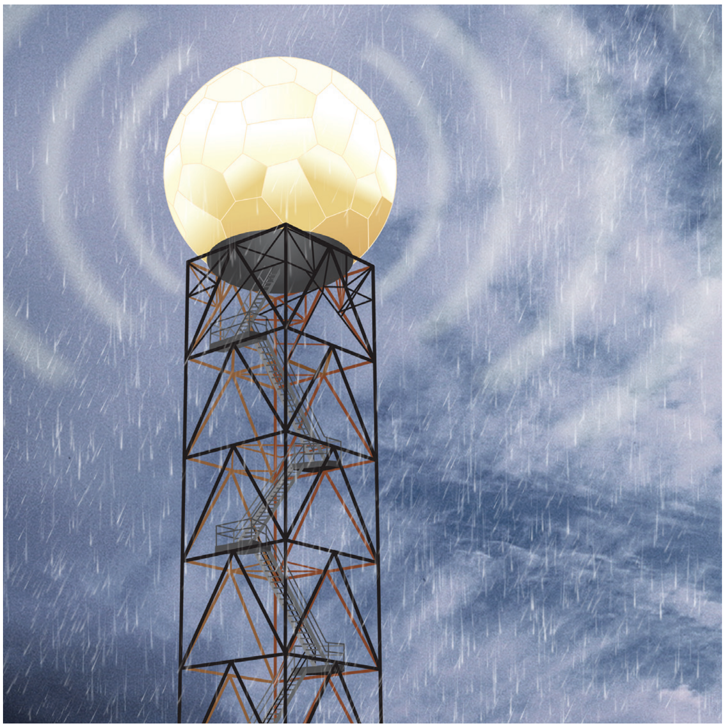
ECMWF Feature article

from Newsletter Number 117 – Autumn 2008

METEOROLOGY

The ECMWF 'Diagnostics Explorer': A web tool to aid forecast system assessment and development



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

doi:10.21957/is14d23143

This article appeared in the Meteorology section of ECMWF Newsletter No. 117 – Autumn 2008, pp. 21-29.

The ECMWF ‘Diagnostics Explorer’: A web tool to aid forecast system assessment and development

Mark Rodwell, Thomas Jung

Many people may be familiar with the various plots available online at ECMWF giving forecasts and forecast verification data. Here, we highlight a new set of online plots that help diagnose in more detail the performance of many aspects of the ECMWF Integrated Forecast System (IFS). These include the data assimilation system, weather forecasts and the model climate. This diagnostics package (produced by the ECMWF Diagnostics Section) is called the “Diagnostics Explorer”. The present contents of the Diagnostics Explorer are summarised in Table 1. There are three main components.

- Data assimilation section includes diagnostics on observation usage and analysis increments.
- Weather forecast section includes diagnostics on forecast error and scale-dependent scores.
- Model climate section includes a wide selection of mean and variability diagnostics for both the atmospheric and coupled models.

Plots for both the data assimilation and weather forecast sections are available as seasonal means of the operational IFS (where they are compared with the same season in the previous year) and also for the tests of the experimental IFS suites (compared to the operational suite).

This article aims to introduce a representative sample of the diagnostics available in the Diagnostics Explorer and to some of the ways these diagnostics can be used. In particular, ‘seamless’ approaches to system diagnosis are highlighted whereby products from the data assimilation, weather forecast and model climate components can be used together to gain a better insight into the IFS. While the examples shown are of interest in their own right, they should primarily be considered as examples of how the Diagnostics Explorer can be used more generally.

IFS Component	Diagnostics
Data assimilation	***Observation space – observation usage*** • Many data sources including 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)

Table 1 Summary of the present diagnostics available on the ‘Diagnostics Explorer’ website.

Assessment and interpretation of weather forecast error

Some of the most common scores used to assess weather forecast skill are based on 500 hPa geopotential heights. One example would be northern hemisphere anomaly correlations as a function of forecast lead time. A different perspective is offered by the Diagnostics Explorer.

Figure 1a shows zonal-mean root-mean-square (rms) errors in geopotential at day 5 as a function of height for the March to May season of 2005. Intelligent shading intervals are designed to cover most of the plot without being dominated by extreme values. Contours are then used, where necessary, to capture these extreme values. In the figure, the most extreme errors are found within the mid-latitude jets at around 300 hPa. Using the Diagnostics Explorer, it is possible to find out how these errors have changed over the subsequent years.

The plots in Figures 1b, 1c and 1d show the change in zonal-mean rms error in day 5 geopotential forecasts between adjacent years. Notice that statistically significant differences are indicated by bold colours (see Box A).

- ***Changes from 2005 to 2006 – Figure 1b.*** There is a reduction in errors in the southern hemisphere and above 100 hPa in the tropics, along with some degradation around 200 hPa in the tropics. While geopotential is not the best choice for examining the tropical atmosphere, consistent results are seen in the Diagnostics Explorer for temperature errors and these are associated with the implementation of a higher vertical resolution around the tropopause.
- ***Changes from 2006 to 2007 – Figure 1c.*** The previous tropical tropopause degradation is reversed – this improvement is associated with changes in the physical parametrization schemes including the introduction of a ‘Monte Carlo’ cloud over-lap scheme (*Morcrette et al.,* 2007). There are improvements in the northern hemisphere and the largest of these are statistically significant.
- ***Changes from 2007 to 2008 – Figure 1d.*** This plot is generally blue – indicating a reduction in medium-range forecast error.

Taking all the years together, it is clear that the general trend has been to reduce forecast errors.

Decreasing rms errors are generally indicative of improved skill but they can also be associated with diminishing “activity” in the model. Clearly, it is important to check for changes in activity as well as error from one model cycle to the next. In addition it is useful to know if changes in error and activity occur at planetary or synoptic scales.

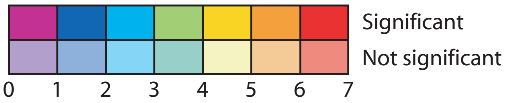
Figure 2 shows forecast error (solid) and forecast activity (dotted) for planetary and synoptic scales for 500 hPa geopotential height in the northern mid-latitudes. This plot is part of the experimental suite comparison of model cycle 32r3 (Cy32r3) against the previous operational model cycle 32r2 (Cy32r2). For lead-times of 1 to 4 days, both planetary-scale error (solid, thick) and synoptic-scale error (solid, thin) are reduced in Cy32r3 (blue) compared to Cy32r2 (red). The blue circles indicate statistical significance.

Statistical significance testing

A

To demonstrate that one model has a better mean score than another, it is essential to show that the difference in scores is large compared to the uncertainty in the estimated means. This is why statistical significance is assessed wherever possible in the Diagnostics Explorer. The assessment is made using the two-sided Student’s t-test and takes account of serial correlation in the data. The significance level used is 5%. Wherever the dates for both timeseries are the same (for example in experimental-suite tests), the more powerful one-sample t-test is performed.

For the analysis increments and forecast error plots, a ‘dual colour palette’ has been developed to display the significant and insignificant differences. In the example in the colour bar below, a value of 3.5 would always be coloured green – a bold green is used if the value is statistically significant and a pale green is used if it is not significant. This dual colour palette draws the IFS developer’s attention away from the insignificant differences that could otherwise cause unnecessary concern. Other plots use cross-hatching to indicate significance.



Significance testing requires access to at least 30 times more data than that stored as averages. It is not feasible to have this much data online at present and so the Diagnostics Explorer does not produce plots ‘on-demand’. Instead, for every season and every experimental-suite test over 7,000 plots are produced to allow the user a lot of flexibility to explore the IFS.

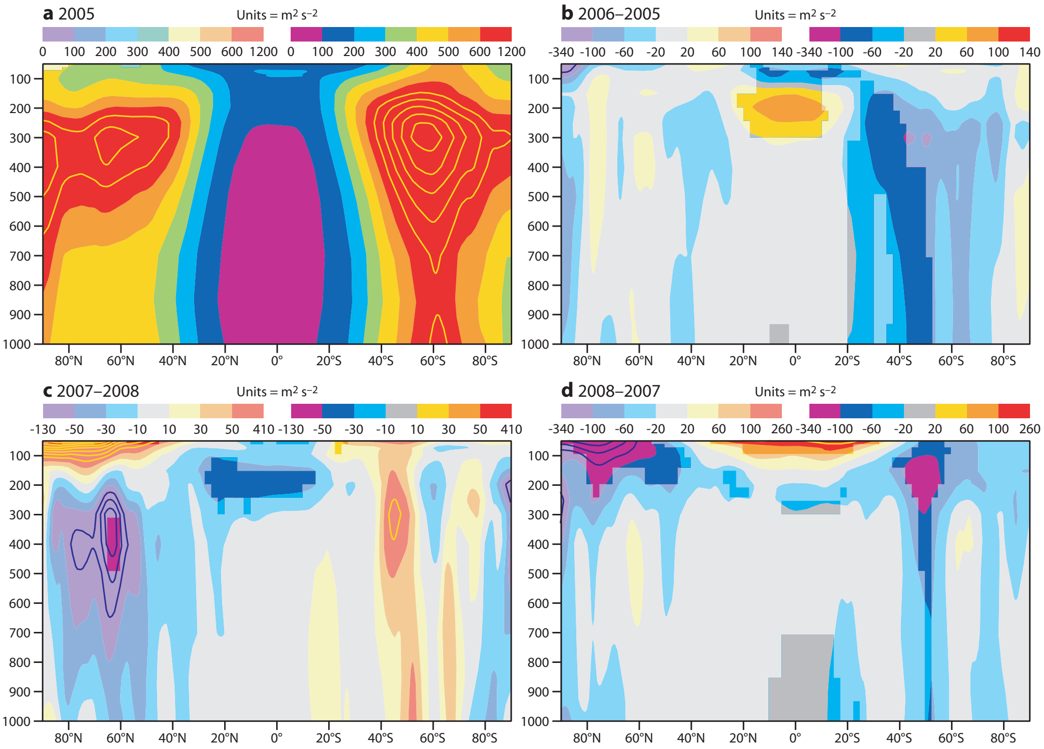


Figure 1 Zonal-mean of rms error of the five-day forecast of geopotential as a function of height for the March to May season for (a) 2005, (b) difference between 2006 and 2005, (c) difference between 2007 and 2006 and (d) difference between 2008 and 2007. Statistically significant values (at the 5% level) are shaded in bold.

Since the dotted curves indicate that Cy32r3 is more active than Cy32r2 at both spatial scales, the reduction in error must be associated with increased skill. By comparing with the observed activity in Figure 2 (dashed), the increase in synoptic-scale activity in the forecast (dotted, thin) is seen to improve the previous under-representation of activity at these scales. However, planetary-scale activity in Cy32r3 (blue, dotted, thick) grows with forecast lead-time and this over-corrects the previous under-estimation.

Forecast skill for a given spectral band is lost when the error curve (solid) meets the activity curves. It can be seen that there is still skill at synoptic scales by day 10 and even more skill at planetary scales.

The question arises as to why the planetary activity increases above the observed level in Cy32r3. A similar plot to Figure 2, but for tropical 200 hPa velocity potential, shows an earlier and more exaggerated increase in planetary activity. This suggests that tropical convection and the forcing of extratropical Rossby waves (*Rodwell & Jung,* 2008) could be involved in this change.

The activity as defined in the scale-dependent plots does not distinguish between changes in transient activity and changes in model bias. The climate runs (see Box B) provide a large amount of data and can therefore be used to distinguish between these two possibilities. Figure 3 shows power spectra as a function of longitude for tropical velocity-potential at 200 hPa from (a) ERA-40 and (b) the climate runs of the atmospheric model Cy32r3. The observations in Figure 3a show a clear peak in power at the 40–60 day timescale over the Indian Ocean and western Pacific. This is the signature of the Madden-Julian Oscillation. The atmospheric model Cy32r3 (Figure 3b) produces, for the first time, sufficient power at these timescales. However it is clear that there is too much power at very long time scales associated with planetary waves. On the other hand, the fact that the IFS also has biases in the tropics is apparent from Figure 4 that shows systematic precipitation errors for Cy32r3.

doi:10.21957/is14d23143

Climate runs

B

For each model cycle, a large set of 13-month long integrations with the atmospheric component of the IFS are carried out in order to investigate the climate of the ECMWF model. Runs were started on 1 November of each of the years 1962–2005 using observed sea surface temperature and sea ice fields as lower boundary conditions. The first model cycle in the climate section of the Diagnostics Explorer is Cy29r2. The runs are carried out using a horizontal resolution of T159 with 91 levels in the vertical (60 levels prior to Cy31r1). The results are diagnosed for the four standard seasons December to February, March to May, June to August and September to November; the first month being discarded to allow the model to spin-up.

The model integrations are compared with observational data from various sources including (re-)analysis, satellite and SYNOP data. The satellite data sets have been compiled and kindly made available by scientists of the Physical Aspects Section.

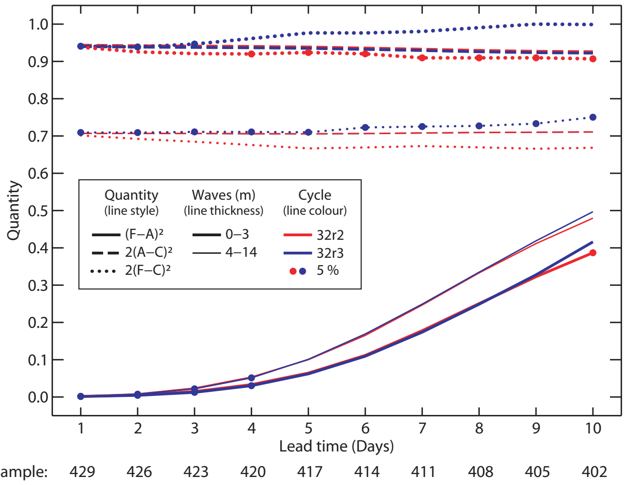


Figure 2 Mean-squared error and mean-squared activity for planetary and synoptic scale variability of the 500 hPa geopotential height in the northern mid-latitudes (35°-65°N) for Cy32r2 (red) and Cy32r3 (blue). Line style indicates the quantity. *Solid lines:* mean-squared forecast error relative to a consistent analysis [(F–A)^2]. *Dashed lines:* mean-squared analysis activity relative to the ERA-40 climatology [2(A–C)^2]. *Dotted lines:* mean-squared forecast activity relative to the ERA-40 climatology [2(F–C)^2]. Line thickness indicates wave-band. *Thick lines:* "planetary variability" using zonal wavenumbers 0–3. *Thin lines:* "synoptic variability" using zonal wavenumbers 4–14. Filled circles on the curve for a particular cycle indicate that the cycle is significantly better than the other cycle at the 5% statistical significance level (using a paired, two-sided t-test). All curves are normalised by the largest value on the plot. Numbers at the bottom of the figure indicate the sample size for each lead-time. The sample includes forecasts from two research experimental suites and the experimental suite run by the operations department.

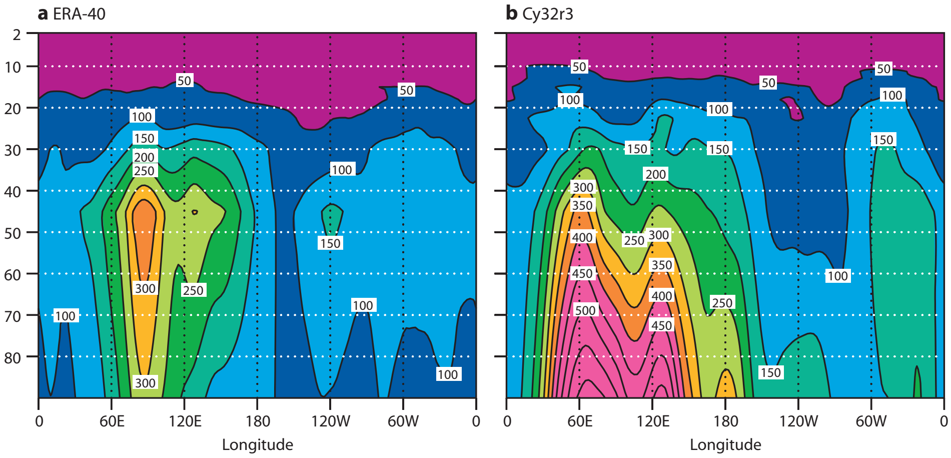


Figure 3 Average power spectra of tropical (5°S–5°N) velocity potential anomalies at 200 hPa (m^{2}s^{-1}) as a function of longitude for (a) ERA-40 and (b) Cy32r3. Results are based on all December–February seasons for 1962–2005. Anomalies have been computed by removing the mean annual cycle.

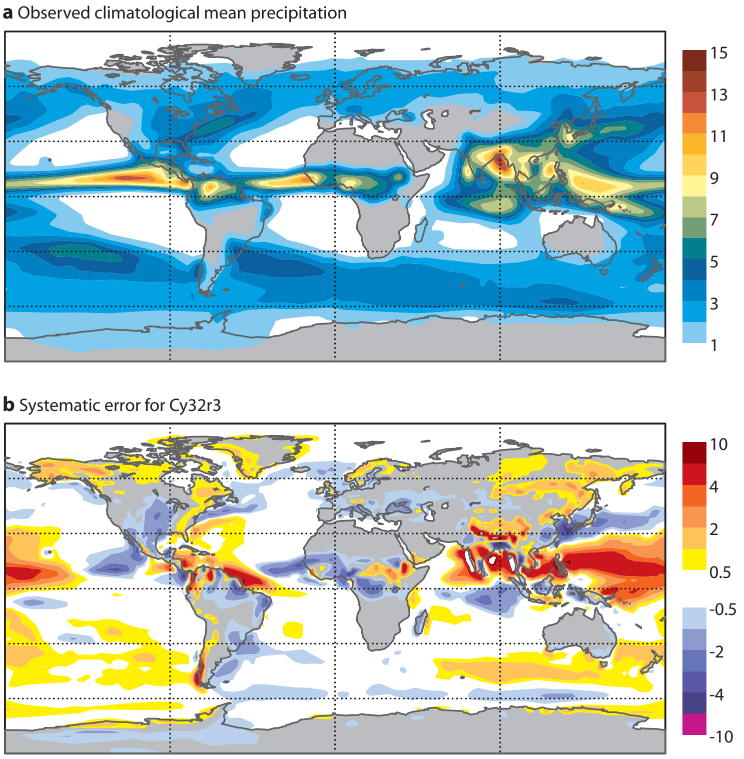
Representation of the Indian summer monsoon

The precipitation bias of Cy32r3 can be seen in Figure 4 to extend over the Indian Peninsular. This represents an excessively strong Indian summer monsoon. A realistic representation of the Indian summer monsoon by numerical models is crucial given the large number of people directly affected by it and its potential for affecting the climate in distant locations. In addition to the excessive rainfall in Cy32r3, the low-level monsoonal winds over the Arabian Sea are also too strong. *Rodwell & Hoskins* (1996) show that these two features are intimately related but it is difficult to determine from the climate runs what comes first: the excessive rainfall or the excessive monsoon inflow. In order to shed more light on the possible origin of this error it is helpful to determine how early the strong monsoon inflow develops within the forecast.

Figure 5 shows mean 850 hPa horizontal wind errors at day 5 from the medium-range weather forecasts for (a) Cy32r3 and (b) the then operational cycle Cy32r2 for the period 11 June to 1 August 2007. Both cycles have winds that are too strong over the Arabian Sea but Cy32r3 has the strongest winds (another plot in the Diagnostics Explorer shows that this difference is statistically significant with a large magnitude of around 3 ms–1). It is clear that the particularly excessive monsoon inflow in Cy32r3 starts to occur even in the medium-range.

The difference in the mean analysis increments for winds at 850 hPa between Cy32r3 (experimental suite) and Cy32r2 (operational suite) is shown in Figure 6 using data from 1 June to 1 August 2007. It can be seen that for Cy32r3 the observations have a bigger impact than in Cy32r2 in slowing down the excessive strong low-level jet produced by the first guess. Such an early appearance of the increased mean wind error strongly indicates that the cause is local to the Arabian Sea (and not caused by excessive monsoonal precipitation).

A more detailed investigation of the analysis increments suggests that it is particularly at the lowest-most levels, where the moisture transport peaks, that the first guess produces too strong winds. Our analysis therefore suggests that the excessively strong Indian summer monsoon in Cy32r3 has its origin in problems with simulating the vertical structure of the low-level monsoonal inflow over the Arabian Sea. This could be associated with the (otherwise beneficial) change in vertical diffusion parametrization at Cy32r3. Changes to the vertical diffusion and convection scheme made in Cy33r1 have led to a moderate reduction of the overly active Indian summer monsoon.


a Observed climatological mean precipitation

b Systematic error for Cy32r3

Figure 4 (a) Observed climatological mean precipitation (mm day^{-1}) from GPCP data for the June-August season for 1979–2001 along with (b) the corresponding systematic errors for Cy32r3 for 1963-2005.

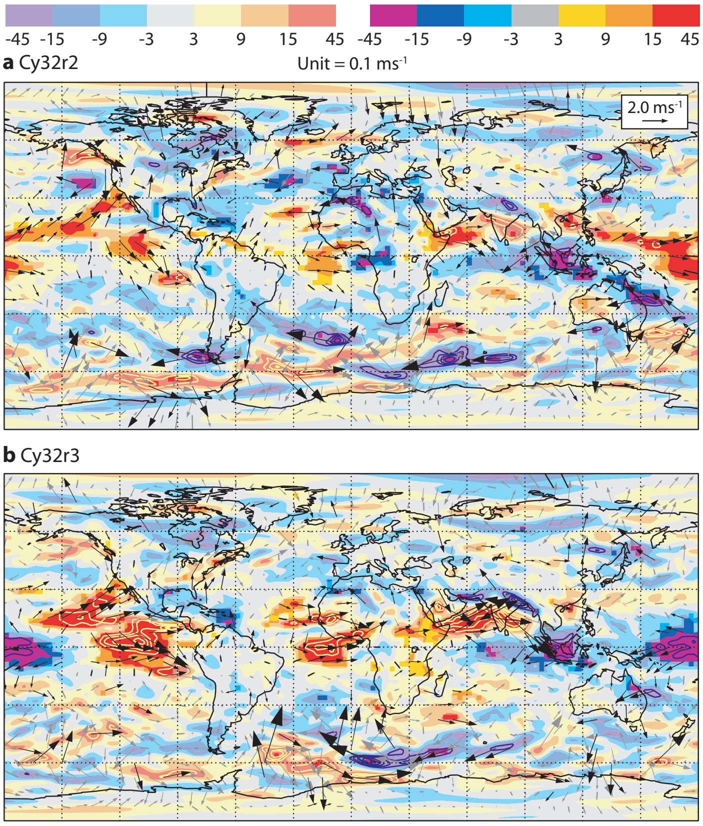


Figure 5 Mean systematic errors of the zonal wind component (shading) and horizontal winds (arrows) at 850 hPa for day 5 forecasts with (a) Cy32r2 and (b) Cy32r3. Results are based on all 00 UTC forecasts starting between 11 June and 1 August 2007.

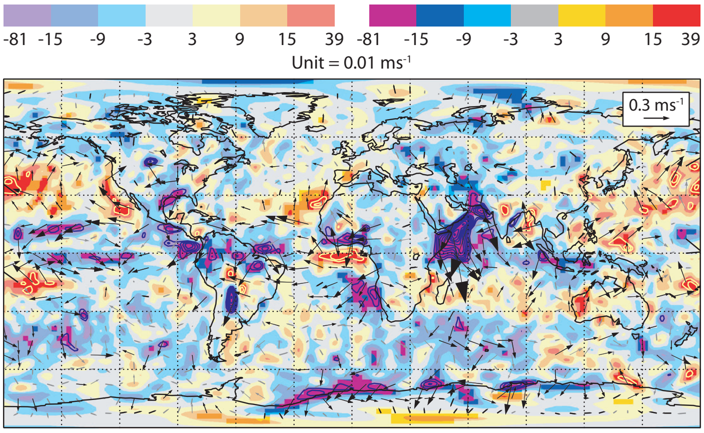


Figure 6 Change in mean analysis increments for the meridional wind component (shading) and horizontal winds (arrows) at 850 hPa from Cy32r2 to Cy32r3. Results are based on all 00 UTC and 12 UTC analyses made between 1 June and 1 August 2007.

Understanding the errors in the Hadley Circulation

The investigation of the Indian summer monsoon clearly highlights the power of a seamless approach to weather and climate – a feature that is central to the philosophy of the Diagnostics Explorer. We now consider how the data assimilation component of the Diagnostics Explorer (observation usage and analysis increments) can be used to provide a better understanding of the mean errors in the Hadley Circulation.

As an example, Figure 7a shows the zonal-mean day 2 errors in temperature and meridional circulation averaged over December to February (DJF) 2007/8. The dominant branch of the Hadley Circulation (in the northern hemisphere in DJF) is consistently forecast to be too weak. This has been a long-standing issue for the IFS. In addition, there is a temperature error in the tropics with the lower-troposphere too cool, the mid- troposphere too warm and the upper-troposphere/ lower-stratosphere too cool. The analysis increments (Figure 7b) show that these temperature and meridional circulation discrepancies exist very early in the forecast. This indicates that the problem has a local (tropical) explanation. Figure 7c shows the analysis increments at 500 hPa. The cooling increment at this level can be seen to occur over much of the tropics, particularly over the Indian Ocean/western Pacific region.

doi:10.21957/is14d23143

Mean analysis increments show where the observations are consistently different from the model's first guess. However, they do not tell us whether it is the model or the observations that are (most) at fault.

The Diagnostics Explorer contains a section on observation usage within the data assimilation system. These plots are in "observation space" so, for example, satellite brightness temperature observations are compared with brightness temperatures simulated by the model. A brightness temperature does not reflect a temperature at a single height in the atmosphere but rather a weighted integral of temperature over an atmospheric layer. The 'AIRS' Satellite channel 215 "sees" temperatures within the 700–300 hPa layer with a maximum weighting at around 500 hPa. Figure 7d shows "first-guess departures" (observation minus first guess) for this channel. Comparison with Figure 7c shows that these departures strongly support the tropical cooling increments at 500 hPa. The magnitude of the first-guess departure can be as large as 0.9 K and generally has a value of around 0.1 K.

Having identified one set of observations that support the systematic analysis increments, it is now important to quantify the magnitude of the likely residual bias in these observations (i.e. the bias after the observation has been bias-corrected). Figure 7e shows the variational bias correction (*McNally et al.,* 2006) applied to this data by the data assimilation system. The magnitude of this correction is typically of order 0.05 K. If these corrections do account for most of the observation bias then one could conclude that residual observation bias is even smaller and not responsible for the mean analysis increment. This would then highlight model error as the more likely cause for the cooling increment. A word of caution is appropriate, however, since Figure 7f shows that the number of AIRS channel 215 observations used within the data assimilation system is generally smaller in the regions of larger mean first-guess departures. This drop in observation usage is associated with cloud screening of infrared data. The "model error" conclusion would be stronger if other observations could be found to back-up the analysis increments. One such set of observations are the 'AMSUA' channel 5 microwave brightness temperatures. These observations can "see through" the tropical clouds and the observation count plots show that the AMSUA data are actually used more than the AIRS within the data assimilation. With the additional support of a few radiosonde stations, the "model error" conclusion appears to be quite robust.

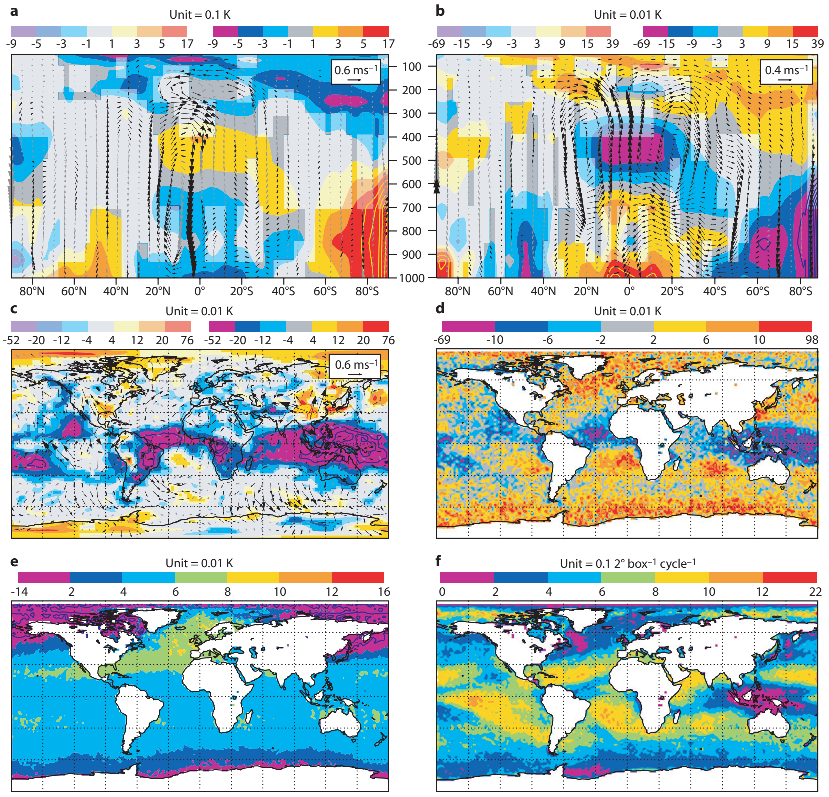


Figure 7 Results highlighting the nature of Hadley Circulation systematic error during December to February 2007/08. (a) Zonal-mean Day 2 forecast error of temperature (shaded) and meridional circulation. (b) Zonal-mean analysis increment of temperature and meridional circulation. (c) 500 hPa analysis increment of temperature and horizontal wind. (d) First-guess departures (observation minus first guess) from AIRS satellite channel 215 infrared brightness temperature. (e) Variational bias correction applied to the AIRS channel 215 brightness temperature within the data assimilation system. (f) Average number of AIRS channel 215 observations used per 2° grid-box in each data assimilation cycle.

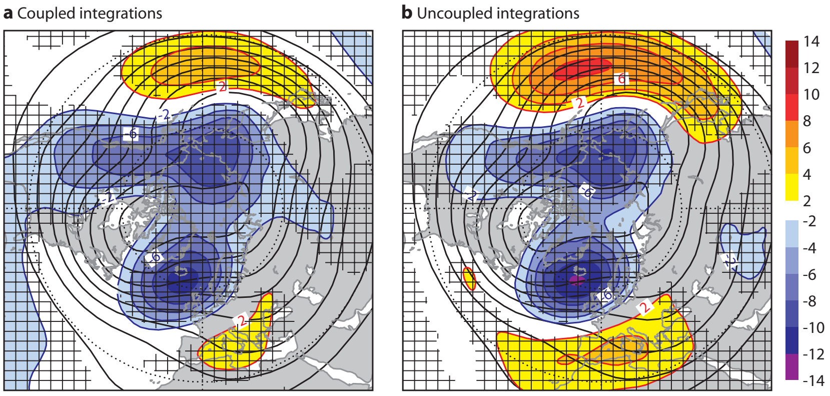
Diagnosis of seasonal forecast bias

ECMWF runs a coupled atmosphere-ocean model to make predictions several months in advance. In the current system, called System 3 (*Anderson et al.,* 2007), the atmospheric component (based on Cy31r1 at T159L62) is coupled to the Hamburg Ocean Primitive Equation Model (HOPE). A large set of hindcasts were carried out with System 3 to allow post-processing (bias correction) of operational forecasts. Moreover, diagnostic runs (hindcasts) were carried out by members of the ECMWF Seasonal Forecast Group; for these the atmospheric component of System 3 is run in uncoupled mode by prescribing observed sea-surface temperature and sea ice fields. It can be argued that these diagnostic runs provide an estimate of the upper limit of seasonal predictability with the current system and give the opportunity to investigate the impact that atmosphere-ocean coupling has on systematic model errors.

Hindcasts with the coupled and uncoupled version of System 3 have been diagnosed and the results are available on the Diagnostics Explorer. To give an example of what can be learnt from these results, Figure 8a shows systematic error in 500 hPa geopotential height for the coupled atmosphere-ocean model. Evidently, systematic errors are quite substantial taking values of similar magnitude to those of the observed seasonal mean anomalies that System 3 aims to predict. In the North Atlantic region a cyclonic bias stands out, which is associated with an underestimation of the observed frequency of Euro-Atlantic blocking events (*Jung,* 2005). Also an anticyclonic bias is prominent in the North Pacific region. One might speculate that these errors are due to a drift of the coupled system, particularly in the tropics, which could lead to the erroneous generation of stationary Rossby waves over the northern hemisphere. The fact that the run with prescribed sea-surface temperature anomalies (Figure 8b) produces similar biases, however, suggests that the origin of this error lies in the atmospheric component of System 3. By looking at similar diagnostics for more recent model cycles, the Diagnostics Explorer reveals that recent model changes led to substantial reductions in the size of systematic errors in 500 hPa geopotential height over the North Pacific and North Atlantic.

ECMWF training courses

One of the duties of ECMWF is to assist its Member States and Co-operating States in the training of forecasters and scientists in numerical weather forecasting through an extensive educational programme. In spring 2008 the Diagnostics Explorer was used for the first time in the Predictability, Diagnostics and Seasonal Forecasting module of the NWP Course to introduce diagnostics techniques and to discuss the performance of the ECMWF forecasting system at time scales of hours (analysis), days (numerical weather forecasting) and several months (seasonal forecasting). After an introduction to the Diagnostics Explorer, the students were asked to use it to answer a set of questions. In this way the students learnt, amongst other things, about aspects of the nature of forecast error and its growth, how to assess year-to-year changes in forecast error and how to diagnose a complex data assimilation system. Given positive feedback from the students, it was decided that use of the Diagnostics Explorer will be an integral part of future training courses.


a Coupled integrations

b Uncoupled integrations

Figure 8 Mean systematic error of 500 hPa geopotential height fields (shading in dam) for December to February in 1982–2005 for (a) coupled and (b) uncoupled integrations of System 3. Statistics for the models are based on five ensemble members. Also shown are climatological mean fields from ERA-40 (contours). Mean errors that are significantly different from zero at the 5% level are hatched.

Outlook

Recently, work on the Diagnostics Explorer has focused on the development of the diagnostics software (with technical help from the Metview Team) and uploading the plots to the web (with the help of Claude Gibert from the Meteorological Operations Section). Results are now available for all components of the IFS: the data assimilation system, the wave model, the forecast model and the coupled atmosphere-ocean model. What are the plans for future developments of the Diagnostics Explorer?

The incorporation of new diagnostics, aimed at helping to understand the origin of forecast error on time scales from hours to many months, is an ongoing activity. For example, diagnostics of the vorticity balance in the atmospheric model (*Rodwell & Jung,* 2008), including the Rossby wave source, will soon be added. Furthermore, it is planned to include the results of special experiments designed to address issues of particular concern. For example, model climate sensitivity to increasing resolution. With the introduction of Seasonal Forecast System 4, it is also planned to incorporate diagnostics of the ocean data assimilation system in a fashion similar to the one already used for the atmosphere. Finally, it is planned to extend the use of Diagnostics Explorer by making the software available to all scientists at ECMWF.

We hope that, with the Diagnostics Explorer and its further developments, the Diagnostics Section can make a contribution to future improvements of all components of the IFS.

Online access to the Diagnostic Explorer by Member States will be considered in the near future, subject to interest and resources.

Further Reading

Anderson, D., T. Stockdale, M. Balmaseda, L. Ferranti, F. Vitart, F. Molteni, F. Doblas-Reyes, K. Mogensen & **A. Vidard,** 2007: Seasonal Forecast System 3. *ECMWF Newsletter No. 110,* 19–5.

Bechtold, P., M. Köhler, T. Jung, F. Doblas-Reyes, M. Leutbecher, M. Rodwell, F. Vitart & **G. Balsamo,** 2008: Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Q. J. R. Meteorol. Soc.,***134,** 1337–1351.

Jung, T., 2005: Systematic error of the atmospheric circulation in the ECMWF forecasting system. *Q. J. R. Meteorol. Soc.,***134,** 1337–1351.

McNally, A., T. Auligné, D. Dee & **G. Kelly,** 2006: A variational approach to satellite bias correction. *ECMWF Newsletter No. 107,* 18–23.

Morcrette, J.-J., P. Bechtold, A. Beljaars, A. Benedetti, A. Bonet, F. Doblas-Reyes, J. Hague, M. Hamrud, J. Haseler, J.W. Kaiser, M. Leutbecher, G. Mozdzynski, M. Razinger, D. Salmond, S. Serrar, M. Suttie, A. Tompkins, A. Untch & **A. Weisheimer,** 2007: Recent advances in radiation transfer parametrizations. *ECMWF Tech. Memo. No. 539.*

Rodwell, M.J. & **B.J. Hoskins,** 1996: Monsoons and the dynamics of deserts. *Q. J. R. Meteorol. Soc.,* **122,** 1385–1404.

Rodwell, M.J. & **T. Jung,** 2008: Understanding the local and global impacts of model physics changes: An aerosol example. *Q. J. R. Meteorol. Soc.,* **134,** 1479–1497.

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.

doi:10.21957/is14d23143