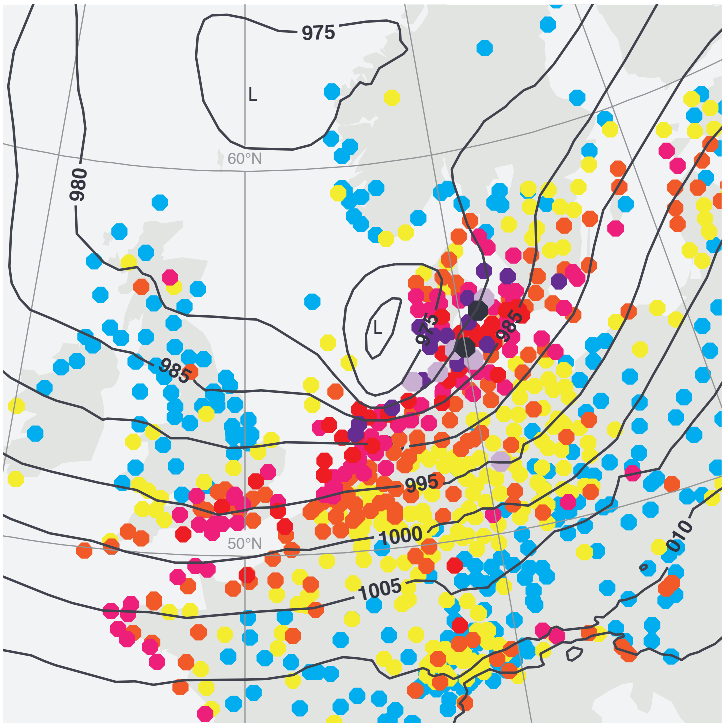
ECMWF Feature article

from Newsletter Number 137 – Autumn 2013

METEOROLOGY

An evaluation of recent performance of ECMWF's forecasts



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

doi:10.21957/hi1eektr

This article appeared in the Meteorology section of ECMWF Newsletter No. 137 – Autumn 2013, pp. 15–18.

An evaluation of recent performance of ECMWF's forecasts

Alan Thorpe, Peter Bauer, Linus Magnusson, David Richardson

Since the end of 2009 the anomaly correlation coefficient (*ACC*) measure of the skill of ECMWF's high-resolution forecasts (HRES) for the northern hemisphere extra-tropics has been essentially constant – that is, not showing an increase in *ACC* values for over 3 years. A number of new model cycles have been introduced in this period. Pre-operational testing of the new cycles demonstrated increases in skill resulting from the improvements to the data assimilation, model resolution and representation of physical processes, as regularly reported in the *ECMWF Newsletter* and on the ECMWF website. Why, then, does the operational HRES skill not reflect these improvements? If the interpretation of this period – hereafter referred to as a pause – were to be that the skill of the ECMWF HRES has reached a plateau then this is not in accord with our scientific expectations! Consequently the pause is worthy of some investigation and comment.

As well as changes to the forecasting system, there are other factors that influence forecast skill, including the potential predictability of the atmosphere and the characteristics of the particular skill measure being used. In this short contribution we consider the role of these factors in explaining the apparent pause in skill improvement.

Interpreting the ACC time series

As shown in Figure 1, the HRES pause in *ACC* over the last three years was preceded by a year (calendar year 2009) when the *ACC* increased rapidly. Note that there were three new model cycles introduced during 2009 and then in January 2010 the decrease in horizontal mesh size to 16 km was introduced. As the usual *ACC* scores are presented in the form of 12-month centred running averages some of the apparent increase in skill during the second half of 2009 was associated with the resolution change early in 2010.

It is important to place the period since 2009 into the context of the longer-term trends for HRES since, say, 2001. (Note that longer-term changes in the scores over the past 30 years are shown in *Janoušek et al*., 2012 and contributing factors are discussed in *Magnusson & Källén*, 2013.) The time series from 2001 to early 2013 given in Figure 1 appears to follow a trend of an *ACC* increase of about 7.5% per decade but with two significant 'anomalies' over that period: the first from mid-2006 to mid-2007 and the second more significant one from mid-2009 to mid-2011. In both these anomalies the *ACC* was well above the trend line. In fact the monthly values show that the winters 2009/10 and 2010/11 had particularly high values. It is difficult from this information alone, until more time has elapsed, to distinguish between the 'pause' picture and the 'upward trend plus anomalies' picture of the last three years.

An illuminating way to try to isolate the effect of the potential predictability of the atmosphere is by examining reforecasts using a fixed forecasting system so that the model developments do not contribute to skill changes. At ECMWF the ERA-Interim forecast system is used for this purpose. It utilises the Integrated Forecasting System (IFS) version used in operations from December 2006 to June 2007 although the horizontal resolution used is lower (80 km) than that of the then current operational model (25 km).

If one looks at *ACC* in Figure 1 for ERA-Interim there is a trend since 2001 of about 1.5% increase per decade that is most likely due to an improvement in the observational network. But superimposed on that trend are anomalies in the two periods mentioned previously for the HRES *ACC*, with the mid-2009 to mid-2011 period being highly anomalous and large. An interpretation of this is that during that period the atmosphere exhibited high potential predictability which was realised in the higher-than-trend *ACC* scores of HRES. These higher *ACC* scores in effect produced the pause by starting the three-year period of the pause with an (anomalous) high; see also *Andersson & Richardson* (2011) for discussion of high values of ACC during 2010. Between mid-2011 and mid-2012 the ERA-Interim *ACC* values dropped, alongside the apparent pause in HRES progress, which could be interpreted as a decrease of potential predictability. The difference in *ACC* between HRES and ERA-Interim shown in Figure 1 has increased in this period, indicating the improvements of the HRES relative to the fixed ERA-Interim system.

doi:10.21957/hi1eektr

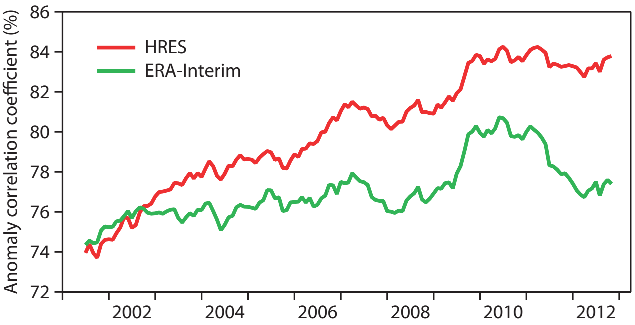


Figure 1 Time series since 2001 of the ACC for the 500 hPa geopotential height in the northern hemisphere extra-tropics at day 6 for HRES and ERA-Interim forecasts. A 12-month centred running-mean has been used.

The period 2007 to 2012 has also been examined by *Langland & Maue* (2012) who draw similar conclusions about the link between skill and atmospheric variability, based on *ACC* values for ECMWF and other centres. They highlight that there is a good correlation between the *ACC* skill variations and the AO (Arctic Oscillation) index with higher than normal *ACC* values being associated with a strong negative phase of the AO (i.e. pressure is high in the Arctic and low in mid-latitudes). This situation occurred in the second anomalous period referred to here.

The relationship between RMSE and ACC

It is sometimes said that the *ACC* favours large anomalies relative to the climatology as much as (or more than) actual forecast skill improvements. However, the root-mean-square error (*RMSE*) does not depend on the size of climatological anomalies but it does depend on the accuracy of the overall level of 'activity' that a given forecasting system exhibits; in other words systems that are relatively inactive (say compared to observed analysis variations) can produce smaller *RMSE* values irrespective of their real forecast skill. Consequently it is important to consider both *ACC* and *RMSE* when evaluating the performance of an NWP system.

The definitions of *RMSE* and *ACC*, and a way to mathematically relate the two using the concept of the atmospheric activity, are shown in Box A, following the approach in the '*ECMWF User Guide'*, Appendix A-2 (*Persson*, 2013). The *RMSE* does not explicitly depend on climatology (see equation (A1)) as it refers only to the contemporaneous values of the forecast and analysis; it is perhaps the simplest measure of forecast skill. Unlike *RMSE*, the *ACC* depends explicitly on climatology (see equation (A2)). In the operational verification at ECMWF, the centred version of *ACC* (in which the area-mean value is subtracted from each term) is used. For fields where the bias can be considered low, the difference between centred and un-centred *ACC* is small.

The activity of the model can be defined as the standard deviation of the difference between the forecast and climate (i.e. the forecast anomaly). For the atmosphere the activity depends upon the corresponding difference between the analysis and climate (i.e. the analysis anomaly). If the forecasting system is such that the activity of the forecast is similar to that of the atmosphere on average, the activity can be thought of as a measure that relates to the atmospheric state and not the modelling system. For a poorer forecasting system the forecast anomalies may be too low relative to the analysis anomalies, that is the model is under-active. From the definitions of *RMSE* and *ACC* (equations (A3) and (A4)), one can infer that whilst *ACC* values are largely unaffected by such under-activity the *RMSE* values will be reduced making forecasts appear to be more accurate.

Now assume that the activity in the model is the same as that in the atmosphere – this will be referred to as *ACT* (see equation (A5)). The *ACC* can now be specified in terms of *RMSE* and *ACT* (see equation (A6) which is duplicated below).

$$ACC \approx 1 - \frac{RMSE^2}{2ACT^2}$$

From this equation one can see that if the activity level was approximately constant then lower/higher values of *RMSE* go along with higher/lower values of *ACC*. However, *ACC*, *RMSE* and *ACT* vary in time so it is helpful to differentiate the above equation with respect to time so that the linkage between these time variations can be seen.

$$\frac{\mathrm{d}ACC}{\mathrm{d}t} \approx \frac{RMSE^2}{ACT^3}\frac{\mathrm{d}ACT}{\mathrm{d}t} - \frac{RMSE}{ACT^2}\frac{\mathrm{d}RMSE}{\mathrm{d}t}$$

From this tendency equation one can see that increases/decreases of *ACT* contribute in the same way as decreases/increases in *RMSE* to the overall change in the *ACC*. In principle, more complex behaviour is also possible. For example, during a period when the activity level is increasing and, say, the rate of change of *RMSE* is relatively small, an increasing *ACC* is likely to be seen. This may be why it is sometimes said colloquially that 'ACC likes big anomalies'. In general if *RMSE* were to be constant, greater/lesser levels of activity imply greater/lesser values of *ACC*.

RMSE – a different story

In Figure 2 we show the values of *RMSE* at 500 hPa for HRES at day 6. Now, if we examine the *ACC* pause from the end of 2009 we see that the *RMSE* values have largely continued to fall, except perhaps for the last 6 months. From an *RMSE* viewpoint one would not characterise the last 3 years as being a pause in the progress of forecast accuracy. In fact during the period from summer 2011 to now the monthly *RMSE* values have been noticeably lower than corresponding months in previous years. In this context it is interesting to note that *Langland & Maue* (2012) point out that an examination of the RMSE measure of HRES skill shows that during these periods of negative AO index (and high *ACC* values) the forecast skill, as measured by monthly values of the *RMSE*, was actually lower than at other times.

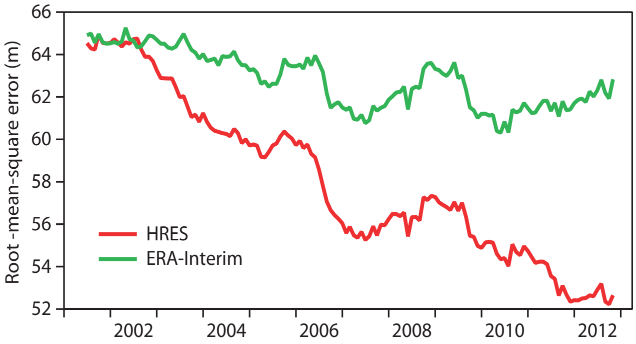


Figure 2 Time series of RMSE for the 500 hPa geopotential height in the northern hemisphere extra-tropics at day 6 for HRES and ERA-Interim. A 12-month centred running-mean has been used.

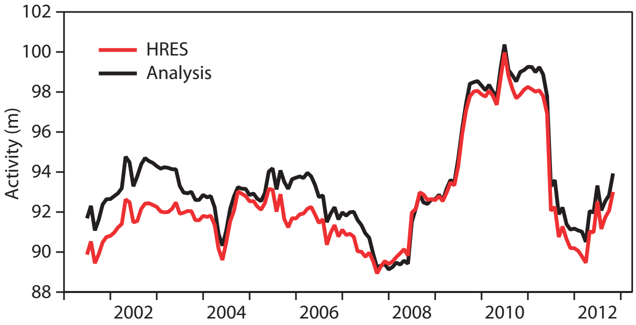


Figure 3 Activity for the analysis and forecast at day 6 in terms of the standard deviation of the 500 hPa geopotential height anomalies in the northern hemisphere extra-tropics. A 12-month centred running-mean has been used.

To interpret the variations of *RMSE* and *ACC* more quantitatively using the tendency equations we first need to consider the activity of the forecast and analysis. For the tendency equation it is assumed that the model and analysis have the same level of activity. Figure 3 shows that to a good approximation the assumption is valid (though the activity in the forecast is a little lower than in the analysis) and so either of these time series can be taken as *ACT*. The assumption is also valid for ERA-Interim which has a good level of activity.

Typical recent values for the 500 hPa geopotential height for day-6 forecasts for HRES, taken from Figures 1 to 3, are *ACT*=94 m, *RMSE*=53 m and *ACC*=83%. Using these typical recent values of *ACT* and *RMSE* in the tendency equation allows the coefficients to be estimated as:

$$\frac{\mathrm{d}ACC}{\mathrm{d}t} \approx 0.34\frac{\mathrm{d}ACT}{\mathrm{d}t} - 0.6\frac{\mathrm{d}RMSE}{\mathrm{d}t}$$

So a 1% change in ACC can occur if (over the same timescale) there is either a change in *ACT* of about 2.9 m or in *RMSE* of about 1.7 m. For the less skilful ERA-Interim system (*ACT*=94 m, *RMSE*=62 m and *ACC*=78%) the same change in *ACT* (2.9 m) would account for a 1.3% change in ACC which is larger than that for HRES.

Figures 1 and 2 show that, as an overall trend, the *RMSE* decreases and *ACC* increases with time for HRES as the model and data assimilation improve. Indeed for much of the period (and this is true more generally) there is an extremely good (anti-) correlation between ACC and RMSE.

For the two-year period from mid-2009 to mid-2011, the atmosphere experienced a period of unusually high activity. Looking on the month-to-month anomalies in the activity (not shown), December to February 2009/10 and December 2010 stands out as much more active than normal. But the occurrence of larger amplitude anomalies does not imply a larger day-to-day variability in the atmosphere. In fact the opposite might be the case as large anomalies tend to be more persistent (e.g. we know that the AO was persistently in a negative phase during this period). Consequently, in this period that includes the first half of the pause, the relationship between *ACC* and *RMSE* is less straightforward.

During the rapid increase in *ACT* during the second half of 2009, the *RMSE* was decreasing and from the tendency equation we can see that both these tendencies contributed to the significant increase in *ACC* at that time. In mid-2011, when the *ACT* level decreased very rapidly, we see from Figure 2 that the *RMSE* also decreased rapidly. From the tendency equation we would expect that such a combination might be consistent with a rather small decrease in the *ACC*, which is in fact what happened – see Figure 1. It appears that the balance of terms in the tendency equation during the growth and decay phases of this anomaly in *ACT* was different.

In the second half of the pause, mid-2011 to end of 2012, there has been a period when the time variations of *ACC*, *RMSE* and *ACT* have returned to much smaller values and the tendency equation has been satisfied by all three terms being relatively small.

If we compare the *ACC* and *RMSE* for HRES relative to ERA-Interim reforecasts we see that there is little difference between *ACC* and *RMSE* in terms of the progress over the last three years – see Figure 4.

During the period when the *ACT* anomaly is developing (as the pause begins), the increase in *ACC* is larger for ERA-Interim (Figure 1). This is expected from the tendency equation because there are inherently larger values of *RMSE* for ERA-Interim as it is an older, lower resolution version of the forecasting system.

Relationship between *RMSE, ACC* and atmospheric activity

A

The definition of the root-mean-square forecast error, *RMSE*, in terms of the values of the forecast, f, and the analysis, a, is:

$$RMSE^2 = \overline{(f-a)^2} = \overline{(f-c)^2} + \overline{(a-c)^2} - 2\overline{(f-c)(a-c)} \quad \text{(A1)}$$

The definition of the un-centred ACC (*Wilks,* 2006) is:

$$ACC = \frac{\overline{(f-c)(a-c)}}{\sqrt{\overline{(f-c)^2}\,\overline{(a-c)^2}}} \quad \text{(A2)}$$

Here the overbar indicates a regional or global average and $\overline{(f-c)^2}$ and $\overline{(a-c)^2}$ are the squared standard deviations of the forecast anomalies and analysis anomalies from the climate, c, respectively; they are measures of the 'activity' in the forecast and analysis which will be denoted by ACT_f^2 and ACT_a^2.

Equations (A1) and (A2) can now be written as:

$$RMSE^2 = ACT_f^2 + ACT_a^2 - 2\overline{(f-c)(a-c)} \quad \text{(A3)}$$

$$ACC \approx \frac{\overline{(f-c)(a-c)}}{ACT_f\ ACT_a} \quad \text{(A4)}$$

If we assume that the activity in the model is approximately equal to that in the atmosphere (i.e. the analysis) we can define the quantity *ACT*:

$$ACT_f^2 \approx ACT_a^2 = ACT^2 \quad \text{(A5)}$$

Combining equations (A3), (A4) and (A5), we obtain:

$$ACC \approx 1 - \frac{RMSE^2}{2ACT^2} \quad \text{(A6)}$$

Equation (A6) shows the relation between *ACC*, *RMSE* and *ACT*.

Conclusions regarding the pause

In conclusion, the perception of a pause in forecast skill since late 2009, derived from looking at *ACC* values, is probably erroneous. It is a reminder that the interpretation of the fluctuations in *ACC* and *RMSE* (and their correlation) can be complex and for *ACC* it relies heavily on large-scale climate variability such as the phase of the Arctic Oscillation (AO). We need to examine both the *ACC* and *RMSE* values of the HRES together with the *ACT* level and not just any one of these quantities in isolation.

Over the past decade there has mostly been a good anti-correlation between *ACC* and *RMSE* values. The exception is the two-year period in the first half of the pause from mid-2009 to mid-2011 when there was a very large positive anomaly in activity *ACT* level. In this period of the three-year pause there was a very large positive anomaly in the activity level that coincided with the AO being in a prolonged negative phase. During such anomalous periods for activity one can expect a variety of behaviour of *ACC* and *RMSE* such that neither one of these quantities on their own is sufficient to deduce what is happening to forecast skill. In the second half of the pause period the time variations of *ACC*, *RMSE* and *ACT* and have all been relatively small.

The use of a fixed-reference forecast system to create a reforecast dataset is a valuable way to take account of the impact of atmospheric variability on the forecast scores. Consequently, results from ERA-Interim allow the period of high potential predictability during the pause to be accounted for. Comparing with ERA-Interim shows clearly the continuing trend of improvements of the operational HRES during the pause, as a result of the changes to the data assimilation and model introduced during the period.

This article has reviewed the recent skill of the ECMWF forecasts in the context of the long-term evolution of the scores. The overall trend of skill improvement is modulated to some extent by the variability of the atmosphere itself, a sensitivity which affects different skill measures in different ways. This sensitivity is why ECMWF regularly maintains and reviews a range of verification measures alongside the headline *ACC* score together with the atmospheric variability. The use of a fixed reference system (such as ERA-Interim) has now become an important component of the evaluation strategy, providing a benchmark for the IFS and the ability to distinguish between forecasting system improvements and variations in atmospheric predictability. It is essential to maintain such a fixed system (as close as possible to the operational model), running in near-real time, with at least a decade of reforecasts.

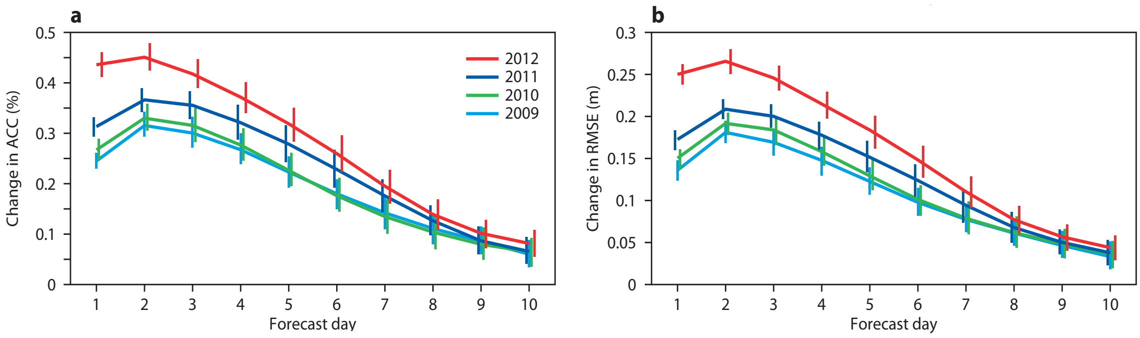


Figure 4 Relative change of (a) ACC and (b) RMSE for the 500 hPa geopotential height in the northern hemisphere extra-tropics for HRES forecasts, normalized by ERA-Interim reforecasts, as a function of forecast lead time, for the years 2009, 2010, 2011 and 2012.

Further reading

Andersson, E. & **D. Richardson,** 2011: Forecast performance 2010. *ECMWF Newsletter* **No. 226**, 10–11.

Janoušek, M., A.J. Simmons & **D. Richardson,** 2012: Plots of the long-term evolution of operational forecast skill updated. *ECMWF Newsletter* **No. 132***,* 11–12.

Langland, R.H & **R. Maue,** 2012: Recent northern hemisphere mid-latitude medium-range deterministic forecast skill. *Tellus A,* **64,** 17531, http://dx.doi.org/10.3402/tellusa.v64i0.17531

Magnusson, L. & **E. Källén,** 2013: Factors influencing skill improvements in the ECMWF forecasting system. *Mon. Weather Rev.*, **141,** 3142–3153.

Persson, A., 2013: ECMWF User Guide. http://www.ecmwf.int/products/forecasts/guide/user_guide.pdf

Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences*, Volume 100, Third Edition (International Geophysics), Academic Press, ISBN-10: 0123850223, 704 pp.

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

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