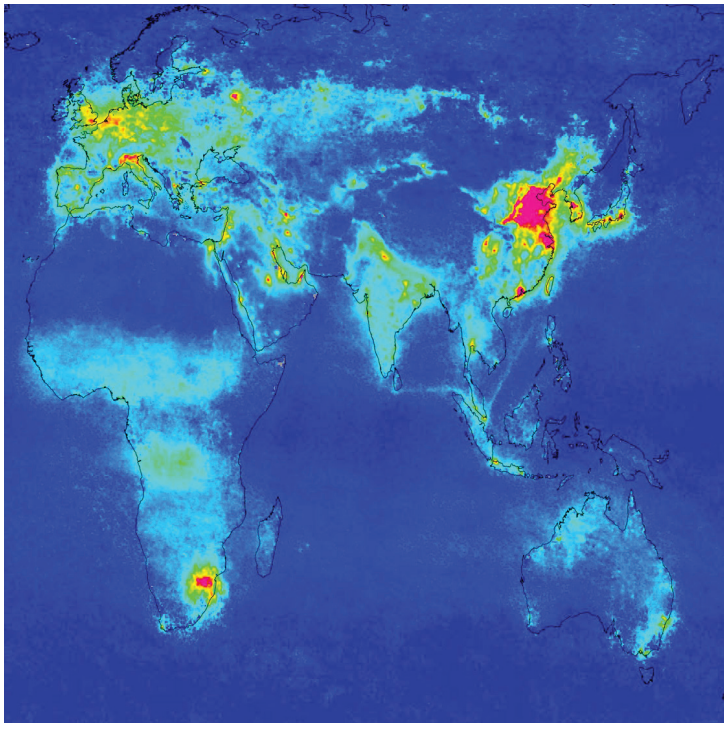
ECMWF Feature article

from Newsletter Number 138 – Winter 2013/14

METEOROLOGY

Have ECMWF monthly forecasts been improving?



NO2 pollution map courtesy of KNMI/IASB/ESA

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

doi:10.21957/z8ma9pkw

This article appeared in the Meteorology section of ECMWF Newsletter No. 138 – Winter 2013/14, pp. 18–23.

Have ECMWF monthly forecasts been improving?

Frédéric Vitart, Franco Molteni, Roberto Buizza

Monthly forecasts (32-day forecasts) have been produced routinely at ECMWF since March 2002, and operationally since October 2004. In the current configuration, the monthly forecasts are generated by extending the 15-day ensemble integrations to 32 days twice a week (at 00 UTC on Mondays and Thursdays). Forecasts are based on the medium-range/monthly ensemble forecast (ENS) which is part of ECMWF's Integrated Forecasting System. ENS includes 51 members run with a horizontal resolution of T639 (about 32 km) up to forecast day 10, and T319 (about 65 km) thereafter. Initial perturbations are generated using a combination of singular vectors and perturbations generated using the ECMWF ensemble of data assimilations, and model uncertainties are simulated using two stochastic schemes.

The climatology (re-forecasts) used to calibrate the real-time forecasts is computed using the re-forecast suite that includes only 5 members of 32-day integrations with the same configuration as the real-time forecasts, starting on the same day and month as the real-time forecast over the past 20 years. The re-forecasts are created a couple of weeks before the corresponding real-time forecast. This strategy for re-forecasts is different to the one used for seasonal forecasting where the model version is frozen for a few years and the re-forecasts are created only once.

An extract of the results published in a recent article (*Vitart*, 2013) is presented hereafter. They document that, on average, the skill of monthly forecasts for weeks 2 to 4 has significantly improved over the past decade.

Monthly forecast skill: how do we measure it?

The skill of the monthly forecasts is routinely evaluated by scoring the 51-member real-time forecasts, mainly against analyses, using a range of measures. For instance, Figure 1 shows skill scores of 2-metre temperature anomalies based on all the real-time forecasts since October 2004, when the monthly forecasts became operational. The skill score is the area under the Relative Operating Characteristic (ROC), which is a measure of the capability of the monthly forecasts to discriminate between occurrence and non-occurrence of events (in this case the event is '2-metre temperature anomaly above the upper tercile of the climatological distribution'). With this measure 1 indicates a perfect forecast and 0 a forecast with the same skill as climatology.

Figure 1 shows a drop of skill with increased time range as expected. For the 12–18 day forecast, the ROC area exceeds 0.7 over large portions of the northern extra-tropics. One week later (i.e. the 19–25 day forecast), the northern extra-tropics still display some skill in predicting 2-metre temperature anomalies, but the highest skills scores are in the tropics. At days 26–32, the skill in the northern extra-tropics is low, although larger than climatology, while in the tropics the skill is still positive. An issue with this type of verification is that it mixes forecasts which have been produced using different versions of the IFS since 2004.

A methodology for evaluating the evolution of the monthly forecast skill scores over the past 10 years could be to compute the skill scores of the real-time forecasts for each season or each year. Figure 2 shows an example of the evolution of the ROC area of 2-metre temperature for days 12–18 since winter 2004. It can be seen that the forecasts at this time range consistently outperform persistence of the previous week's forecast (i.e. using the forecast for days 5–11). However, a major issue with this methodology is that the monthly forecast skill scores are strongly dependant on the large-scale circulation that was predominant during a season. For instance, Figure 2 shows that the skill of the monthly forecasts has decreased since winter 2009–2010. However, the winter 2009–2010 was exceptionally predictable (e.g. *Jung et al*., 2011) with a persistent strong negative North Atlantic Oscillation (NAO) pattern. It is likely that the higher skill score in 2010 is due to this exceptional condition rather than to a degradation in the model performance after 2010. Low frequency variability associated with El Niño Southern Oscillation (ENSO) events can also impact the skill scores for the extended-range forecasts in the tropics and extra-tropics. This makes it difficult to identify trends from a time series of skill scores of real-time forecasts.

Another option for assessing the evolution of the monthly forecast skill scores based on the monthly re-forecasts is discussed in details in the next section.

 doi:10.21957/z8ma9pkw

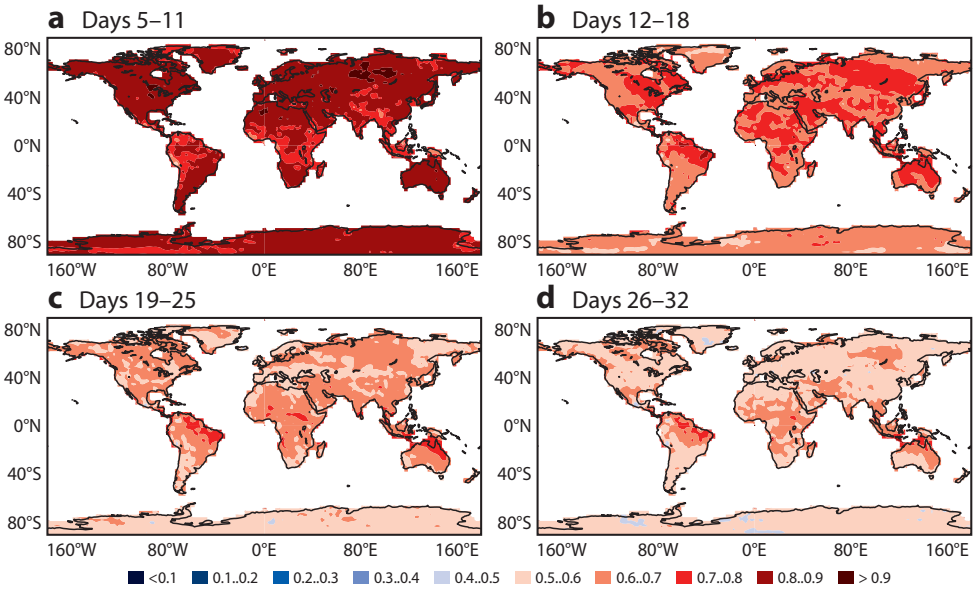


Figure 1 Area under the Relative Operating Characteristic (ROC) curve for the probabilistic prediction of 2-metre temperature anomalies in the upper tercile for weekly periods: (a) days 5–11, (b) days 12–18, (c) days 19–25 and (d) days 26–32. This plot has been produced using all the real-time monthly forecasts since October 2004.

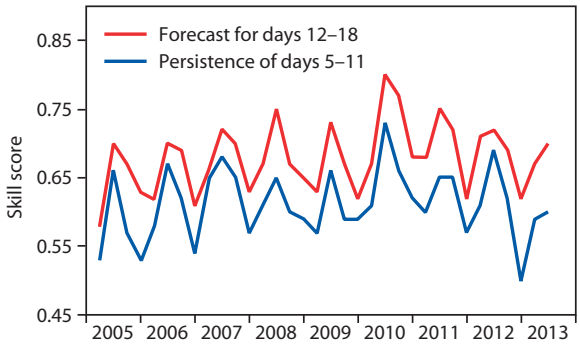


Figure 2 Evolution of the skill scores of the real-time forecast and the corresponding persistence forecast based on probabilities of the previous week The skill score is the area under the Relative Operating Characteristic (ROC) curve for the probabilistic prediction of 2-metre temperature anomalies in the upper tercile over the northern extra-tropics (north of 30°N) calculated for each season since winter 2004.

	March 2002	October 2004	February 2006	March 2008	January 2010	Novem-ber 2011	June 2012
Frequency	Every 2 weeks	Once a week					Twice a week*
Horizontal resolution	100 km days 0-32			50 km days 0-10 80 km days10-32	30 km days 0-20 60 km days 10-32		
Vertical resolution	40 levels Top at 10 hPa	62 levels Top at 5 hPa					
Ocean/Atmosphere coupling	Every hour from day 0			Every 3 hours from day 10			
Re-forecast period	Past 12 years			Past 18 years			Past 20 years
Re-forecast size	5 members						
Initial conditions	ERA-40			ERA-Interim			

* Only for real-time forecasts. The frequency of re-forecasts is still once a week.

Figure 3 Evolution of the main changes in the ECMWF monthly re-forecasts since 2002.

doi:10.21957/z8ma9pkw

Methodology for assessing the evolution of the skill scores in the re-forecasts

As shown in Figure 3, the number of re-forecast years has been changing since 2002, but all the re-forecasts since 2002 have the period 1995–2001 in common. The starting days of the re-forecasts may vary from one year to another, but this should not have a significant impact on the skill scores averaged over a complete year or a season. The scores can be compared for re-forecasts covering the same years and seasons: i.e. all the re-forecasts from 1995 to 2001 that were produced each year between April of a given year until March of the following year. For instance, the scores of 2006 will refer to the scores of all the re-forecasts from 1995 to 2001 that were produced between April 2006 and March 2007 (4 April, 11 April, 18 April..... 27 March 1995–2001) using the IFS versions that were operational between April 2006 and March 2007.

Please note that averages have been computed over a period starting from April to March of the following year to ensure a consistency in the model versions used for a complete winter and a complete summer.

As mentioned above, an advantage of this methodology is that it ensures that all the re-forecasts cover the same seasons and years. There are however two weaknesses of this approach that is worth mentioning – these are associated with differences in ensemble size and changes to the model. For more detail see Box A.

Despite weaknesses in the methodology, the approach taken provides a valuable assessment of the time evolution of monthly forecasts' scores for various aspects of the ECMWF monthly forecasts, as will be shown.

Weaknesses of the methodology used to assess changes in skill scores with time **A**

Firstly, the ensemble size of the re-forecasts is limited to only 5 members instead of 51 members for the real-time forecasts. This can be an issue when considering probabilistic forecasts of rare events for which skill scores are very sensitive to ensemble size. *Weigel et al.* (2008) faced the same issue when they scored the ECMWF re-forecasts produced in 2006 and used a correction of the probabilistic skill score which takes into account the ensemble size. It is worth noting that this is less an issue for the present study than that carried out by *Weigel et al.* because the goal here is to assess the evolution of the monthly forecast skill scores during the period 2002–2012 rather than evaluate the skill of the monthly forecasting system.

Secondly, the model may have changed more than once during the period that has been used to compute the skill scores. This makes the attribution of the variation of skill scores to a specific change in the model physics more difficult. An alternative would be to run a large set of re-forecasts covering the same period with the various versions of the IFS model. But this is too expensive to be done systematically and impossible to be done for old versions of IFS which are no longer supported in the current ECMWF operating systems. Apart from the change of the reanalysis from ERA-40 to ERA-Interim in March 2008, all the re-forecasts have been initialised from the same dataset. Therefore this verification will not take into account possible improvements due to changes in the ECMWF data assimilation from 2002 to 2012.

Evolution of average skill over the northern extra-tropics

The classical ranked probability skill score (RPSS) is a measure of the degree to which a forecast outperforms a reference forecast, in this case climatology. However, a disadvantage of the RPSS is its strong negative bias for small ensemble size. Therefore, a de-biased version of the RPSS, the so-called discrete ranked probability skill score (*Weigel et al*., 2008), has been used to assess the skill evolution of the re-forecasts of 2-metre temperature anomalies produced since 2002. This measure has the advantage of being insensitive to the unreliability due to small ensemble sizes.

Figure 4 displays the evolution of the discrete RPSS of 2-metre weekly-average temperature anomalies since 2002, for three forecast weekly periods: days 12–18, days 19–25 and days 26–32. Though there is a drop in the probabilistic skill score between days 12–18 and days 19–25, the monthly forecasts still display better skill than climatology (positive RPSS). These results also suggests that there have been improvements in the RPSS scores of 2-metre temperature anomaly re-forecasts over the northern extra-tropics for all three time ranges (days 12–18, days 19–25 and days 26–32) since 2002. The values of the discrete RPSS for days 26–32, although still very low, are now close to the values for the previous week (days 19–25) re-forecasts that were produced in 2002. The skill scores of days 19–25 have also improved in time almost linearly and get close to the skill scores of days 12–18 in the early years of the ECMWF monthly forecasts.

doi:10.21957/z8ma9pkw

According to Figure 4, the forecasts for weeks 3 and 4 are now more skilful than 10 years ago and therefore the current monthly forecasts are more likely to produce useful early warnings of cold or heat waves. For instance, Figure 5 shows the prediction at various time ranges of 2-metre temperature anomalies during the cold wave over Europe in March 2013. It is impressive that the 32-day ensemble forecasts predicted cold anomalies over Europe three weeks in advance. Figure 6 shows an example of prediction of a summer heat wave in Southern Europe.

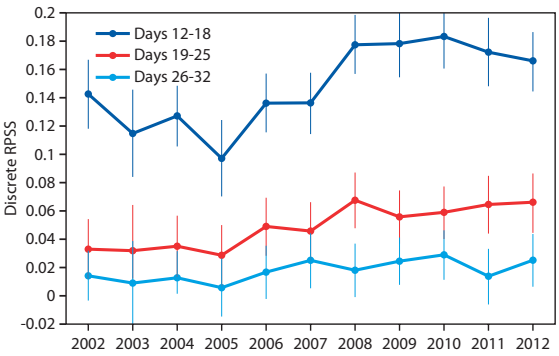


Figure 4 Evolution of the discrete ranked probability skill score (RPSS) of 2-metre temperature weekly mean anomalies over the northern extra-tropics (north of 30°N) since 2002 for days 12–18, days 19–25 and days 26–32. Only land points have been scored. The RPSS has been computed from terciles and for all the ECMWF re-forecasts for the extended boreal winter (October to March).

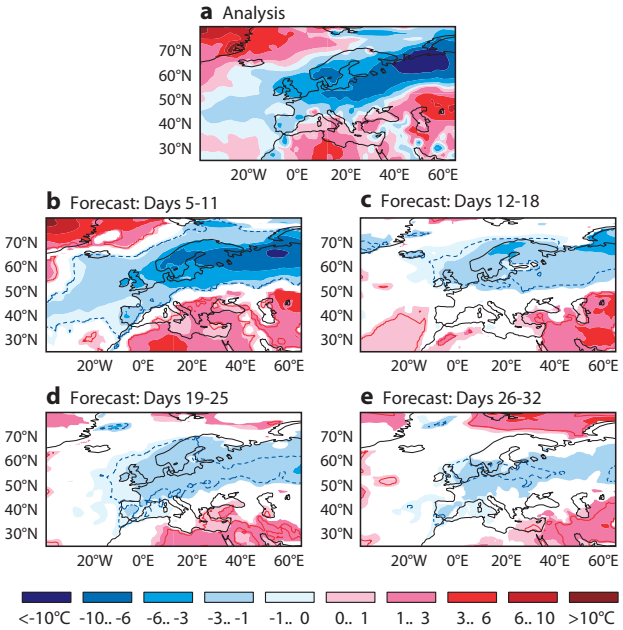


Figure 5 Weekly mean 2-metre temperature anomaly ensemble mean forecasts verifying on the 18–24 March 2013 for the time ranges days 5–11, days 12–18, days 19–25 and days 26–32. The top panel shows the verification computed from ERA-Interim.

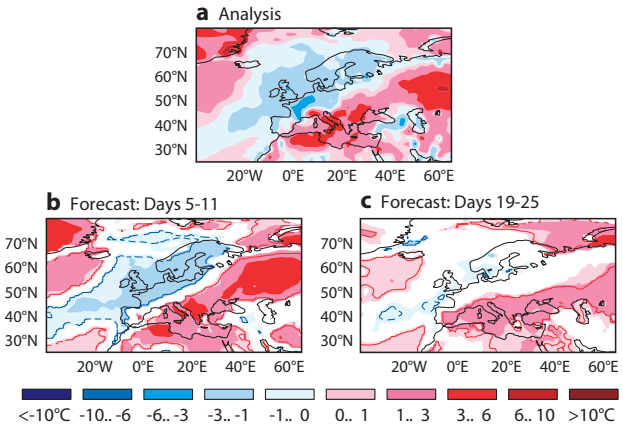


Figure 6 Weekly mean 2-metre temperature anomaly ensemble mean forecasts verifying on the 9–15 July 2012 (top panel) for the time ranges days 5–11, days 12–18, days 19–25 and days 26–32. The top panel shows the verification computed from ERA-Interim.

doi:10.21957/z8ma9pkw

Evolution of the predictive skill of the Madden-Julian Oscillation (MJO)

The Madden-Julian oscillation (MJO) is a main source of predictability in the tropics on time scales exceeding one week but less than a season (*Madden & Julian*, 1971). It is characterised by an eastward propagation of convective rainfall from the Indian Ocean to the western Pacific.

For convenience the MJO has been split into eight phases starting with enhanced rainfall over the western Indian Ocean which moves slowly eastwards across the Indian Ocean (phases 2 and 3). The rainfall then crosses the 'maritime continent' of Indonesia and surrounding countries (phases 4 and 5) and arrives in the western Pacific before dying out in the central Pacific (phases 6 and 7). The MJO then continues its eastward propagation in the upper atmosphere over the western hemisphere and Africa (phases 8 and 1). Typically an MJO event lasts between 30 and 60 days.

The Wheeler and Hendon index (WHI, see *Wheeler & Hendon*, 2004) has been applied to all the model re-forecasts and to ERA-Interim over the period 1995–2001 to evaluate the skill of the monthly forecasting system in predicting MJO events and to produce composites for the eight phases of the MJO.

Principal Component Analysis is a method of identifying the characteristic spatial patterns of data set by a much smaller number of 'new' variables. It identifies the underlying structure of the data and extracts the principal components that account for most of the variation in the data. For the MJO the two principle components (PC1 and PC2) are such that:

- Enhanced convection occurs over the maritime continent when PC1 is positive and over the western hemisphere and Africa when PC1 is negative.
- Enhanced convection occurs over the Pacific Ocean when PC2 is positive and over the Indian Ocean when PC2 is negative.

To evaluate the skill of the monthly forecasting system to predict the MJO, a linear bivariate correlation is performed between the time series of PC1 and PC2 from the forecast ensemble-mean time series for different lead times and the corresponding time series computed from ERA-Interim.

Figure 7 shows the evolution of the MJO bivariate correlation skill score from 2002 until 2012 between the ensemble mean re-forecasts and ERA-Interim. In this figure, the three lines show the forecast day in which the bivariate correlation reached 0.5, 0.6 and 0.8. If we consider the MJO bivariate correlation of 0.6 as a limit of MJO prediction skill, the ECMWF monthly forecasting system displayed skill to predict the MJO up to about 15 days in 2002. In 2012, the limit of 0.6 was reached around day 25, suggesting an averaged gain of about 1 day of lead-time per year. The bivariate correlation of 0.5 is now reached beyond day 30 instead of day 22 in 2002. For the bivariate correlation of 0.8, the gain has been of about 5 days over the 10-year period. The difference of MJO skill scores between 2002 and 2012 is statistically significant for the three thresholds (bivariate correlations of 0.5, 0.6 and 0.8) within the 5% level of confidence.

The evolution of the amplitude error of the MJO, calculated from each individual ensemble member and then averaged, does not display an improvement as regular as for the forecast skill scores. According to Figure 8, forecasts produced a too weak MJO in the early years of the monthly forecasting system, with the amplitude about 30% too low beyond forecast day 20. There has been a clear improvement between 2006 and 2008. In 2008, when Cy32r3 was used operationally, the MJO was even slightly too strong. Since 2008, the amplitude of the MJO displays a trend towards weaker MJOs, with amplitudes in the recent years only about 10% weaker than in the ERA-Interim analyses.

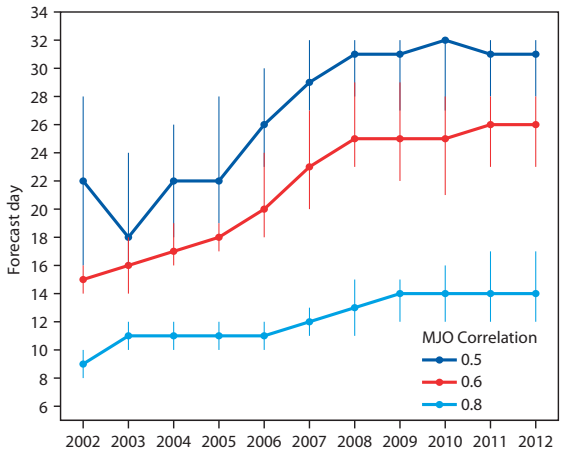


Figure 7 Evolution of the MJO skill scores (bivariate correlations applied to the WHI) since 2002 as indicated by the days when the MJO bivariate correlation reaches 0.5, 0.6 and 0.8. The MJO skill scores have been computed on the ensemble mean of the ECMWF re-forecasts produced during a complete year. The vertical bars represent the 95% confidence interval computed using a 10,000 bootstrap re-sampling procedure.

doi:10.21957/z8ma9pkw

Using reanalysis data, *Cassou* (2008) showed that there is a link between the MJO and North Atlantic Oscillation (NAO). The probability of a positive phase of the NAO (i.e. the difference of atmospheric pressure at sea level between the Icelandic low and the Azores high) is significantly increased about 10 days after the MJO is in Phase 3 (Phase 3 + 10 days), and significantly decreased about 10 days after the MJO is in Phase 6 (Phase 6 + 10 days). The probability of a negative phase of the NAO is decreased (increased) about 10 days after the MJO is in Phase 3 (Phase 6).

Let us now focus on evaluating whether the MJO teleconnections on the northern extra-tropics have improved by comparing the re-forecasts produced each year from 2002 until 2012 with ERA-Interim. This is based on the 500 hPa geopotential height composites 10 days after an MJO in Phase 3 with an amplitude larger than a standard deviation. Only the re-forecasts covering the extended boreal winter season are considered (from October to March).

According to Figure 9, the MJO teleconnections (10 days after an MJO in Phase 3) are more realistic over the northern extra-tropics in 2011 (middle panel) than in 2002 (left panel) compared to ERA-Interim (right panel). The re-forecasts produced in 2011 simulate a stronger positive NAO anomaly than in 2002. However, the impact of the MJO on the NAO is still underestimated in the 2011 re-forecasts compared to ERA-Interim. On the other hand, the ECMWF forecasting system overestimates the positive 500 hPa geopotential anomaly over the northern Pacific. The same conclusions are valid for the composites of 500 hPa geopotential height 10 days after an MJO in Phase 6 (not shown). The improved MJO teleconnections are likely to impact the monthly forecast skill scores in the northern extra-tropics, and in particular the skill of the model to predict the NAO.

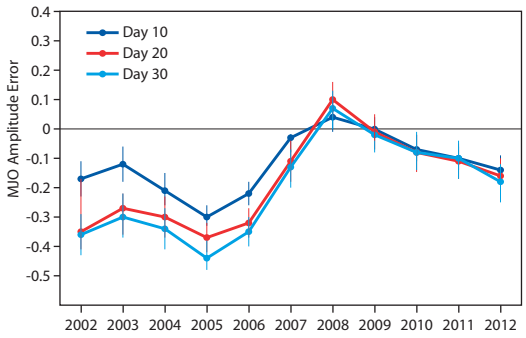


Figure 8 Amplitude error of the re-forecasts relative to the mean MJO amplitude obtained from ERA-Interim analyses. Negative (positive) numbers in the top panel indicate that the MJO simulated by IFS is weaker (stronger) than in the ECMWF reanalysis. The vertical bars represent the 95% confidence interval computed using a 10,000 bootstrap re-sampling procedure.

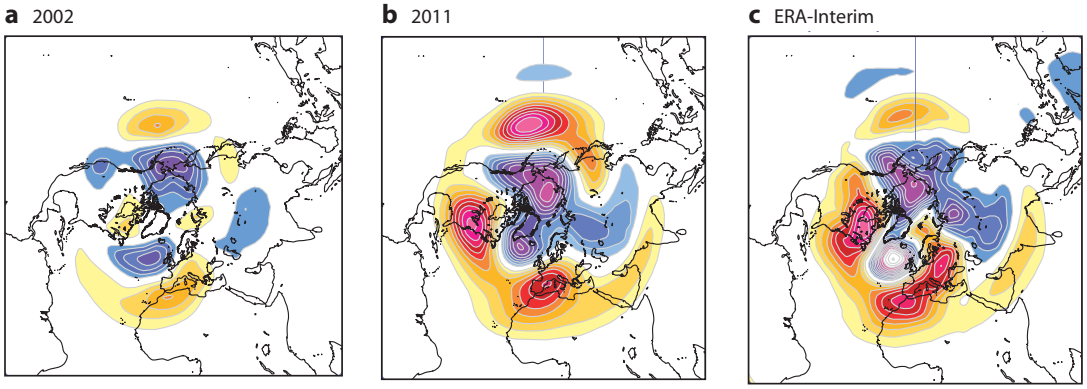


Figure 9 MJO Phase 3 10-day lagged composites of 500 hPa geopotential height anomaly over the northern extra-tropics for all the October to April re-forecasts that were produced in (a) 2002, (b) 2011 and (c) ERA-Interim. Red and orange colours indicate positive anomalies. Blue colours indicate negative anomalies. The lowest contour is at 10 metres and the contour interval is 5 metres.

Evolution of the predictive skill of the North Atlantic Oscillation (NAO)

The prediction of the NAO is of particular importance for the prediction of European weather. An NAO index has been constructed by projecting the daily 500 hPa height anomalies over the northern hemisphere onto a pre-defined NAO pattern based on an EOF (Empirical Orthogonal Function) analysis – a technique used to study possible spatial patterns of variability and how they change with time. The NAO pattern was defined as the first leading mode of EOF applied to the reanalysis of monthly mean 500 hPa height during the 1950–2000 period produced by NCEP (National Centers for Environmental Prediction). NAO skill scores have been produced for each year from 2002 until 2012 by applying the NAO index to the re-forecasts and to ERA-Interim, and by computing the linear correlation between the ensemble-mean re-forecasts and ERA-Interim.

Let us focus on extended winter cases (from October to March). Figure 10 shows that there has been improvement in the prediction of the daily values of the NAO with a gain of about 4 days of lead time for a correlation of 0.5, 3 days for a correlation of 0.6 and 2 days for a correlation of 0.8. As for the MJO, the improvement in the prediction of the NAO cannot be attributed to a single change of the ECMWF forecasting system. The difference of NAO skill scores between 2002 and 2011 are statistically significant within the 5% level of confidence.

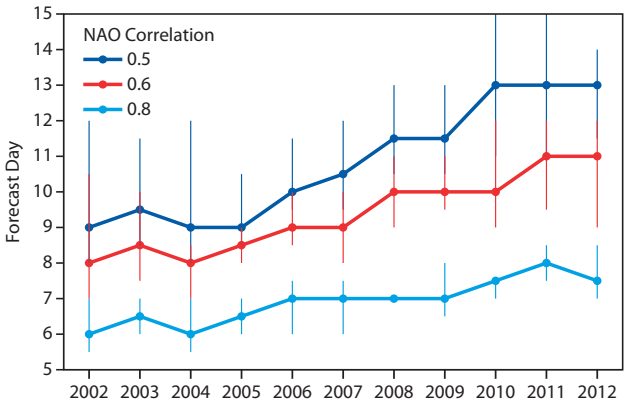


Figure 10 Evolution of daily NAO skill scores since 2002 as indicated by the days when the NAO index correlation reaches 0.5, 0.6 and 0.8. The daily NAO skill scores (correlations applied to the NAO index) have been computed on the ensemble mean of the ECMWF re-forecasts produced from October to March 1995–2001 and ERA-Interim. The vertical bars represent the 95% confidence interval computed using a 10,000 bootstrap re-sampling procedure.

Concluding, have ECMWF monthly forecast been improving?

This study has shown that the skill of the ECMWF monthly forecasts has improved since 2002, the time when ECMWF started producing monthly forecasts. The improvements in the skill scores are particularly high for the prediction of the Madden-Julian Oscillation (MJO), an important source of predictability at the sub-seasonal time scale. Over the northern extra-tropics, the prediction skill of the North Atlantic Oscillation (NAO) and of 2-metre temperature anomalies have also increased, particularly for days 12–18. *Vitart* (2013) shows that a large portion of the improvements in the NAO skill scores can be attributed to the improvements in the prediction of the MJO. For 2-metre temperature, the skill of the 19–25 day forecast in 2012 is getting closer to the skill that the 12–18 day forecast had in 2002. Similar improvements are visible at upper levels, for example in the prediction of the NAO pattern.

The improvements in the monthly re-forecast skill scores reported in this study are likely to be an underestimation of the improvements in the real-time forecasts since this study does not take into account improvements in the generation of atmospheric initial conditions, except for the change from ERA-40 to ERA-Interim in 2008. These improvements are due to a combination of model improvements, better initial conditions (associated with better data assimilation schemes, model improvements and the use of new observing systems), and improvements in the design of more reliable ensemble systems (e.g. thanks to improvements in the simulation of model uncertainties).

Recent changes of the ECMWF medium-range/monthly ensemble forecast (ENS) will help to further increase sub-seasonal forecast skill. In November 2013, three major configuration changes have been implemented affecting the ENS: the atmospheric model is coupled to a new version of the ocean model and from day 0 instead of from day 10, land-surface initial conditions are perturbed, and the vertical resolution has been increased with the introduction of 91 instead of 62 vertical levels and the rise of the top of the atmosphere from 5 to 0.01 hPa (model cycle Cy40r1). Future changes will include the implementation of a sea-ice model instead of persisting sea-ice and of a higher-resolution, ¼° ocean model instead of the current 1° model.

doi:10.21957/z8ma9pkw

It is also worth mentioning that, as part of ECMWF's contribution to the Weather Research Programme (WWRP) and World Climate Research Program 'Sub-seasonal to Seasonal prediction' (S2S) project (http://www.wmo.int/pages/prog/arep/wwrp/new/S2S_project_main_page.html), ECMWF will extend its existing TIGGE (Thorpex Interactive Grand Global Ensemble experiment) archive to include sub-seasonal forecasts from other operational centres. This initiative will provide a unique uniform archive of S2S forecasts that will help scientists and developers to understand the sources of S2S predictability, and make it possible to compare the performance of monthly forecasts of different systems.

Further reading

Cassou, **C.**, 2008: Intraseasonal interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation. *Nature*, doi:10.1038/nature07286.

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

Madden, R.A. & **P.R. Julian,** 1971: Detection of a 40-50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci*., **5,** 702–708.

Vitart, F., 2013: Evolution of ECMWF sub-seasonal forecast skill scores. *Q. J. R. Meteorol. Soc*., in press.

Weigel, A., D. Baggenstos, M.A. Liniger, F. Vitart & **C. Appenzeller,** 2008: Probabilistic verification of monthly temperature forecasts. *Mon. Weather Rev*., **136,** 5162–5182.

Wheeler, M.C. & H.H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Weather Rev*., **132,** 1917–1932.

doi:10.21957/z8ma9pkw