Factors Influencing Skill Improvements in the ECMWF Forecasting System

LINUS MAGNUSSON AND ERLAND KÅLLEN
European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

(Manuscript received 31 October 2012, in final form 8 February 2013)

ABSTRACT
During the past 30 years the skill in ECMWF numerical forecasts has steadily improved. There are three major contributing factors: 1) improvements in the forecast model, 2) improvements in the data assimilation, and 3) the increased number of available observations. In this study the authors are investigating the relative contribution from these three components by using the simple error growth model introduced in a previous study by Lorenz and extended in another study by Dalcher and Kalnay, together with the results from the ECMWF Re-Analysis Interim (ERA-Interim) forecasts where the improvement is only due to an increased number of observations. The authors are also applying the growth model on “lagged” forecast differences in order to investigate the usefulness of the forecast jumpiness as a diagnostic tool for improvements in the forecasts.

The main finding is that the main contribution to the reduced forecast error comes from significant initial condition error reductions between 1996 and 2001 together with continuous model improvements. The changes in the available observations contributed to a lesser degree, but the authors note that all the ERA-Interim forecasts are from the satellite era and here the focus is on the midtroposphere in the extratropics. Regarding the jumpiness in the forecasts, this is mainly a function of the error in the initial conditions and is therefore an insufficient tool to investigate improvements in the full forecasting system.

1. Introduction
The forecast skill in the European Centre for Medium-Range Weather Forecasts (ECMWF) numerical weather prediction system has steadily improved over the past 30 years. The quality of a 6-day forecast in 2010 is about the same as the quality of a 3-day forecast was in 1980 for the northern extratropics. This substantial increase in quality has been achieved through a series of improvements in the forecasting system. Three major contributing factors are the following:

- The forecast model has improved. Increases in spatial resolution as well as more accurate descriptions of physical processes such as radiation, turbulence, and clouds have all contributed.
- The assimilation of observation data has improved substantially leading to a reduction in the initial state error.
- More observations have become available, in particular satellite observations have increased dramatically since the beginning of the satellite era in 1979.

In this paper we will investigate the relative contributions of the factors above to the improvement in forecast quality. A first assessment and interpretation of the skill in terms of error growth rates and error saturation was made by Lorenz (1982). The initial error was assumed to grow exponentially (because of the chaotic properties of the system; Lorenz 1963; referred to as chaotic growth hereafter) while the nonlinear evolution of the chaotic and model error growth leads to error saturation toward the time limit of atmospheric predictability. For a perfect model only initial state errors will lead to forecast error growth. Because of the chaotic nature of atmospheric dynamics such errors will lead to saturation after a finite period of time.

A refinement of the Lorenz (1982) growth study was suggested by Dalcher and Kalnay (1987) and Savijärvi (1995). They included an initial linear error growth that can be associated with model errors. The initial error growth due to inaccuracies in the initial state generally has an exponential behavior while error growth resulting from model inaccuracies can be associated with linear error growth. The Lorenz (1982) error growth model only deals with the initial state error exponential growth while Dalcher and Kalnay (1987) and Savijärvi (1995) also take the model error contribution into account.

Corresponding author address: Dr. Linus Magnusson, ECMWF, Shinfield Park, Reading, RG29AX, United Kingdom.
E-mail: linus.magnusson@ecmwf.int

DOI: 10.1175/MWR-D-12-00318.1

© 2013 American Meteorological Society
When comparing forecasts with analyses, the absolute magnitude of the initial state error can be found through extrapolation of the error curves back to initial time. Assuming that the error grows monotonically over the whole forecast range, an assumption well supported by forecast error data, we can estimate the initial state error reduction by extrapolating from short-range forecast errors. Another way of estimating the initial state error is through a diagnosis of data assimilation components. We will show that the initial state error reduction can be associated with particular events in the development of new data assimilation techniques, while the forecast error reduction in the medium range appears to have a smoother time evolution.

A further aspect of forecast errors and forecast reliability is so called jumpiness. If two forecasts verifying at the same time but originating from two different initial times are very different but forecasts from initial times before and after the jump are similar, the forecasts appear to be less reliable. The forecast jumpiness has decreased over time in the ECMWF forecast system. We will analyze the main reason for this and conclude where future efforts should be spent to further reduce the jumpiness. One way of dealing with jumpiness is to assess the spread of an ensemble forecasting system. If the spread of the ensemble is reduced after a jump then the subsequent forecasts should be more reliable. It is beyond the scope of this study to investigate this matter further, but Bengtsson et al. (2008) have shown that also the forecast spread in an ensemble system follows the same general rules as the forecast error growth in a deterministic forecasting system.

This investigation excludes the tropical area between 20°S and 20°N. One justification for doing this is that the geopotential height field variability is generally small in this area and, therefore, does not significantly affect the error estimates. Another reason is that tropical dynamics are very different from midlatitude dynamics, short-term error growth is much faster as it is coupled to convective motion systems that have relatively short lifetimes. Error saturation for these types of systems is reached on time scales much shorter than the ones found for large-scale flow in midlatitudes. On the other hand there is more potential predictive skill in the tropics on longer time scales because of the close coupling between ocean surface temperature anomalies and atmospheric circulation patterns (Shukla 1998). To further study the short time-scale error saturation in the tropics one should focus on wind information as the wind field is dynamically governing the height field variability in the tropics. Currently high-quality wind observations are scarce in the tropics. Future space-based observations of wind profiles (Stoffelen et al. 2005) may make it possible to investigate short-term tropical predictability with a methodology similar to the one used here.

This study focuses on error evolution in the short and medium range, from initial time to about two weeks ahead. Beyond two weeks there may also be predictable signals even if the medium-range forecast error appears to have saturated. These signals are coupled to particular flow events such as the Madden–Julian oscillation or blocked flow regimes over the North Atlantic–European area and links between the atmosphere and slower parts of the system such as oceans and land surfaces (Hoskins 2013). In this study we are only considering running annual means of errors and error growth. This time filtering smooths out the effect of blocking events and other phenomena that may contribute to an enhanced predictive skill on time scales of a month or so. Predictive skill resulting from couplings to the ocean and land surface may give rise to long time-scale fluctuations in error levels, this feature is clearly seen in our data but we do not analyze this further.

A comprehensive evaluation of the ECMWF forecasting system was made by Simmons et al. (1995) and Simmons and Hollingsworth (2002). We will build on their study by extending the analysis results to cover also the last 10 years and relate our findings to known improvements in the ECMWF forecasting system, which could not be easily identified 10 years ago. Furthermore, we will show results from recent experiments using the ECMWF reanalysis system to determine forecast error growth rates. Here the analysis and forecast systems are frozen in time and any changes in forecast quality can only be attributed to changes in the observing system or changes in the predictability of the flow state of the atmosphere. If the forecast skill systematically improves with time this is very likely due to an improved observing system. An interesting question is how much the investments made in the observing system have contributed to forecast skill improvements versus investments made in forecast model and data assimilation technique developments. This aspect will be discussed in this paper.

2. The growth model

To study the properties of error growth, we will use the simple error growth model originally presented in Lorenz (1982) and extended in Dalcher and Kalnay (1987):

$$\frac{dE}{dt} = (aE + \beta) \left( 1 - \frac{E}{E_\infty} \right). \quad (1)$$

The model describes the growth of $E$, the difference (perturbation) between two realizations of the atmosphere.
(either the difference between the model and the real atmosphere or two model simulations). Because of the chaotic behavior of the system, the growth of a small perturbation is proportional to its amplitude. The growth rate corresponds to the parameter $\alpha$, which is related to the leading Liapunov exponent of the system. If the volume of the model phase space is finite (the case for the atmospheric system), the perturbation amplitude is limited. Averaged over many realizations, the perturbation amplitude approaches a climatological value $E_{\infty}$, determined by the volume occupied in the phase space by the model and the analysis.

For the imperfect model case (in the presence of model error), forecast errors will occur even with perfectly known initial conditions ($E = 0$). Because of this, Dalcher and Kalnay (1987) extended the growth model by adding a linear growth rate $\beta$, in order to represent the model error.

The resulting growth model is a second-order polynomial in $E$ where the growth rate for small $E$ is close to $aE + \beta$ and zero for $E = E_{\infty}$. The solution to the differential in (1) yields an expression of the forecast error as function of the forecast length. The value at $t = 0$ is the analysis error or the amplitude of the initial perturbation ($E_0$).

The growth model in (1) does not include any skewness in the error growth. A skewness in the model was introduced in Stroe and Royer (1993) and the two versions were compared in Simmons et al. (1995). The conclusion was that the two models agreed for the medium range, while they started to diverge for longer forecast ranges, and therefore, for this study (1) is sufficient. The growth model assumes exponential growth of initial state errors. Correlations between analysis errors and forecast error will show up as superexponential growth for short lead times while the presence of model biases will lead to subexponential error growth in an RMS sense; both phenomena lead to uncertainties in the interpretation of the growth model given by (1).

In this study we will apply the model for the root-mean-square error (RMSE) of the 500-hPa geopotential height ($z_{500}$) and for the Northern and Southern Hemispheres, defined here as $20^\circ$–$90^\circ$N and $20^\circ$–$90^\circ$S, respectively. (The hemispheres will hereafter be referred to as NH and SH, respectively.) We will use operational forecasts from 1 January 1981 to 1 June 2012. Before 1986 forecast data are only available for every 24-h time step, therefore, we will only apply the growth model parameter estimation from 1986 for the operational forecasts. In addition to the operational data we will use forecasts initialized from ECMWF Re-Analysis Interim (ERA-Interim; Dee et al. 2011). ERA-Interim (both analysis and forecast) uses the ECMWF data assimilation system and forecast model from 2006, but with the spectral resolution $T_255$ and 60 vertical levels. The forecast scores based on the ERA-Interim system are calculated from 1981 to 2012.

Before 1986 the operational forecasts are verified against the operational analysis from the system present at that time. The forecast errors after 1986 are in respect to the analysis originating from ERA-Interim. This choice reduces the correlations between analysis and forecast errors for short lead times compared to if we had used the operational analyses for verification. For all time series a 12-month running mean is applied. This excludes variations in predictive skill on the monthly and seasonal time scales, conditioned by, for example, the occurrence of the Madden–Julian oscillation (Vitart and Molteni 2010).

### 3. Improvement of the operational forecast

Since the operational start in 1979, the ECMWF forecast model and the data assimilation system have been continuously developed. Table 1 lists major developments affecting the high-resolution forecasting system. Among the important upgrades is the introduction of four-dimensional variational data assimilation (4D-Var) at the end of 1997 and subsequent changes in the use of data in the assimilation were undertaken (Simmons and Hollingsworth 2002). One important change here was the upgrade of the usage of raw microwave radiances from the Television Infrared Observation Satellite (TIROS) Operational Vertical Sounder (TOVS) and Advanced TIROS TOVS (ATOVS) satellite-borne instruments in the year 2000. A major change in the model physics took place in 2007 when changes to the convection scheme and the vertical diffusion were introduced (Bechtold et al. 2008). A comprehensive description of

<table>
<thead>
<tr>
<th>Year</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>T213/L31</td>
</tr>
<tr>
<td>1991</td>
<td>Semi-Lagrangian scheme</td>
</tr>
<tr>
<td>1996</td>
<td>3D-Var</td>
</tr>
<tr>
<td>1997</td>
<td>4D-Var</td>
</tr>
<tr>
<td>1998</td>
<td>T319/L31</td>
</tr>
<tr>
<td>1999</td>
<td>T319/L50</td>
</tr>
<tr>
<td>2000</td>
<td>T511/L60</td>
</tr>
<tr>
<td>2006</td>
<td>T799/L91</td>
</tr>
<tr>
<td>2007</td>
<td>New convection scheme</td>
</tr>
<tr>
<td>2010</td>
<td>T1279/L91</td>
</tr>
<tr>
<td>2011</td>
<td>Ensemble of data assimilations</td>
</tr>
</tbody>
</table>
the changes between 2005 and 2008 is given in Jung et al. (2010).

Figure 1 shows the RMSE as a function of forecast lead time for the NH (Fig. 1a) and SH (Fig. 1b). The error is calculated as an annual average for three different years; 1990 (blue), 2000 (red), and 2010 (green). From these curves the improvement of the forecast quality over the last 20 years is evident. The rate of the improvement for the NH has been about 1 day decade$^{-1}$ in terms of the time range of predictive skill: a 6-day forecast from 2010 has the same error as a 4-day forecast from 1990. For the SH the rate of improvement was even higher between 1990 and 2000. Between 1990 and 2000 we see a clear reduction of the short-range forecast error, especially for the SH (see below).

Studying the shape of the curves in Fig. 1, we see an increase in the error growth in the beginning of the forecast, with a maximum in the growth rate after 4–6 days, followed by a slower error growth and signs of approaching the saturation level after 10 days. Figure 2 shows the rate of change for the error ($dE/dt$) as a function of the error amplitude $E$ (crosses) for the same data as in Fig. 1. The unit for error growth is meters per hour. Each cross represents the growth linearly differentiated between two consecutive lead times.

For each of the datasets presented in Fig. 2, a second-order polynomial has been fitted (solid lines), representing the simplified growth model in (1). From the coefficient of the polynomial, the parameters $a$, $b$, and $E_\infty$ can be obtained. In the plot, $E_\infty$ corresponds to the value of $E$ when $dE/dt = 0$ and $b$ is $dE/dt$ when $E = 0$.

For both the NH and SH the data from all 3 years yield parabolas that appear to fit the data well. One question mark is the parabola for year 2000 and the SH, which overestimates the error growth for short lead times compared to the data.

Figure 3 shows the time series of the forecast error (blue) after 24 h (Figs. 3a,b) and after 144 h (Figs. 3c,d). To compare the results with a frozen model and data
assimilation system, the ERA-Interim forecasts are also included in the figure (green). The operational forecasts from before 1986 have been verified against the analysis from the same system, not ERA-Interim, which explains the jump in the curve in 1986. Note that the ERA-Interim forecasts are verified against its own analysis, presumably leading to an underestimation of the short-range forecast errors due to correlations between analysis and forecast errors.

For the 24-h operational forecasts, we see a monotonic decrease of the error. The most pronounced reduction of the error, especially for the SH, appears around year 1996, 1997, and 2000, which coincides with the introduction of three-dimensional variational data assimilation (3D-Var) (1996), 4D-Var (1997), and an increased use of satellite data and the upgrade of the model resolution to T511 (2000). After 2001, a continued but not as marked reduction of the 24-h error for the operational forecast (blue) is present. Comparing the data from before and after 1986, noting the discontinuity most clearly seen in the SH, we see the large influence of the choice of the verifying analysis.

For the error after 144 h, only a slight reduction of the error is present between 1986 and 1995 for the NH. For the SH the error even increased. From 1995 and onward a rapid decrease is found. These properties will be discussed below. After 2002 the operational forecast error at +144 h is lower than the ERA-Interim error while the short-range forecast error only falls below the ERA-Interim error after 2005. This can be explained by the fact that we verify both forecasts with ERA-Interim analyses. For the ERA-Interim forecasts, the error reduction is much less than for the operational forecasts. However, the reduction in the 144-h error for the SH after 1986 is about 10 m for ERA-Interim compared to 45 m for the operational forecast. The error reduction for the ERA-Interim forecasts must be due to improvements in observing system such as new satellites.

For the error after 144 h, only a slight reduction of the error is present between 1986 and 1995 for the NH. For the SH the error even increased. From 1995 and onward a rapid decrease is found. These properties will be discussed below. After 2002 the operational forecast error at +144 h is lower than the ERA-Interim error while the short-range forecast error only falls below the ERA-Interim error after 2005. This can be explained by the fact that we verify both forecasts with ERA-Interim analyses. For the ERA-Interim forecasts, the error reduction is much less than for the operational forecasts. However, the reduction in the 144-h error for the SH after 1986 is about 10 m for ERA-Interim compared to 45 m for the operational forecast. The error reduction for the ERA-Interim forecasts must be due to improvements in observing system such as new satellites.

Figure 4 (NH) and Fig. 5 (SH) show the time series of the growth model parameters from 1986 to 2012. The coefficients have been obtained by fitting the second-order polynomial to the daily data with a 12-month running mean applied. The parameters for the operational forecasts show a year-to-year variability that can be due
to different flow patterns in the atmosphere leading to different error growths. The value of $\alpha$ is varying between $0.013$ and $0.018$ s$^{-1}$ for the NH and between $0.010$ and $0.016$ s$^{-1}$ for the SH. Despite the large variability there is a sign of an upward trend, which could correspond to a higher degree of realism in recent model versions. A variability in $\alpha$ is also present for ERA-Interim where the forecasting system is kept constant. This indicates that the year-to-year variability in the atmosphere has an impact on the parameter fitting and that different atmospheric flow situations yield different error growth rates.

For the model error term $\beta$, we see a steady decrease since 1996. The variability before 1996 could be due to uncertainties in the parameter estimations. For ERA-Interim a long-term reduction of $\beta$ for the SH is present, which is unexpected for a system that uses the same model for the whole period. This could be an artifact of a decrease in correlation between analysis and forecast errors due to an increased amount of satellite observations.

The saturation level $E_s$ shows a decrease from 1997 up to the present. This parameter is dependent on the model activity, but also includes a component of systematic model error (bias), which has decreased over the years (not shown). To further investigate the model activity, an independent measure of the error saturation level is also plotted in Fig. 4c. Here the saturation level is calculated as $\sqrt{\langle f - \bar{f} \rangle^2 + \langle a - \bar{a} \rangle^2}$, where $\bar{f} - \bar{e}$ is the time-averaged forecast anomaly with respect to climate and $\bar{a} - \bar{e}$ is the time-averaged analysis anomaly with respect to climate. This diagnostic relies on the forecast and the outcome (analysis) being uncorrelated for long lead times [see the appendix in Simmons et al. (1995) for details]. The climate is defined from the ERA-Interim daily climatology. Comparing the two estimates we see that the independent evaluation yields higher values of asymptotic variability. The difference increases in later years. One explanation to the difference between the methods is the uncertainty in the parameter estimation for the growth model. If the error growth parabola is skewed to the left, the asymptotic limit will be underestimated. One reason for such a skewness could be a slower than estimated approach to the asymptotic limit of RMSE caused by predictive skill coupled to earth-system
components such as the oceans, the land surface, and the upper stratosphere.

The asymptotic limit \( E_\infty \) increases for the SH from 1986 to the mid-1990s for both estimation methods. This increase could explain the increased 144-h forecast error for the SH over this period seen in Fig. 3d.

To investigate the separate contributions from the initial state and the model to the error growth in the early forecast range, the growth \( (dE/dt) \) has been calculated between 12 and 24 h lead time. For this time range the error growth is not influenced by saturation. Figure 4d shows the time series of the error growth (red, solid) together with the contribution from model errors \( \beta \) (blue) and chaotic growth \( \alpha E \) (green). The contribution from chaotic error growth is a product of the error in the analysis, which determines \( E \) at short lead times, and the value of the chaotic error growth parameter \( \alpha \). For the NH the we see a continuous reduction in total error growth over the years, with the error growth decreasing from 0.38 m h\(^{-1}\) around 1990 to less than 0.14 m h\(^{-1}\) in 2011. The largest contribution to the reduction comes from a decrease in the chaotic error growth due to the reduction in the analysis error between 1996 and 2001 and a larger contribution from the model error term since 2002. The same reduction is present for the SH, but with an even larger contribution from the reduction in the analysis error.

As a consistency check of the diagnostics, the sum of both contributing parts is included in Fig. 4d (red, dotted). We see that the sum somewhat overestimates the error growth from \( dE/dt \), which could be due to sub-exponential error growth for short lead times. Around 1999 we see an anticorrelation between the components for the NH (the feature is not present for the SH). It is difficult to say whether this is related to any change in the model system or if it is an artifact of the diagnostic method. It is discussed in Simmons and Hollingsworth (2002) that the summer of 1999 was unusual in terms of error growth in the NH. A similar behavior is also found for the ERA-Interim data (at least for \( \alpha \)), which indicates that the increase in the chaotic growth parameter was at least partly due to the flow situation.

To investigate the sensitivity of the results regarding the trends in \( \alpha \) and \( \beta \) to the values of the RMSE itself (whether the position on the parabola affects the parameter estimate), we have made an additional parameter fitting that only uses data pairs that have an RMSE between 20 and 85 m (NH), an error range covered by all 12-month running means from 1986 to 2012. The results
regarding the trends are in line with the results above, although noisier than using all data points. In a second sensitivity test we estimated $\alpha$ and $\beta$ from three data pairs with small errors, with the linear approximation $(\alpha E + \beta)$. We restricted the data points to have errors over 20 m in order to sample the same error range for whole time series. The long-term trends for the parameters were also here in line with the results above (not shown).

In this section we have diagnosed the improvements of the RMSE for the operational forecasts. By comparing the improvements in the operational forecasts and ERA-Interim forecasts we found that the major contribution for the error reduction in the latest 30 years comes from model and data assimilation improvements (Fig. 3). By using the error growth model, we have separated the growth due to chaotic growth of initial errors and model error. We have found that the initial error had the fastest reduction between 1996 and 2001 and the model improvements played the most important role during the period 2002–10 (Figs. 4d and 5d).

4. Implications for differences of successive forecasts

One desirable property of a forecasting system is a good consistency between successive forecasts verifying at the same time but with different forecast lead times. If successive forecasts differ and “jump” between different solutions, the forecasters have difficulties in interpreting the results. The “jumpiness” could be interpreted as an indication of uncertain forecasts. The method of estimating the forecast uncertainty from successive forecasts was evaluated in Buiizza (2008), with the conclusion that an ensemble prediction system was superior estimating the uncertainties compared to the jumpiness between successive forecasts. The inconsistency between successive forecasts is also used as a diagnostic tool to evaluate the forecast system. In this section we will investigate the usefulness of this diagnostic tool by comparing the growth of the difference between two forecasts issued with a 24-h lag but valid at the same time.

When comparing two forecast simulations that originate from the same model, the model error term $\beta$ in (1) should vanish (“perfect model” comparison). The inconsistency between the forecasts is therefore a function of the chaotic growth, the asymptotic limit and the initial difference, which is the analysis innovation between the initial times of the forecasts (in this case 24 h).

Figure 6 shows the evolution of the difference between two forecasts with a lag of 24 h from 1986 to 2012 for the NH (Figs. 6a,c) and SH (Figs. 6b,d). Figures 6a,b show the difference between 24-h forecasts and the analysis (0-h forecast) valid at the same time, while Figs. 6c,d show the difference between 120- and 144-h forecasts. Here, we see clear similarities between the two evaluated time intervals in the evolution of the lagged forecast differences; for both a clear reduction appears around the year 2000, the same period when we saw a clear decrease in the initial error in the previous section.

One feature in the difference between the 120- and 144-h forecast is a peak around 1992. This was connected to several problems with the introduction of a higher-resolution semi-Lagrangian version of the forecasting system in September 1991 (Simmons et al. 1995).

If we compare the result for the 120–144-h difference in Fig. 6c with the 144-h forecast error in Fig. 3c for the NH and Fig. 6d with the result in Fig. 3d for the SH, we see a faster decrease in the forecast error than in the lagged-forecast difference. While the decrease in the forecast error is the combined effect of the decrease in the initial error and model error, the decrease for the forecast difference is only dependent on the initial error. For the SH (Fig. 3d) we see the increase in the difference between 120 and 144 h between 1986 and 1995, similar to that for the forecast error. This is related to the increase in the model activity.

Figure 7 shows the growth of the difference between the lagged forecasts from Fig. 6 for the NH and the year 2010. Here we can see that the data do not fit the parabola perfectly because the data are skewed to the left. Because the growth model is not skewed, the $\beta$ parameter is expected to be somewhat overestimated and $\alpha$ underestimated.

Figure 8 shows the growth model parameters for the lagged forecast differences for the NH. Here we see that $\beta$ is not zero. This could be an artifact of the skewness, as discussed above. For $\alpha$, the values show no clear trend and the values are between 0.016 and 0.020. The $E_\infty$ parameter increased from 1986 to 1993. The peak of the saturation level is connected with the problem of the semi-Lagrangian scheme discussed above. Around 1999, the asymptotic limit of the difference decreases and since then it has been fairly constant. Also, here the growth model seems to underestimate the saturation level compared to the forecast anomaly estimation, which for the forecast difference is defined as $\beta = \sqrt{2\alpha^2 - \alpha^4}$ (we compare two integrations from the same model), but both methods show similar signals. The results are similar for the SH but with an even stronger signal for the increase in $E_\infty$ between 1987 and 1992 (not shown).

In this section we have investigated the forecast jumpiness as a diagnostic tool for monitoring forecast system improvements. By studying the terms in the error growth model, the jumpiness is a function of the size of
24-h forecast errors and the realism of the model (chaotic growth and model activity). An improved realism in the model could thus lead to an increased jumpiness between subsequent forecasts. Therefore, this is not a useful tool to monitor improvements in the forecast model.

5. Discussion

The ECMWF forecasting system has been and is continuously being improved. In this report we have investigated the decrease in the forecast error with a simple error growth model. The model contains parameters representing the chaotic (exponential) growth, model error (linear) growth and the error saturation parameter. The forecast error is then a function of these parameters together with the initial state (analysis) error.

While the chaotic growth and the error saturation are dependent on the nature of the atmospheric system, the model error and the analysis error are targeted when we aim to reduce the forecast error. We must, however, also be aware that a more realistic model could have a faster chaotic growth and a higher error saturation level than a less realistic model. Thus, when optimizing the forecasting system we have to make sure that model realism is maintained when scoring parameters improve.

For short-range forecasts, a strong reduction of the ECMWF forecast error was found around the year 2000. Around this time the use of satellite data were increased by the introduction of 4D-Var together with

![FIG. 6. Mean RMS difference between two forecasts with a lag of 24 h from 1986 to 2012: (a),(b) 0–24 and (c),(d) 120–144 h for (a),(c) NH and (b),(d) SH.](image)

![FIG. 7. Difference growth for 24-h lagged forecasts for the NH and the year 2010.](image)
other changes in the forecasting system. The reduction was largest for the Southern Hemisphere. Comparing the evolution of the error after 12 and 144 h we see a more continuous reduction in the 144-h than in the 12-h forecast errors. The 144-h error is a combination of model errors accumulated and amplified (due to chaotic error growth) over the integration and initial errors that have been amplified since the initialization. Improvements in the initial conditions are not only due to improvements in the data assimilation system but also model improvements that lead to an improved quality in the first guess for the data assimilation.

The forecast errors for the ERA-Interim, which uses a frozen data assimilation and model system but includes changes in the observing system, show a weaker trend for the forecast error compared to the operational forecast error. This indicates that changes in the observing system seem to play a relatively small role for improving forecast quality during the investigated period. However, we have to bear in mind that the ERA-Interim period used here (1981–present) is in the satellite era, even if the data were not fully used in the operational forecasts until the year 2000. Improvements in the observing system are a necessary requirement for improving forecast quality. Developments in the forecasting system are needed for exploiting the full potential of the observations. In this paper we have not discussed the tropics but we believe that the forecasting system will gain substantially from a further evolution of the observing system, in particular better wind observations.

Using this type of simple growth model has some limitations. For the forecast error investigation, around the year 1999 we found, simultaneously, a clear decrease in the model error contribution and an increase from the analysis error contribution to the error growth. We suspect this correlated change is due to an ambiguous parameter fitting. In Simmons and Hollingsworth (2002), it is claimed that the analysis error initially amplifies superexponentially and, therefore, also contributes to the linear parameter. In any case, the initial growth of the analysis error is a source of uncertainty in the fitting of the growth model parameters.

Another limitation of the growth model is the symmetry of the parabola; the growth model has no skewness. In reality, the growth of the error is somewhat skewed. It could be due to sub/superexponential error growth for early lead times and/or a delayed approach toward the asymptotic error limit due to an enhanced predictability resulting from couplings to slow parts of the geophysical system. An example of such a coupling is the atmosphere–ocean interaction associated with El Niño/La Niña events and its effects on midlatitude flows. For this type of study, the skewness could have an artificial effect on the growth model parameters if the range of RMSE changes in time. We made a sensitivity test where we limited the range of RMSEs to between 20 and 85 m and found similar results to those using all data pairs, indicating robustness in our results.
One quality measure often used to diagnose the forecasting system is the consistency between subsequent forecasts. If the forecasts “jump” between different solutions the forecasting system is judged to be unreliable. Here we show that the mean difference between two forecasts issued with 24-h difference, is a function of the analysis error, and not directly a function of the model error. With the model error contribution missing, the jumpiness is unsuitable for diagnosing the reliability of the forecast system.

The results from this type of error growth model also have implications for the design of an ensemble prediction system. A well-calibrated ensemble system should have a one-to-one match between the RMSE of the ensemble mean and the standard deviation of the ensemble members, for all forecast lengths (Bengtsson et al. 2008). To fulfill this condition, all growth model parameters for the evolution between two ensemble members need to be equal to the development of the forecast error. It means that the variability of the model (which determines $E_x$), needs to be the same as the variability of the real atmosphere. Furthermore, one needs to include some model error simulation in the ensemble forecasting, in order to obtain a nonzero $\beta$ parameter. This fact motivates the use of stochastically perturbed physics (SPPT) and spectral backscatter schemes in the ECMWF ensemble prediction system (Palmer et al. 2009). The model has to have a chaotic growth rate similar to the atmosphere and finally the amplitude of the initial perturbation needs to match the error in the analysis.

In this study we have focused on errors of the 500-hPa geopotential height. However, the model improvement is also evident for near-surface parameters (e.g., precipitation; Rodwell et al. 2010). For near-surface variables the scope for further improvements is still large, mainly due to remaining issues with systematic model errors. We have excluded the tropics in our error analysis, the main reason being large uncertainties in the estimation of forecast error. Hopefully a future improved atmospheric reanalysis will provide the opportunity to investigate the tropical forecast improvements better than we are able to do at present. An additional difficulty for the tropics is the multiple time scales (Shukla 1998; Peña and Kalnay 2004) affecting the forecast error growth rather than the single dominating scale (baroclinic instability) in the extratropics. To account for multiple time scales, the error growth model has to be modified. Finally, this study only discusses the average performance of the forecasting system, but a recent study by Rodwell et al. (2013) focused on special cases of large forecast errors.

The contribution today to the error growth at short lead times is of the same order of magnitude for the model error and analysis error contributions. Therefore, we need to improve both the data assimilation and the forecast model in order to continue to improve the forecast system. New techniques for data assimilation (weak-constraint, long-window 4D-Var, Fisher et al. (2011); and ensemble of data assimilations, Isaksen et al. (2010)) as well as resolution increases and model physics developments will also be central research and development areas in future numerical weather prediction research.

Acknowledgments. We would like to acknowledge Adrian Simmons for valuable comments and Martin Janousek for the calculation of the scores. We would also like to thank Anabel Bowen for help with the preparation of the figures for this report and the three anonymous reviewers whose suggestions helped us improve this paper.

REFERENCES


——, 1982: Atmospheric predictability experiments with a large numerical model. Tellus, 34, 505–513.


