspective and it describes boundary-layer aspects which tend to be decoupled from the free atmosphere during winter under very stable conditions over sea ice and snow.

The ACC of $T2m$ forecasts of the different TIGGE systems over the Arctic are shown Figure 6. The relatively poor skill of some model systems in the short range when verified against ERA-Interim (Figure 6) compared to their own analysis (Figure 7) shows that $T2m$ analysis fields are relatively poorly constrained by the observations in the Arctic leaving them especially prone to systematic model error. In the medium range, after about 5 days into the forecast when errors have had time to grow, the analysis used for verification plays a smaller role, at least for the EPS from CMC, ECMWF, NCEP and UKMO.

Generally, the deterministic skill of $T2m$ forecasts is lower than those for $Z500$ in the short range and medium range (cf. Figures 6 and 1). This might be explained by the fact that $Z500$ is more strongly influenced by rather predictable planetary waves and synoptic systems, whereas more unpredictable boundary-layer processes impact on $T2m$ as well. On the other hand, the skill of $T2m$, albeit small, becomes comparable to that for $Z500$ from about day 10. One possible explanation for this feature is that the lower boundary conditions (including sea ice and snow) start to provide a source of skill for $T2m$ (much more so than for $Z500$). However, it is also possible that systematic $T2m$ differences between the forecast model and the ERA-Interim climatology play an increasingly important role for longer lead times.

Deterministic errors of $T2m$ forecasts in the NH midlatitudes, expressed in terms of ACC, are comparable to those in the Arctic (not shown).

3.3. *Analysis uncertainty*

In the previous section it was suggested that analysis uncertainty plays an important role when it comes to verifying $T2m$ forecasts. In the following, therefore, $T2m$ analysis uncertainty will be explored in some more detail.

The spatial structure of the $T2m$ analysis uncertainty for different times of the year, expressed in terms of the mean of the

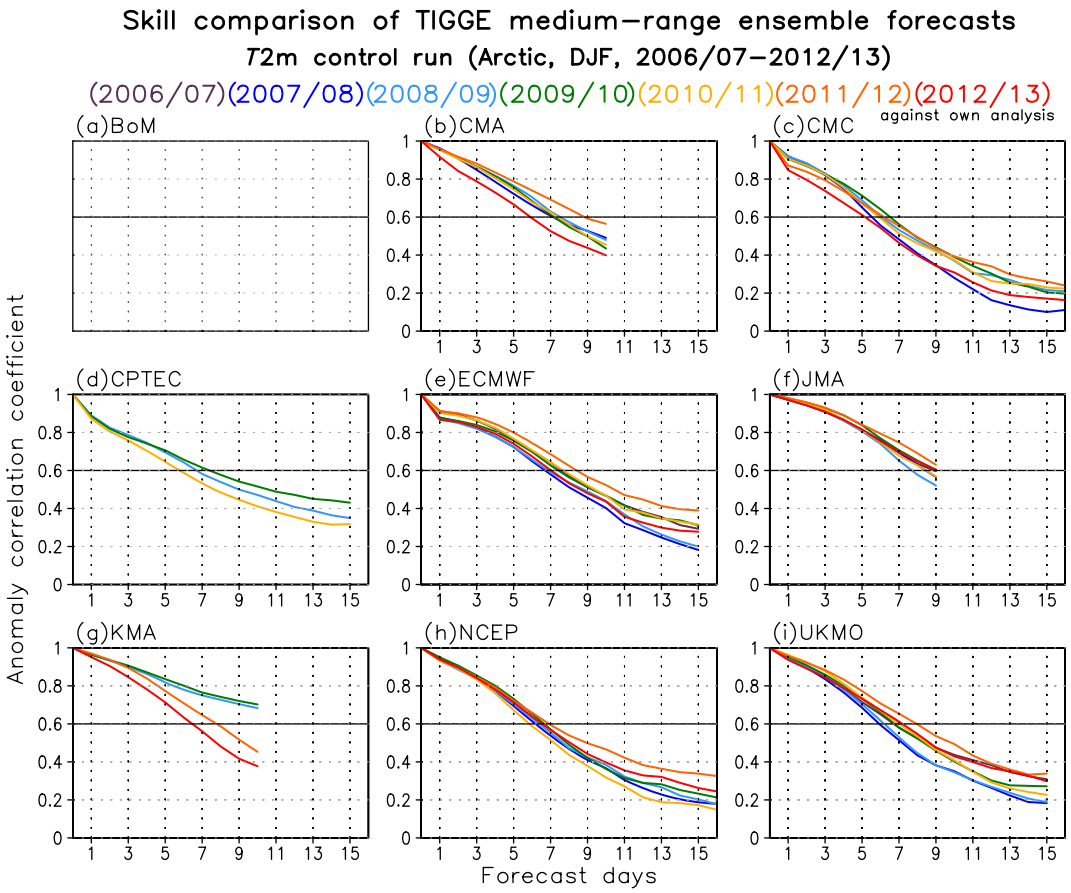

Figure 7. As Figure 6, but for verification against their own analysis data. BoM is not included because analysis fields for 2 m temperature are not available from the TIGGE archive.

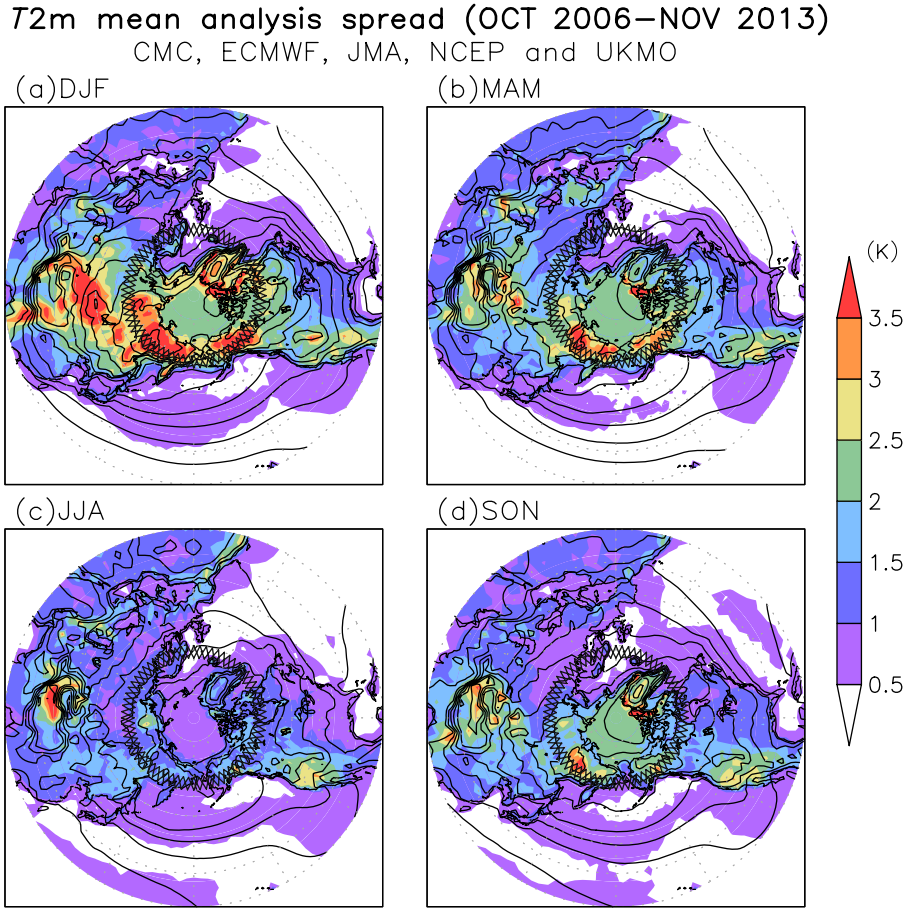

Figure 8. Analysis uncertainty for *T*2m over the Northern Hemisphere for (a) December–February, (b) March–May, (c) June–August, and (d) September–November during the period from October 2006 to November 2013, measured in terms of the mean of the daily standard deviation for operational analyses from five leading NWP centres: CMC, ECMWF, JMA, NCEP, and UKMO. The hatching in each panel shows the latitude belt 60–65°N, indicating the boundary between the midlatitudes and the Arctic as used in this study.

daily standard deviation of different TIGGE analysis products, can be inferred from Figure 8. The largest $T2$m uncertainty, which can amount up to 2–4 K, can be found over NH land regions and over the Arctic Ocean. More specifically, the annual cycle suggests that analysis uncertainty is largest over sea ice and snow. There is also enhanced $T2$m analysis uncertainty along the Arctic coast in Russia and Alaska during the boreal winter and spring. Overall this picture is consistent with the fact that areas covered by snow and ice are generally rather poorly observed and that there are large uncertainties associated with the parametrization of stable planetary boundary layers (Holtslag *et al.*, 2013).

Interestingly, the uncertainty of $T2$m analysis fields is relatively small over the Arctic Ocean during boreal summer (Figure 8(c)). Here it is hypothesized that this has mostly to do with that fact that near-surface temperature in summer is forced to stay relatively closely to the melting temperature of ice.

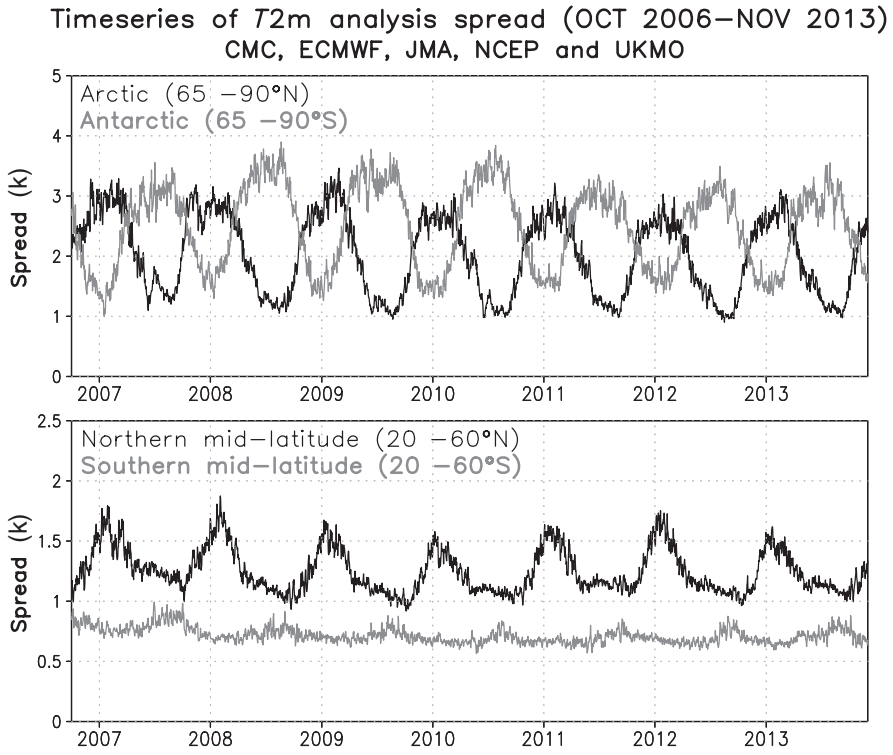

Figure 9. Time series of daily analysis uncertainty for $T2$m over (a) the polar regions and (b) midlatitudes during the period from October 2006 to November 2013, measured by spread among operational analyses from CMC, ECMWF, JMA, NCEP, and UKMO.

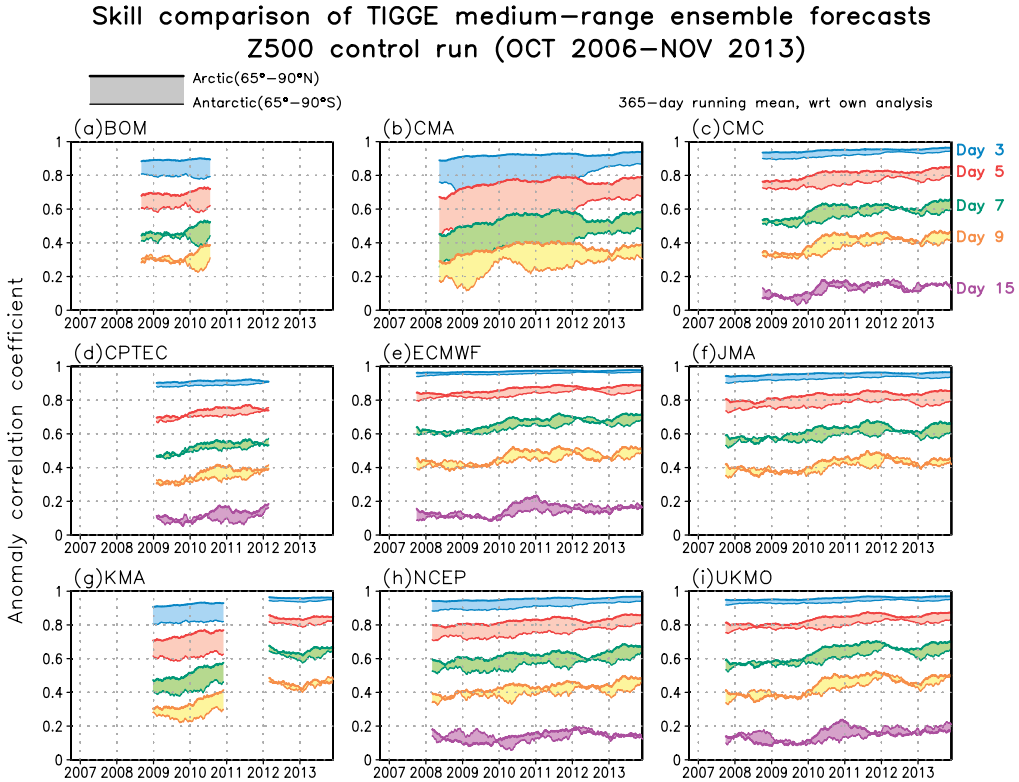

Figure 10. Smoothed time series of daily anomaly correlation coefficients of 500 hPa geopotential height control forecasts over the Arctic (north of 65°N, bold lines) and the Antarctic (south of 65°S, thin lines) at lead times of 3, 5, 7, 9, and 15 days during the period from October 2006 to November 2013 for the nine global forecasting systems. The forecasts were verified against their own analyses. Smoothing is applied using 365-day running means.

Time series of the $T2$m analysis uncertainty for the period October 2006 to November 2013 is shown in Figure 9 for different regions over the NH and SH. The strong annual cycle in $T2$m analysis uncertainty described above is also evident for Antarctica, with slighty larger values than the Arctic during all seasons. For both polar regions, uncertainty plateaus are found during winter and summer with rapid changes during autumn and spring.

For the period considered here, there is no strong evidence for a reduction in analysis uncertainty of $T2$m, neither for the Arctic nor the Antarctic.

For the NH, $T2$m analysis uncertainty is larger in the polar regions than in midlatitudes during all seasons except summer. For the SH analysis, uncertainty is always smaller in midlatitudes, albeit less pronounced during austral summer. Finally, the SH midlatitudes show a much weaker annual cycle in $T2$m analysis uncertainty than the NH. One possible explanation for this differences lies in the fact that $T2$m is strongly constrained by sea surface temperature and that a larger fraction of oceanic areas in the SH is covered by oceans.

4. Conclusions and discussion

In this study, data from different global forecasting systems, which contribute to the TIGGE database, have been used to assess the deterministic and probabilistic skill of state-of-the-art global weather forecasting systems in the polar regions.

It turns out that the forecast skill for parameters such as $Z500$ and $T2$m in the Arctic is comparable to that found in the NH midlatitudes. However, relative differences between the quality of different forecasting systems appear to be amplified in the Arctic. Furthermore, analysis uncertainty in the Arctic is much more of an issue than it is in the midlatitudes, especially when it comes to near-surface parameters over snow- and ice-covered surfaces.

For the 7-year period considered here (2006/2007–2012/2013), most of the changes in forecast skill can be explained by flow-dependent growth of forecast error. However, there are some notable exceptions such as associated with the implementation of the UKMO system at KMA in 2010. Our results highlight the importance of analysing data from reforecast datasets as well, in order to draw meaningful conclusions when it comes to interpreting year-to-year changes in forecast skill.

In order to keep this article conclusive but at the same time concise, the emphasis has been put on the Arctic. However, it seems worthwhile also to provide at least some insight into the differences between the Arctic and Antarctic. Time series of daily anomaly correlation coefficients of $Z500$ control forecasts over the Arctic and the Antarctic are shown in Figure 10 for various lead times. Smoothing has been accomplished by employing a 365-day running mean filter. The most obvious thing to notice is that deterministic $Z500$ forecasts for the Arctic are slightly more skilful than for the Antarctic. Furthermore, there is no obvious evidence for a change in this difference of the 7-year period considered here.

This study can be seen as an important contribution to a much larger effort that will be necessary to develop a comprehensive understanding of the performance of numerical prediction systems in the polar regions. Further work will be required to verify more user-relevant parameters such near-surface winds, temperature and sea ice drift using a wider range of appropriate verification techniques.

Acknowledgement

We are indebted to our numerous colleagues who made the TIGGE project happen.

References

Adams N. 2004. A Numerical Modeling Study of the Weather in East Antarctica and the Surrounding Southern Ocean. *Weather Forecasting* **19**: 653–672.

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

Bromwich DH, Monaghan AJ, Manning KW, Powers JG. 2005. Real-time forecasting for the Antarctic: An evaluation of the Antarctic Mesoscale Prediction System (AMPS). *Mon. Weather Rev.* **133**: 579–603.

Bromwich DH, Otieno F, Hines K, Manning KW, Shilo E. 2013. Comprehensive evaluation of polar weather research and forecasting performance in the Antarctic. *J. Geophys. Res.* **118**: 274–292.

Emmerson C, Lahn G. 2012. 'Arctic opening: Opportunity and risk in the high north'. Chatham House–Lloyd's Risk Insight Report. http://www.chathamhouse.org/publications/papers/view/182839 (accessed 29 August 2014).

Ferranti L, Klinker E, Hollingsworth A, Hoskins BJ. 2002. Diagnosis of systematic forecast errors dependent on flow pattern. *Q. J. R. Meteorol. Soc.* **128**: 1623–1640.

Hamill TM, Bates GT, Whitaker JS, Murray DR, Fiorino M, Galarneau TJ Jr, Zhu Y, Lapenta W. 2013. NOAA's second-generation global medium-range ensemble reforecast data set's second-generation global medium-range ensemble reforecast dataset. *Bull. Am. Meteorol. Soc.* **94**: 1553–1565.

Holtslag A, Svensson G, Baas P, Basu S, Beare B, Beljaars A, Bosveld F, Cuxart J, Lindvall J, Steeneveld G, Tjernström M, Van De Wiel B. 2013. Stable atmospheric boundary layers and diurnal cycles: Challenges for weather and climate models. *Bull. Am. Meteorol. Soc.* **94**: 1691–1706.

Jung T. 2005. Systematic errors of the atmospheric circulation in the ECMWF forecasting system. *Q. J. R. Meteorol. Soc.* **131**: 1045–1073.

Jung T, Leutbecher M. 2007. Performance of the ECMWF forecasting system in the Arctic during winter. *Q. J. R. Meteorol. Soc.* **133**: 1327–1340.

Jung T, Vitart F, Ferranti L, Morcrette J-J. 2011. Origin and predictability of the extreme negative NAO winter of 2009/10. *Geophys. Res. Lett.* **38**: L07701, doi: 10.1029/2011GL046786.

Jung T, Gordon N, Klebe S, Bauer P, Bromwich DH, Day J, Doblas-Reyes F, Fairall C, Hines K, Holland M, Iversen T, Lemke P, Mills B, Nurmi P, Renfrew I, Smith G, Svensson G, Tolstykh M. 2013. 'WWRP Polar Prediction Project implementation plan'. WWRP/PPP No. 2. http://polarprediction.net/en/documents/ (accessed 29 August 2014).

Langland RH, Maue RN. 2012. Recent northern hemisphere mid-latitude medium-range deterministic forecast skill. *Tellus A* **64**: 17531.

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

Park Y-Y, Buizza R, Leutbecher M. 2008. TIGGE: Preliminary results on comparing and combining ensembles. *Q. J. R. Meteorol. Soc.* **134**: 2029–2050.

Powers J, Manning KW, Bromwich DH, Cassano J, Cayette A. 2012. A decade of Antarctic science support through AMPS. *Bull. Am. Meteorol. Soc.* **93**: 1699–1712.

Simmons AJ, Hollingsworth A. 2002. Some aspects of the improvement of skill of numerical weather prediction. *Q. J. R. Meteorol. Soc.* **128**: 647–677.

Weigel A, Liniger M, Appenzeller C. 2007. The discrete Brier and ranked probability skill scores. *Mon. Weather Rev.* **135**: 118–124.

Wilks DS. 2011. *Statistical Methods in the Atmospheric Sciences*. Elsevier: Amsterdam.