Diagnostics at ECMWF

Mark Rodwell[1] **and Thomas Jung**[1]

[1] *ECMWF, Shinfield Park, Reading RG2 9AX, United Kingdom*

ABSTRACT

Here, some insight is given into the strategies and recent work of the Diagnostics Group at ECMWF. Diagnostics within the context of operational NWP must be targeted primarily at understanding forecast error. Hence the function of the Diagnostics Group lies somewhere between forecast verification and model development. Three examples that highlight both strategy and work are presented here. The first example is the use of the 'initial tendencies' approach that could enable model developers to harness the power of data assimilation to identify errors and test solutions. The second example highlights the the need to quantify forecast error as a function of spatial scale with, in particular, more attention given to smaller scales than currently done. The final example emphasises and addresses the need to better monitor and diagnose the prediction of weather parameters such as precipitation.

1 Introduction

Figure 1 shows the spatial anomaly correlation coefficient (ACC) for Northern Hemispheric 500 hPa geopotential height (Z500) over the period 1980–mid 2009. The blue circles show monthly-means of daily values at a forecast lead-time of 1-day (D+1). There is some variability in these values but close inspection indicates that it is associated with the annual cycle rather than more random fluctuations. Indeed, the 12-month running-mean values (red) display a fairly smooth upward trend. This implies that the remaining error is getting smaller although the key issue is not the absolute size of the error but rather its magnitude relative to its uncertainty. For example, coincident with the improving trends has been an astronomic (literally) increase in observation data. Since 1996 there was a 100-fold increase in the volume of satellite data assimilated at ECMWF. Has this led to a better estimation of the 'truth' and thus a more precise estimation of forecast error?

The green squares show monthly-mean ACC at D+5. There is more variability in scores than at D+1 and this has a strong 'random' component as well as an annual cycle. The 12-month running-mean values (pink) still show an upward trend but it is less smooth than at D+1. Although the ultimate achievable level of skill at D+5, in the presence of chaos, is unlikely to be 100%, it would appear that larger future reductions in error are possible at D+5 than at D+1. Hence our uncertainty in the truth is less of an issue at longer lead-times. However, sampling uncertainty associated with the flow-dependence of potential skill, the growth of interactions between the resolved flow and parametrized diabatic processes, and general chaos are more important at D+5. Again, therefore, the issue of optimising 'signal-to-noise' is relevant.

Two important questions arise from this discussion

- Is there an optimal lead-time for the diagnosis of model error?
- Can sensitivity to sampling uncertainty be minimised?

The key deterministic forecast target at ECMWF is a one day gain per decade in the lead-time at which the ACC of extratropical Z500 falls to 60%. It is interesting to see what spatial scales are associated with

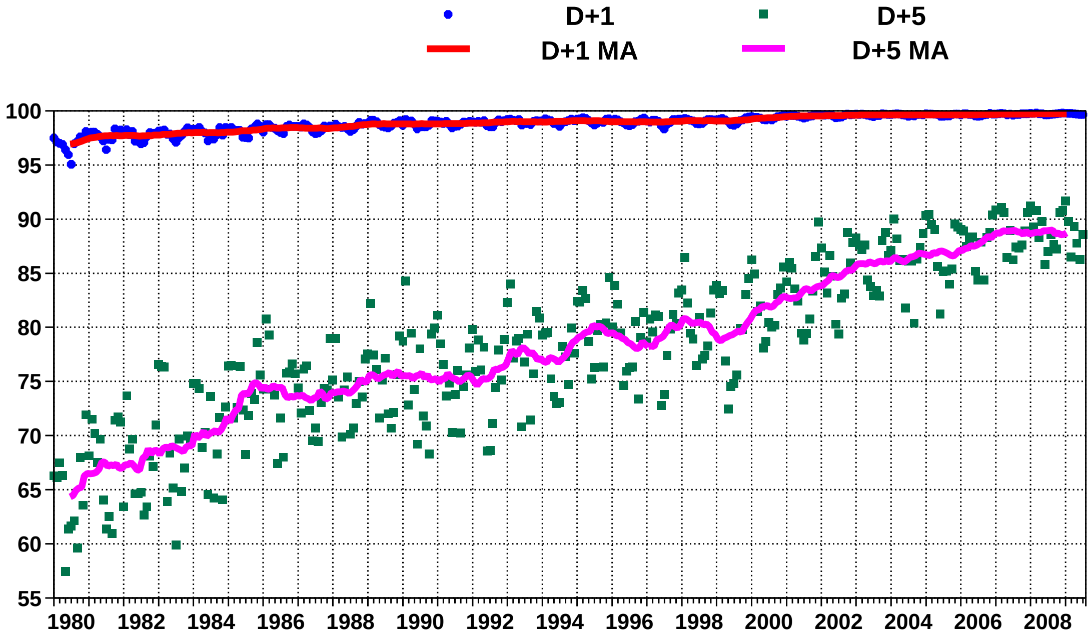

Figure 1: Northern Hemisphere monthly-mean spatial anomaly correlation coefficients (×100) *and 12-month running means of 500 hPa geopotential heights at lead-times of 1 and 5 days.*

error in Z500. The power spectra in Fig. 2 show annual-mean temporal variance of D+1 Z500 error as a function of total wavenumber. The maximisation of error at wavenumbers 5–15 reflects the dominant spatial scales of Z500. In the past, much diagnostic work has also focused on these large-scales. With ECMWF's deterministic forecast resolution recently being raised to T_L1279, it is clear that

- More verification and diagnosis of error at smaller scales is required.

This could involve explicit separation of scales in a variable as in Fig. 2 or it could involve the use of parameters (such as precipitation for example) that naturally have smaller spatial scales.

Figure 3 shows how Northern Hemisphere winter 'blocking' frequency changed for two recent updates to the forecast model. Results are based on model integrations initiated on 1 November for the years 1963–2002 and run at resolution T_L159, with prescribed sea-surface temperature, over the subsequent December–February season. Fig. 3(a) shows that updates incorporated into model cycle 33R1 increased Euro-Atlantic and Pacific blocking frequency so that it is generally within the bounds of observed uncertainty, as deduced from ERA-40 re-analyses (grey shading). This was a welcome result since blocking has traditionally been a difficult flow-type to represent in models.

Quantities such as blocking frequency and many others including energy flow diagrams, tropical wavenumber-frequency diagrams, extratropical synoptic activity, the magnitude and timescale of ENSO or the Madden-Julian Oscillation represent useful metrics with which to compare models or model cycles but they do not indicate why one model is better than another or, indeed, why a subsequent model cycle (35R3) apparently became worse in terms of blocking frequency (*c.f.* Fig. 3b). While the Diagnostics Group at ECMWF does calculate metrics such as blocking frequency, it is clear that

- Further diagnostics are required that delve deeper, and permit a fuller understanding of the root-causes of forecast error.

Much of the strategy of the Diagnostics Group at ECMWF is developed around the questions and state-

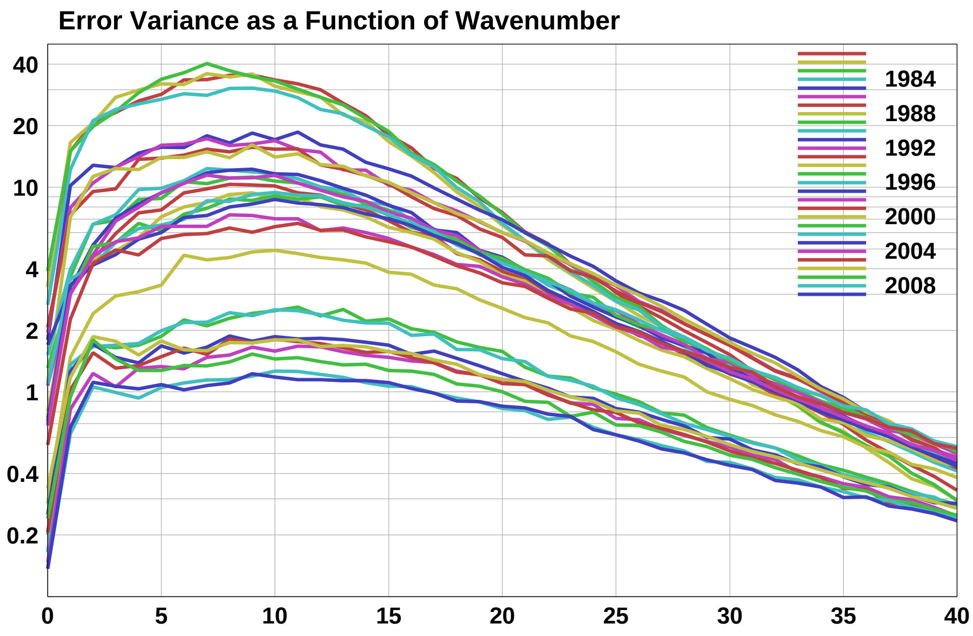

Figure 2: Temporal variance of D+1 error in 500 hPa geopotential heights as a function of total wavenumber. Modified from plot of Adrian Simmons.

ments itemised above. It will be clear from the above introduction that Diagnosis, in the context of operational NWP, lies between forecast verification and forecast system development. The dividing lines between these tasks can be blurred and here I will stray into both areas in order to highlight the continuity (or 'seamlessness') of work that is required to achieve more accurate forecasts.

2 The 'Diagnostics Explorer'

The Diagnostics Group at ECMWF has developed a web-based 'Diagnostics Explorer' to help researchers identify and investigate forecast errors. The aim is to present, as seamlessly as possible, diagnostics of the entire data assimilation and weather forecasting system, and metrics of the model climate. Other sections within ECMWF produce diagnostics related to their specific field of interest but the diagnostics in the Explorer are unique in giving the over-view of the entire system. The contents of the Diagnostics Explorer are listed in Table 1 and documented further in Rodwell and Jung (2008a). Figure 3 is based on plots available on the Diagnostics Explorer. In the subsequent sections, present or future content of the Explorer are discussed.

3 Forecast error

Figure 4 shows 500 hPa temperature errors averaged over all operational 0 UTC forecasts made at ECMWF for the season December–February 2008/9. The four plots (a–d) show these mean errors for the forecast lead-times of 1, 2, 5, and 10 days, respectively.

At Day 1 (Fig. 4a), there is a uniform and statistically significant warm error over much of the tropics. (5% significance is indicated by the use of the bold colours, insignificance by the use of the pale colours in the 'dual colour pallet'). Generally there is also a cool error over the northern mid-latitudes. By Day 2 (Fig. 4b), the mean errors have got stronger (note the change in shading interval) although there is

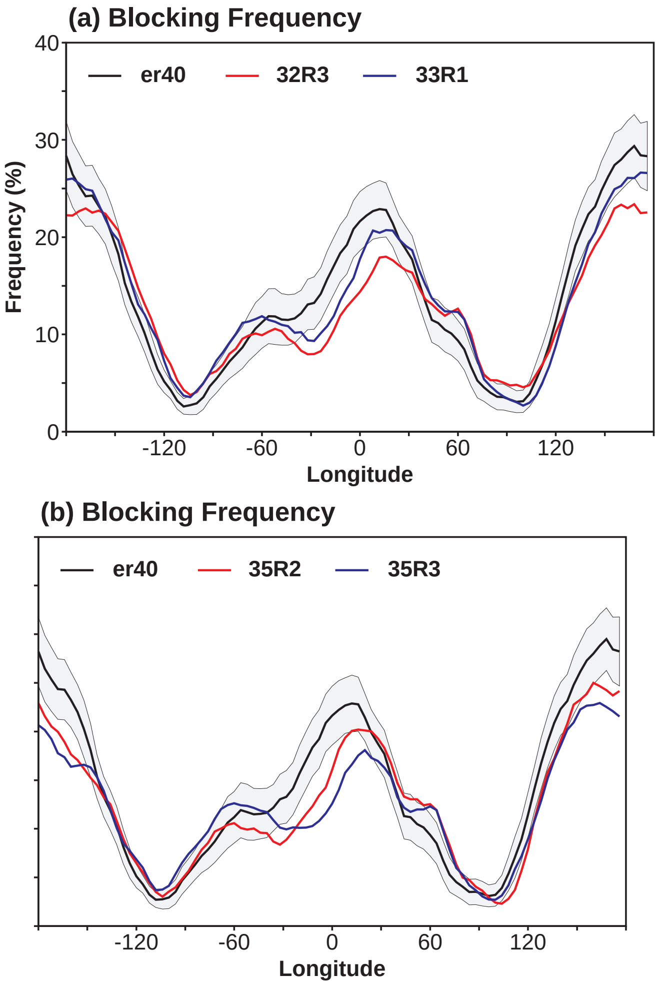

Figure 3: Frequency of Northern Hemisphere blocking during December–February 1963–2002 based on analyses and climate simulations with model cycles as indicated. Analysis are from ERA40 for the period 1963-2002 and operational analyses thereafter). The grey shading indicates the 95% confidence interval about the observed-mean frequency.

no visible increase in the area that is statistically significant. Through Days 5 and 10 (Fig. 4c,d), the maximum values of mean errors continue to grow but the uniform pattern of tropical error seen at day 1 is replaced by a more complex pattern with a decreasing area over-which the mean error is statistically significant.

An interpretation of these results is that by days 5 and 10, interactions, teleconnections and loss of predictability have confused a simple investigation of the root causes for the mean forecast error. Statistical significance actually increases as the lead-time *decreases*. Taken to the ultimate extreme, one might expect that the optimal lead-time to use when searching for physical parametrization deficiencies (relevant to NWP) would be at timestep 1 of the forecast. (see e.g., Klinker and Sardeshmukh, 1992). In fact, timestep 1 introduces other problems associated with sampling the diurnal cycle so here the focus will be on the first few timesteps (Rodwell and Palmer, 2007). Since the ‘first few timesteps’ occur within the data assimilation window, it is appropriate to next discuss data assimilation.

IFS Component	Diagnostics
Data Assimilation	**Observation space – observation usage** • Many data sources including radiosonde and satellite • Data count, first-guess departures (mean, rms), bias corrections
	Model space – analysis increments • Prognostic and other parameters • Mean, standard deviation, rms • 21 pressure levels and zonal means
Weather Forecast	**Forecast error** • Prognostic and other parameters • Mean, standard deviation, rms • 21 pressure levels and zonal means
	Scale-dependent error and activity • Several parameters, levels and regions • All spatial scales and selected spatial scales
Climate of atmospheric model and coupled model	**Seasonal-means of error** • Several diagnostics including geopotential height, winds, velocity potential, Hadley and Walker circulations, ocean waves
	Seasonal-means of variability • Blocking • ENSO teleconnections • Empirical Orthogonal Functions • Planetary and synoptic activity • Power spectra • Tropical waves (including Madden-Julian Oscillation)

http://intra.ecmwf.int/plots/d/inspect/_dir_diagnostics/Diagnostics/

Table 1: Products within the on-line 'Diagnostics Explorer': A 5D view of the IFS. All diagnostics are produced for operational forecasts (seasonal means) and experimental cycles ('E-suites'). Some diagnostics are produced for research experiments. 'Initial Tendency' diagnostics will be added. The aim is a seamless and efficient diagnosis of the entire forecasting and data assimilation system for the purpose of monitoring progress and informing development decisions. Other ECMWF Sections produce more detailed diagnostics for their particular IFS component.

4 Data assimilation: Observations and analysis increments

In the data assimilation process, the aim is to produce an 'analysis' that is as close to the observations as possible but also being (approximately) a valid model state. This analysis is then used as the initial conditions for a weather forecast. The data assimilation starts with a 'first guess' forecast initiated from a previous analysis. The 'analysis increments' are what is added to this first guess in the process of

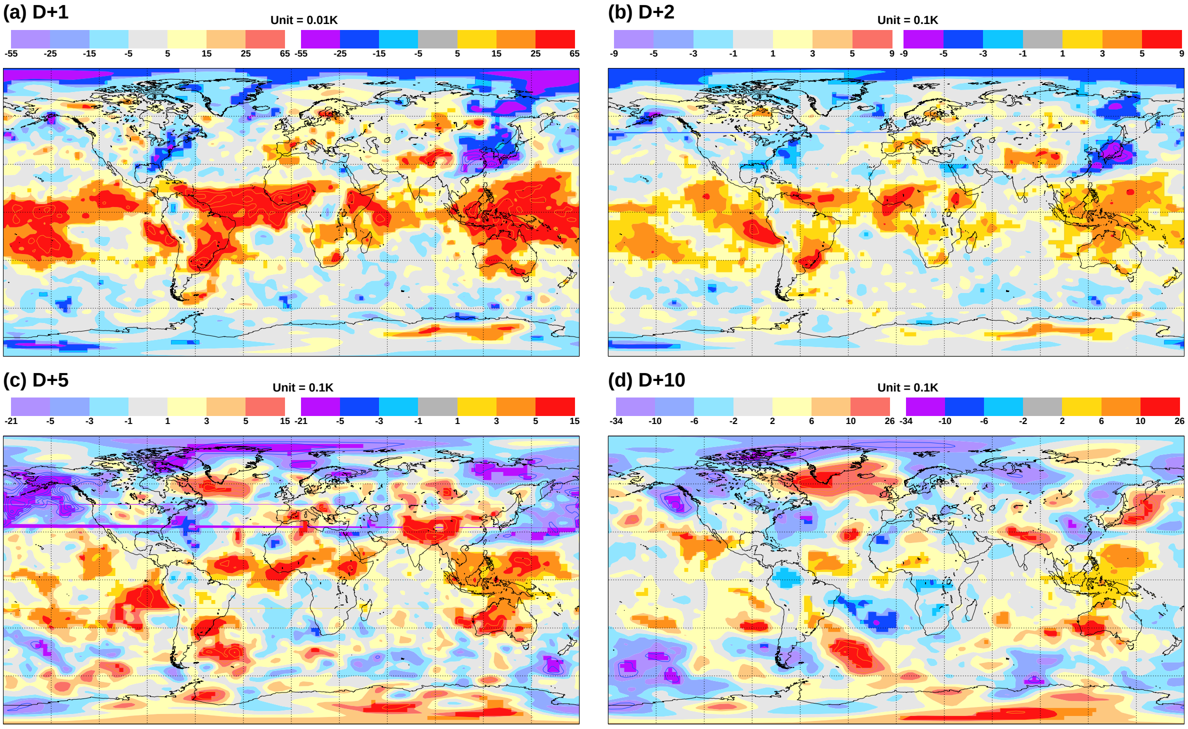

Figure 4: Mean forecast error for temperature at 500 hPa averaged over all forecasts initiated at 0 UTC and verifying within the season December–February 2008/9. The panels show the mean forecast error for a selection of forecast lead-times. (a) At a lead-time of one day (D+1). (b) D+2. (c) D+5. (d) D+10. Bold colours indicate that the mean forecast error is statistically significantly different from zero at a significance level of 5%. Contours are used to extend the colour shading scheme where necessary. The contour interval is the same as the shading interval.

arriving at the analysis. If the model used to make this first guess forecast has a bias (but that the observations are initially assumed to be unbiased; see later), then the analysis increments (averaged over sufficient data assimilation cycles) will be in the sense of correcting this model bias. Fig. 5a shows the operational analysis increments for 500 hPa temperature for the same December–February 2008/9 season as used for the forecast error results (Fig. 4). In the tropics, where the Day 1 forecast error indicated an erroneous warming by the model (Fig. 4a), the analysis increment shows a compensating cooling increment. Similar correspondence is apparent in the extratropical regions too.

Such temperature increments will only occur if there are ‘supporting’ observations. These observations do not need to be direct observations of temperature since any observable quantity that can also be derived from the model state has the potential to influence the analysis. For example, one could consider as such a quantity the brightness temperature as observed by the ‘AIRS’ infrared satellite channel 215. This brightness temperature represents a weighted mean of temperatures between about 700 hPa and 300 hPa; with the weight maximising at around 500 hPa. Using these weights, it is possible to derive the brightness temperature from the model state and thus make a comparison between the observed value and that predicted in the first guess forecast. In essence, the data assimilation process iteratively modifies the model state in order to minimise the observation minus first guess difference for all such derived (and underived) quantities (subject to other constraints). Fig. 5b shows the mean observation-minus-first-guess for this brightness temperature. The pattern agreement between the analysis-minus-first-guess (Fig. 5a) and the observation-minus-first-guess (Fig. 5b) indicates that AIRS channel 215 is one source of observations that support the increments.

The average volume of AIRS channel 215 observations assimilated during a data assimilation cycle

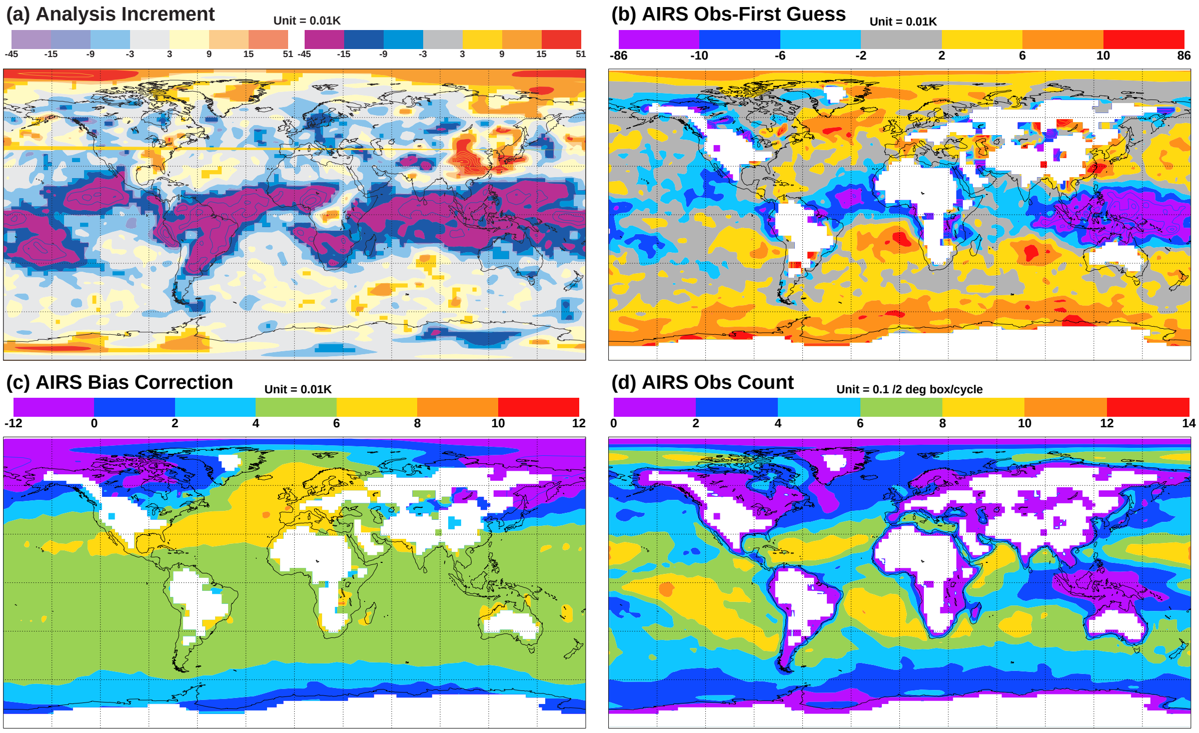

Figure 5: (a) Mean analysis increment of temperature at 500 hPa. (b) Mean "first guess departure" (observation minus first guess forecast) for the "AIRS" satellite channel 215. The weighting function for this channel maximises at about 500 hPa. (c) The mean bias correction applied to the AIRS observations. (d) The mean number of AIRS observations per 2^o grid-box per assimilation cycle. The plotted means are based on all 0 and 12 UTC data assimilation cycles within the season December–February 2008/9. The bold and pale colours in (a) have the same interpretation as in Fig. 4.

is indicated in Fig. 5(d). Note the lack of data usage over land areas and reduced usage over the cloud-affected Indonesian warm-pool region. Other observations fill these gaps. For example 'AMSUA' microwave channel 5 and radiosondes also provide information on 500 hPa temperatures and are not affected by clouds.

Figure 5(c) shows the bias correction applied to this AIRS data by the data assimilation system. The correction is deduced through a large-spatial-scale 'fitting' of the observations to the first guess forecast and is known as variational bias correction, 'VarBC'. It is possible that VarBC could mis-attribute some systematic model error to observation bias. However, this is difficult in practice since some data (AMSUA microwave channel 14, radiosonde and 'radio occultation' data) are considered accurate enough to not need bias correction and, in addition, the ∼6 million observations assimilated during each cycle of the data assimilation system are thought to provide ∼20 independent vertical modes of information. While it is vital to diagnose observation usage, Fig. 5(b) demonstrates that VarBC does not remove all systematic differences between the first guess and the observations – and the remaining differences are more likely to be attributable to model error.

Hence it can be argued that model error is optimally diagnosed within the data assimilation system and that data assimilation should be used as a tool within model physics development. This proposed strategy is consistent with the fact that data assimilation is itself beginning to explicitly represent model error (see, e.g., Trémolet, 2007, on developments in 'weak constraint 4D variational data assimilation'). It is clear that there are synergies here that could be harnessed to advance the performance of forecasting systems.

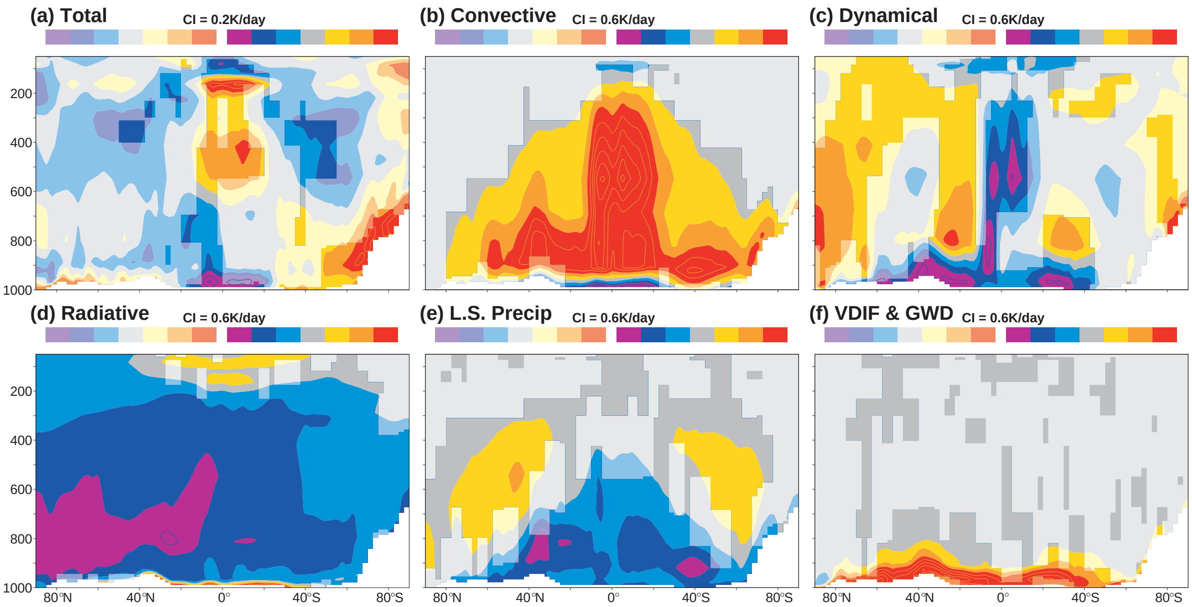

Figure 6: Temporal and zonal-mean initial temperature tendencies for processes as indicated. Average is over December 2008, 4 forecasts per day, with tendencies accumulated between T+1 and T+7 hr. Based on model cycle is 33R1 run at T_L159 L91. Values are accumulated on model levels. The y-axis shows approximate pressure in hPa. Bold and pale colours have the same interpretation as in Fig. 4.

5 Initial tendencies

It has been argued that analysis increments are indicative of model error but how can this error be attributed to a particular process within the model? Here some results are presented that are based on 4D variational data assimilation experiments using a 6-hr assimilation window for the month of December 2008. From the resulting analyses, short forecasts have been initiated. Figure 6 shows zonal-mean temperature tendencies integrated over leadtimes 1–7 hr and over the four consecutive forecasts made each day and then averaged over the month of December. Figure 6 (b–f) show these ‘initial tendencies’ for individual physical process (and the dynamics). The balance between individual processes highlights, for example, how the radiation, Fig. 6(d), (and its impact on the vertical diffusion of surface sensible heat fluxes, Fig. 6(f)) destabilises the vertical profile, while the convection, Fig. 6(b), acts to restore equilibrium. The balance between all processes is not complete, however, since Fig. 6(a) indicates a residual, net (total) tendency with, for example, a warming in the tropical mid troposphere and a warming/cooling dipole above. There is also an interesting cooling on the extratropical upper troposphere.

Figure 7(a) shows the zonal mean analysis increment of temperature based on the corresponding set of the 6-hr window data assimilations. Notice how similar this is to (minus) the total initial tendency, Fig. 6(a). If the upper-tropospheric dipole does represent model error, then the individual process tendencies suggest that it is likely to be associated with convection, dynamics and/or radiation since these are the processes that have strong magnitude at this ellevation. Consistent with this reasoning, Fig. 7(b) shows how the analysis increment is changed with the inclusion of the new ‘McICA’ radiation scheme (which incorporates a Monte Carlo approach to cloud overlap). The change acts to reduce mean increments (*c.f.* Fig. 6a and b) and also reduces the root-mean-square of increments in the mid and upper troposphere, Fig. 6(c).

Initial tendencies have been applied in a number of other contexts including the assessment of climate models (Rodwell and Palmer, 2007), the impacts of model aerosol changes (Rodwell and Jung, 2008b), the over-active Asian Monsoon in the ECMWF forecast model (Rodwell and Jung, 2008a) and, more

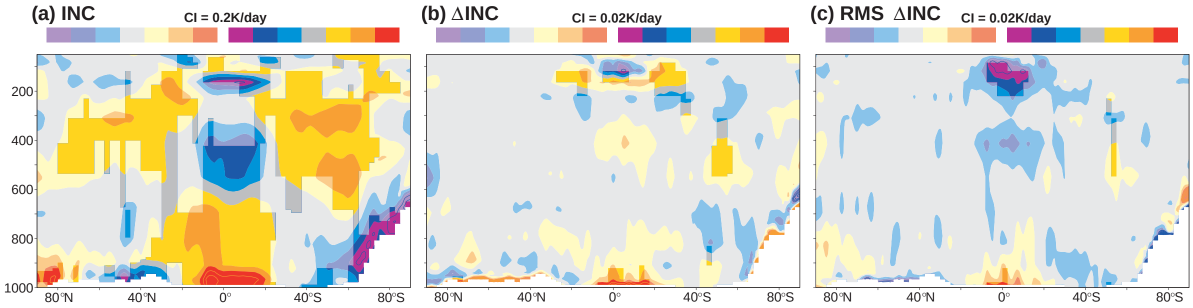

Figure 7: Temporal and zonal-mean analysis increments of temperature. Average is over December 2008, 4 data assimilation cycles per day. (a) Total increment for model cycle is 33R1. (b) Change in mean increment when the McICA radiation scheme is introduced. (c) Change in RMS of increments when the McICA radiation scheme is introduced. Data assimilation (outer-loop) uses resolution T_L159 L91. Values are accumulated on model levels. The y-axis shows approximate pressure in hPa. Bold and pale colours have the same interpretation as in Fig. 4.

recently, the physics of the Madden-Julian Oscillation.

6 Scale dependent error

Figure 8(a) shows (solid) mean-square-error in Z500 for the northern mid-latitudes (35–65^{o}N) for March–May 2008 (blue) and 2009 (red). It is clear that 2008 was better predicted than 2009 at all lead-times. Moreover (as indicated by the 5% significance circles), the difference was sufficiently large that it could not be accounted for simply by uncertainties in sampling (from the same distribution). This suggests that there was either a degradation in the forecast system or that the circulation in 2009 was different (sampled from a different distribution). If the circulation was different, it could have been inherently harder to predict (perhaps with more synoptic activity) or it could have involved flow-types that the forecast system has particular problems with. Clearly this degradation in error could have represented a serious issue for ECMWF.

The dotted and dashed curves in Fig. 8(a) show ‘activity’ in the forecast and analysis, respectively. The activity is quantified as (twice) the mean-squared anomaly from climatology. It can be seen that 2009 was actually less active than 2008. As an aside note that, at the limit of no predictability, the error curve should match the activity curves; hence there is substantial skill remaining in the forecasts even at D+10.

It is possible to linearly decompose mean-squared error (or activity) into different spatial scales (based on total wavenumber for the case of the globe, or zonal wavenumber for the case of latitudinal bands). Figure 8(b) shows the contribution to the total error and activity for the zonal wavenumber bands 0–3 (thick; representing planetary waves) and 4–14 (thin; representing synoptic scales). It can be seen that 2009 did have more synoptic activity than 2008 and this probably explains the increased synoptic error in 2009 (as indicated by the solid thin curves). However, the striking observation is that there was much less planetary-scale activity in 2009 and yet planetary-scale error was significantly larger. It is this planetary-scale contribution that is statistically significant and that dominates the total changes seen in Fig. 8(a).

Fig. 8(c) shows the same errors and activity based on the ERA-Interim reanalyses. Since ERA-Interim uses a fixed (older) forecast cycle (31R2) and fixed model resolution (T_L255), it is evident that this issue with the planetary waves is not the result of recent system updates.

The ECMWF forecast system would appear, therefore, to have difficulties with the planetary wave pattern experienced in 2009. Figure 9(a) and (b) show March–May Z500 anomalies for 2008 and 2009,

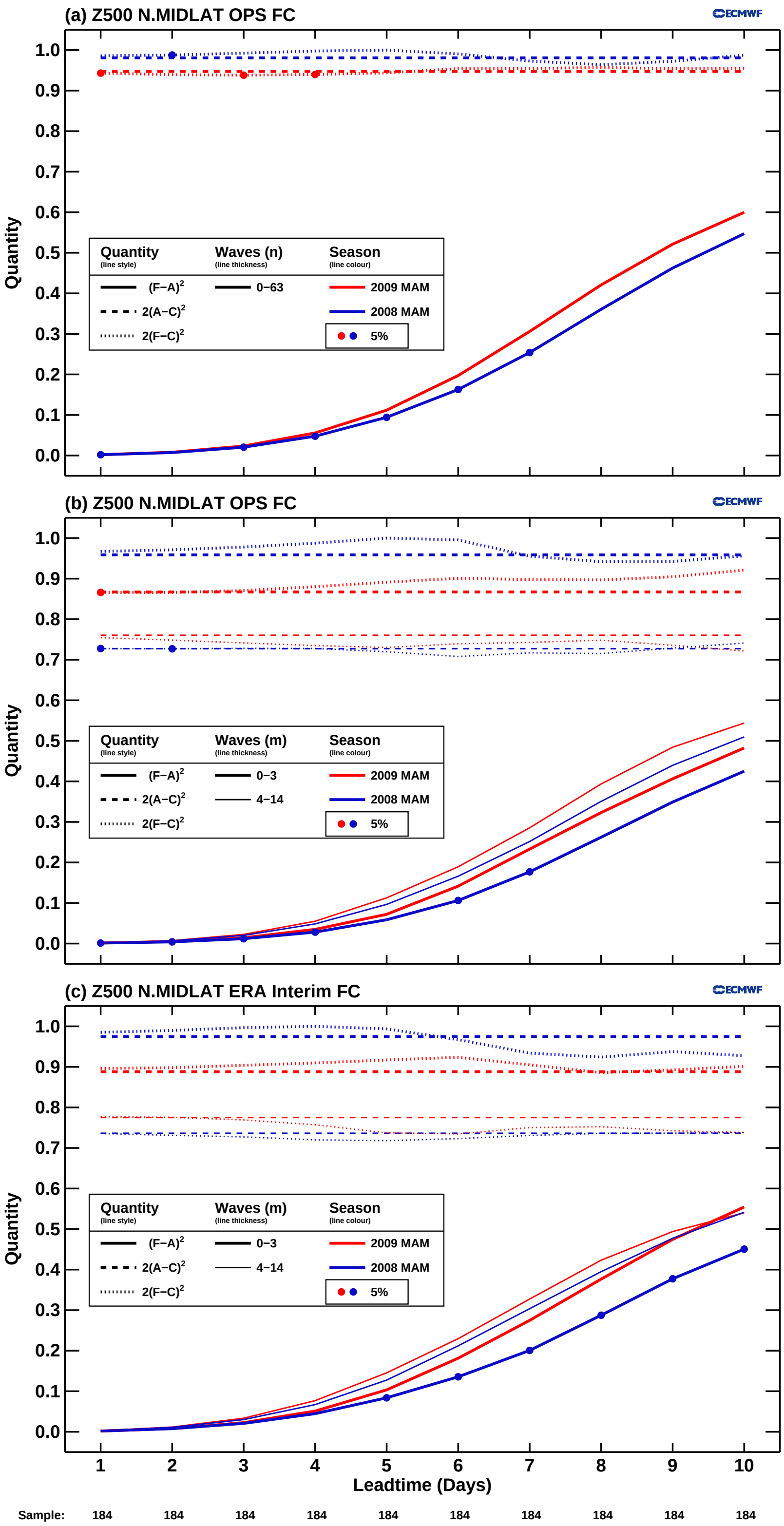

Figure 8: Mean-squared error (solid), analysis 'activity' (dashed) and forecast 'activity' (dotted) in northern mid-latitude (35–65^{o}N) Z500 for March–May, 2008 (blue) and 2009 (red). 'Activity' is defined in the text. (a) Zonal wavenumbers 0–63 from operational analyses and forecasts. (b) Zonal wavenumbers 0–3 (thick) and 4–14 (thin) from operational analyses and forecasts. (c) as (b) but using re-analyses and re-forecasts made within the ERA-Interim re-analyses project. Dots highlight the year that is statistically significantly best at the 5% level. All data are normalised by the largest value represented.

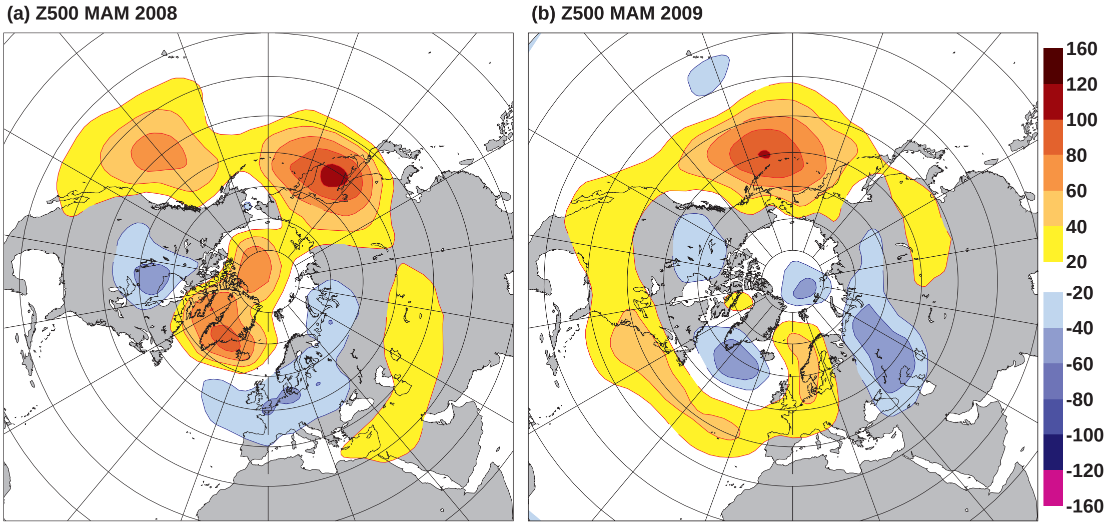

Figure 9: March–May mean 500 hPa geopotential height anomaly relative to 'ERA-40' 1962–2001 climatology. (a) 2008. (b) 2009. The contour interval is 20m.

respectively. It can be seen that large-scale circulation anomalies were different for the two years with a 'filled-in' Aleutian Low and a strengthened North Atlantic Oscillation in 2009. Further diagnostic work is required to understand why the forecast system has difficulties with this large-scale anomaly circulation pattern.

7 SEEPS: A new score for the verification and diagnosis of precipitation

Contours in Fig. 10 show (a) observed (*i.e.* analysed) and (b) D+4 forecast Z500 verifying at 12 UTC on 23 August 2008. The correspondence is indicative of the improvements in large-scale skill over recent years. However, it is clear that Z500 is not sufficient to characterise the entire flow. Precipitation (shaded), for example, is rather poorly predicted over Europe in this example. This emphasises the need to monitor other aspects of the forecast; for example aspects of direct relevance to the user community and aspects representative of diabatic processes. Since precipitation is user-relevant and a consequence of diabatic processes, it would appear to be a natural choice. Recent work in the Diagnostics Group (Rodwell et al., 2010) has focused on developing a new approach to the verification of precipitation that should be particularly useful for monitoring progress and for guiding development decisions, as well as for the initial diagnosis of forecast error. The approach, or score, is called here 'Stable Equitable Error in Probability Space' (SEEPS). It is a three-category error score that incorporates four key principles:

1. Error measured in 'probability space' (Ward and Folland, 1991). The climatological cumulative distribution function (see later) is used to transform errors into probability space. This allows the difficult distribution of precipitation to be accommodated in a natural way and reduces sampling uncertainty associated with extreme (possibly erroneous) data.

2. Equitability (Gandin and Murphy, 1992). By applying the equitability constraints, a forecast system with skill will have a better expected score than a random or constant forecast system. In addition, scores from different climate regions can be readily combined.

3. Refinement (Murphy and Winkler, 1987). A constraint is devised to encourage a forecast system to predict all possible outcomes; thereby promoting a better distribution of forecast categories.

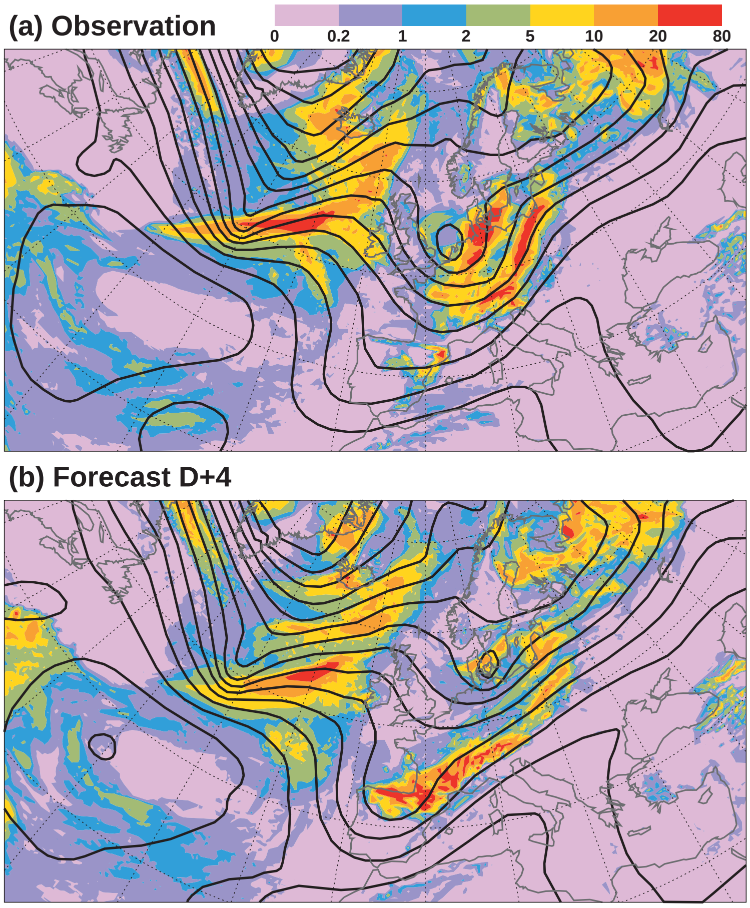

Figure 10: 500 hPa geopotential height field (Z500, contoured with interval 50m) and 24-hour accumulated precipitation (shaded, mm). (a) 'Observations': analysed Z500 and short-range (D+0–D+1) forecast precipitation centred at time 12 UTC on 23 August 2008. (b) Forecast: D+4 forecast Z500 and D+$3\frac{1}{2}$–D+$4\frac{1}{2}$ forecast precipitation verifying at the same time.

4. Reduction of sensitivity to sampling uncertainty, for sufficiently skillful systems, by ensuring that all perfect forecasts are accorded zero error.

Figure 11 shows the cumulative distribution function (cdf) based on a climatology of station observation data for Balmoral, Belgium in October. It can be seen that the climatological probability of dry weather at this location and in this month is $p_1 = 0.45$. In SEEPS, the three categories ('dry', 'light precipitation' and 'heavy precipitation') are defined by the climatological probabilities p_1, p_2, p_3. Experimentation suggests that $p_2/p_3 = 2$ is a good way to define 'light' and 'heavy' precipitation. For Balmoral in October, this gives a probability for light precipitation of $p_2 = \frac{2}{3}(1 - p_1) = 0.37$. The precipitation amount corresponding to the probability $p_1 + p_2 = 0.45 + 0.37 = 0.82$ is 5mm. Hence heavy precipitation is defined as (24 hour) accumulations greater than 5mm. Since the values of p_1, p_2, p_3 are dependent on location and month of the year, the definitions of 'light' and 'heavy' precipitation also vary. In this way, the SEEPS score adapts to the local climate and assesses the salient aspects of local weather.

The SEEPS error matrix, $\{s_{vf}\}$, is given in Table 2, where f is the forecast category and v is the verifying observation category. The sample-mean SEEPS score, $\tilde{S}$, is then calculated using

$$\tilde{S} = \sum_{v,f} \tilde{p}_{vf} s_{vf} \quad , \tag{1}$$

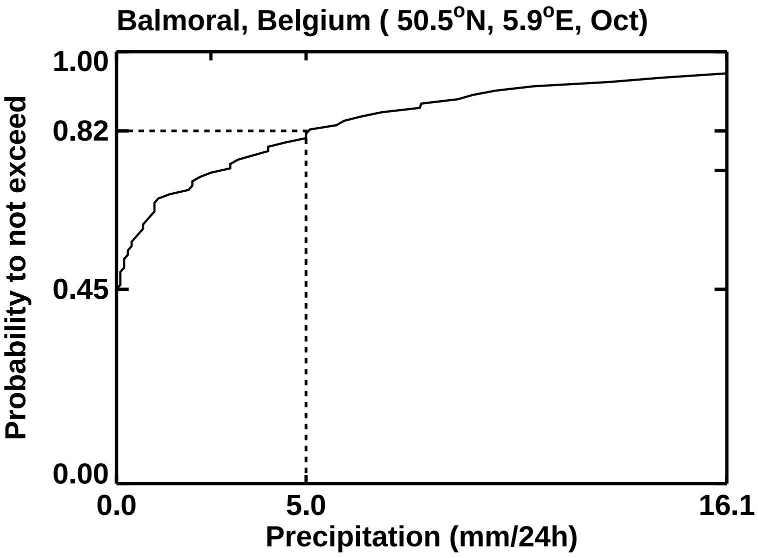

Figure 11: Cumulative distribution of 24-hour precipitation (12–12UTC) for Balmoral, Belgium in October based on 1980–2008 observations. The extreme right of the graph corresponds to the 95th percentile of the distribution. Dotted lines indicate the sub-division of the wet days in the ratio 2:1.

				Obs	
Prob			p_1	p_2	p_3
Cat				v	
			1	2	3
FC	f	1	0	$\frac{1}{1-p_1}$	$\frac{1}{p_3}+\frac{1}{1-p_1}$
		2	$\frac{1}{p_1}$	0	$\frac{1}{p_3}$
		3	$\frac{1}{p_1}+\frac{1}{1-p_3}$	$\frac{1}{1-p_3}$	0

Table 2: Error matrix for SEEPS error score. f is the forecast category and v is the verifying observation category. p_1, p_2 and p_3 are the climatological probabilities of ‘dry’ conditions, ‘light precipitation’ and ‘heavy precipitation’, respectively.

where $\{\tilde{p}_{vf}\}$ is the sample joint distribution.

The Gerrity sequence of skill scores (Gerrity, 1992) is derived by taking the mean of $n-1$ 2-category skill scores that are asymptotically equivalent (*i.e.* for large sample size) to the ‘Peirce Skill Score’ (Peirce, 1884). In a similar way, the SEEPS error score can also be written as the mean of two 2-category error scores of the form shown in Table 3 (where p and q are the climatological probabilities of categories 1 and 2, respectively. The first 2-category score in the mean uses category 1 as dry weather $(p = p_1)$ and category 2 as ‘light’ *or* ‘heavy’ precipitation $(q = p_2 + p_3)$. The second 2-category score uses category 1 as dry weather *or* ‘light’ precipitation $(p = p_1 + p_2)$ and category 2 as ‘heavy’ precipitation $(q = p_3)$. Interesting, the 2-category score defined by the error matrix in Table 3 is also asymptotically equivalent to (1-) the Peirce Skill Score. The only difference between SEEPS and (1-) the Gerrity Skill Score is that SEEPS is less sensitive to sampling uncertainty for sufficiently skillful forecast systems. However, this difference is important because it enables sample-mean SEEPS scores to reflect more precisely the

			Obs	
		Prob	p	q
			v	
		Cat	1	2
FC	f	1	0	$\frac{1}{q}$
		2	$\frac{1}{p}$	0

Table 3: Error matrix for a 2-category score from which SEEPS can be constructed. f is the forecast category and v is the verifying observation category. p, q are the climatological probabilities of categories 1 and 2, respectively.

true skill of the forecast system.

Other desirable attributes, common to both SEEPS and the 3-category Gerrity score, are penalties for 'hedging' (whereby forecasts for 1 category are altered to another category with no physical insight) and the promotion of refinement and discrimination.

7.1 Precipitation errors identified by SEEPS

SEEPS can be used as a first step in the diagnosis of forecast error. Fig. 12(a) shows observed 24-hour accumulated precipitation (in mm) on 16 December 2008, and Fig. 12(b) shows the corresponding D+4 forecast precipitation. (D+4 is chosen because of ECMWF's mandate to improve medium-range forecasts). Notice that large parts of northern Europe were predicted to have drizzle but were actually 'dry' (pink). Since this region is generally wet in December (Fig. 12c) and an incorrect forecast for a likely category is strongly penalised by SEEPS, the differences in precipitation categories (*c.f.* Fig. 12d and e) lead to relatively large SEEPS scores (Fig. 12f). This partly explains why the mean European score for this forecast was one of the worst in 2008. Verification at the dry/wet boundary has important physical significance because of the existence of positive feedbacks with latent heating. From the users' perspective, of course, drizzle is also of great relevance. Hence it is desirable that SEEPS can highlight this error.

Clearly, SEEPS can also identify other forecast errors such as a failure to predict heavy large-scale precipitation and the incorrect positioning of convective cells.

Note that the scores in Fig. 12(f) are plotted with variable sizes to indicate their relative weight within an area-mean score – deduced to take account of the heterogeneous observation density.

7.2 Extratropical-mean SEEPS scores

Area-mean scores have been produced, taking the station network density into account, for the period 1995-2008. Plots for the Extratropics (north of 30^{o}N *and* south of 30^{o}S), based on ∼2000 station observations per day, are shown in Fig. 13. Figure 13(a) shows the annual mean scores based on the 12UTC operational forecasts as a function of leadtime. The colours indicate the years. There is a general progression to lower errors over these 14 years. The black curve shows the most recent year (2008).

The 70% confidence intervals plotted in Fig. 13(a) show the degree of uncertainty in the annual means. They are deduced from the daily scores taking autocorrelation into account following the methodology

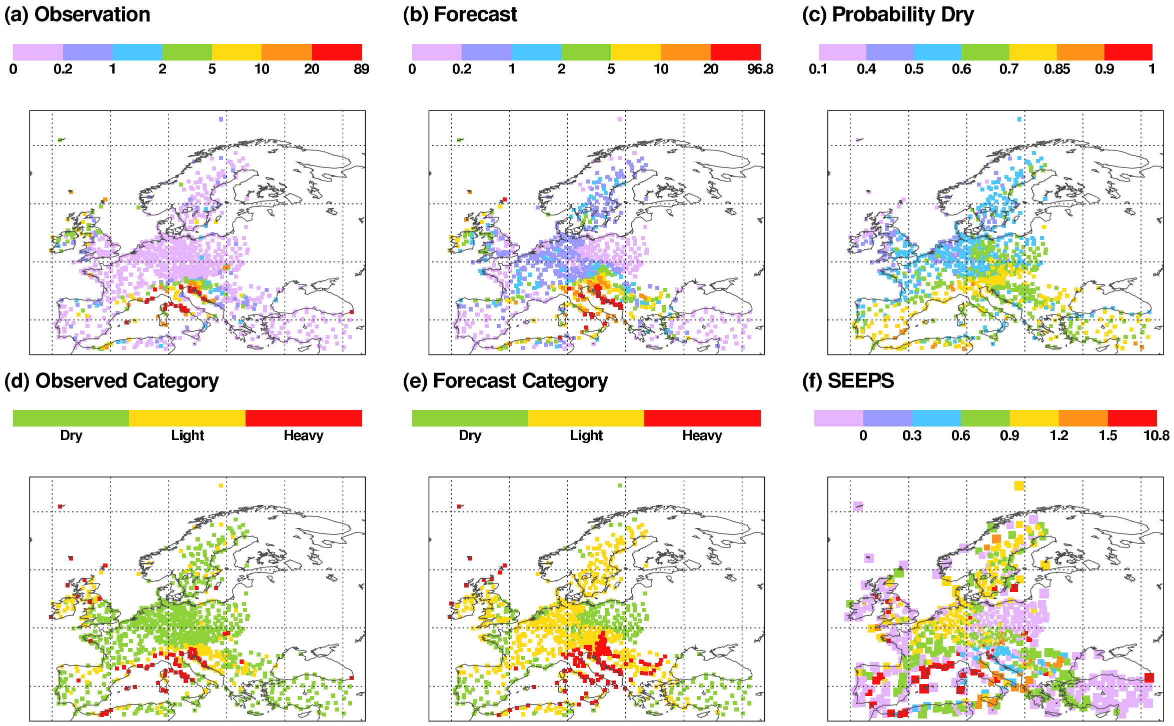

Figure 12: (a) Observed precipitation accumulated over 24 hours 2008/12/15 12UTC to 2008/12/16 12UTC. (b) Forecast precipitation accumulated over leadtimes 72 to 96 hours and valid for the same period as the observations. (c) Probability of a 'dry' day in December, based on the 1980–2008 climatology. (d) Observed precipitation category. (e) Forecast precipitation category. (f) SEEPS. Units in (a) and (b) are mm. Squares in (f) are plotted with areas proportional to the weight given to each station in the area-mean score.

of von Storch and Zwiers (2001). If one mean lies within the confidence interval of another, then there is no significant difference. If confidence intervals just touch, then mean scores are significantly different at the 14% level, assuming equal variances. It can be seen that it is generally not possible in year y to demonstrate that forecasts are better than in the previous year $y-1$: it takes a few years for improvements to become unequivocal.

It can be seen that by D+10, the SEEPS score is tending towards 1. This is one of the desirable features associated with equitability: by construction, expected SEEPS scores for all stations and all months of the year lie between 0 and 1 and this makes the aggregation of all the stations within an area a meaningful and useful concept (despite sub-regions having very different climates).

Fig. 13(b) shows (light green) daily SEEPS scores at D+4 for the same operational forecasts. The general improvement over the years is clearly apparent when a 365-day running mean is applied (black). The 31-day running mean (dark green) highlights a seasonal cycle in SEEPS scores. This feature is common to many precipitation scores and reflects the fact that large-scale precipitation is generally easier to predict than convective precipitation.

Fig. 13(c) shows the annual-mean of the leadtime at which the SEEPS score for each daily forecast first reaches a value of 0.6. The value of 0.6 was chosen because it corresponds approximately to the present annual-mean score at D+4. The red curve relates to the operational forecast data shown in Fig. 13(a) and (b). The gains in leadtime amount to ∼2 days over the 14-year period. The graph is annotated to show when the model's resolution was changed during this period and also to show when one key model cycle (25R4) was introduced. This model cycle had many updates that could have directly affected the forecast of precipitation. However, there were 40 packages of updates applied to the operational

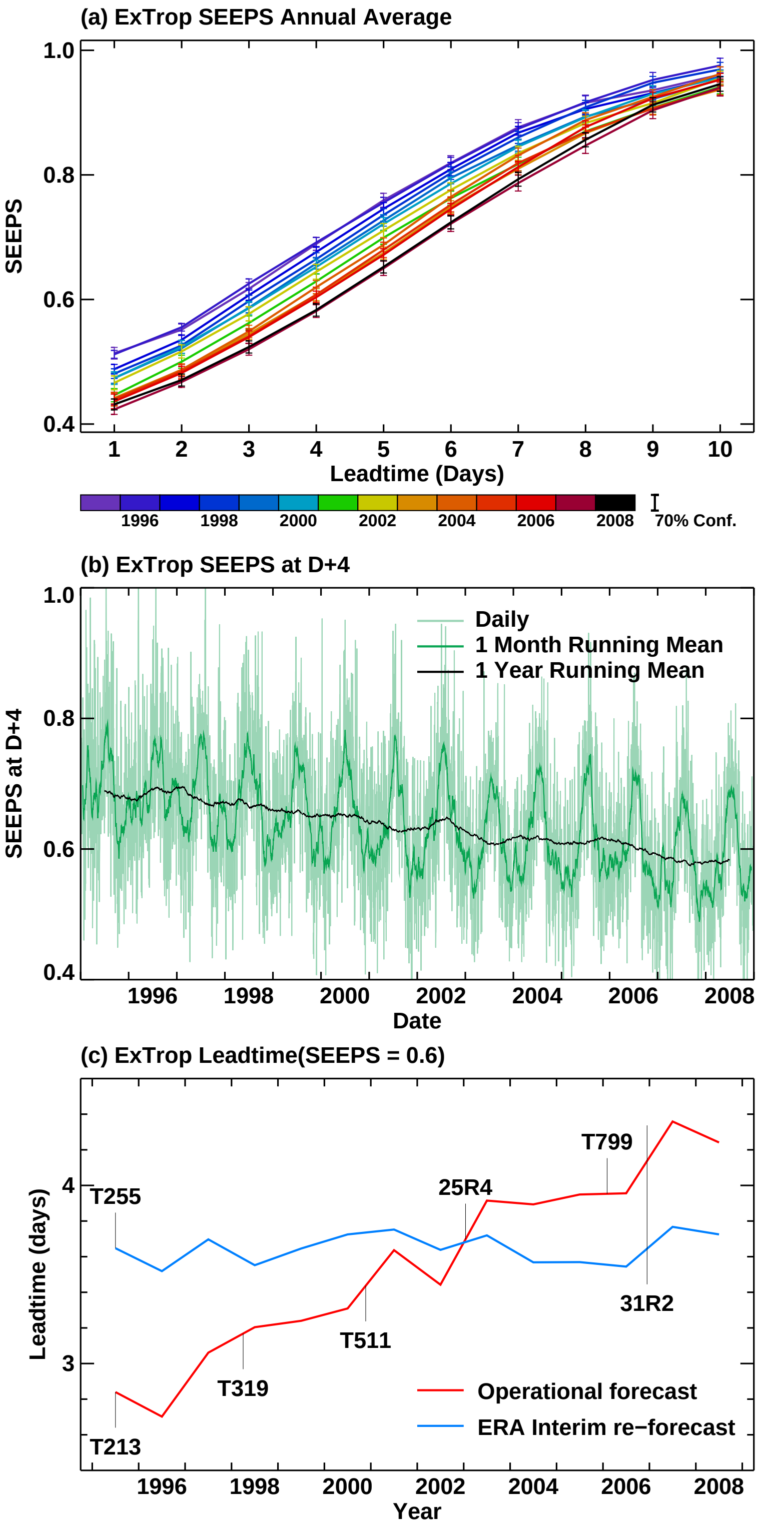

Figure 13: Extratropical-mean SEEPS results. (a) Annual-mean of daily operational scores as a function of lead-time. 70% confidence intervals for these annual means are indicated. (b) Time-series of operational scores at D+4 with running means as indicated. (c) Annual-mean lead-time at which the score rises to 0.6 based on the operational forecasts and on the forecasts made during the production of the ERA-Interim re-analysis, as indicated. The extratropical average is over the combined region north of 30^{o}N and south of 30^{o}S, taking account of observation density.

data assimilation and forecasting system over this period and many of these will have contributed to the improvement.

The blue curve in Fig. 13(c) shows comparable results for re-forecasts made within the ERA-Interim re-analysis project. ERA-Interim is based on a single model cycle (31R2) and a single model resolution (T255). The date that this cycle was first used in the operational forecast system (12 December 2006) is also indicated on the graph. The differences between the red and blue curves at this date highlight the impact of resolution. The flatness of the ERA-Interim SEEPS curve is striking. It indicates that the increase in available sources and volume of data used to initialise the forecast (a 100× increase over this period) has had almost no lasting impact on the prediction of precipitation. Instead, the lasting improvements in the extratropical operational scores must be due to improvements to model physics, increases in model resolution, and to the way the data assimilation system has improved to better use the available observations. New data sources will target more directly the hydrological cycle so the conclusions from the 1995–2008 period may not hold in future.

8 Discussion

In order for Diagnostics to continue to promote forecast system development, it must adapt to the new reality; of ever more skilful large-scale medium-range prediction, higher model resolution and an increased emphasis on weather parameters and severe weather.

The issue of uncertainty in diagnosis is increasingly important. Examples of how two different sources of uncertainty can be reduced have been discussed.

- By diagnosing error earlier-on in the forecast, before interactions between the physical processes and the resolved flow have had time complicate the picture.
- By designing scores that are less sensitive to sampling uncertainties.

It is also important to benefit from the synergies that derive from a more seamless diagnosis of the forecast system. In this context, the benefits of diagnosing model error at short lead-times (using ‘initial tendencies’) will be enhanced when model error is explicitly represented in data assimilation.

Recent efforts have focused on the development of an equitable precipitation score that can be used to monitor and guide system development and also be used as a tool in the diagnosis of forecast error. The scope for opening-up new avenues of diagnostic research with this score is becoming apparent. For example, systematic precipitation score differences are emphasising the need to predict particular synoptic situations better – such as depressions over the Mediterannean.

References

Gandin, L. S. and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361–370.

Gerrity, J. P., 1992: A note on Gandin and Murphy’s equitable skill score. *Mon. Wea. Rev.*, **120**, 2709–2712.

Klinker, E. and P. D. Sardeshmukh, 1992: The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements. *J. Atmos. Sci.*, **49**, 608–627.

Murphy, A. H. and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454.

Rodwell, M. J. and T. Jung, 2008a: The ECMWF 'diagnostic explorer': A web tool to aid forecast system assessment and development. ECMWF Newsletter 117, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

Rodwell, M. J. and T. Jung, 2008b: Understanding the local and global impacts of model physics changes: An aerosol example. *Quart. J. Roy. Meteor. Soc.*, **134**(635), 1479–1497.

Rodwell, M. J. and T. N. Palmer, 2007: Using numerical weather prediction to assess climate models. *Quart. J. Roy. Meteor. Soc.*, **133**(622 A), 129–146.

Rodwell, M. J., D. S. Richardson, and T. D. Hewson, 2010: A new equitable score suitable for verifying precipitation in NWP. *Quart. J. Roy. Meteor. Soc.*, p. In Review.

Trémolet, Y., 2007: Model error estimation in 4d-var. *ECMWF Technical Memorandum*, **520**, available at http://www.ecmwf.int/publications/.

von Storch, H. and F. W. Zwiers, 2001: *Statistical Analysis in Climate Research*. Cambridge University Press. 484 pp.

Ward, M. N. and C. K. Folland, 1991: Prediction of seasonal rainfall in the north Nordest of Brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol.*, **11**, 711–743.