TECHNICAL MEMORANDUM

637

Developments in Diagnostics Research

M.J. Rodwell, T.Jung, P. Bechtold, P. Berrisford, N. Bormann, C. Cardinali, L. Ferranti, T. Hewson, F. Molteni, N. Wedi, M.A. Balmaseda, G. Balsamo, M. Bonavita, R. Buizza, M. Dahoui, A. Garcia-Mendez ,M. Leutbecher, P. Lopez, Y. Trémolet and F. Vitart

Research Department

Presented to the 39th Session of the SAC,
4–6 October 2010

October 2010

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

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, omission and for loss or damage arising from its use.

Abstract

Over the years, data assimilation and forecast schemes have evolved into very complex systems. For example, the ECMWF data assimilation system now handles a very large range of space and surface-based meteorological observations, combining these with prior information of the atmospheric state and using a comprehensive linearised forecast model to ensure that the observations are incorporated in a dynamically consistent way. Meanwhile the ECMWF forecast model itself represents ever more physical processes, with ever increasing complexity, and is integrated at ever higher resolution. Recognising that any analysis or forecast is actually a probabilistic one, ECMWF continues to incorporate representations of uncertainty into all its forecast components. Efficient and coordinated diagnosis of such a complex system is a necessity.

Diagnostics has always been an active area of research and development at ECMWF. Presently, ECMWF has a small nucleus of scientists dedicated to diagnostics, who have a broad understanding of the global circulation and an overview of the assimilation and forecasting systems. Their unifying role will become ever more important as progress leads, inevitably, to more fragmentation and specialisation within the overall task of producing forecasts. However diagnostic work is not only about getting an overview, because all researchers need to produce ever more detailed diagnostics within their own particular fields. These trends raise the question of whether we will need to enhance communication, coordination and collaboration across the traditional section boundaries in future. In order to address this issue, and to produce this report, a ‘Working Group on Diagnostics’ (WGD) has been established at ECMWF.

In this report, a diagnostics framework is introduced which highlights key diagnosis areas and their links. From each area, one or two ‘strategic’ diagnostic tools (or techniques) are highlighted. These tools are ‘strategic’ in that they are likely to be developed further and will inform future system developments. In order to demonstrate the utility of across-section collaborative work, these tools have been used to collectively address a long-standing problem for the ECMWF forecasting systems: that of the over-active Asian monsoon. This case-study provides a structural blue-print for future across-section projects focused on other forecasting problems and on the assessment of new model cycles.

In discussing the above, together with future diagnostic requirements, this paper shows where ECMWF stands at present in terms of diagnostic work and proposes a strategy that will ensure it meets the challenges of the future.

1 Introduction

1.1 What do we mean by ‘Diagnostics’?

To some, ‘Diagnostics’ is about a knowledge of the real world (Lorenz energy cycles, equatorial wave dispersion diagrams, etc). For others, Diagnostics is about the in-depth analysis of a particular case-study (the analysis of the UK’s ‘October storm’ of 1987, for example). For a third group, Diagnostics is the investigation of observation, analysis, or forecast-system error. Although seemingly very different, all these interpretations embody the concept of understanding. This understanding goes beyond simple metrics of the circulation or the monitoring of scores. While the Diagnostics work of an operational forecast centre such as ECMWF will fall most naturally into the latter ‘diagnosis of error’ category, metrics of the real world often form a good basis for assessing this error. In addition, single forecast failures often necessitate the need for the diagnosis of case-studies and these often feature in the ‘daily report’. Nevertheless, there is an emphasis on the ‘diagnosis of error’ and this will be somewhat different from what some readers will be familiar to.

1.2 Why do we need diagnostic research?

Standard geophysical textbooks contain many examples of diagnostic techniques (predominantly circulation metrics) and one may ask whether there is a need for further diagnostic research. In fact the need is great. A few key reasons are that

- The increasing accuracy of forecast systems (and observations) means that residual errors are smaller than ever before. More precise diagnostic tools are required to quantify these errors.
- The volume of observations assimilated has increased ‘exponentially’ over recent decades and will continue to increase in the future. Sophisticated tools are required to assess data quality and the redundancy of information in order to prioritise efforts on the most promising observations.
- Forecast models represent increasing numbers of physical (and micro-physical) processes. The scope for interactions between these processes is increasing and new diagnostic tools that can identify remaining model errors in the face of such interactions are required.
- As resolution increases in the future, the model will move into a ‘grey zone’ where non-hydrostatic effects begin to be important and convective processes begin to be explicitly resolved. Diagnostics that target this grey zone (and beyond) will become necessary.
- The growing use of probabilistic forecasts has led to the inclusion of uncertainty estimates in all aspects of the forecast system. This brings with it the increasing need for careful diagnosis of the relationship between predicted uncertainty (ensemble spread) and actual error.
- 4D-Var incorporates both model and observational aspects, and tools are increasingly required to diagnose their respective contributions.

1.3 The Working Group on Diagnostics

This paper discusses the present state of diagnostics at ECMWF, recent advances, and strategies for the future. In order to address these topics, a Working Group on Diagnostics (WGD) was recently established at ECMWF. The WGD comprises representatives from all sections in the Research division and the Meteorological Operations section. It has been led by one of the dedicated diagnostics research scientists. The idea being that these representatives act as conduits of information between the WGD and their respective sections. The first task of the WGD was to define clearly its roles and these can be summarised as the:

1. Over-sight of collaborative projects
    - Diagnosis of existing assimilation or forecasting problems
    - Diagnosis of major new model cycles
2. Strategic coordination of diagnostic developments
    - Highlighting opportunities for new diagnostic tools of common interest
    - Making existing diagnostic tools more widely usable
    - Ensuring sufficient computing and storage resources for diagnostics
3. Across-section communication of information and results
    - Through representatives to their sections
    - Using a central diagnostics web page
    - By coordination with the special topics of the OD/RD meeting
    - With seminars on tools and collaborative projects

The next sub-section highlights the areas within the overall task of forecasting where there is scope for diagnosis and, therefore, diagnostic tools. Diagnostics cannot always be unified in a seamless way, but they can sometimes all be applied to the same forecasting problem. Hence the following sub-section will introduce a key and long-standing problem of the ECMWF forecast system. This problem (that of the over-active Asian monsoon) will form the basis of a thread, or story, that provides a unifying theme in this paper.

1.4 The scope for Diagnostics at ECMWF

Figure 1 presents a schematic diagram of the forecasting system (deterministic, ensemble, atmosphere-only or coupled, and at all lead-times). For example, the model and observations are combined in the analysis. The word ‘analysis’ refers to the process of data-assimilation (deterministic or ensemble) and also the analyses themselves. The analysis is the ‘coal-face’ where model and observations ‘collide’ and the scope for diagnosis is immense and growing. The left yellow box summarises, in broad strokes, these possibilities for diagnosis. For example, it is possible to monitor the quality of the observations through a comparison with the first-guess (FG) forecast and with other observations. Through data denial experiments and through adjoint techniques it is possible to diagnose the impact of different sets of observations on the quality of the forecast. Mean analysis increments can indicate systematic model error associated with ‘fast’ processes. An essential component of 4D-Var is the linearised model, and the natural place to diagnose its ability to approximate the full non-linear model is within the data assimilation. Complementary diagnostics are possible within the re-analysis, particularly because of its focus on long timeseries and climate-related trends. Comparison, within the ensemble data assimilation, of analysis uncertainty and analysis error gives useful information about the representation of stochastic processes within the model.

The analysis is used to initiate the forecast. The word ‘forecast’ refers to the forecast process and also the forecast products themselves. The corresponding yellow box again summarises the scope for diagnosis when comparing the forecast with the analysed and/or observed ‘truth’. For example, the impact of physics changes on forecast error can be diagnosed. Dynamical teleconnections and their interactions with the physics can also be assessed. Errors associated with the coupled (ocean-atmosphere) processes can be diagnosed here too. Hindcasts initiated from the re-analyses (which involve a single model cycle) offer scope for distinguishing flow-dependent changes in error from forecast system improvements. As with the ensemble data assimilation, discrepancies between ensemble forecast spread and error can give

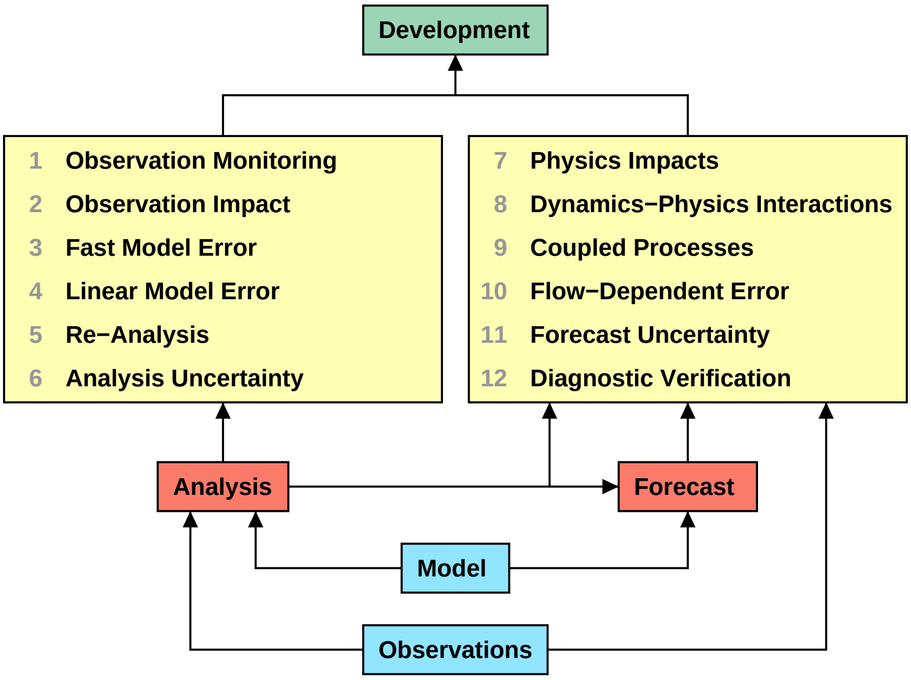

Figure 1: Diagnostics framework showing the key aspects involved in forecasting. 'Forecasting' includes deterministic and probabilistic forecasts at all ranges, and both atmosphere and ocean components. The 'model' includes both non-linear and linear versions. Possibilities for diagnosis are summarised in the yellow boxes. These include diagnosis of the analysis (data assimilation) and forecast. There is a direct correspondence between the labels in these yellow boxes and the subsections of section 2.

information about the representation of stochastic processes and aid in diagnosing reasons for probabilistic forecast 'busts'. 'Diagnostic Verification' includes aspects traditionally thought of as 'Forecast Verification', but which can provide useful, additional, understanding of forecast error. For example, tracking statistics can lead to useful understanding of systematic errors in the speed and intensity of cyclonic features, and will become increasingly important in attempts to diagnose problems with rare, mobile, severe weather events. Conditional sampling based on poor weather verification scores is helping to focus diagnostic attention towards key forecast errors.

Within each diagnosis area of the framework presented in Fig. 1 there will be several diagnostic approaches and tools. It is not useful to discuss them all here in great detail. Instead, a few 'strategic' tools, that will be developed further, that will be key to future developments, and that demonstrate the breadth and depth of ECMWF's diagnostic work will be discussed and applied. Wherever possible, the application will be to the key issue of the over-active monsoon.

1.5 A key forecasting problem – the over-active Asian summer monsoon

The over-active Asian monsoon has been a long-standing issue for ECMWF's forecasting systems. The fact that it is a long-standing issue suggests it is a hard problem to solve, perhaps involving interactions over multiple timescales, perhaps with a remote, or seemingly unconnected root cause. Hence it makes a good candidate for a collaborative project spanning many sections within ECMWF (and beyond: ECMWF also has a collaborative project with the UK Met. Office). Figure 2 introduces the problem by showing results that span a large range of diagnostic areas, including observations, analyses, medium-, monthly-, and seasonal-ranges, and research experimentation.

The top row (Fig. 2a–c) shows observed 1991–2007 GPCP (Adler et al., 2003) precipitation climatologies for June (left), July (middle) and August (right), respectively. Each month displays broadly the same pattern of precipitation with maxima centred over the west coasts of India and Burma/Myanmar and on the Equator. Precipitation over land areas generally maximises in July. All panels below the top

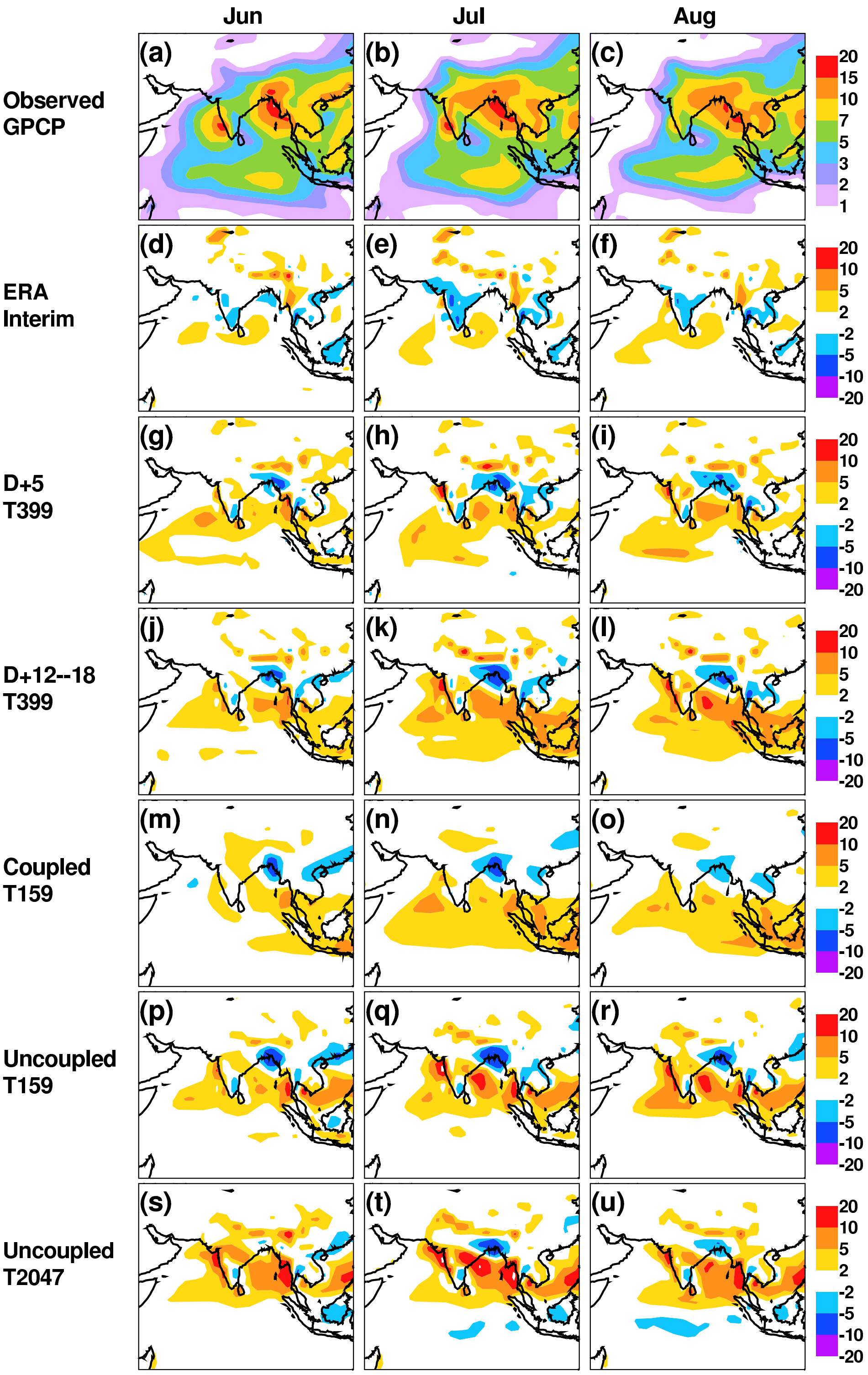

Figure 2: Asian summer monsoon precipitation climatologies (in mm day^{-1}) based on the 17 years 1991–2007. (a)–(c) 'Observed' GPCP for June, July and August, respectively. All other fields are anomalies relative to GPCP. (d)–(f) ERA Interim based on the precipitation accumulated over first 12 hours of the 0 and 12 UTC forecasts using model cycle 31R2. (g)–(i) At a lead-time of 5 days based on hindcasts (5 members once a week) using model cycle 35R2 with persisted SST. (j)–(l) as (g)–(i) but at lead-times 12–18 days with coupled model from day 10. (m)–(o) Climatology of the coupled seasonal-forecast model started on 1 May (months 2, 3 and 4, respectively) using model cycle 36R1. (p)–(r) As (m)–(o) but for the uncoupled model at T_L159 started from 1 November (months 8, 9 and 10, respectively). (s)–(u) As (p)–(r) but at the much higher resolution of T_L2047.

row show differences with this GPCP data for the same three months and the same set of years.

Short-range (12h) precipitation forecasts from ERA-Interim re-analyses (Fig. 2d–f) are sometimes used as an alternative for observed precipitation but note that the Indian Peninsula is somewhat drier than GPCP, with more precipitation over the seas surrounding the Peninsula and over the Himalayas.

At the longer (medium-range) lead-time of 5 days (Fig. 2g–i) re-forecasts using the recent cycle 35R2 at resolution T399 demonstrate how the over-prediction of precipitation over the ocean gets worse and begins to encroach over the land regions. Note the intense localised precipitation bias on the western coast of India, particularly in July. Note also that a dry bias has developed over Bangladesh. Averaged over the lead-times 12–18 days (Fig. 2j–l), the biases seen at day 5 generally become even stronger.

Figure 2(m)–(o) show results from coupled model integrations initialised on 1 May (for the years 1991–2007) using cycle 36R1 at atmospheric resolution T159. In general, precipitation biases over the ocean are reduced but Bangladesh remains dry.

Figure 2(p)–(r) show results for atmosphere-only 13-month simulations initialised on 1 November using cycle 36R1 at resolution T159. Comparison with the coupled (Fig. 2m–o) and D+12–18 uncoupled (Fig. 2j–l) results demonstrate that coupling reduces the precipitation biases over the oceans and that, without this coupling, biases continue to increase with lead-time. The implication is probably that coupling sweeps an atmospheric problem under the oceanic 'carpet'.

The final row (Fig. 2s–u) show uncoupled results identical to the previous row but at a much higher resolution (T2047). These simulations were conducted under the 'Athena' project that utilised surplus computing power in the United States. It can be seen that biases actually get worse. Clearly, without any special tuning or physics adjustments, raw resolution is not a solution to the over-active monsoon.

In as much as the forecast panels are more yellow and red than they are blue and purple, this problem is referred to here as the 'over-active' Asian summer monsoon.

In section 2, the diagnostic tools of each of the areas highlighted in Fig. 1 will be discussed and, where appropriate, applied to the Asian monsoon problem. It should be noted that the monsoon project is a 'work-in-progress', and used here primarily as a convenient vehicle to demonstrate our diagnostic tools, and as a means of establishing a blue-print for future collaborative projects. A summary of the monsoon project, together with initial diagnostics of the new cycle 36R4 are given in section 3. A discussion of lessons learnt and proposed future strategies is presented in section 4.

2 Diagnostic tools

This section discusses ongoing diagnostic research activities in the 12 areas identified in Fig. 1. As discussed above, some consistency is achieved by applying diagnostic tools to the 'over-active Asian monsoon' problem wherever possible.

2.1 Observation Monitoring

Observation monitoring is a key component of the ECMWF diagnostic system. Many diagnostics are based on departures of the First Guess (FG) and analysis from the observations (so called 'innovations' or 'departures'), as calculated during the assimilation process. Statistics on these provide a powerful basis for the characterisation of observation, assimilation or model aspects.

The generation and plotting of observation statistics requires a high degree of flexibility, provided by the OBSTAT-tool developed at ECMWF. Such flexibility is important to allow diagnosis from many perspectives: temporal, geographical, vertical column, land and sea, data usage flags, observation angles, etc. Currently, all satellite observations presented to the analysis system are monitored and published on the web[1].

[1] http://www.ecmwf.int/products/forecasts/d/charts/monitoring/satellite

Monitoring of observation departures plays a role upstream and downstream of the data assimilation process. Its use can be broadly grouped into three areas:

- Characterisation of observations and their observation operators, including observation biases (Geer et al., 2010), performance of quality control (Krzeminski et al., 2009), bias correction (Auligné and McNally, 2007), and the specification of observation errors (Desroziers et al., 2005; Bormann and Bauer, 2010; Bormann et al., 2010), etc. ECMWF frequently provides important input to the calibration/validation of new satellite missions (Lu et al., 2010).
- Monitoring of the temporal stability of the observations to ensure that only consistently good quality data is used in operations. Data anomalies are detected automatically, and these may trigger corrective action. The routine monitoring statistics are of great interest also externally for data providers and other NWP centres.
- Highlighting of model problems which will be apparent, for instance, through consistent systematic non-zero FG departures for several observations (Healy, 2008). In the future, this will be extended to comparisons between all assimilated observations and forecasts at different ranges.

An example of relevance to the monsoon problem is the monitoring of Indian radiosonde data. For many years, temperature reports from these radiosondes have generally been blacklisted in the ECMWF operational data assimilation because of their apparent inconsistency and low quality, as inferred from FG departures. See for example in Fig. 3 the monthly-mean (solid) and standard-deviation (dashed) of first-guess departures for Thiruvananthpuram, southern India prior to 2009. As part of a programme to update the Indian radiosonde network, the sonde-type used at Thiruvananthpuram was recently changed. First-guess departures in Fig. 3 from early 2009 onwards show a remarkable improvement. This result confirms the previous working hypothesis that the main problem was with the radiosondes, and not with the model (first-guess). It also justifies removing the blacklisting from Indian radiosonde temperatures as the sondes are updated. Presently temperatures from 10 Indian radiosondes are assimilated at ECMWF. This assimilation will allow residual model errors to be better diagnosed in this critical monsoon region.

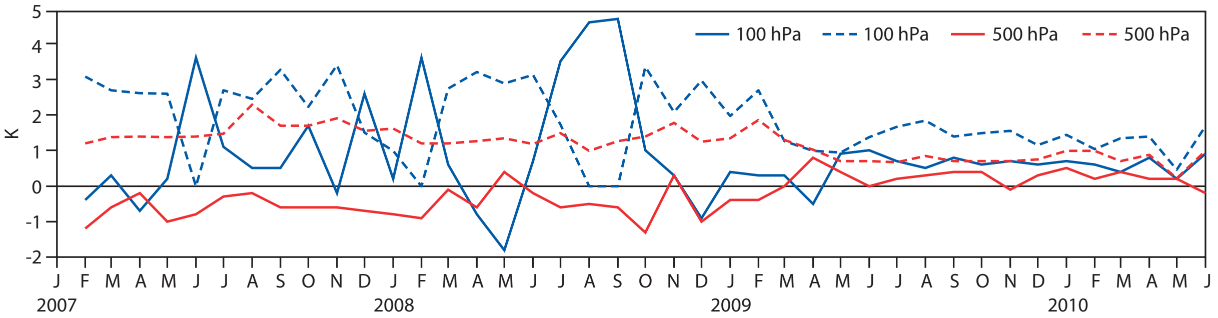

Figure 3: Monthly-mean (solid) and standard-deviation (dashed) of first-guess departures (radiosonde observed temperature minus first-guess) at 500 hPa (red) and 100 hPa (blue) for Thiruvananthpuram, southern India (77°E,8°N). Units are K.

Of particular diagnostic importance is the ability to cross-check observations from different observing systems that observe similar aspects of the atmosphere. This helps, to disentangle biases seen in FG departures into contributions from observations or the FG.

For example, monitoring of mean FG-departures for the Tropical Rainfall Measuring Mission (TRMM) Microwave Imager (TMI) highlighted a 46-day oscillation with a peak-to-peak amplitude of up to 3 K in brightness temperature, linked to the precession of the equator crossing time of the TRMM orbit. FG departures for similar instruments (SSMI and AMSR-E) do not show the same characteristics, so the bias must come from the TMI instrument itself. Further analysis shows that the cause is most likely solar heating of the main reflector of TMI. This reflector is not a perfect reflector, so the instrument measures a combination of Earth emission and the physical temperature of the reflector. Geer et al. (2010) estimate that the variation of the reflector temperature, as the satellite moves in and out of the Earth's shadow, is up to 70 K. Based on this physical understanding of the bias, it is possible to correct for it in the data assimilation, giving more accurate results.

The Indian summer monsoon provides an example where FG departures suggest model problems. Figure 4(a) shows surface wind vectors as diagnosed from ASCAT scatterometer data together with SSMI channel 3 brightness temperatures, that are positively correlated with lower-tropospheric ($\sim$850 hPa) humidity. An important component of the Indian monsoon circulation is a strong low level southwesterly wind over the Arabian Sea. The humidities and surface winds in this oceanic region are crucial for sustaining the monsoon precipitation. Figure 4(b) shows atmospheric motion vector (AMV) winds at around 950 hPa based on infrared and visible imagery, and also shows radiosonde specific humidity observations at 850 hPa. This AMV wind information is complementary to the ASCAT observations and shows a similar monsoon circulation. The radiosonde data is mainly land-based and thus disjoint from the SSMI microwave observations.

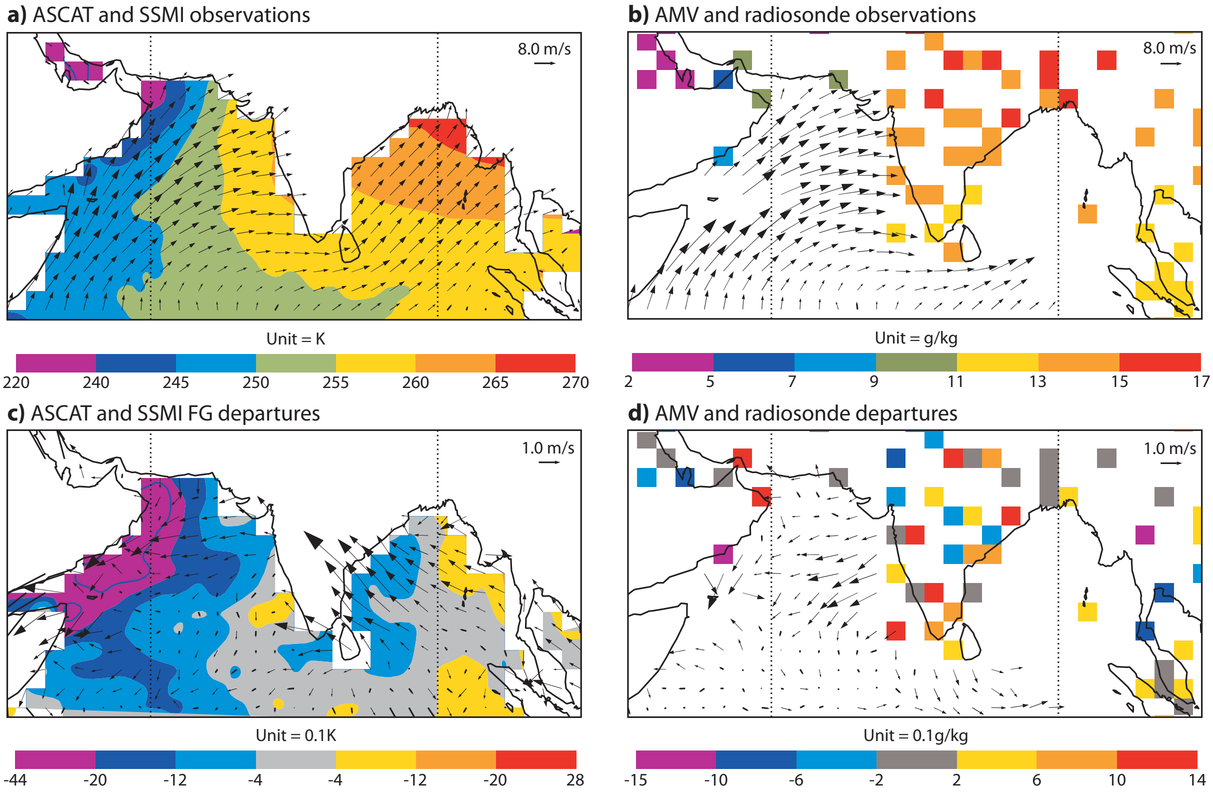

Figure 4: Mean June–August 2009 observation diagnostics (a) SSMI channel 3 microwave brightness temperature (positively correlated with specific humidities at $\sim$850 hPa) and surface winds as diagnosed from ASCAT scatterometer data. (b) Radiosonde specific humidities at 850 hPa and atmospheric motion vector winds at $\sim$950 hPa. (c) As (a) but for FG departures (observation minus first guess) averaged over all 0UTC and 12UTC 'DCDA' analyses. (d) As (b) but for but for FG departures. Note that gross observation biases have been accounted for by the variational bias correction scheme (Dee and Uppala, 2009) and that no vectors have been plotted in (b) and (c) over the Bay of Bengal due to relatively poor data coverage (where there were no observations within a 2^o grid-box for more than 70% of the 12-hour assimilation windows).

The mean FG departures for the ASCAT surface winds and SSMI lower-tropospheric humidities are shown in Fig. 4(c) while mean FG departures for the AMV winds and radiosonde humidities are shown in Fig. 4(d). The ASCAT and AMV wind departures both indicate that observed southwesterly winds are weaker than in the FG, and thus point to a model problem. The mean observation departures for SSMI indicate a drier atmosphere (below 700) close to the Arabian Peninsula than in the FG. At present, there is little other observational data to compare with the SSMI over the ocean (AMSR-E uses the same instrumentation and cannot be considered independent). The radiosonde data, on the other hand, show higher observed humidities over the southern Indian Peninsula than in the FG.

Another example where FG departures point to model biases and help to characterise these are stratospheric temperatures. Various observations provide stratospheric temperature information (*e.g.*, radiosondes, bending angles from GPS-RO, satellite radiances from AMSU-A, AIRS, and IASI) and therefore highlight the system's performance for stratospheric temperatures. Consistent biases in mean FG departures or consistent bias corrections are frequently found for these observations, pointing to model biases in the stratosphere.

For many satellite observations, the bias corrections analysed with the variational bias correction (*e.g. Dee, 2004)* also provide an important diagnostic. Ideally, these bias corrections should correct for observation or observation operator biases. However, in practice observation bias corrections can also correct for forecast model biases or biases resulting from quality control errors (*e.g.* cloud contamination, see Auligné and McNally, 2007). To identify the former again requires careful cross-examination of bias corrections and FG departures for different observation types. The latter implies a sub-optimal use of the observations. Monitoring of observation bias corrections will become even more complex with the further development of weak constraint 4D-Var (Trémolet, 2007a), to ensure an adequate separation between observation and model bias in the assimilation.

Development of observation-based diagnostics will continue, in step with the extended use of observations in the assimilation system. For instance, enhanced characterisation of existing observations and model aspects will be possible with upcoming satellite observations such as the ADM-Aeolus mission which will provide unprecedented wind profile information from space. Also the growing use of satellite data in cloud and rain-affected areas will continue to challenge the ability of the forecast model to represent these adequately, with growing demands on the specification of the background and observation errors and biases involved in their assimilation. Finally, assimilated observations (including satellite radiances) are expected to be increasingly used for forecast verification, especially for humidity, complementing analysis-based verification.

2.2 Observation impact

Data assimilation systems provide an estimate of the atmospheric state by combining meteorological observations with a prior background first-guess forecast, taking into account estimated observation and background error.

Figure 5 shows the mean analysis increments (analysis minus 12-hour FG forecast) of specific humidity and wind vectors at 850 hPa averaged from June to August 2009. In general agreement with the ASCAT and AMV FG departures (Fig. 4c,d), the assimilation of observations tends to decrease the westerly component of the wind and, over the Bay of Bengal, increase its southerly component. In agreement with the SSMI FG departures (Fig. 4c), moisture is removed to the east of the Arabian Peninsula and, in agreement with the radiosonde FG departures (Fig. 4d), it is added over the southern Indian Peninsula.

The influence in the analysis of each observation can be computed during the assimilation process. In particular, the 'degree of freedom for signal' (hereafter DFS, Tukey, 1972; Velleman and Welsch, 1981; Wahba et al., 1995; Purser and Huang, 1993) quantifies the number of statistically independent directions constrained by each observation (Cardinali et al., 2004). Results can be gathered together – for example by observation type. The DFS depends on the assigned accuracy of the observations and background as well as on the model itself that is used as a space and time propagator. The DFS is also affected by the number of assimilated observations – the more observations from a specific instrument that are assimilated, the larger the DFS of that instrument will be. Figure 6 (green bars) shows the DFS, as a percentage of the total, over the monsoon region 20^{o}S–45^{o}N and 35^{o}E-110^{o}E and for June 2009. The observation types providing the most information are SSMI (23%), IASI (15%), AMSR-E (10%), AMSU-A (10%), AIRS (8%) and GPS-RO(6%). All the remaining data types contribute less than 5% of the total DFS.

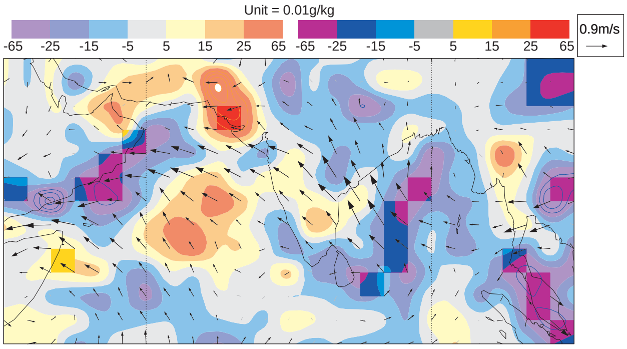

Figure 5: Mean June–August 2009 analysis increment in 850 hPa specific humidity (shaded) and horizontal wind averaged over 0UTC and 12UTC 'DCDA' analyses. A statistical test has been performed and bold colours and black vectors indicate that the mean value is significantly different from zero at the 5% level.

Recently, adjoint based observation sensitivity techniques have been used to measure the observation contribution to the forecast error (see, for example, Baker and Daley, 2000; Cardinali and Buizza, 2004; Morneau et al., 2006; Zhu and Gelaro, 2008). The observation impact is evaluated with respect to a scalar function representing the short-range forecast error (here a global dry energy norm). As with the observation influence in the analysis, the observation forecast error contribution (FEC) is computed for each measurement assimilated, and can also be gathered by observation type. (Cardinali, 2009, see also SAC paper 2009). The DFS and FEC are different but related quantities. They are, in fact, both functions of the background and observation accuracies and the model. FEC, additionally, depends on the forecast error. In an optimal and unbiased system it is expected that the FEC should be similar to the DFS, since the information extracted from the observations during the assimilation procedure should be propagated by the model into the forecast. Loss of forecast impact of a particular observation type with respect to the DFS can be then attributed to model errors such as model bias.

Figure 6 (black bars) also shows the FEC as a percentage of the total for the same observation groupings. It can be seen that, for some observation types, FEC is smaller than the DFS. These include AMSR-E, SSMI and HIRS, IASI, AIRS and SCAT (scatterometer) data. Channel comparisons for the sounding observation types indicate that the channels most affected are the ones providing information on temperature and, to a certain extent, humidity fields below 700 hPa (not shown). The loss of impact on forecast error is believed to be caused by the presence of model error, either random or systematic, in the lower

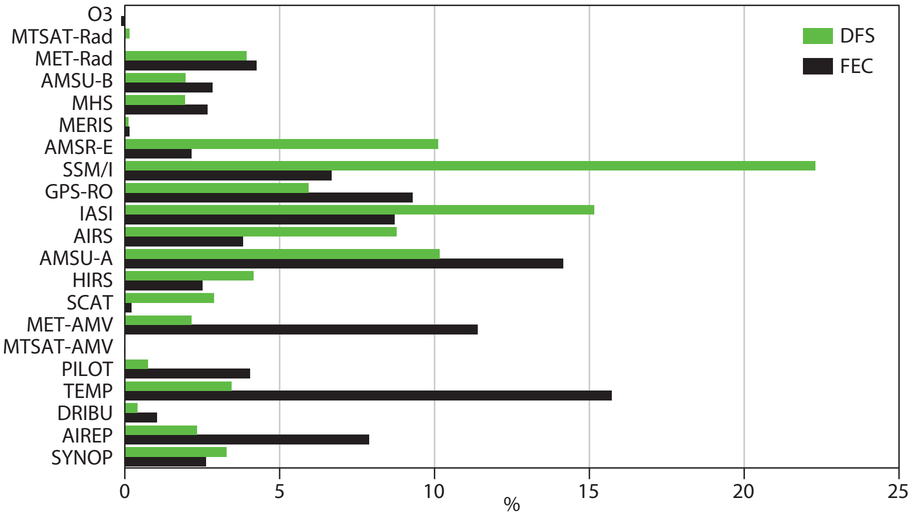

Figure 6: Degree of freedom for signal (DFS, green) and forecast error contribution (FEC) to D+1 forecast error (black) for different observation types assimilated in the extended Indian summer monsoon region (20^{o}S–45^{o}N and 35^{o}W–110^{o}W).

troposphere.

Maps of FEC for different observation types are very useful to highlight areas where observations lead to forecast improvement. Over the monsoon region, the SSMI, SCATT and AMV show a beneficial impact over the Indian Ocean (not shown). To investigate the changes caused by the assimilation of these observations, Observation System Experiments (OSEs) can be performed.

Over the years, OSEs have been the traditional tool for estimating data impact (Bouttier and Kelly, 2001; Lord et al., 2004; English et al., 2004). Usually OSEs are performed by removing subsets of observations from the assimilation system and then comparing analyses and forecasts against a control experiment that includes all the observations. The value of the observation is, in this case, assessed by comparing analysis performance and comparing forecast skill using different statistical indices. Several independent experiments need to be performed for quite long periods, generally a few months. Here results are discussed in terms of analysis differences between assimilations for June 2009 with and without SSMI ('all-sky' Geer et al., 2010) data.

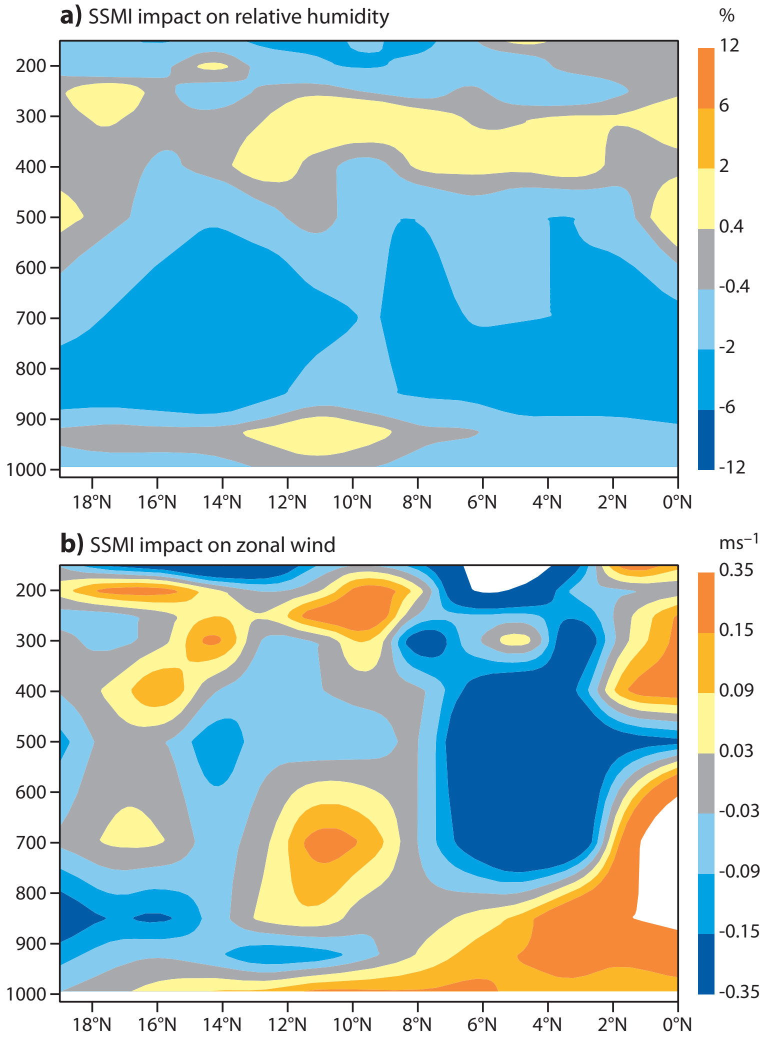

Figure 7: Zonal average (50^o–71^{o}E) cross-section of the mean differences between analysis experiment with and without assimilation of SSMI data for June 2009: (a) relative humidity (%) and (b) u component of wind ($m s^{-1}$).

Figure 7(a) shows a latitude–height cross-section, averaged over the Arabian Sea (50^{o}E–71^{o}E), of analysis differences in relative humidity (experiment with SSMI minus experiment without SSMI). The assimilation of these radiances is seen to decrease the relative humidity in the troposphere (below 500hPa) by up to 6%. The inclusion of SSMI data appears also to strengthen analysed surface winds and reduce them aloft by up to 0.35 ms^{-1} (Fig. 7b). The surface results may be partly due to a 'direct' link via the observation operator, which is also dependent to some degree on changes in surface emissivity (that can be brought about by the action of the surface wind on waves and sea-spray). The results aloft indicate the indirect impact of humidity on wind via interactions within the forecast model. Other OSEs performed with and without AMV (Radnoti et al., 2010) and SCAT observations indicate that these observations also lead to a similar magnitude of zonal wind reduction (not shown).

The previous section (2.1) highlighted systematic differences between some observation types and the model. In particular, the data depicts a drier and less intense monsoon circulation. Here it is shown that these differences are reflected in the mean analysis increments. Moreover, differences between the DFS and FEC clearly indicate a loss of information during the model propagation, mainly, on humidity and temperature, supporting a model error problem and pointing to a tropospheric bias.

2.3 Fast model error

In this section the focus is on the diagnosis of model errors that develop very early on in the forecast. At these timescales, errors originating from different physical processes have had little time to interact with each other, particularly remotely via the resolved flow but also possibly through direct local interaction.

Figure 8(a) shows mean 850 hPa forecast error at a lead-time of 5 days for specific humidity and horizontal wind. Errors in both fields are large relative to the mean field (*c.f.* Fig. 4a,b). For example, mean wind errors of 3ms^{-1} over the Arabian Sea compare to mean total winds of $\sim$10ms^{-1}, and mean humidity errors of 1gkg^{-1} compare to mean total humidities of $\sim$13gkg^{-1}. Local statistical significance is indicated by the bold shading and black vectors. In general, the low-level flow is too westerly and too zonal over the Arabian Sea and Bay of Bengal, with too much humidity off the Arabian Peninsula and in the Bay of Bengal. Figure 8(b) shows the mean errors at a lead-time of 1 day. In general, the same features are evident – they are weaker in magnitude but equally statistically significant. This growth of error magnitude with lead-time also points to a model problem rather than to observation bias.

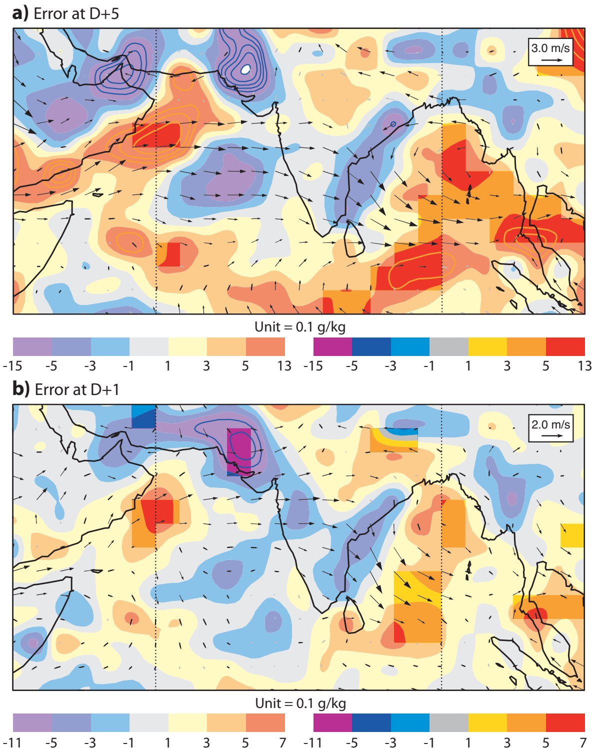

Figure 8: Mean June–August 2009 0UTC forecast error in 850 hPa specific humidity and horizontal wind relative to operational analyses. (a) At a lead-time of 5 days. (b) At a lead-time of 1 day. Units for q are 0.1gkg^{-1}. A statistical test has been performed and bold colours and black vectors indicate that the mean value is significantly different from zero at the 5% level.

In general, above the planetary boundary layer, patterns of systematic error evolve more with lead-time than indicated in Fig. 8(a,b) and they also tend to loose their statistical significance (Rodwell and Jung, 2010). Maintaining statistical significance in error diagnostics in the face of improving assimilation and forecast systems can be thought of as key to the whole error diagnosis problem. Without a reasonable signal-to-noise ratio, one cannot be sure that the error is real, let alone identify its cause. These results give credibility to the assertion that errors are best diagnosed at very short lead-times (or indeed during the first time-step in the work of Klinker and Sardeshmukh, 1992), before the interactions via the dynamics have taken place and before the state of the system has drifted from the 'true state' towards the model's manifold. With the introduction of 4D-Var, recent work (Rodwell and Palmer, 2007; Rodwell and Jung, 2008b) has suggested that the optimal period over-which to diagnose error is the data assimilation window. In this case, mean forecast error (as estimated through the difference between the forecast and a verifying analysis) is the same as (minus) the mean analysis increment. (Note the correspondence, except for sign, between the mean analysis increment (Fig. 5) and the D+1 forecast error Fig. 8b).

Mean forecast error represents an imbalance in model process tendencies. Figure 9(a–d) show specific-humidity tendencies at 850 hPa due to the explicit dynamics, vertical diffusion (including surface fluxes) and gravity-wave-drag, convection, and large-scale precipitation, respectively. (Note that it is best to avoid the first model time-step as this is structurally different, and so tendencies shown in Fig. 9 are accumulated over the lead-times 1–13hr and over the 0UTC and 12UTC forecasts). The fact that their sum (Fig. 9e) is so similar to (minus) the analysis increment (Fig. 5) confirms that these are the dominant processes acting on specific humidity. Hence, if model error is the cause for the analysis increments, one or more of these processes is at fault. Future work on the monsoon will look more closely at processes within the planetary boundary layer and will hopefully provide a fuller understanding of this model problem.

More generally, it can be seen that the diagnosis of analysis increments and initial tendencies provides an approach to model development that is very complementary to the traditional 'bottom-up' methodology where individual parametrizations are improved in isolation. The analysis increments / initial tendencies approach has several useful attributes:

- It assesses model error at states close to the true state.
- Errors are more readily identified (and more statistically significant) before interactions have taken place.
- Possible causes for the error are suggested.
- It can be used to prioritise model development work.
- It provides process-oriented metrics (useful in the climate projection context, Rodwell and Palmer, 2007)

The future challenge for the development of 'Initial Tendencies' is to get it to fully connect with the model development process. To help this approach become more widely adopted at ECMWF, better user-support and an easy-to-use way of 'driving' this diagnostic tool will be provided in future. Note, however, that an interesting poor-man's approach to initial forecast error is already being used as a complementary tool in the development of the ECMWF land-surface scheme. In this approach, 36-hour forecast errors in 2m temperature have been deduced for a model with modified land-surface, but initialised-by and verified-against the operational analysis. If such forecast 'errors' are reduced almost everywhere (as they have been) then there is less need to undertake a full assimilation experiment to confirm the benefits of the modification.

In the medium-term, ECMWF aims to introduce 'Weak-constraint 4D-Var' data assimilation into the Troposphere (it is already present in the Stratosphere). Weak-constraint 4D-Var takes account of model

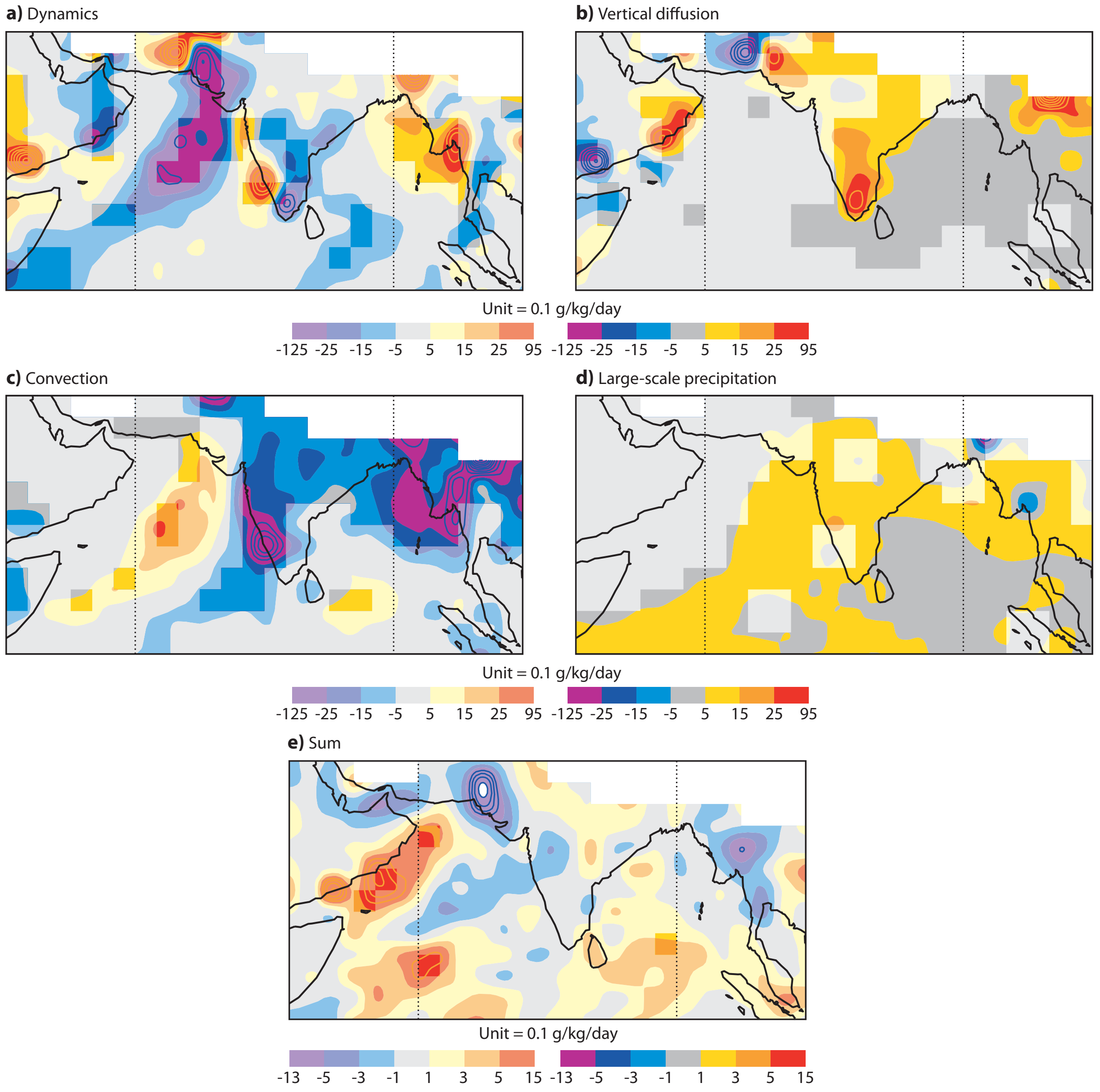

Figure 9: Mean July 2009 model process tendencies between forecast lead-times 1 and 13 hours in 850 hPa specific humidity ($g\,kg{-}1\,day^{-1}$) averaged over 0UTC and 12UTC forecasts from: (a) the explicit dynamics, (b) vertical diffusion (including surface fluxes) and gravity wave drag, (c) the convection, (d) the large-scale precipitation, and (e) the total over all the above process tendencies. A statistical test has been performed and bold colours indicate that the mean value is significantly different from zero at the 5% level. Regions where the 850 hPa surface dipped below the surface (orography) sometime during the month have been blanked-out.

‘drift’ within the data assimilation window (the $\mathscr{F}$-term of Trémolet, 2007b). In future, therefore, it will be important to use ‘$\mathscr{F}$’ instead of (or in combination with) any residual mean analysis increment when diagnosing model error.

The ‘seamless’ traceability of forecast error signals from long lead-times to shorter lead-times, initial tendency errors and, ultimately, to comparisons with observations is highly useful. The main motivation behind the development of the on-line ‘Diagnostics Explorer’ (Rodwell and Jung, 2008a) was to make this seamless traceability readily accessible to all. At present, the Explorer contains over 1 million diagnostic plots of the operational, esuite and ERA-interim systems. As discussed above, a good signal-to-noise ratio is essential for error diagnosis and so statistical significance is indicated wherever possible. Note that the Diagnostics Explorer does not produce plots on-demand since complete flexibility would require on-line access to several years of twice-daily 4D fields. Currently, this does not present any great limitation, but the strategy may need to be revised in future.

2.4 Linear model error

The tangent-linear model and its adjoint are essential components of the variational assimilation system. Furthermore, they are used to compute the singular vectors which provide initial perturbations for the EPS. The development of tangent-linear versions of components of the forecast model (say a parameterisation scheme or an observation operator) proceeds in several stages that require a range of diagnostics. Tompkins and Janisková (2004) and Lopez and Moreau (2005) describe some of the diagnostics that are being used routinely at ECMWF to monitor and help further develop the tangent-linear model. These are briefly summarised here and then an additional new diagnostic based on the computation of singular vectors is described.

Most nonlinear parameterisation schemes require some modifications before they are suitable for linearisation. Strong nonlinearities, for instance those arising from switches, require regularisation. Similarly, highly-nonlinear functions require smoothing and some perturbations need to be artificially reduced or even set to zero. Otherwise, the tangent-linear model would permit excessive growth and would poorly approximate finite differences of nonlinear integrations. Other modifications may be required for reasons of computational efficiency. The modified nonlinear scheme is sometimes referred to as simplified scheme to acknowledge the differences from the full nonlinear scheme since some physical processes included in the latter scheme are often not represented in the linearised version. As a standard procedure in this step, the errors of medium-range and extended-range forecasts with the simplified nonlinear scheme are compared to the errors obtained with forecasts using the full nonlinear scheme.

In the next step, the tangent-linear and adjoint versions of the simplified nonlinear scheme are developed. A prerequisite before proceeding to further tests, is to check the numerical correctness of the tangent-linear and adjoint code.

The main diagnostic technique for the linearised model is an analysis of nonlinear residuals of a first order Taylor approximation of finite difference. The residuals can be studied for single time steps for individual parameterisation schemes and for the evolution of the entire model over the 12-hour assimilation window. These tests can include a range of perturbation amplitudes and can test whether the residual is sensitive to the perturbation sign (for details see *e.g.* Lopez and Moreau, 2005).

The analysis of discrepancies between linear and nonlinear computations has recently been extended by comparing analysis increments produced by low-resolution minimisations using the linearised model with those evolved with the full high-resolution non-linear model in 4D-Var trajectories (Lopez, 2010). Performing this comparison in observation space permits the inclusion of observation operators in the diagnostic.

The systematic diagnosis of nonlinear residuals from the first-order Taylor formula can also exhibit spurious growth in the full nonlinear forecast model (see Sec. 4.7.2 in the Head of Research Department's progress report at the 37th Session of the SAC in 2008).

As a final test and before starting a meteorological evaluation of a new tangent-linear scheme in 4D-Var, the leading global singular vectors are computed for a wide range of meteorological situations. The singular vectors are computed for an optimisation time of 12 h and at the resolution of the 4D-Var inner loop (T_L255, 91 levels). The typical growth of the leading singular vector at this resolution and optimisation time is about one order of magnitude in terms of the total energy norm. However, if the tangent-linear version of a parameterisation is used that is insufficiently regularised the growth can be significantly increased. As an example, Figure 10 shows the growth of the leading singular vectors for the tangent-linear model with and without the tangent-linear version of the non-orographic gravity wave drag (GWD) scheme described by Orr et al. (2010). Using the tangent model with moist processes, vertical mixing and orographic GWD, the leading 12-hour singular vector grows typically one order of magnitude. With the non-orographic GWD scheme without regularisations the structure of the leading singular vector changes in many cases and the growth is up to two orders of magnitude larger in some cases. Regularisations of the tangent-linear non-orographic GWD scheme have been developed and refined by adapting those parts of the scheme that are responsible for the excessive growth of those structures identified by the leading singular vectors (see also the Section on tangent-linear and adjoint physics in the 2010 progress report). The leading singular vector computed with the regularised non-orographic GWD scheme does not exhibit the excessive growth.

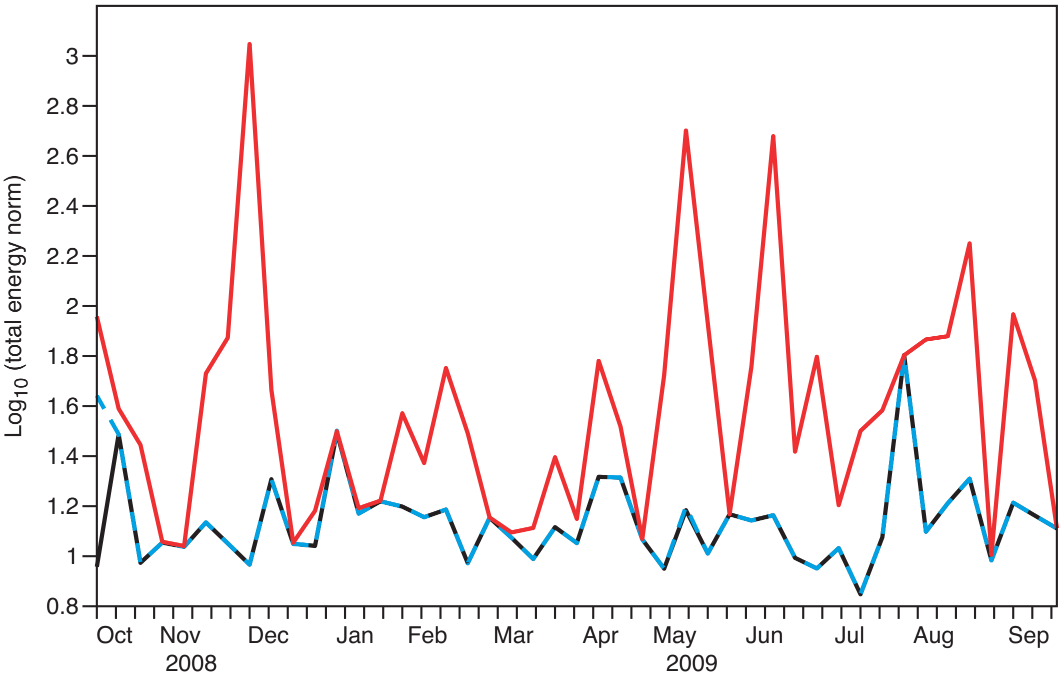

Figure 10: Final total energy norm of the leading 12-hour global singular vector computed with three different tangent-linear models: (i, black) including dynamics, vertical mixing, moist processes and orographic gravity wave drag, (ii, red) same configuration as (i) but in addition with the non-regularised non-orographic gravity wave drag scheme, (iii, blue) as configuration (i) but the regularised non-orographic gravity wave drag scheme. At initial time, the total energy norm of all singular vectors is one. A range of different meteorological situations is sampled by computing the leading singular vector for 45 cases spanning one year.

2.5 Re-Analysis

The current ECMWF reanalysis, ERA-Interim, produced data for 1989 onwards and is now running in near real time. As with any reanalysis it is important to monitor the production of the analyses and forecasts in order to ensure the data is of the highest possible quality. In reanalysis, as compared with NWP, the emphasis is on the temporal consistency of the data and as a consequence many of the diagnostics emphasise the time dimension, (see *e.g.* Dee and Uppala, 2009).

Reanalysis produces various diagnostics aimed specifically at monitoring the assimilation and forecast system. Some are standard diagnostics, such as those produced by OBSTAT (see section 2.1). Others are specific to ERA. These include timeseries of analysis differences (ERA-Interim minus operations), forecast errors (at T+12, 24 and 120) and analysis increments. Atmospheric circulation indices and budgets of mass, moisture, ozone, energy and angular momentum are also diagnosed, both for the purpose of studying the atmosphere scientifically and as a means of further monitoring the data. There are plans to greatly increase the range of these diagnostics and increase the amount displayed on the external web site[2].

Similar issues with analysis increments as discussed in sections 2.2 and 2.3 are found for the Indian

[2] http://www.ecmwf.int/products/forecasts/d/inspect/catalog/research/eraclim/timon

summer monsoon. Here, an example of a diagnostic is given that essentially assesses the temporally varying quality of the analysed divergent winds in the tropics, which is also of relevance to the monsoon. The divergent wind has long been known to be one of the greatest sources of uncertainty in atmospheric analyses (Boer and Sargent, 1985). Furthermore, it is well known that the quality of tropical analyses has a great impact on the perceived skill of tropical forecasts. The diagnostic is based on the hemispheric and vertical integral of the continuity equation. This states that the tendency in hemispheric dry mass should be equal to the hemispheric dry mass convergence.

Assuming hydrostatic balance, the northern hemispheric dry air mass (m_{NH}) can be calculated by integrating the dry density (ρ) by height (z) from the surface (SFC) to the top-of-the-atmosphere (TOA), and by area (A) over the Northern Hemisphere (NH):

$$m_{NH} = \int_{NH} \int_{SFC}^{TOA} \rho \, dz \, dA = \frac{1}{g} \int_{NH} \int_{0}^{1} (1-q) \frac{\partial p}{\partial \eta} \, d\eta \, dA \quad , \tag{1}$$

where q is specific humidity, η is the hybrid model-level coordinate, and p is pressure.

The integrated continuity equation can be written as

$$\frac{\partial m_{NH}}{\partial t} = \int_{NH} \int_{SFC}^{TOA} \frac{\partial \rho}{\partial t} \, dz dA = -\frac{1}{g} \int_{NH} \underline{\nabla} \cdot \int_{0}^{1} \underline{v} (1-q) \frac{\partial p}{\partial \eta} \, d\eta \, dA = \oint_{EQ} \int_{SFC}^{TOA} \rho v \, dz \, dl \quad , \tag{2}$$

where $\underline{v}$ is the horizontal wind. The second equality in (2) embodies the continuity equation. The final equality in (2) embodies the divergence theorem and emphasises how the hemispheric dry mass convergence is dependent on the cross-equatorial mass flux. (In the last term, v is the meridional wind at the equator, EQ, and l is distance along the equator).

The Northern Hemisphere mass tendency can thus be estimated from model level analysis data in two ways. The first 'mass tendency' method uses the third term in (1) – and takes differences between consecutive analyses. This is plotted in Fig. 11a for ERA-40 (black) and ERA-Interim (red). Because the atmosphere is largely hydrostatic, this tendency estimate is probably well constrained by surface pressure observations. Indeed there is very good agreement between the ERA-40 and ERA-Interim curves (the ERA-Interim curve lies on top of the ERA-40 curve). In addition, the globally-integrated mass tendency is very small, as it should be (not shown).

The second method uses the third term in (2), and is plotted in Fig. 11b (note the different y-axis scale). The 'mass convergence' estimate is an order of magnitude larger than the 'mass tendency' estimate – and non-zero averages over long periods in the mass convergence are clearly erroneous. Equation (2) shows that the 'mass convergence' estimate depends on the analysed density and meridional wind at the equator. The implication is that these may not be so well constrained by the observations. Indeed, the divergent wind is not well observed globally. Graversen et al. (2007) point this out and furthermore note that if mass transports are wrong, this also leads to erroneous heat and momentum fluxes.

The fact that discrepancies are much smaller in ERA-Interim than in ERA-40 is probably because observations are better able to constrain the cross-equatorial mass-flux within 4D-Var – due to enhanced internal consistency between temperature and wind fields – and because ERA-Interim uses a more recent model cycle. That the mass convergence improves with time in ERA-Interim (Fig. 11b) indicates reduced observation bias and/or increased observation coverage with time in the tropics. Nevertheless, this budget study highlights continued uncertainties in analysed tropical winds and/or densities.

There is potential to do further work on the subject of mass conservation and the continuity equation. For example, the evaluation of the dry mass convergence using 6 hourly analyses involves temporally sub-sampling this field which can lead to large errors (see, *e.g.* Haimberger et al., 2001). These errors can be avoided by accumulating the forecast convergence in the IFS at every time step, which would also enable the treatment of the continuity equation in the forecast to be studied. The diagnostic could also be applied to ECMWF's operational analyses and could be used to compare the realism of ECMWF's analysed tropical winds with those of other forecasting centres.

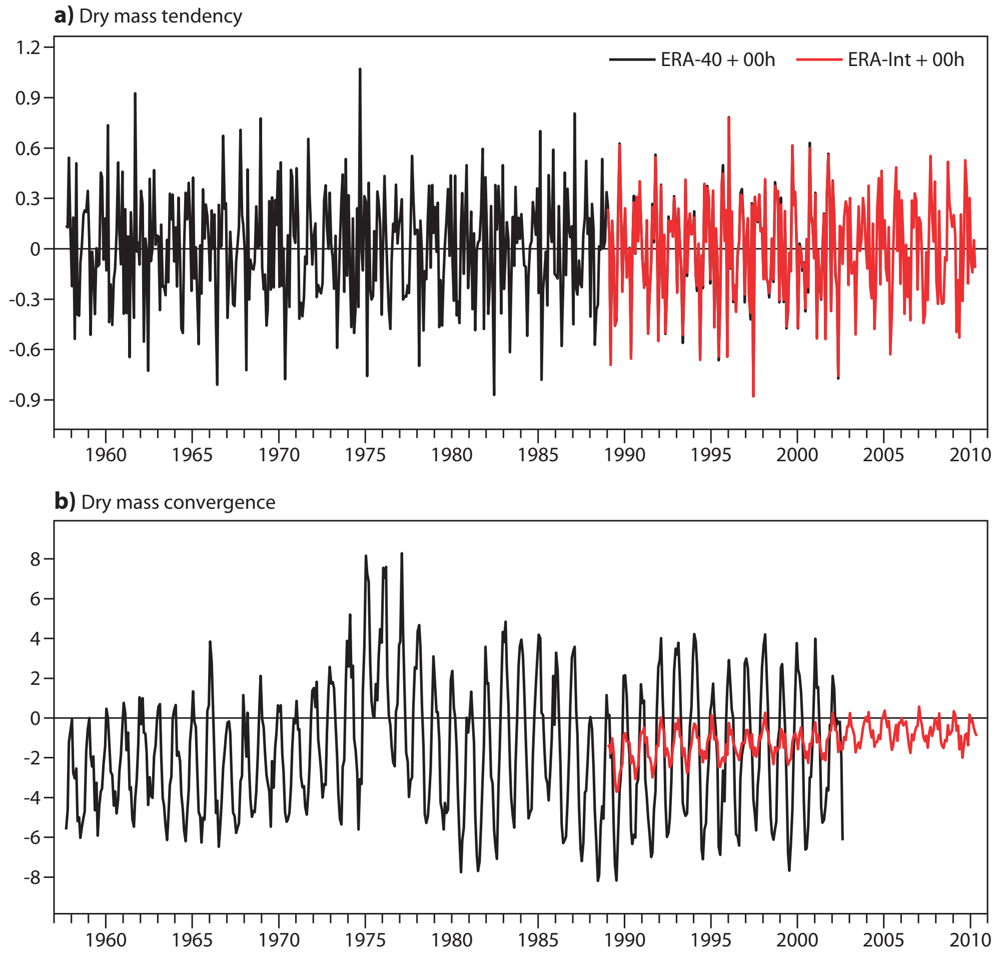

Figure 11: Time series of the Northern Hemisphere dry mass budget ($kg\,m^{-2}day^{-1}$): (a) mass tendency and (b) mass convergence.

2.6 Analysis uncertainty

With the introduction of Cycle 36R2, ECMWF runs operationally an ensemble of variational data assimilations (EDA Isaksen et al., 2010) based on the explicit perturbation of the observations, sea-surface temperature and model physics tendencies. The background states are implicitly perturbed because they are evolved from the perturbed analysis fields and include the model error parametrization term (‘Stochastically Perturbed Parametrization Tendencies’ method, SPPT Palmer et al., 2009). Standard deviations of the EDA analyses and background forecast fields provide estimates of the analysis and background errors (Fisher, 2003). These error estimates can then be used, respectively, for the initialisation of the ECMWF ensemble prediction system (EPS, Buizza et al., 2008), and to provide flow-dependent background-error variances to the operational deterministic 4D-Var (in replacement of the current quasi-static background error estimates derived from the ‘randomisation’ technique (Fisher and Courtier, 1995; Fisher, 2003)). The estimation of analysis and forecast errors through the EDA variances is affected by the small ensemble size (currently 10 members) and by systematic errors arising from approximations in the representation of observation error covariances and model error, and typically lead to an under-dispersive ensemble (Fisher, 2007). In this context the capacity to diagnose the magnitude, spatial distribution and temporal evolution of the EDA under/over-dispersiveness can give us insights into the accuracy of the prescribed observation error covariances and the behaviour of the model error parameterizations.

A useful diagnostic is the spread-error plot, an example of which is given in Fig. 12. These are obtained by binning the EDA spread (here the standard deviation of the 9-hour forecast) into deciles and plotting the mean spread for each bin against the corresponding root-mean-square error (here the error of the

9-hour forecast relative to the deterministic 4D-Var analysis – considered to be the best estimate of the truth). For a reliable EDA, these curves should lie on the diagonal (dashed black line). Fig. 12 shows two such curves based on the monsoon region [45^{o}E-105^{o}E, 0^{o}N-30^{o}N] (green) and on the South Atlantic sub-tropical anticyclone region [60^{o}W-0^{o}E, 30^{o}S-0^{o}N] (red) during August 2008. The relative flatness of both curves indicates that the spread is not a perfect estimator of error in these regions.

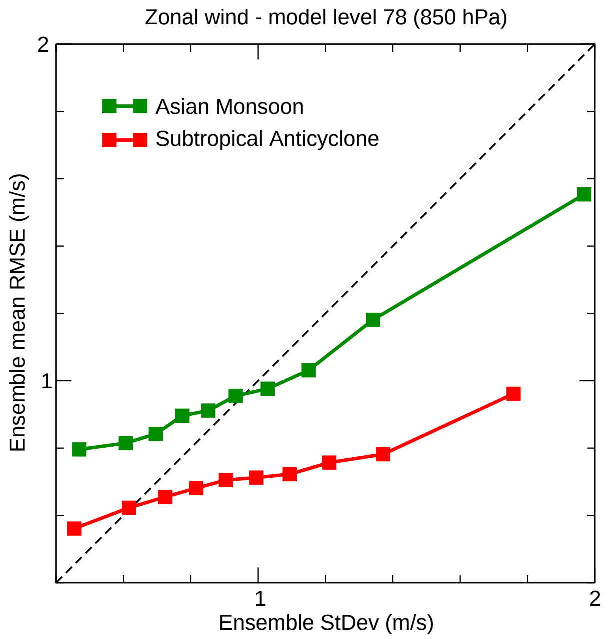

Figure 12: Spread-Error diagram for zonal wind (ms^{-1}) at model level 78 (about 850 hPa) for the Asian monsoon region [45^{o}E-105^{o}E, 0^{o}N-30^{o}N] (green), and the South Atlantic subtropical anticyclone region [60^{o}W-0^{o}E, 30^{o}S-0^{o}N] (red). The spread is of the 9-hour EDA forecasts (binned into deciles). The corresponding RMSE is of the EDA 9-hour ensemble mean forecast relative to the deterministic 4D-Var analysis. Results are based on data from August 2008.

Because these curves condense a lot of information, their main use is for monitoring the EDA as a whole. However, they could help in formulating sensitivity experiments to understand, for example, the reasons for differences between regions – which could include model error, assumed observation and background error covariances, data coverage etc.

2.7 Physics impacts

The model physics comprises processes such as radiation, moist convection, clouds and grid-scale microphysics, boundary-layer turbulence, and surface processes. Whenever possible each physical process (scheme) is evaluated both separately and in the context of the entire IFS model framework. The specific modelling ‘tools’ used to evaluate the individual processes against observational data include:

- Diagnostic or prognostic off-line versions of the different schemes.
- A single-column model version of the IFS that is forced by observed or analysed large-scale tendencies and/or surface fluxes.
- So-called ‘DDH’ files of high temporal resolution operational model output for a predefined set of locations.
- A climate diagnostics package, consisting of a 4-member ensemble of 1-year integrations at resolution T$_L$159, that is optimised to allow for quick testing and turnaround.

Each scheme is compared against relevant observational datasets. For example, radiation and cloud schemes are compared against radiative fluxes and cloud products from ground based stations (ARM, BSRN) and satellites (CERES, Cloudsat/Calipso) (*e.g. Ahlgrimm and Köhler, 2010; Morcrette et al., 2008*). For the future, it is also planned to include datasets that can address aerosol-cloud-radiation interactions.

The surface scheme is compared against data from flux towers, surface stations and measurements of river discharge over specified basins. As discussed in section 2.3, short-range forecast error is also used to assess changes to the surface scheme.

The convection and boundary-layer turbulence schemes within the single column model can initially be compared against data collected during specific field campaigns, *e.g.* GATE and TOGA-COARE (Willett et al., 2008; Bechtold et al., 2004) for convection and GABLS, BOMEX, ASTEX, DYCOMS (*e.g. Zhu et al., 2005*) for the boundary layer. However, due to strong interaction with clouds, radiation and the dynamics it is imperative to assess, at an early stage, the convection and boundary-layer schemes within the full three-dimensional model.

Increasingly popular datasets used for physics validation include satellite imagery from geostationary satellites (infrared and water vapour bands) and precipitation data from both space-borne platforms (*e.g.* TRMM) and ground-based radars (OPERA and NEXRAD data sets). For the radar data comparison is mainly between derived rain rates (and not radar reflectivities) and model output. For geostationary imagery, a model-to-satellite approach has been chosen where, for given spectral bands, the radiation that would be measured by the satellite is deduced from the predicted model state.

An example of the model-to-satellite approach is given in Figs. 13(a,b) which show, for 31 July 2009 at 00UTC, infrared brightness temperatures in the 10.3μm band from MTSAT and the IFS analysis, respectively. These brightness temperatures approximately reflect cloud-top temperature. In general, cloud top temperatures appear to be reasonably well predicted by the combination of forecast model and observation operator. Fig. 13(c) shows monthly mean D+5 errors in 00UTC brightness temperatures for July 2009 based on the operational forecasts. Since MTSAT is not continuously archived at present, the analysis is used here as a surrogate for the observations primarily to demonstrate the diagnostic approach. However, consistent with Fig. 2(g), these mean ‘errors’ clearly indicate too much convection over the Indian Ocean (cold brightness temperature errors) and too little cloud and convection over the Bangladesh region (warm brightness temperature errors).

2.8 Dynamics-physics interactions

In this section the focus is on the diagnosis of atmosphere-only forecasts beyond the short-range. When diagnosing medium-range and longer forecasts, challenges arise from the fact that different dynamical and physical processes have time to interact (possibly nonlinearly), which makes it increasingly difficult to distinguish cause and effect.

Most of the diagnostic research in this area is based on the application and development of techniques that can be summarised by the expression ‘targeted numerical experimentation’. By routinely applying a hierarchy of models with varying degrees of complexity (but within the framework of the same IFS cycle) the aim is to better understand the impact of model changes. The most important techniques, with strong strategic relevance, will be discussed in the following.

2.8.1 Sensitivity experiments

The monitoring of the climate of the model plays a central role in the diagnostics work of ECMWF. For each new model cycle, long integrations with the IFS are carried out and various aspects of the model climate and its variability are assessed (see also section 2.7). These experiments are augmented by additional sensitivity experiments in which the impact of selected physics changes are studied in further detail. A comprehensive study has been recently completed (Jung et al., 2010a). Furthermore,

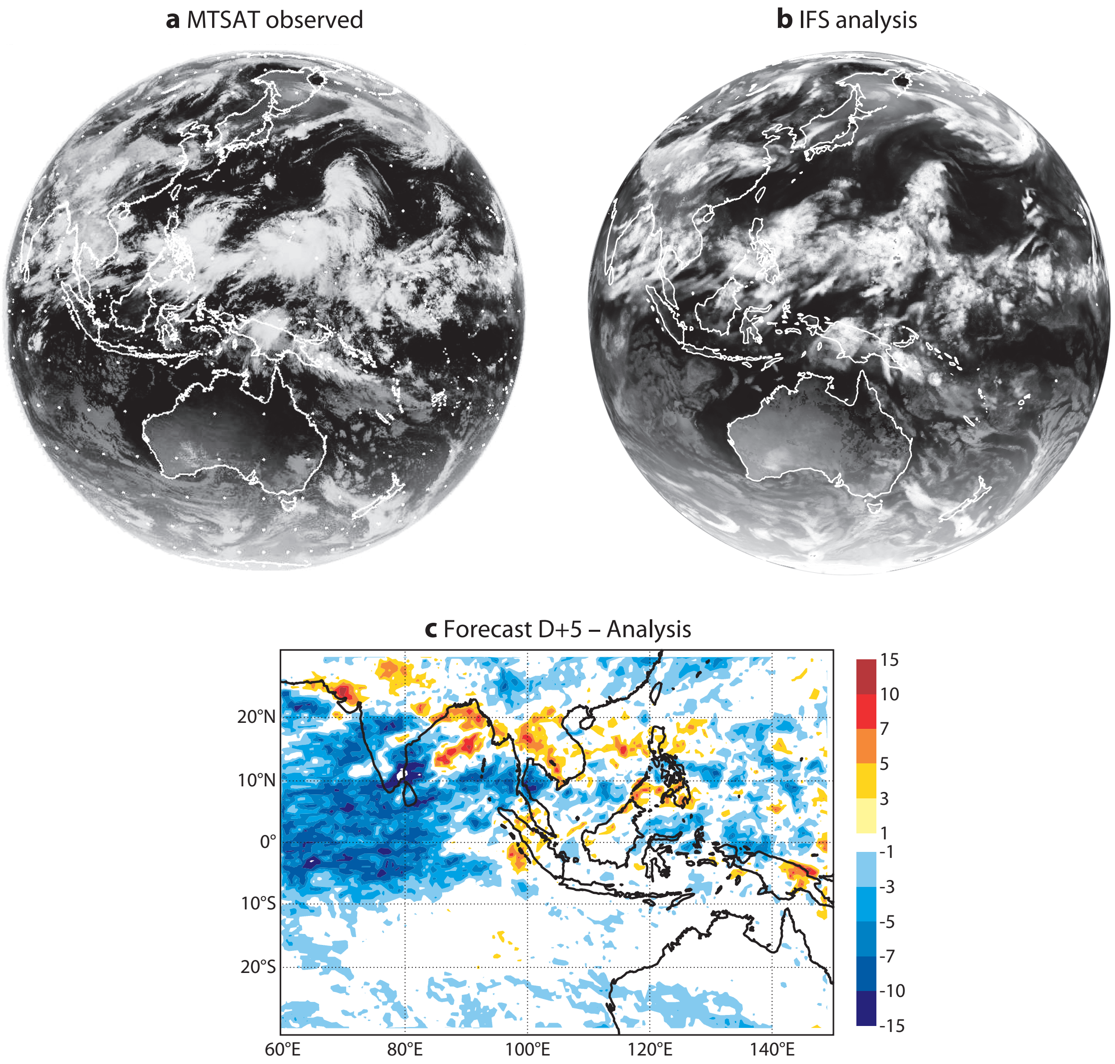

Figure 13: Infrared brightness temperatures for 31 July 2009 00UTC (a) as observed by MTSAT and (b) from the ECMWF analysis. Also shown are (c) monthly mean brightness temperature differences (K) for July 2009 at 00 UTC between the D+5 ECMWF forecasts and the verifying analysis fields.

the sensitivity of the ECMWF model climate to resolution has been studied in unprecedented detail (T_L159, T_L511, T_L1279 and T_L2047) in the framework of the Athena project (dedicated use of the NSF Cray XT4 *Athena* supercomputer in the US). In the past, most of the focus in these investigations has been on synoptic and planetary-scale aspects of the atmosphere; severe weather has been considered only indirectly. However, ongoing activities at ECMWF and the partner institutions in diagnosing the results from the Athena project will concentrate on severe weather (*e.g.*, extratropical wind storms, tropical cyclones, and tips jets). The diagnostics developed in this framework will be incorporated in the standard diagnostics software packages.

A key-element for the comprehensive assessment of these experiments is the publication of the plethora of generated plots on the Diagnostics Explorer in a way that is readily accessible to scientists within ECMWF (Rodwell and Jung, 2008a).

From these seasonal integrations it is very difficult to infer the origin of systematic model error and to understand exactly how physics changes influence the model climate. In order to shed some light on the underlying mechanisms it has been found useful to study the influence of a particular physics change over a range of time scales (from hours to months) (Rodwell and Jung, 2008b; Jung et al., 2010a). This 'seamless diagnostic perspective' is illustrated in Fig. 14, which shows how the new convection scheme, introduced in cycle 32R3 (Bechtold et al., 2008), influences mean 500 hPa geopotential height (Z500) fields as early as D+1. This suggests that some of the improvements in the Northern Hemisphere atmospheric circulation seen with the introduction of the new convection scheme are local (*i.e.* involving dynamics-physics interactions in the extratropics) rather than remotely forced from the tropics (see Jung et al., 2010a, for details).

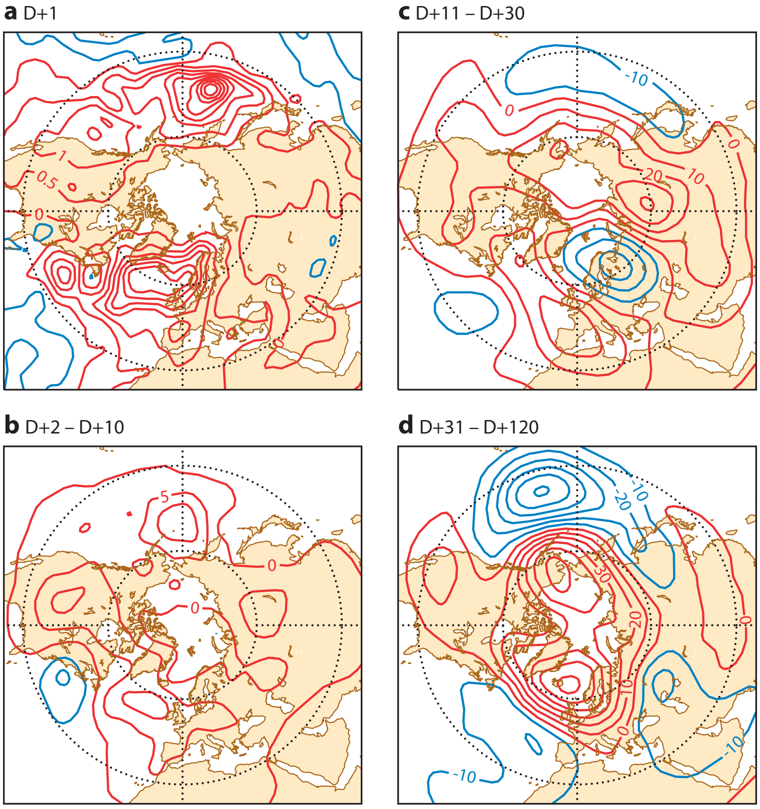

Figure 14: Mean difference in 500 hPa geopotential height (in m) between experiments with the 'new' (cycle 32R3) and 'old' (cycle 32R2) convection scheme: (a) D+1, (b) D+2–D+10, (c) D+11–D+30 and (d) D+31–D+120. Contour interval is (a) 0.5 m, (b) 5 m and (c)–(d) 10 m. From: Jung et al. (2010a).

2.8.2 Idealised simulations on the sphere: Held-Suarez and aquaplanet experiments

A substantial reduction in complexity is achieved by considering dry 'Held-Suarez' model climate simulations (Held and Suarez, 1994) designed to evaluate dynamical cores of atmospheric general circulation models independently of the physical parametrizations. Global Held-Suarez simulations on the sphere display two symmetric zonal mean zonal flow jets enclosing a simplified tropical regime – enforced by a simple relaxation to a prescribed equilibrium temperature on the sphere and the addition of a prescribed momentum dissipation (Held and Suarez, 1994). For example, Fig. 15(a) illustrates the decrease of zonal mean temperatures in the equatorial stratosphere with increased horizontal resolution. This change suggests that the similar change seen in the full model (Fig. 15b from the Athena results) can be explained as primarily a response of the dynamics to resolution. There is also a direct dynamic cooling response in the equatorial troposphere (Fig. 15a), but this appears to be 'masked' when the physics is allowed to interact (Fig. 15b). This separation of dynamical responses from the total response is an important aspect of this diagnostic tool.

Held-Suarez simulations also exhibit substantial extratropical variability. Hence the Held-Suarez setup may also be used to explore tropical-extratropical interaction with and without elaborate physical parametrizations or moist processes (Wedi, 2010).

A lesser reduction of complexity is achieved with aquaplanet simulations. Aquaplanet models consist of a dynamical core and the physical parametrizations of radiation, boundary layer turbulence, convection

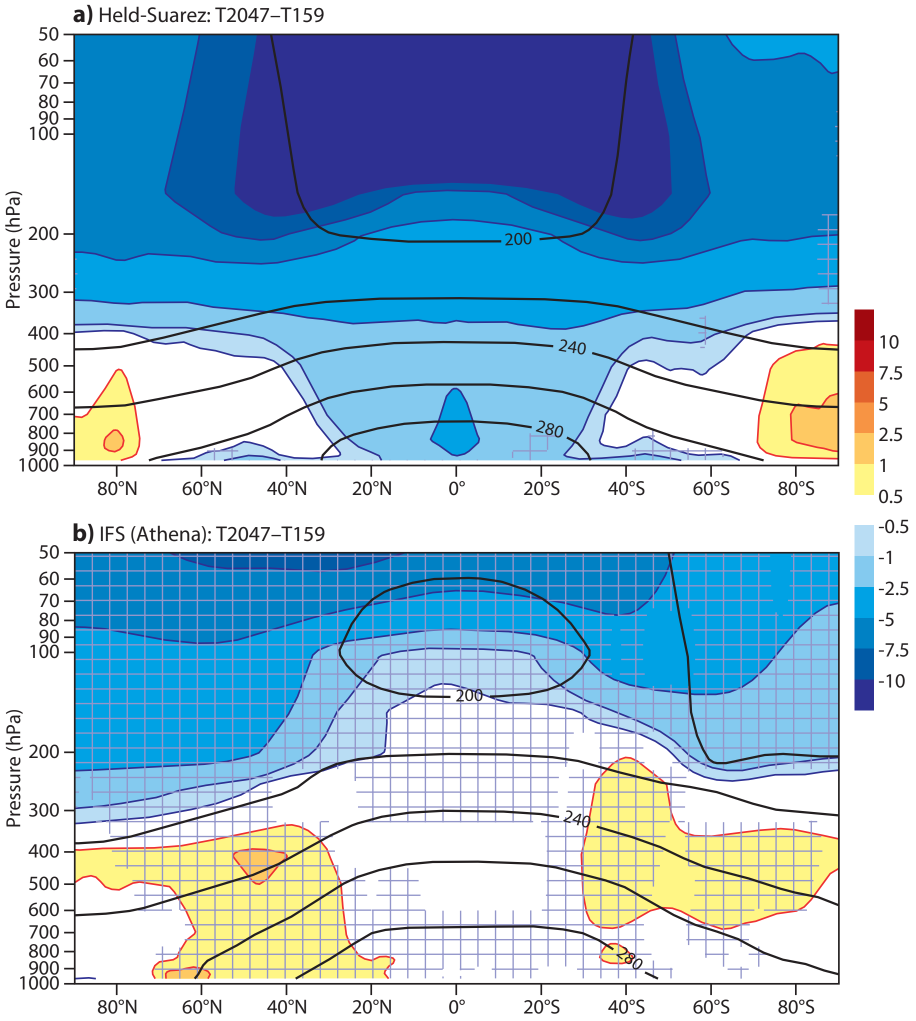

Figure 15: Mean zonal average temperature differences (K) between runs at T_L2047 and T_L159 during the September–November (SON) season. (a) Dry Held-Suarez simulation (100 days of integration). (b) Atmospheric component of the ECMWF model (1989–2007).

and clouds. Global simulations are made with an underlying flat sea surface and the model is forced by a family of sea-surface temperature (SST) distributions and prescribed external forcings. The Aquaplanet Intercomparison Project (APE) (Neale and Hoskins, 2000) documents the inter-comparison of various models for given SST distributions in a forthcoming ATLAS[3].

The application of both Held-Suarez and aquaplanet simulations at every relevant model cycle will support the early diagnosis of changes in model behaviour, while substantially saving computational cost, in particular when combined with the reduced-radius approach described in the following section.

2.8.3 *Idealised simulations on the sphere: reduced-radius experiments*

An alternative diagnostic strategy can be exploited for the 3D global atmosphere, where the planetary radius is suitably reduced to capture non-hydrostatic phenomena without incurring the computational cost of actual simulations of weather and climate at non-hydrostatic resolution $\Delta x \leq \mathscr{O}(2)$ km (Wedi and Smolarkiewicz, 2009). In other words, the size of the computational domain is reduced without changing the depth or the vertical structure of the atmosphere. Here the underlying assumption is that the essential flow characteristics remain unchanged when the separation of horizontal and vertical scales is reduced (Kuang et al., 2005). The usefulness of this strategy is illustrated in Wedi and Smolarkiewicz

[3] http://www.met.reading.ac.uk/ mike/APE

(2009) and Wedi et al. (2009) with a set of canonical flow problems including horizontally and vertically propagating spherical acoustic-waves, ‘local-scale’ orographically forced gravity waves, and ‘global-scale’ planetary Rossby waves in Held-Suarez simulations. In Wedi and Smolarkiewicz (2009) the test cases were conducted for two very different global dynamical cores – the non-hydrostatic IFS and the multi-scale anelastic research code EULAG (Prusa et al., 2008).

An interesting example is the ability to simulate propagating waves at internal critical levels, *i.e.* the height at which the background flow (in the direction of wave propagation) is equal to the horizontal phase speed of the wave. Waves attenuate as they approach their critical level and momentum carried by these waves is transferred to the mean flow at and below that height. This scenario together with convective and shear instabilities forms “critical layers”, which are an important aspect of mesoscale orographic flows and of circulation features such as the quasi-biennial oscillation (Wedi and Smolarkiewicz, 2006). Figure 16(a–c) illustrate the EULAG, non-hydrostatic and the hydrostatic IFS results, respectively, for the case of non-linear flow past a three-dimensional hill in the presence of a critical level (Grubišić and Smolarkiewicz, 1997). The critical level is a preferred location for internal wave breaking, with the resulting flow locally non-linear and non-hydrostatic. In the case of the non-hydrostatic simulations the forming homogeneous mixed layer acts as a perfect reflector to all incoming waves. In contrast the hydrostatic solution (Fig. 16c) shows also a wave response above the critical layer. The simulations indicate a sensitivity to both the horizontal and vertical resolution as well as the model equations used, thus complementing the diagnostic tools for the verification of the dynamical core of the IFS.

A future focus is targeting idealised moist simulations with resolved and unresolved convection in the small-planet environment. This will help explore the ‘grey-zone’ boundaries of hydrostatic and non-hydrostatic regimes as well as resolved and unresolved convection regimes in the broader context of physics-dynamics coupling and interaction. The small planet approach is thus complementary to research projects investigating scale interactions of the tropical atmosphere (*e.g.* CASCADE) and to ultra-high resolution limited-area studies performed in the member states.

2.8.4 Relaxation experiments

The relaxation technique is a well known diagnostic tool, and has been used at ECMWF in the past (Hasler, 1982; Ferranti et al., 1990). The basic idea is to artificially suppress the development of forecast error in a given region, by relaxing back to a set of analyses, and to then investigate the consequences in other regions. In its current implementation (Jung et al., 2008, 2010c,b; Jung and Rodwell, 2010) this involves adding an extra tendency term to the model of the form:

$$-\lambda(\mathbf{x}-\mathbf{x}_{ref}) \quad , \tag{3}$$

where the model state vector is represented by $\mathbf{x}$, and the analysis fields towards which the model is drawn by $\mathbf{x}_{ref}$. The e-folding timescale of the relaxation, λ^{-1}, is a function of the variable, location and height.

The method is illustrated in Fig. 17, which shows Z500 anomalies for DJF 2009/10 from ERA-Interim, the control integration (prescribed SST and sea ice), the experiment with tropical relaxation and the experiment with relaxation of the Northern Hemisphere stratosphere. Given that none of the experiments manages to reproduce the negative phase of North Atlantic Oscillation (NAO) suggest that internal (extratropical) atmospheric dynamics explain the unusual circulation during the winter of 2009/10 making it difficult for seasonal forecasting systems to predict the onset of the negative phase of the NAO. Interestingly, tropical relaxation managed to reproduce the negative phase of the NAO for the winter of 2005/06 (Jung et al., 2010c) showing that the relaxation method is capable of discriminating between different origins of extratropical circulation anomalies.

The relaxation approach is a powerful tool to understand possible remote origins of forecast errors (Jung et al., 2010b) and circulation anomalies (Jung et al., 2010c). Applications of this technique are

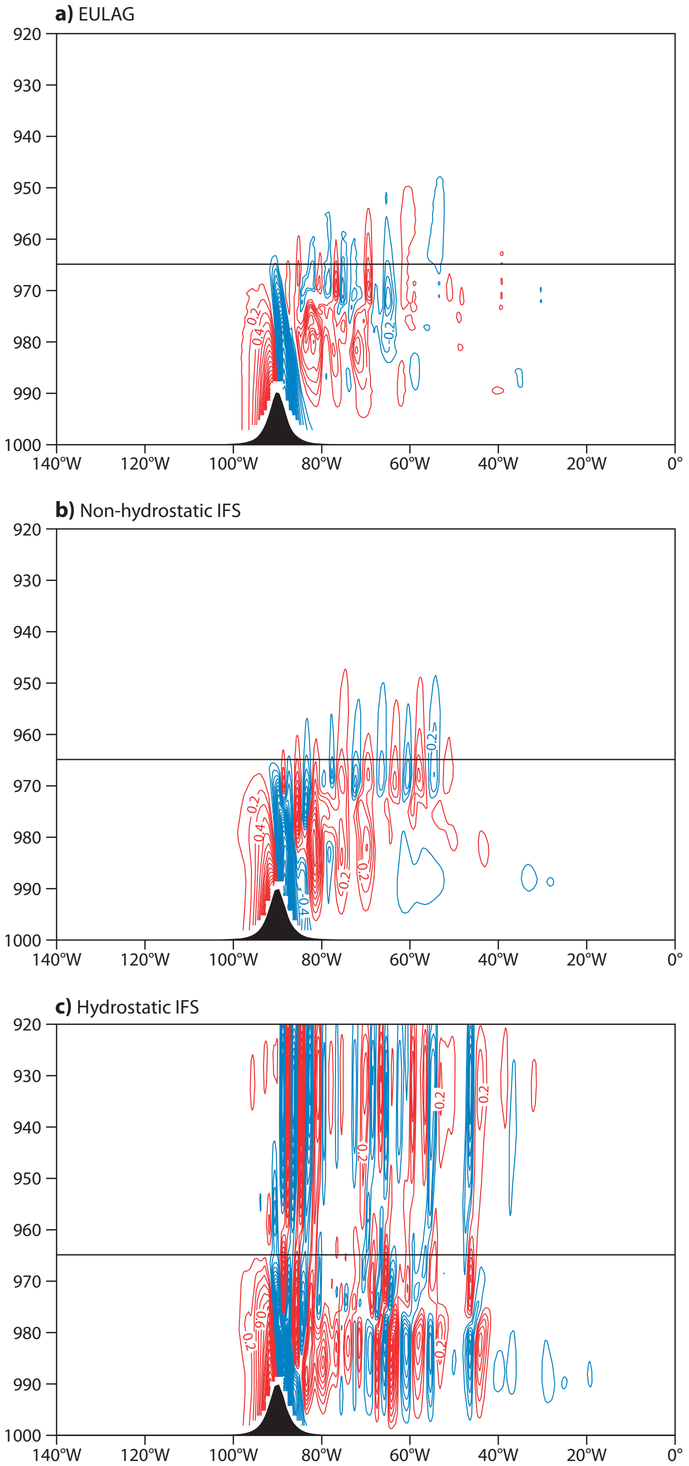

Figure 16: Cross section of vertical velocity for the case of non-linear flow past a three-dimensional hill in the presence of a critical level (indicated by the thin horizontal line). (a) EULAG,(b) non-hydrostatic and (c) hydrostatic version of the IFS. All integrations were carried out on a 'small planet' with a radius of about 20 km (and about 250 m horizontal and 35 m vertical resolution). The contour interval is 0.2 m s^{-1}. The vertical axis is pressure in hPa. For details see Grubišić and Smolarkiewicz (1997).

particularly well suited to understanding larger-scale aspects of the atmospheric circulation – for example the remote impacts of the over-active monsoon. While this excludes to some degree the application to severe weather events, the relaxation is a powerful approach to diagnose more persistent, large-scale *high-impact* events, such as the MJO, Euro-Atlantic blocking, and the NAO, whose simulation and pre-

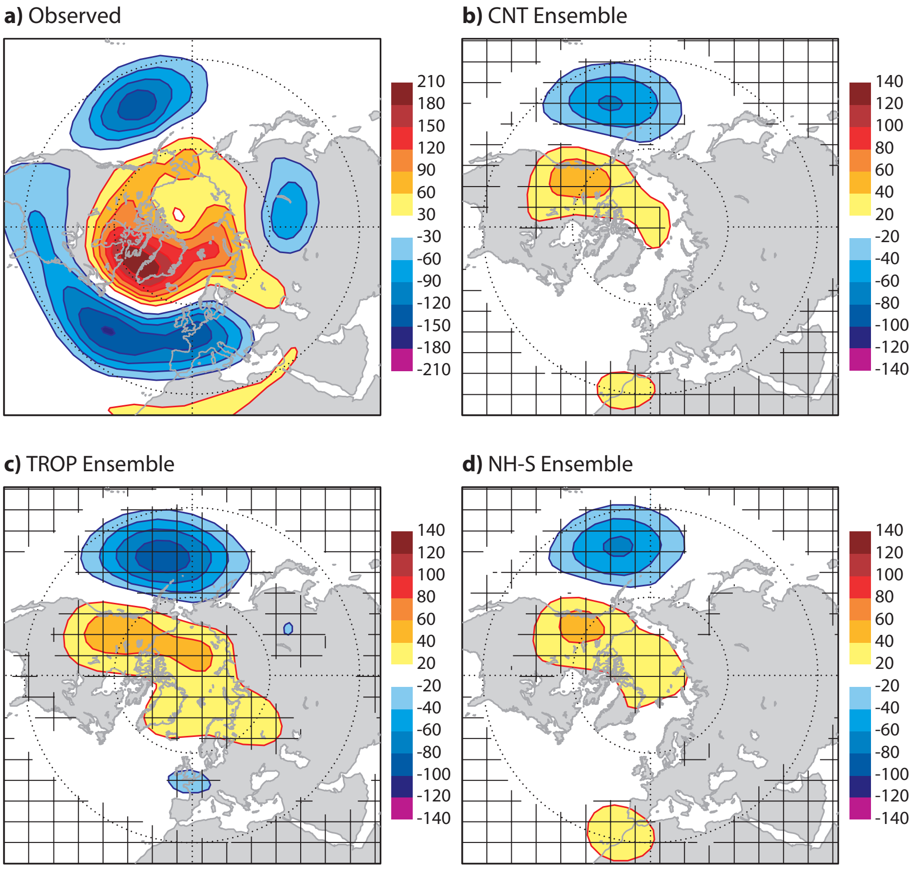

Figure 17: Geopotential height anomalies at the 500 hPa level (in m) for the period 1 December 2009 to 28 February 2010: (a) ERA Interim and ensemble forecasts (b) without relaxation, (c) with tropical relaxation and (d) with stratospheric relaxation. All forecasts were started on 1 November. Statistically significant differences (at the 95% confidence level) in (b)–(d) are hatched.

diction still pose serious challenges. For the future it is planned to further develop this relaxation code on the IFS (*e.g.* scale-dependent relaxation and relaxation of specific humidity) and to apply it in a wider context.

2.9 Coupled processes

This section deals with the evaluation of errors in the coupled ocean-atmosphere system on the timescales relevant to the monthly and seasonal forecasts. The errors coming from the each individual component are analysed separately as well as in the coupled system.

2.9.1 Diagnosis of ocean data assimilation

There are major changes taking place in ECMWF's coupled forecasting system at the moment, with the NEMO ocean model replacing the old HOPE model, and with the adoption of the NEMOVAR assimilation system. NEMOVAR is currently being implemented for 3D-Var assimilation of temperature and salinity profiles, ocean currents and satellite altimeter data (Mogensen et al., 2009). In addition, it is forced with ERA-Interim surface fluxes of sensible-heat, moisture, momentum and (penetrative) radiation. Furthermore, SSTs are strongly relaxed to mapped SST observations (with a timescale of ∼2 days) and weakly relaxed to salinity climatology (with a timescale of ∼2 years). With this approach, mean analysis increments should reflect systematic errors in the representation of surface fluxes and/or the ocean model (*i.e.* not the entire coupled model) and so diagnostics should be complementary to those obtained from the NWP assimilation.

An example of NEMOVAR ocean analysis increments is given in Fig. 18, which shows mean ocean potential-temperature increments at the equator. In the Eastern Pacific increments remove heat below the thermocline and add heat above while, in the western Pacific, the opposite happens. This has the effect of increasing the tilt of the thermocline. These increments could partly be reflecting ocean model error. However, these increments are common to other ocean models forced by the same surface winds. In addition, a joint diagnostic project with the UK Met. Office has highlighted large differences in seasonal-means of analysed zonal winds at 850 hPa in this region (up to 2ms^{-1} – *i.e.* 25% of the total). Hence it is quite likely that the surface momentum fluxes used in the assimilation also play an important role in generating these ocean increments.

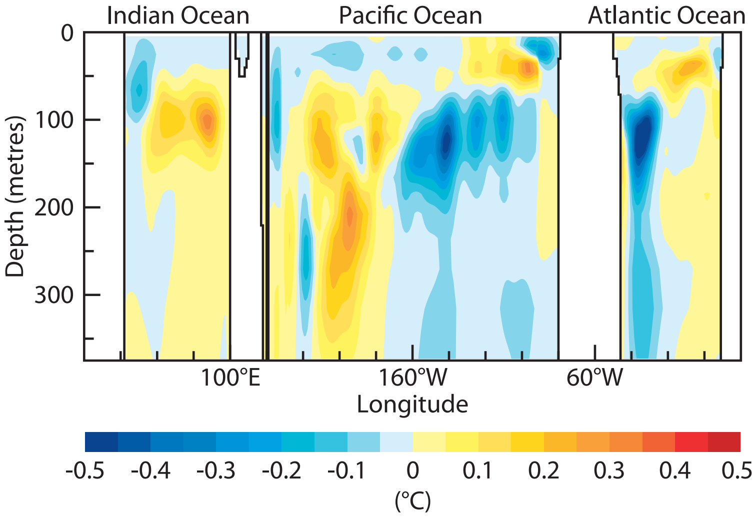

Figure 18: Equatorial cross-section of temporal mean, 10-day window, potential temperature increments (^{o}C) from NEMOVAR 3D-Var assimilation. The averages have been computed for the period of 1989 to 2006.

Research diagnostics also include the comparison with independent (non-assimilated) observations and with other datasets such as the World Ocean Atlas 2005 (WOA05). As with NWP (section 2.2), the ocean data assimilation is also assessed by conducting Observing System Experiments (OSEs). Indeed, the ocean data assimilation system provides a good environment to diagnose errors in surface fluxes and the ocean model. Although discussed here, section 2.9.1 clearly fits well within the left-hand yellow box of Fig. 1.

2.9.2 Diagnosis of large-scale variability and teleconnections in coupled and SST-forced hindcasts

In long-range forecasts performed with coupled models, errors may originate either from an incorrect prediction of SST variability, or from an inability of the coupled model to reproduce the correct atmospheric response to SST anomaly.

In order to investigate the origin of seasonal forecast errors, a 41-member, uncoupled, seasonal hindcast ensemble forced with observed SST is produced each quarter. It allows a quick assessment of whether a particular seasonal-mean atmospheric anomaly could have been forecast if the correct SST had been predicted. Results from those uncoupled seasonal hindcasts are accessible on the internal web, and are systematically used during the joint meeting of the Operations and Research Divisions when discussing the seasonal forecast performance.

In addition, a diagnostic package has been developed to assess the correspondence between patterns of large-scale variability and teleconnections simulated by the model and those found from observational datasets and re-analyses. The package uses regression and EOF analysis to compute a set of climate variability patterns which are relevant to seasonal and inter-annual time-scales. The diagnosed fields can be classified into three main areas:

- Variability/covariability of SST and heat/water fluxes
- Atmospheric response to SST variability
- Dominant modes of atmospheric variability/covariability

The similarity between patterns derived from observational and GCM datasets is quantified with the use of 2-dimensional 'Taylor' diagrams. Distances in Taylor diagrams are used to define metrics which can be used as 'figures of merit' in the evaluation and comparison of GCM simulations. The package can be used to compare the variability generated by different model cycles, or to evaluate how variability and teleconnection patterns differ in SST-forced versus coupled simulations.

An example of the output of this diagnostic package is presented in Fig. 19 and Fig. 20. These figures highlight the links between SST and monsoon precipitation variations in (a) ERA-Interim, (b) the uncoupled hindcasts using system 3, (c) the coupled forecasts using system 3, and (d) in a 5-member, 20-year hindcast set with the new coupled model (IFS-36R1/NEMO). Figure 19 shows covariances with the first principal component of precipitation within the red box indicated. The patterns for the observations and coupled simulations (Fig. 19a,c,d) look rather similar and incorporate a clear east-west rainfall dipole. The new coupled model (Fig. 19d) appears to produce a stronger dipole than that of the re-analysis and coupled System 3 (Fig. 19a,c), although the percentage of variance explained by the first PC is in close agreement with the re-analysis value (16.2% versus 15.5%). The precipitation pattern from the uncoupled simulations (Fig. 19b) is strikingly different and emphasises a north-south tripole. In the uncoupled hindcasts, the first PC of South Asian rainfall explains a significantly larger proportion of variance (21%) than the first PC of ERA Interim, while the second PC (associated with stronger east-west rainfall gradients) accounts for about 11% of the variance.

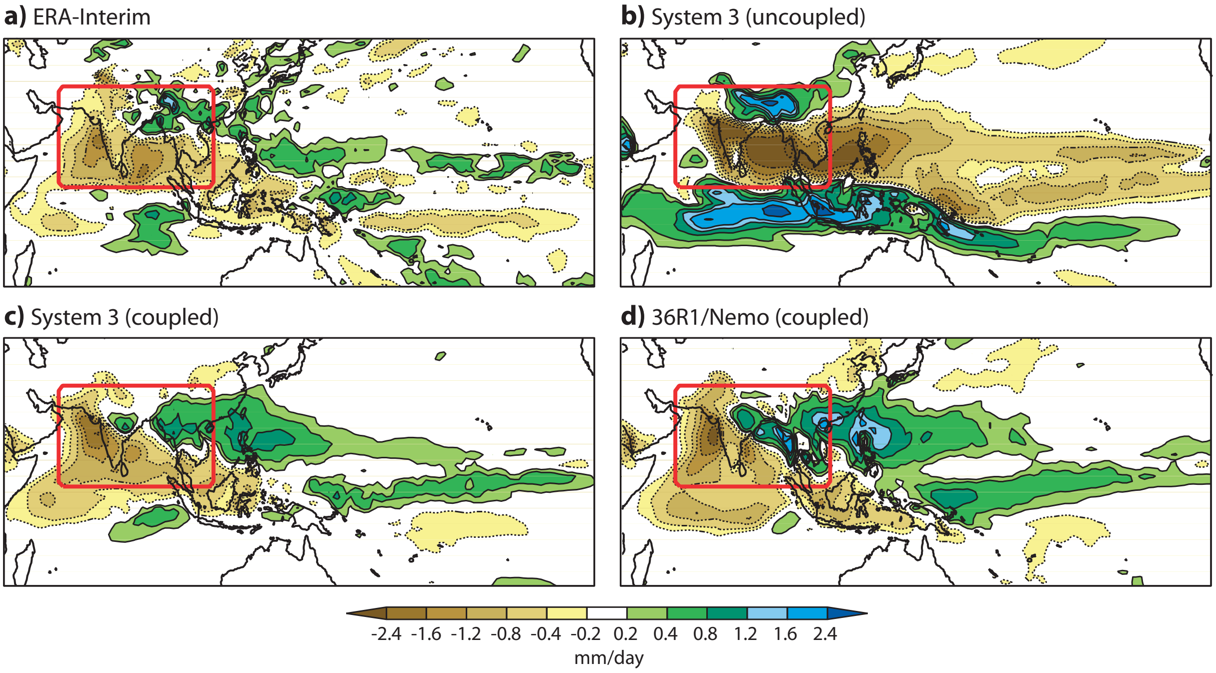

Figure 19: Covariances of precipitation over the whole domain shown with the first principal component of precipitation in the red box indicated. Results are based on (a) ERA-Interim, (b) the uncoupled hindcasts using system 3, (c) the coupled forecasts using system 3, and (d) a 5-member, 20-year hindcast set with the new coupled model (IFS-36R1/NEMO).

In Fig. 20, the covariance of SST with South Asian rainfall PC1 is displayed. Both coupled models (Fig. 20c,d) show stronger covariance with El-Niño – La-Niña SST anomalies than that observed

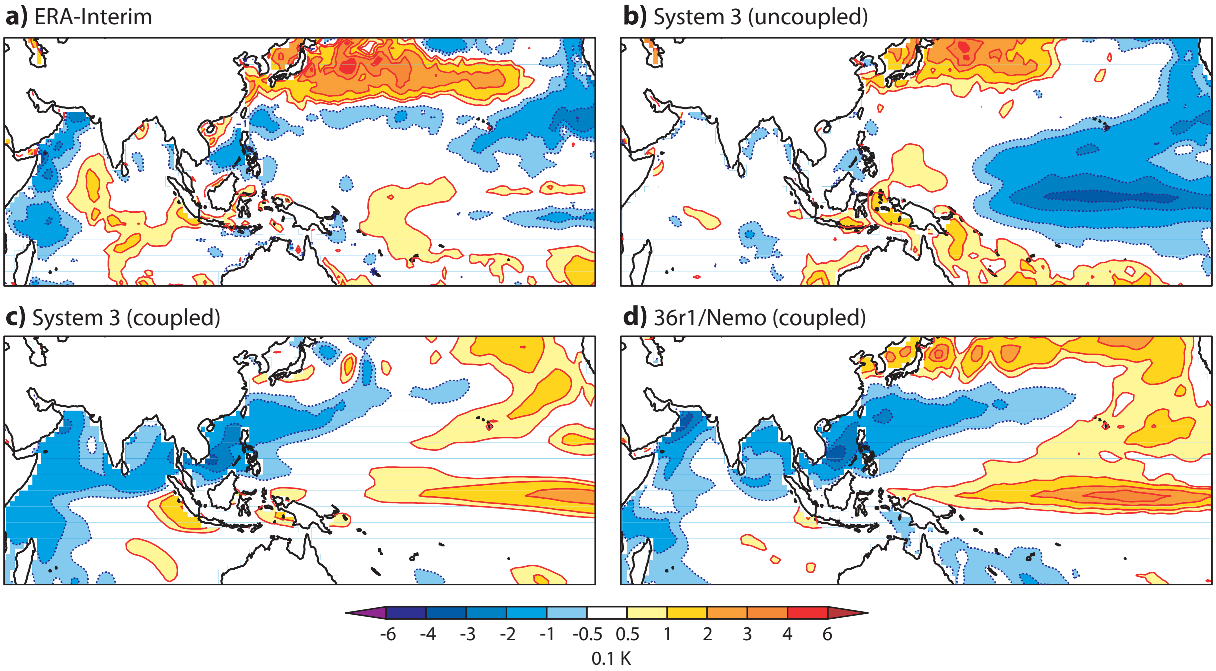

Figure 20: The covariance of SST with the first principal component of precipitation as in Fig. 19. Results are based on (a) ERA-Interim, (b) the uncoupled hindcasts using system 3, (c) the coupled forecasts using system 3, and (d) a 5-member, 20-year hindcast set with the new coupled model (IFS-36R1/NEMO).

(Fig. 20a). The more westerly position of the maximum of equatorial SST anomalies in the IFS-36R1/NEMO system is a consequence of its cold bias in the Pacific cold tongue. The very different precipitation pattern for the uncoupled model is connected to a rather different covariance with SST (Fig. 20b), but it is interesting to note that the uncoupled simulations suggest an important role for extratropical SST anomalies over the northwestern Pacific – as also indicated by the observations.

These differences underline the importance of coupling for getting the correct variability of monsoon precipitation. In the context of this paper, it emphasises the importance of a variety of diagnostic tools that can assess the effect of air-sea interactions and coupled processes on the modelled variability and, consequently, on the pattern and intensity of predicted anomalies.

2.9.3 Diagnosis of the mixed-layer model

In section 1.5, it was shown that the monsoon is less over-active in coupled simulations (Fig. 2q) than it is in uncoupled simulations (Fig. 2n). This can be explained through mixed-layer diagnostics presented in Takaya et al. (2010). They show how Indian Ocean SST (particularly around 10^oS) becomes cold relative to SST analyses over the first 10 days during July experiments. In effect the over-strong monsoon surface winds in the atmospheric model lead to increased latent heat fluxes, but these are moderated (in the medium-range) if the ocean is able to respond. Hence the monsoon in the coupled model is provided with less moisture than it is in the uncoupled model. Very similar arguments apply for the weakening of tropical cyclones when they are allowed to interact with the ocean (Bender et al., 1993; Wada, 2009). The cooling effect is actually underestimated by the mixed-layer model when compared with TMI observations and this may be explained by the fact that the implemented mixed-layer model does not represent upwelling.

2.9.4 Sensitivity experiments with the coupled system

Numerical experiments in which specific regions of the ocean are relaxed towards observed conditions have been used recently with the IFS/HOPE system to assess the impact of SST biases in the Gulf Stream region (Balmaseda et al., 2010). Regional SST relaxation is going to be implemented in the IFS/NEMO and IFS/high-resolution-mixed-layer systems.

Experimentations designed to evaluate the impact of the land surface initial conditions (Ferranti and Viterbo, 2006; Weisheimer et al., 2009) have shown that initial memory in the land surface is an important predictability source. More recently, coupled simulations using observed sea-ice concentration ((Balmaseda et al., 2010)) have shown its impact on the atmospheric circulation at seasonal time-scales.

2.9.5 Other diagnostic tools

A seasonal forecast suite developed in collaboration with the Ensembles project and Meteorological Operations section provides a comprehensive set of basic diagnostics and verification statistics (such as biases, timeseries of atmospheric and oceanic indices, deterministic and probabilistic scores) for any seasonal forecast experiment. This suite is instrumental in the development of the new seasonal forecast System 4. Results from this suite are accessible on the internal web.

Seasonal-mean teleconnections and transient Rossby-wave packet diagnostics (Grazzini and Lucarini, 2009) also highlight links to circulation features such as over the North Atlantic / European region.

The MJO is a major potential source of predictability in the Tropics on time scales exceeding one week. It has a significant impact on the Asian monsoon, it affects the development of El-Niño events and impacts on tropical cyclogenesis. Hence dedicated diagnostics are routinely used to assess the quality of the MJO predictions in every new model cycle (Vitart et al., 2007). An example will be presented in section 3.

The evaluation of the coupled system climate is crucial for any further development of the seasonal forecast system. During the development process, several sets of coupled hindcasts using the most recent atmospheric model cycle have to be produced and analysed. Ideally diagnostics of the coupled and uncoupled systems should be based on the same experimental set-up. Efforts are going to be directed towards making the system more accessible to other research sections.

2.10 Flow-dependent error

Hindcasts, made using a constant IFS cycle within the ERA-Interim project, can help distinguish flow-dependent changes in operational forecast skill from those associated with cycle changes. For example, ECMWF's operational northern hemispheric Z500 errors were substantially larger during the March–May (MAM) season of 2009 than they were in the previous year. It is important to understand whether MAM 2009 was simply less predictable or whether a problem had been introduced into the IFS in the intervening period.

This example is discussed with the help of a diagnostic tool that separates error and 'activity' into different spatial scales (*e.g.* 'synoptic' and 'planetary'). Figure 21(a) shows results based on the operational forecasts for the northern mid-latitudes as a function of forecast lead-time. Each curve has three attributes: colour, thickness and style. Red curves show MAM 2009 and blue curves show MAM 2008. Thick curves relate to planetary scales and thin curves to the synoptic scales. Solid curves show the mean-squared forecast error. Dashed curves show (2×) the mean-squared analysis anomaly (from the long-term climatology: the 'observed activity'). Dotted curves show forecast activity. Coloured circles on any curve indicate a statistically significantly better result compared to the other year. Importantly, because squared quantities are shown, the planetary and synoptic scales can be added to give the total (error or activity). In the limit of no predictability, the error curves would reach the activity curves.

The poor scores for MAM 2009 can partly be explained by an increase in synoptic-scale error (thin, solid). This is consistent with increased synoptic activity (thin, dashed). However, planetary-scale errors (thick, solid) were also larger (statistically significantly so) even though planetary activity (thick, dashed) was actually less in 2009.

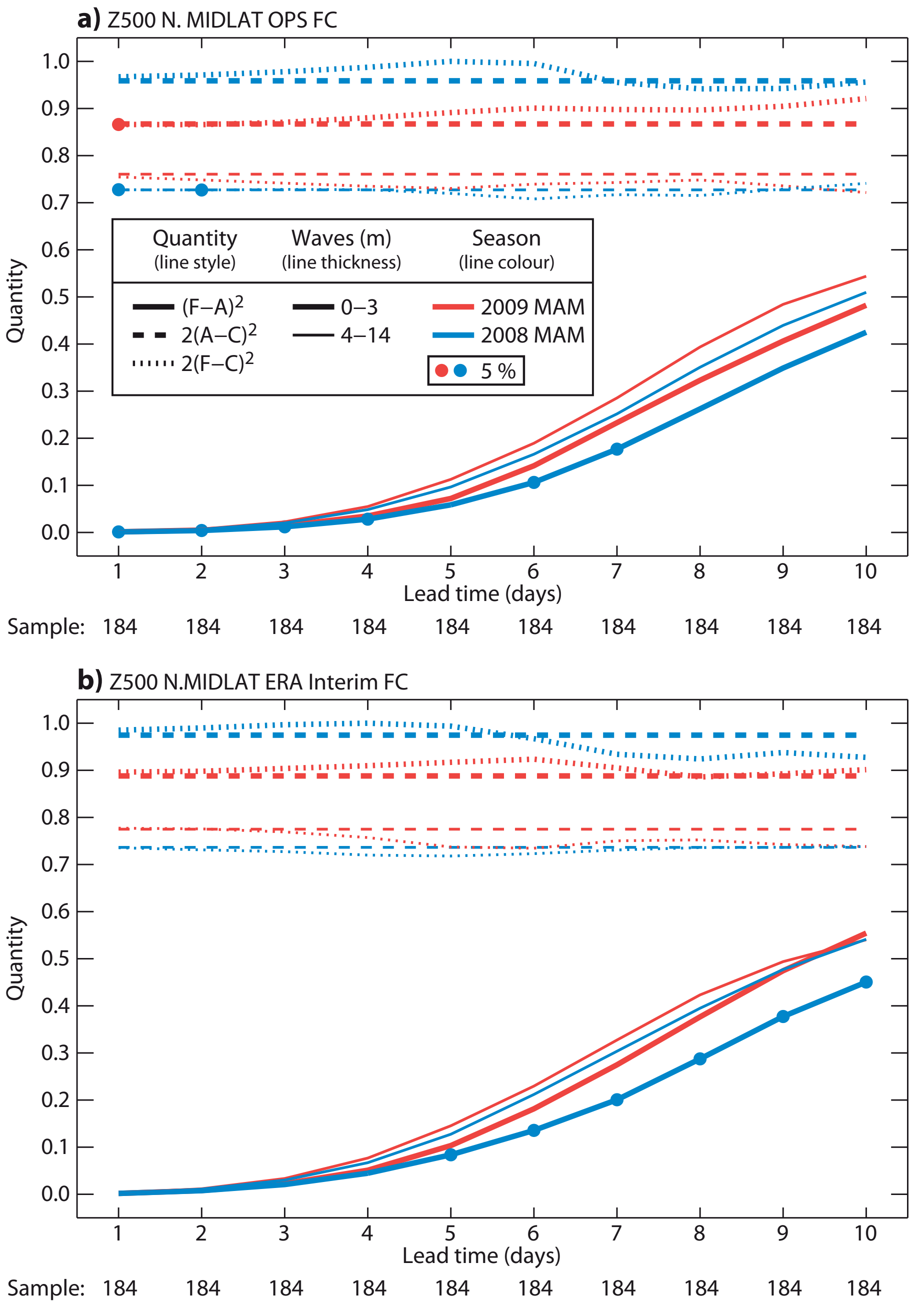

Figure 21: Scale dependent error and activity for the northern midlatitude 500 hPa geopotential heights in the March–May (MAM) season 2009 versus MAM 2008: (a) Operational forecasts, (b) ERA-Interim hindcasts. Line style indicates the quantity: solid for mean-squared error, dashed for mean-squared activity of the analysis, dotted for mean-squared activity of the forecast. Activity is defined as twice the mean squared anomaly of a given field relative to climatology. Line thickness indicates the group of waves included: thick for planetary waves with zonal wavenumbers 0–3, thin for synoptic-scale waves with zonal wavenumbers 4–14. Line colour indicates the year: red for MAM 2009, blue for MAM 2008. Dots are placed on a curve if that year is statistically significantly better than the other year. Errors for different scales can be added linearly. Similarly activities can be added to give the total activity over zonal wavenumbers 0-14. The error tends to the activity at the limit of no predictability.

The corresponding results based on the ERA-Interim hindcasts show very similar signals (Fig. 21b) and this leads to the (re-assuring) conclusion that a recent model cycle update is not the reason for the poorer scores in MAM 2009.

Further work is required to understand why the IFS had particular problems predicting the larger-scale planetary wave pattern in MAM 2009. The relaxation experiments like those discussed in section 2.8.4

may be helpful in this regard.

2.11 Forecast uncertainty

The evaluation of ensemble forecasts differs from that of deterministic forecasts because the former provide a probability distribution that is not better known a posteriori than it was known a priori (Talagrand and Candille, 2009). Any meaningful evaluation has to accumulate forecast-verification pairs over sufficiently large samples in order to enable the comparison of like objects, *i.e.* predicted distributions and empirical distributions of forecast errors estimated from observations or analyses.

A versatile diagnostic for ensemble forecasts that assesses reliability is the relationship between ensemble spread and the ensemble mean error (Talagrand et al., 1997). It is used in a number of ways at ECMWF to assess the characteristics of the EPS by looking at different variables in the model state space and all forecast lead times. The diagnosis of ensemble prediction systems at ECMWF focuses on how inconsistencies between ensemble spread and the ensemble mean error are linked to deficiencies in the formulation of the sources of uncertainty in initial conditions and the forecast model.

The basic form of the spread-error relationship considers simple averages of the variances over a region and a set of forecasts. The more sophisticated form of the spread-error relationship considers the joint distribution of ensemble standard deviation and ensemble mean error for a fixed lead time (Talagrand et al., 1997; Leutbecher and Palmer, 2008). The spread-reliability is determined from the conditional distribution of the ensemble mean error for given ensemble standard deviation σ_{ens}. Statistical consistency, *i.e.* reliability, requires that the standard deviation of this conditional distribution is equal to σ_{ens}, the value of the ensemble standard deviation on which the distribution of the error is conditioned. An example of this conditional spread-error diagnostic has already been discussed in Section 2.6 on analysis uncertainty. Depending on the variable and lead time it may be necessary to account for analysis or observational uncertainty. This can be done following the method of Saetra et al. (2004) and work on using this method in the verification of ensemble forecast at ECMWF is in progress.

The spread-error diagnostic has been extended to consider different spatial scales. Jung and Leutbecher (2008) use waveband filters to focus on planetary, synoptic and sub-synoptic scales. They showed that the over-dispersion of the ECMWF EPS prior to model cycle 32r3 in the early forecast ranges is particularly prominent in the synoptic scales in the mid-latitudes.

Often a ‘forecast bust’ is thought of as a deterministic forecast that is far from the truth. Here, probabilistic forecast busts have been defined to occur when the ensemble spread is small, but the mean error is large. In such a situation, the probabilistic forecast may be over-confident. One such probabilistic bust is evident in the operational D+5 forecast verifying on 27 October 2008 (see Fig. 22). On this date, the ensemble spread, S^2_{pf} (black, ‘pf’ stands for ‘perturbed forecast’), suggests slightly increased uncertainty in the forecast, but the squared error of the ensemble mean, E^2_{em} (red) is much larger (n is the ensemble size and $(n-1)/(n+1)$ is a scaling to make the mean squared error quantitatively comparable with the mean spread). Occasionally, even for a perfectly calibrated forecasting system, the verifying analysis will fall outside the ensemble, but it is worth examining whether this represents something more systematic.

Figure 23 shows a composite of five such ‘probabilistic busts’ in the D+5 ensemble forecast. Four of the busts are in spring and one is in autumn. Although results are preliminary, the suggestion is that anomalies over the US, panel (a), appear to lead, 5-days later, to a planetary wave-like feature down-stream, (b), and enhanced forecast errors over Europe (d) that were not predicted by the ensemble spread (c). If these results are representative of a systematic issue (more cases will establish this), then this study may be highlighting model problems with the representation of mesoscale convective systems and mesoscale convective vortices over the US, for example (see Schumacher and Davis, 2010) that may not, therefore, be explicitly accounted-for in the singular vectors used to perturb the ensemble.

Also shown in Fig. 22 is the correlation between the daily time-series of ensemble spread and error of the ensemble-mean. Although a perfectly calibrated ensemble forecast system would still have a correlation

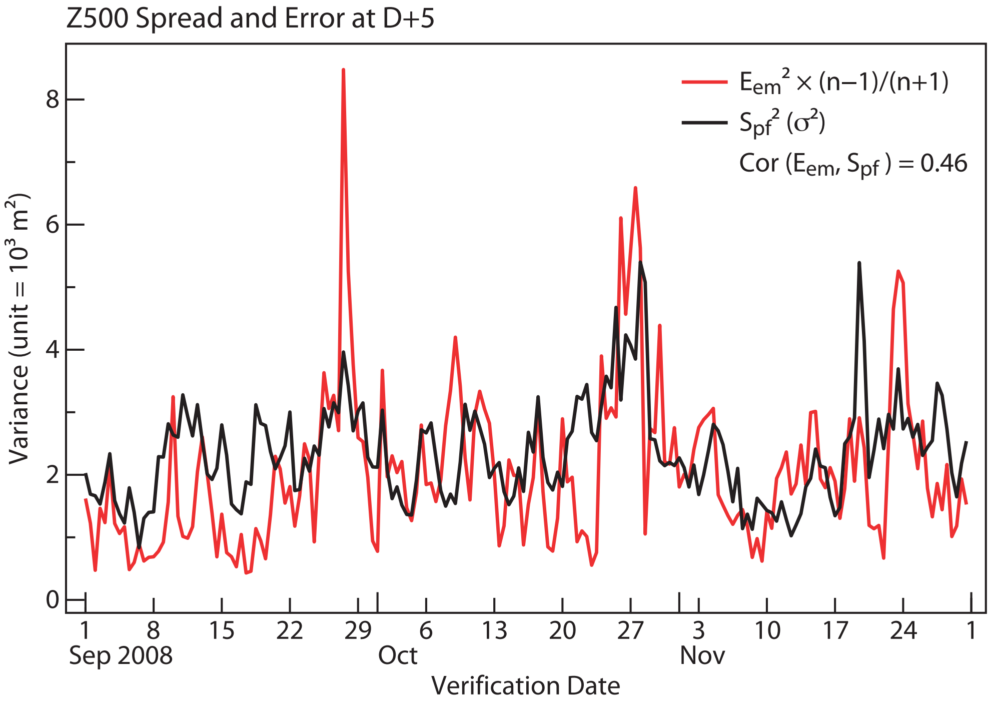

Figure 22: Time-series of D+5 ensemble spread (black) and squared ensemble mean forecast error (red) for 500 hPa geopotential heights over Europe (unit = $10^3\,m^2$). The error is scaled so that, in a perfectly calibrated system, they should have the same expected (long-term-mean) value as the spread.

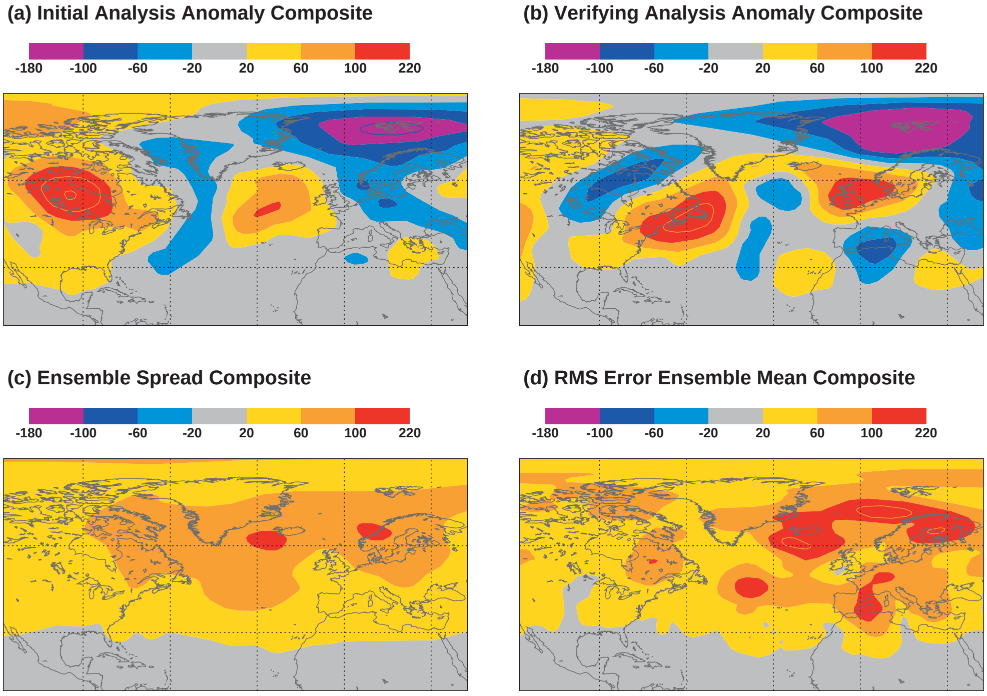
(a) Initial Analysis Anomaly Composite
(b) Verifying Analysis Anomaly Composite
(c) Ensemble Spread Composite
(d) RMS Error Ensemble Mean Composite
-180
-100
-60
-20
20
60
100
220

Figure 23: Composites of probabilistic busts in D+5 ensemble forecasts over Europe. (a) Initial analysis anomaly. (b) Verifying analysis anomaly. (c) Ensemble spread at D+5. (d) Root-mean-squared error of the ensemble mean at D+5. The composite comprises forecasts verifying on the dates 20070404, 20070424, 20080306, 20080417, 20080927.

less than 1, it is worth monitoring this correlation in future. Assessment of the spread-error relationship on the daily timescale may, for example, highlight a possibility that, although spread is well-calibrated to error in the long-term-mean, there may be a compensation occurring, with too much spread in situations of reasonably predictable flow and too little spread in situations that are (presently) less predictable.

A proper score that is particularly suited for identifying probabilistic forecasts busts in the sense discussed above is the Ignorance score computed for the Gaussian density with the mean and variance predicted by the ensemble. Work to use it routinely at ECMWF to assess ensemble forecasts has started. The Ignorance score is particularly efficient in quantifying the skill (or lack of skill) in flow-dependent variations of uncertainty as demonstrated by Leutbecher (2009).

2.12 Diagnostic verification

Some aspects of verification provide considerable insight into forecast system error. Two such aspects are discussed below.

2.12.1 Verifying the tracks of cyclonic features

Adverse and severe weather events in both the tropics and extra-tropics commonly relate to the passage of cyclonic features in the lower troposphere. So if the handling of these cyclonic features in forecasts and model analyses can be compared, one can effectively verify forecasts of inclement weather 'by proxy'. Whilst such an approach implicitly relies on the integrity of model analysis fields it has the major advantage of circumventing observation-related problems. Importantly, such an approach can also incorporate severe weather events over data sparse areas, notably oceans, hugely increasing the dataset size. This tackles the primary hurdle for severe weather verification, namely small samples, and so can markedly increase result robustness. Thus we have a verification strategy that, through the use of innovative feature-track-related diagnostics, can directly address ECMWF's severe weather forecasting goals. As discussed below, the verification ties in with new cyclone-related products, and so is dual-purpose, providing diagnostic information not just for ECMWF but also for its product users.

The main tools required here are sophisticated automated algorithms for identifying and tracking the said cyclonic features. Having imported and extended pre-existing Met Office code (see Hewson and Titley, 2010), ECMWF now has such tools at its disposal. Indeed these are now providing products for forecasters in real time (see Hewson, 2009b), and a verification component has recently been built in. Compared to some other trackers our methodology has the distinct advantage of being able to follow features across a wide range of scales, with smaller frontal waves and polar lows being successfully tracked alongside larger hurricanes and extra-tropical depressions. Figure 24(a) shows an example of a number of features automatically identified in an analysis field, including a chain of frontal waves (orange spots), that brought heavy rainfall to the British Isles.

To illustrate some of the diagnostic capabilities of cyclone track verification a 6 week period from winter 2009/10 was examined. Figure 25 shows the median position error, as a function of lead time, for all features tracked forward from T+0 in various model forecasts. This tells us that the deterministic forecasts provide a minimum error feature position forecast up to day 3.5, after which the EPS feature mean takes over as the most accurate. Meanwhile the EPS control run is a little less accurate than the deterministic, showing, importantly, that even going from 32km to 16km resolution can improve feature trajectories. Note also that the (unperturbed) control run has a lead time gain of about half a day in the early medium range over a perturbed EPS member (black line). The fact that other cyclone attributes, in addition to position, are stored in a database when tracking software is run will enable other diagnostics pertaining to model handling of cyclonic features, such as biases in deepening rates, to be examined in detail in due course. They may also provide pointers to why errors and biases occur (see Hewson, 2002).

One of the cyclone database attributes stored is the 'maximum wind 1km above the earth's surface within a 300km radius of the feature point'. A related product is a 'storm track strike probability' plot - for this the maximum wind attribute is used to place all features tracked in EPS members into three categories of severity (Fig. 24b shows an example for the 'all features' category). Plots of this type show the probability that a cyclonic feature, attaining the given severity threshold, will pass within 300km

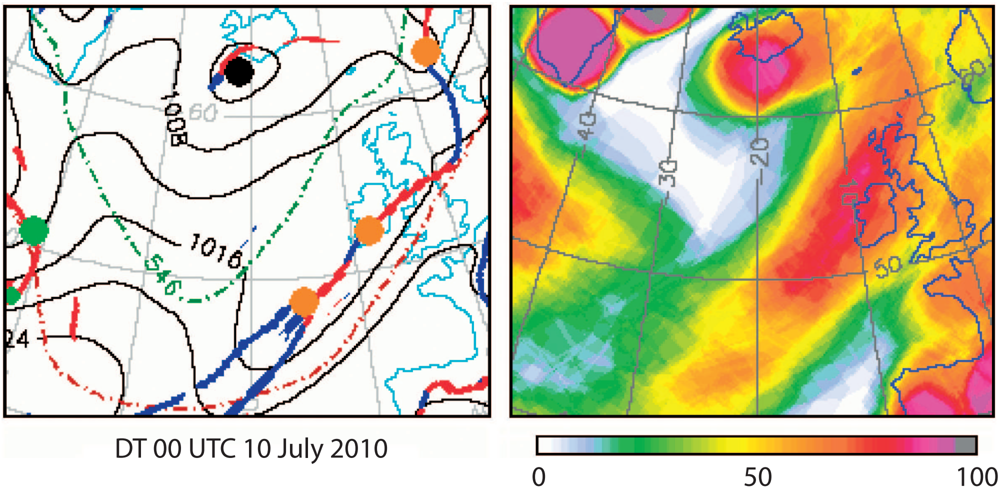

Figure 24: (a) Objectively identified synoptic features in model analysis for 00UTC 10 July 2010. Mean sea level pressure contours are black, with 1000-500hPa thickness lines of 546 and 564 dm in colour (dash-dot). Objective warm and cold fronts are red and blue respectively. Frontal waves are orange, 'barotropic lows' are black and diminutive frontal waves (Hewson, 2009a) are green. (b) D+4 storm track strike probability product from the ECMWF EPS, for the same validity time, for all features, scale in %. Tracking time window is +/-12h. Red band crossing Scotland corresponds in this case to the forecast track of the frontal waves.

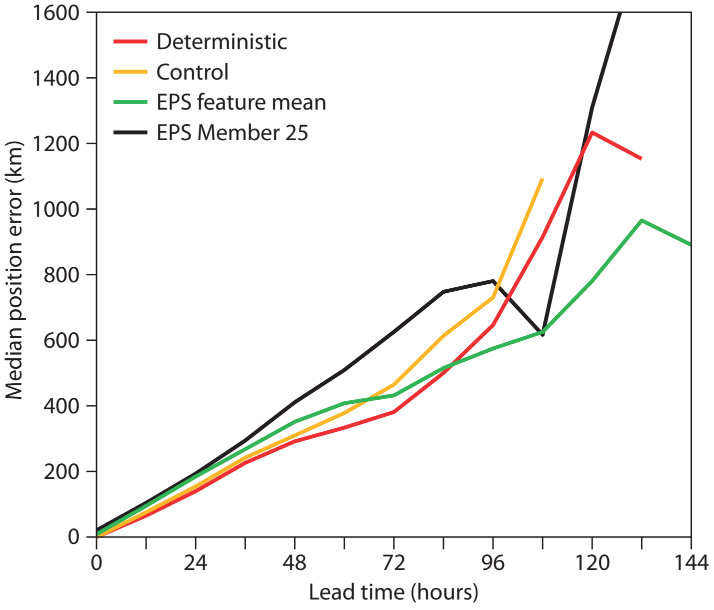

Figure 25: ECMWF cyclonic feature median position error in trial runs of the Jan 2010 model upgrade, over an extended N Atlantic domain (Deterministic forecasts were at T1279 resolution; Control and EPS members were at T639 for these lead times). Verification period 25/11/09-16/1/10. 'EPS Feature mean' denotes the centre of gravity position for surviving features in the EPS runs.

within a 24 hour period, and as such are analogous to the 'cone of uncertainty' formally used with tropical cyclone predictions. The difference is that the new product covers all features at once, not just a single named system. Such products have been verified, again over a 6 week period, using a Brier Skill Score approach, and some results are shown in Fig. 26, for the 'all features' (solid) and 'windstorm' (dashed) categories. To verify 'adverse weather by proxy' in some general sense, one can focus on the 'all features' category, whilst to verify 'severe weather by proxy' - in this case severe winds - one can focus on the 'windstorm' class. The trial version of the new high resolution forecasting system, that was introduced in late January 2009 (red), generally outperformed the then operational system (blue).

This is re-assuring, and tallies with signals from more conventional measures. However the plot also highlights that handling by the EPS of the 'windstorm' category at leads of 2 to 4 days may have been degraded. This is being monitored. This important result would generally be masked out by classical verification measures. Note also that verification can also be applied to other EPS systems, and here the black line shows the performance of the global version of the Met Office MOGREPS ensemble over the same period. This indicates that the ECMWF EPS has a lead time gain over MOGREPS of typically 2 days, even in the early medium range.

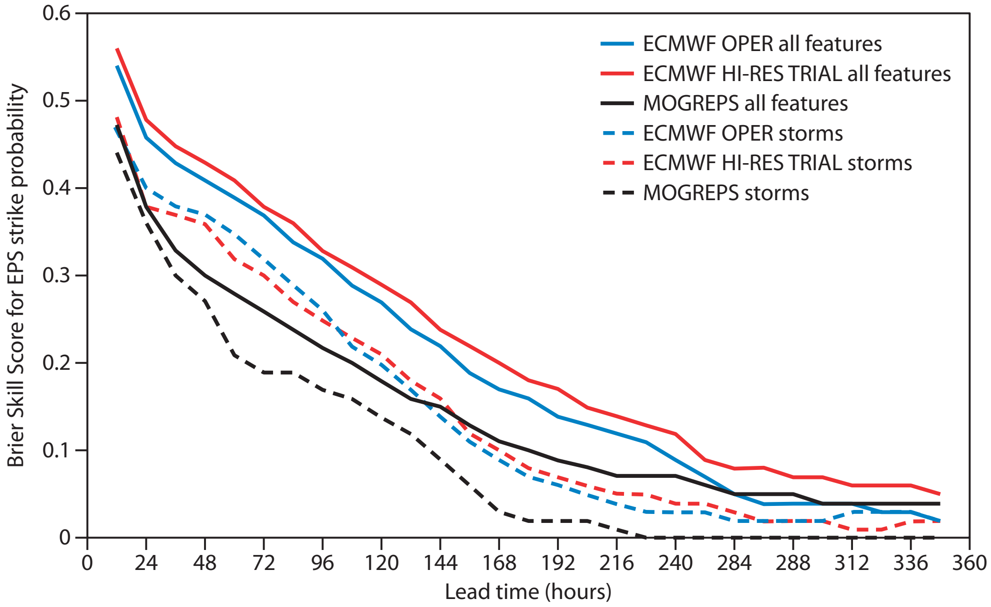

Figure 26: Brier skill scores for forecast strike probabilities for two classes of cyclone: 'all features' (solid) and 'storms' (dashed), for two EPS systems (see legend - 'ECMWF OPER' was at resolution T399 to day 10, then T255; 'ECMWF HI-RES TRIAL' was T639 to day 10, then T319). The 'storms' category denotes those features around which the wind speed, at 1km altitude, within a 300km radius, exceeds 60kts locally. The verification period and domain are as in Fig. 25.

Future plans include classifying cyclonic features according to other measures, such as area-integrated precipitation totals in their vicinity, to address other severe weather aspects. 'Rainstorm' strike probability plots will be one output, and these will probably be verified using rainfall in short range model forecasts. The diagnostic value of this is that it can potentially give insights into precipitation biases and moisture budget issues in the ECMWF model. One problem might be the integrity of the model precipitation 'analysis', and it may be better to use some form of blended model-observation precipitation analysis instead. There is also considerable scope to compute and verify products from experimental runs. For example, one could examine the effect that introduction of a mixed layer ocean model has on storm tracks and the precipitation characteristics of those storms. There is also scope to apply the tracking, routinely, in monthly and seasonal forecasts; at present this is only done for tropical cyclones.

2.12.2 Conditional sampling based on weather verification

It is strategically important for ECMWF to focus more on the prediction of precipitation. As a first step towards this goal, verification measures for precipitation are required. However, precipitation is a difficult quantity to verify owing to its highly skewed distribution. Work has focused (Rodwell et al., 2010) on developing a stable deterministic verification measure that is particularly relevant for assessing and monitoring the deterministic initialisation of the model and the model's deterministic physics. This score is called 'SEEPS' (which stands for 'Stable Equitable Error in Probability Space'). It is a three-category score that assesses the prediction of dry weather and the prediction of precipitation quantity. Essentially, it is (1 minus) the average of two 2-category Peirce Skill Sores, defined by the climatological probability distribution in such a way that they are asymptotically stable in the limit of a perfect forecasting/observation system. Because SEEPS is defined by the local climatological probability distribution of precipitation, it always assesses the salient aspects of the local weather. It has been shown to be able to demonstrate statistically significant differences between two consecutive model cycles, and to show trends in operational scores. Importantly, SEEPS has properties that promote 'refinement' and inhibit 'hedging'. With the emphasis in operational centres such as ECMWF on the diagnosis of error, it is highly beneficial that this error score gives as true a reflection as possible of the 'quality' of the forecast system.

Recent work on the processing of 'SYNOP' data exchanged over the GTS has allowed, for the first time, near-real-time verification of precipitation anywhere over the globe (even when the reporting time does not coincide with whole-day lead-times from the 0 and 12UTC forecasts). For example Fig. 27 shows scores for India (which generally reports at 3, 9 and 15 UTC) at lead-times of 2 and 5 days. In general tropical scores are poorer than those for the extratropics, but at least it can be seen that D+2 (Fig. 27a) has better scores than D+5 (Fig. 27b), and this indicates that tropical trends at constant lead-time will be identifiable in future.

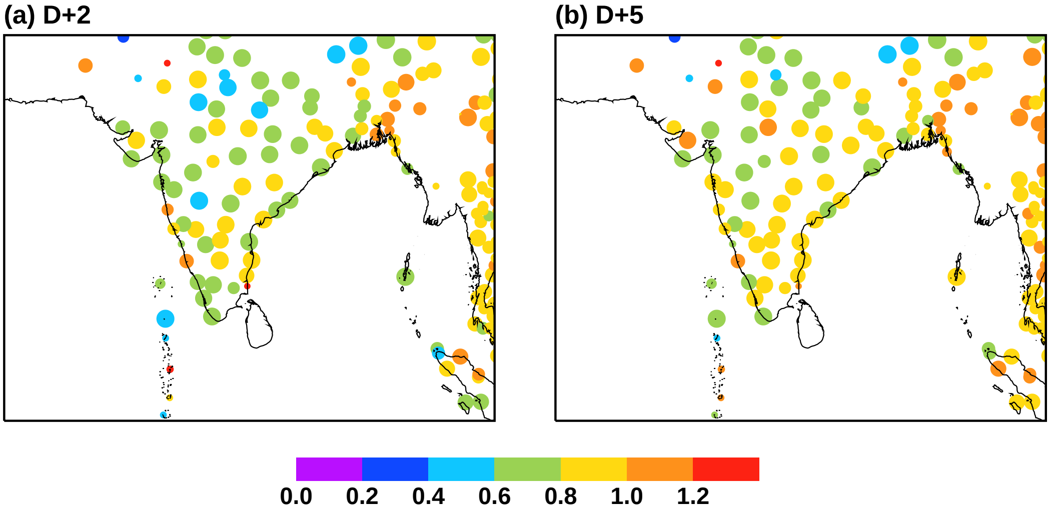

Figure 27: Mean SEEPS precipitation error based on operational deterministic forecasts verifying in June–August 2009. (a) D+2. (b) D+5. An error of zero indicates a perfect categorical forecast system. A mean error of 1 indicates a forecast system with no predictive skill.

Such scores will not only aid top-level decisions about new cycle implementation but also open-up new avenues of diagnostic research. For example, the SEEPS score is highlighting particular problems in the prediction of precipitation associated with depressions over the Mediterranean and this will lead to a re-focusing of attention on our predictive skill for these depressions and the (sometimes severe) precipitation associated with them. Breaking-down the score into contributions from particular categorical errors can also help identify the features that are most problematic and need to be prioritised (for example the over-prediction of light precipitation in the present forecast system).

Future work will focus on 'proper' probabilistic scores. The joint use of deterministic and probabilistic scores will allow a complete assessment of the deterministic model and the uncertainty aspects associated with ensemble data assimilation and stochastic physics.

3 Collaborative projects

The Working Group on Diagnostics has highlighted two classes of collaborative projects: those addressing existing problems, and those addressing new forecast system cycles. In section 3.1 the key findings of the 'over-active monsoon' project are summarised, while in section 3.2 the diagnosis of new model cycle 36R4 (due for implementation in the last quarter of 2010) is discussed.

3.1 Diagnosis of existing problems: Summary of the monsoon project

For several years, our forecast systems have produced too much precipitation over the west coasts of India and Burma/Myanmar and over the seas surrounding the Indian Peninsula. This over-activity generally also applies to the Indian Peninsula itself, particularly in August. The exception to this wet bias is a marked dry region over Bangladesh. This pattern of biases (Fig. 2) gets worse with increasing lead-time, and higher resolution (without parametrization adjustment) does not help. To understand this problem better and to demonstrate many of the diagnostic tools available at ECMWF, a collaborative project on the Asian summer monsoon was established. Here, the results of this project are summarised.

Recently, new observations have been assimilated within this monsoon region. These include Indian radiosonde temperature profiles (diagnostics indicated improved observation quality so that these could be un-blacklisted; Fig. 3) and ASCAT scatterometer surface winds (Fig. 4a). In addition, SSMI radiances (primarily sensitive to total column humidity) are now assimilated in all-sky conditions in 4D-Var in a unified way. Observation impact assessments indicate that the SSMI data, for example, have a strong impact on the analysis (Fig. 6). Increased observation usage within the assimilation has helped in the diagnosis of the monsoon problem.

Comparison between ECMWF and UK Met Office analyses indicate that our estimates of the ‘truth’ are quite different in the monsoon region. For example, the mean analysed zonal winds at 925 and 850 hPa over the Arabian Sea are up to 2ms^{-1} slower than those of the Met Office. The hemispheric mass-budget study of the ECMWF re-analysis provides complementary information that suggests that tropical winds in general remain poorly analysed (Fig. 11).

Examination of differences (in observation space) between observations and the first-guess can help differentiate these estimates of the truth. First-guess departures indicate that the low-level monsoon circulation in the ECMWF first-guess is about 1ms^{-1} stronger than observed (Fig. 4c). Agreement between the first-guess departures of different observations such as ASCAT and atmospheric motion vectors (*c.f.* Fig. 4c,d) all add weight to the suggestion that the first-guess departures (and consequent analysis increments; Fig. 5) reflect a model problem.

Analysis differences between ECMWF and the UK Met office may indicate differences in how strongly each analysis is drawn to the observations. However, there are aspects of the flow that are not well observed. ASCAT, for example, only ‘observes’ winds at the surface. A good model representation of boundary layer processes is required to communicate ASCAT surface departures up to 850hPa. Hence analysis differences may also reflect model differences.

Results also show that ECMWF low-level zonal wind errors over the Arabian Sea (relative to the analysis) grow with time to 3 ms^{-1} by D+5 (Fig. 8a) and (not shown) to about 4 ms^{-1} by D+10. Since this bias is even greater than the likely uncertainty in the analysis, it is further evidence that this aspect of the over-active monsoon reflects a model problem. Similarly, forecast ‘MTSAT’ brightness temperatures grow with lead-time (Fig. 13c). They indicate enhanced convection over India, the tropical and northern Indian Ocean and Arabian Sea, and reduced convection over Bangladesh.

Initial tendency budget studies have indicated which processes could be involved in the initial model error (Fig. 9). The boundary-layer momentum error, for example, is likely to be associated with vertical diffusion, convection or the dynamics itself (or a combination of these). Further investigation of these initial tendency budgets is planned.

The coupled model displays weaker precipitation biases than the uncoupled model (*c.g.* Fig. 2n,q). Partly this may be because an interactive ocean can moderate the erroneously large surface latent heat fluxes of the atmosphere-only model. The act of coupling also has an impact on the variability of monsoon precipitation (Fig. 19). It is likely that these bias and variability changes are not independent of each other. The implication is that better understanding of the monsoon problem can be gained through coupled experimentation and that it is desirable, in general, to assess (atmospheric) physics changes within the context of the coupled model.

Recent work means that it is now possible to monitor daily forecast scores of monsoon precipitation (Fig. 27) as well as assess biases. This should ensure that account is taken of the monsoon problem when development decisions are made in future.

The above collaborative study has gone a long way towards characterising the monsoon problem and indicating possible reasons for it. More work, however, is required to fully resolve the problem. Particular attention will be given to boundary-layer processes, coupled processes, and forecast departures in observation space.

3.2 Diagnosis of new model cycles: Application to cycle 36R4

The monsoon project has motivated developments in our representation of convection in the new cycle 36R4. In particular a reduction in the hydrological cycle through a reduction in the shallow entrainment rate (it becomes relative humidity dependent and consistent with deep convection), and a re-tuning of the deep convective entrainment/detrainment rates - which reduces the upper-tropospheric mass fluxes and model cold bias and improves the upper-level wind field.

A change in the microphysical scheme and the cloud-radiation interaction, that did not directly target the monsoon problem, had an equally positive effect on the Asian monsoon circulation. The microphysical developments comprise a new 5-species prognostic microphysical scheme (cloud liquid water, ice, rain, snow, and cloud fraction) that can interact more with the radiation through the snow content. A new water droplet effective radius for raining clouds has been introduced and the cloud overlap in the radiation becomes dependent on latitude. These microphysical and radiation developments increase the vertical stability and this leads to the beneficial decrease in precipitation.

Both cycle 36R4 and cycle 36R3 have improved the 2m temperatures over land compared through the introduction of a variable leaf area index (Boussetta et al., 2010).

The impact of cycle 36R4 on the Asian summer monsoon is illustrated in Fig. 28 based on T_L511 analysis experiments for the period mid-July to August 2009. As a reasonable proxy for the precipitation error and the spin-up of the monsoon circulation (based on further comparisons with the GPCP2.1 dataset) the difference between the D+10 and D+1 precipitation for cycle 36R3 is depicted in Fig. 28(a). The over-active monsoon is clearly evident, together with the dry error over Bangladesh. Figure 28(b) shows the difference in precipitation rate between cycle 36R4 and 36R3 during the 10-day forecast range for the same period. Clearly cycle 36R4 reduces the precipitation over the Indian Ocean and increases the precipitation over the Bangladesh region. Consistent with the improvement in precipitation, the upper-level RMS wind errors at D+10 (Fig. 28c) are also reduced with cycle 36R4. The overall improvement in terms of the precipitation and wind errors in the region is about 15%.

The relevant components of our diagnostics suite, some of which were discussed in section 2, are used in the assessment of all new forecasting cycles. Here a diagnostic that targets the representation of the Madden Julian Oscillation (MJO) is discussed. The MJO is the dominant atmospheric mode in the Tropics on the intraseasonal time scale, and its prediction is of primary importance for the monthly and seasonal forecast systems. To evaluate the skill of the model to predict MJO events, 5-member ensemble daily integrations from 15 December 1992 until 31 January 1993 are performed for each new version of the IFS using the same configuration as the operational monthly forecasts. The forecasts are then projected onto combined EOFs of zonal wind at 200 hPa, zonal wind at 850 hPa and OLR, as in Wheeler and Hendon (2004) (see Vitart and Molteni, 2010, for more details). Figure 29 shows the linear correlation between observed and ensemble mean first principal component (PC1, left panel) and the second component (PC2, right panel) as a function of the forecast lead time for CY32R3 (black line), CY36R2 (blue line) and CY36R4 (red line). Interestingly, CY36R4 improves the predictability of the MJO substantially by around 3 days, in particular for the second principal component which has the maximum convection over the Indian Ocean. These results are therefore consistent with the

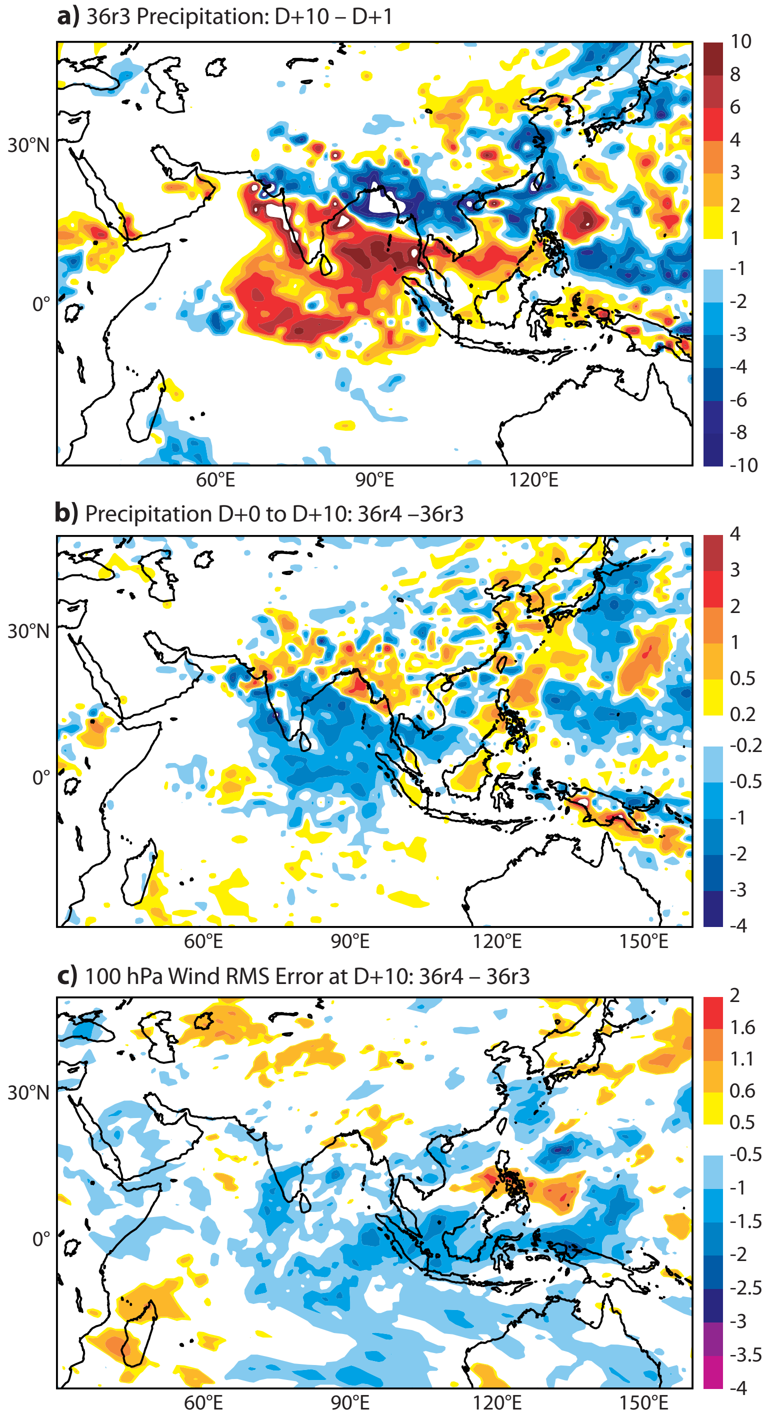

Figure 28: (a) Difference in precipitation (mmday^{-1}): D+10 - D+1 from T_L511 analysis/forecast experiments for the period 16 July to 31 August 2009 for cycle 36R3. (b) Difference in precipitation rate over the 10-day forecast range: 36R4 - 36R3, averaged over the same period. (c) Difference in 100 hPa RMS wind error at D+10: 36R4 - 36R3 (blue implies improvement), averaged over the same period.

improvements seen in Figure 28(b,c). However, there is still a large potential in improving the prediction of the MJO. The current average predictability limit for CY36R4 is around 20 days, so half a cycle, whereas the theoretical prediction limit is expected to be at one full cycle.

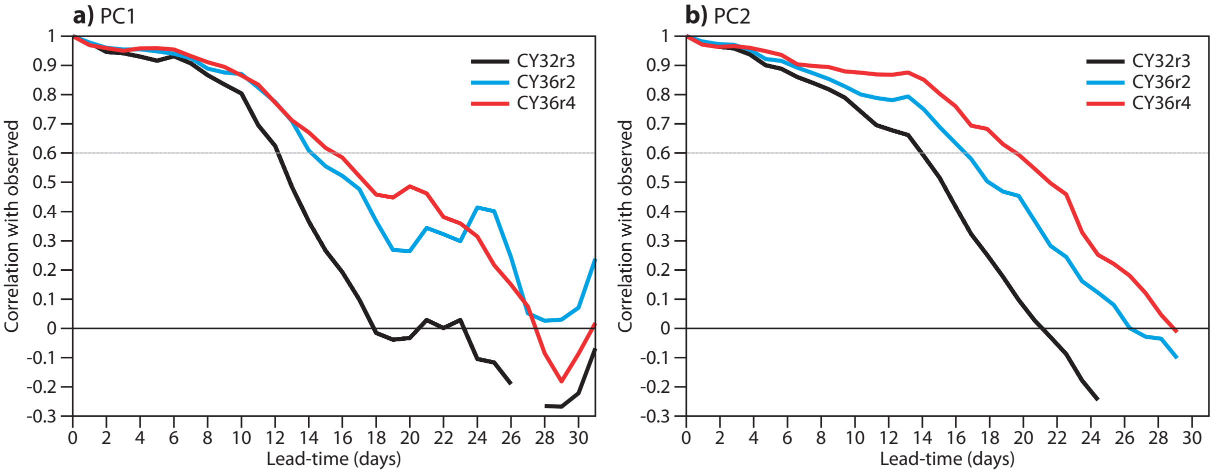

Figure 29: The linear correlation between observed and ensemble mean first principal component (PC1, left panel) and the second component (PC2, right panel) as a function of the forecast lead time for cycle 32R3 (black line), 36R2 (blue line) and 36R4 (red line).

4 Discussion

4.1 The importance of diagnostics research

Diagnostics has several different connotations depending of one's point of view. For operational forecasting, it is primarily about the diagnosis of system error. However this diagnosis can entail the use of circulation metrics and in-depth case-studies which may be more familiar to some readers. Diagnostics research is required to maintain the pace of forecast system improvement. New tools must be able to identify smaller errors, analyse larger observation volumes, contend with more complex models that are run at higher resolution, and address the growing need to quantify uncertainty. These trends in forecast system development cannot be thought of in isolation. For example, physics improvements may lead directly to reduced forecast error, but may also permit a better assimilation of existing (or new) observations, and thus indirectly reduce forecast error through the improvement of initial conditions. Hence there is also a great need for seamless thinking in diagnostics. This involves collaborative diagnostic work, such as using data assimilation to diagnose forecast model error, but it also involves simply enhanced communication.

4.2 The Working Group on Diagnostics: Initial findings

Recently a 'Working Group on Diagnostics' (WGD) has been established to consider, in the broadest terms, ECMWF's diagnostic strategy. ECMWF has never had a written diagnostics strategy and so this has been a useful exercise. The WGD comprises representatives of all Research sections and of the Meteorological Operations section. It has been coordinated by one of the scientists who's work is dedicated to diagnostics research, and has met several times so far. To address the aspects of diagnostic research discussed above, it has formed the view that its roles should include: the over-sight of collaborative projects, the strategic coordination of diagnostic developments, and the across-section communication of information and results. This paper is the first major outcome of the WGD.

In section 1 (Fig. 1) a diagnostic framework was introduced that highlights the various areas of the forecasting task where diagnosis is possible. In section 2 'strategic' tools and techniques from within each diagnosis area were presented. These are tools that will be developed further and that will be central to the future diagnosis of our forecasting system. Some of these have been applied to the collaborative monsoon project and results have been rewarding. Not only has a lot been learnt about the monsoon (see section 3.1 for a summary) but we have begun to develop a blue-print that could be adopted in future collaborative projects. There are significant over-heads in conducting collaborative projects and so this blue-print starts with the careful identification of the topic and of the interested parties. A successful collaborative project requires momentum to be maintained - for example by holding frequent, short meetings, by setting reasonably tight deadlines and by developing a scientific interest in all participants. In addition to specific conclusions about the topic of interest, collaborative projects lead directly to new diagnostic tools and highlight areas where strategic diagnostic developments would be useful in the future.

4.3 Strategies for the future

So where do we go from here? A diagnostics strategy is essential since diagnosis of the forecast system is integral to the overall strategy of the Centre. However, we will not discuss detailed diagnostics strategies for individual sections since this would risk stifling individual innovation. Instead, we present, for discussion, some possibilities for future diagnostics collaboration between sections. These are organised in terms of the three roles of the WGD.

4.3.1 Collaborative projects

There are many possibilities for collaborative projects. Some of these have been discussed above. For example, one project could address the ‘grey zone’ where convection is partially resolved and non-hydrostatic effects become important. Other topics could include the diagnosis of gravity waves, and the diagnosis of model biases such as the stratospheric cold bias. Such projects would be in addition to the more routine collaborative diagnosis of major new model cycles. One key project could focus on improving the prediction of severe weather. It is worth elaborating on this here to emphasise the thought processes that underlie such projects. A project on severe weather would require going back to first principles and systematically identifying severe weather events. The use of re-analyses and re-forecasts would help increase sample size. Having identified the events, we would need to discover whether enough observations associated with severe weather get assimilated, or whether too many are rejected because they are far from the first guess. The first guess may be far away if our physics is unable to represent the key processes involved, and this will need to be assessed. Ultimately, the prediction of severe weather is a probabilistic task and so it will be important to investigate whether the ensemble prediction system is able to adequately represent the chances of a severe weather event. This requires robust verification measures that can assess the delicate balance between ensemble size and ensemble resolution. Hence it is very apparent that severe weather represents a good example of a cross-cutting diagnostic project.

4.3.2 Strategic diagnostic developments

There is also scope for the strategic coordination of diagnostic developments. One possibility may be very apparent to the reader from the project work presented above. This is the issue of the diagnosis of short-range forecast error and it is worth elaborating on this a little. A key issue for physics development is the time taken to diagnose the broader-scale impacts of proposed changes. Until recently, a physics change has generally led to a rather large change in the model’s climate and so fast, low-resolution, climate simulations, that can be run and diagnosed over-night have generally been sufficient. In the future however, as models get better, physics changes should have a smaller impact on model climate. At the same time, increased operational forecast resolution will start to partially resolve some aspects such as convection, and thus low-resolution simulations may not be a suitable starting point from which to diagnose the impact of a physics change on operational forecasts. Finally, the increased number of physical processes represented within the model is greatly increasing the scope for interaction and making it increasingly difficult to identify the cause of any given model climate error. These trends may necessitate the increased use of full-resolution data assimilation experiments to assess errors at much shorter ranges (initial tendency errors, for example). A shift of attention to assessing proposed physics changes within the assimilation could have other benefits. For example, the early assessment of how proposed physics changes interact with assimilation techniques such as variational bias correction could lead to smoother operational implementation. Indeed, correcting model errors very early-on in the forecast would enhance ECMWF's ability to assimilate observational data, and thus produce better analyses with which to initialise the forecast. There can also be synergies since the implementation of weak-constraint 4D-Var into the tropospheric assimilation necessitates a representation of short-range model error. Better diagnosis of short-range model error will also help in the development of stochastic physics and of the ensemble data assimilation system. Clearly therefore, there is justification for the strategic coordinated development of short-range model error diagnostics that can be produced quickly and efficiently from within the data assimilation.

Diagnostics at ECMWF would also benefit from making some tools more widely available. For example, one issue is the desirability for all research scientists to be able to set-up coupled ocean-atmosphere experiments. This would allow, for example, the impact of proposed atmospheric physics changes on the coupled model to be readily diagnosed. The WGD could consider which tools are worth making more widely operable, the need for on-going maintenance and documentation of these tools, and the interface used to drive them. In terms of the governance of diagnostic tools, the WGD could also discuss the use of a common scripting language.

Finally, there is a need to ensure that diagnostics are catered-for in terms of computing resources. For example, higher resolution requires more processing power. Powerful diagnostics require, for example, model-level process tendencies to be archived. The WGD should keep a continual check on these issues.

4.3.3 Communication

In addition to the enhanced communication implicit in the WGD, there are other ways that communication can be maintained and enhanced. A central web-site has been developed that contains the WGD's audit of ECMWF's diagnostic tools. This is likely to be developed further (to contain documentation, for example) and publicised more in future. Other internal web-sites contain many forecast system diagnostics - the Diagnostics Explorer, for example, contains a 5D over-view of the forecasting and data assimilation systems. ECMWF also holds a quarterly joint meeting of the Operations and Research Divisions (the so-called OD/RD meetings). A useful approach for the WGD would be to work alongside the OD/RD meetings by coordinating project work with the OD/RD special topics.

An important aspect to our diagnostics strategy must be to continue to foster and maintain links with institutions outside ECMWF. Some fundamental diagnostics research is well-suited to the academic world since it does not require large computing facilities. ECMWF can gain useful insight from fostering links with this community and, in return, can offer diagnostic access to its operational forecast system. Two-way communication between ECMWF and member states (ECMWF's user community) about diagnostics of forecast system performance is clearly essential but other collaborative diagnostic projects, such as that established with the UK Met Office, are also profitable for both parties. Finally, with seamless ideas gaining increasing acceptance, links to the climate community could be developed further. The NWP and climate communities have different objectives but could use the same diagnostic tools. For example, ECMWF could pass-on its insights into process-oriented metrics. At the same time the climate community could benchmark their models against our operational coupled forecast model in terms of the circulation metrics that they tend to concentrate on.

4.4 Final remarks

This paper has attempted to weave a common thread (the collaborative monsoon project) through a previously somewhat disparate set of diagnostic areas and tools. The work involved has helped unite the various sections within Research and Operations. Meanwhile, discussions within the Working Group on Diagnostics have helped formulate a more explicit diagnostics strategy. The authors of this paper acknowledge the previous suggestions by the SAC, and invite further comments on this strategy.

Acknowledgements

The authors thank Erland Källén, Martin Miller, Tim Palmer and Jean-Noel Thepaut for their useful comments on this manuscript.

References

Adler, R. F., J. Susskind, G. J. Huffman, D. Bolvin, N. E., A. Chang, R. Ferraro, A. Gruber, P.-P. Xie, J. Janowiak, B. Rudolf, S. Schneider, U. Curtis, and P. Arkin, 2003: The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydromet.*, **4**, 1147–1167.

Ahlgrimm, M. and M. Köhler, 2010: Evaluation of trade cumulus in the ecmwf model with observations from CALIPSO. *Mon. Wea. Rev.*. Doi:10.1175/2010MWR3320.

Aulign´e, T. and A. P. McNally, 2007: Interaction between bias correction and quality control. *Quart. J. Roy. Meteor. Soc.*, **133**, 643–653.

Baker, N. L. and R. Daley, 2000: Observation and background adjoint sensitivity in the adaptive observation-targeting problem. *Quart. J. Roy. Meteor. Soc.*, **126**, 1431–1454.

Balmaseda, M. A., F. Ferranti, L.and Molteni, and T. N. Palmer, 2010: Impact of 2007 and 2008 Arctic ice anomalies on the atmospheric circulation: Implications for long-range predictions. *Quart. J. Roy. Meteor. Soc.*, p. In Press.

Bechtold, P., J. P. Chaboureau, A. Beljaars, A. K. Betts, M. Köhler, M. Miller, and J. L. Redelsperger, 2004: The simulation of the diurnal cycle of convective precipitation over land in global models. *Quart. J. Roy. Meteor. Soc.*, **130**, 3119–3137.

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

Bender, M. A., I. Y. Ginis, and Y. Kurihara, 1993: Numerical simulations of tropical cyclone-ocean interaction with a high resolution coupled model. *J. Geophys. Res.*, **98**, 23245–23263.

Boer, G. J. and N. E. Sargent, 1985: Vertically integrated budgets of mass and energy for the globe. *J. Atmos. Sci.*, **15**, 1592–1613.

Bormann, N. and P. Bauer, 2010: Estimates of spatial and inter-channel observation-error characteristics for current sounder radiances for numerical weather prediction. I: Methods and application to ATOVS data. *Quart. J. Roy. Meteor. Soc.*.

Bormann, N., A. Collard, and P. Bauer, 2010: Estimates of spatial and inter-channel observation-error characteristics for current sounder radiances for numerical weather prediction. II: Application to AIRS and IASI data. *Quart. J. Roy. Meteor. Soc.*.

Boussetta, S., G. Balsamo, A. Beljaars, L. Kral, T. Jarlan, and T. Meyers, 2010: Impact of a satellite-derived leaf area index monthly climatology on screen level variables in a global numerical weather prediction model. *Mon. Wea. Rev.*.

Bouttier, F. and G. Kelly, 2001: Observing-system experiments in the ECMWF 4D-var data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **127**, 1469–1488.

Buizza, R., M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 2051–2066.

Cardinali, C., 2009: Monitoring the observation impact on the short-range forecast. *Quart. J. Roy. Meteor. Soc.*, **135**, 239–250.

Cardinali, C. and R. Buizza, 2004: Observation sensitivity to the analysis and the forecast: a case study during ATreC targeting campaign. In: *First THORPEX International Science Symposium*, pp. 6–10.

Cardinali, C., S. Pezzulli, and E. Andersson, 2004: Influence-matrix diagnostic of a data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **130**, 2767–2786.

Dee, D. and S. Uppala, 2009: Variational bias correction in ERA-Interim. *Quart. J. Roy. Meteor. Soc.*, **135**, 1830–1844.

Dee, D. P., 2004: Variational bias correction of radiance data in the ECMWF system. In: *ECMWF Workshop on Assimilation of High Spectral Resolution Sounders in NWP*, pp. 97–112. ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation background and analysis-error statistics in observation space. *Quart. J. Roy. Meteor. Soc.*, **131**, 3385–3396.

English, S., R. Saunders, B. Candy, M. Forsythe, and A. Collard, 2004: Met office satellite data OSEs. In: *Proceedings of the Third WMO Workshop on the Impact of Various Observing Systems on Numerical Weather Prediction, World Meteorological Organisation, WMO/TD*, volume 1228, pp. 146–156.

Ferranti, L., T. N. Palmer, F. Molteni, and E. Klinker, 1990: Tropical-extratropical interaction associated with the 30–60 day oscillation and its impact on medium and extended range prediction. *J. Atmos. Sci.*, **47**, 2177–2199.

Ferranti, L. and P. Viterbo, 2006: The European summer of 2003: Sensitivity to soil water initial conditions. *J. Climate*, **19**, 3659–3680.

Fisher, M., 2003: Background error covariance modelling. In: *Proceedings of the ECMWF Seminar on recent developments in data assimilation for atmosphere and ocean.*, pp. 45–63. ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK.

Fisher, M., 2007: The sensitivity of analysis errors to the specification of background error covariances. In: *Proceedings of the ECMWF Workshop on flow-dependent aspects of data assimilation.*, pp. 27–36. ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK.

Fisher, M. and P. Courtier, 1995: Estimating the covariance matrices of analysis and forecast error in variational data assimilation. *ECMWF Technical Memorandum*, (220).

Geer, A. J., P. Bauer, and N. Bormann, 2010: Solar biases in microwave imager observation assimilated at ecmwf. *IEE Trans. Geosci. Remote Sens.*, **48**, 2660–2669.

Graversen, R. G., E. Källén, M. Tjernström, and H. Körnich, 2007: Atmospheric mass-transport inconsistencies in the ERA-40 reanalysis. *Quart. J. Roy. Meteor. Soc.*, **133**(624), 673–680.

Grazzini, F. and V. Lucarini, 2009: Planetary wave tracking. *Geophys. Res. Abst.*, **11**, 2009EGUGA..11.8429G.

Grubišić, V. and P. K. Smolarkiewicz, 1997: The effect of critical levels on 3d orographic flows: Linear regime. *J. Atmos. Sci.*, **54**, 19431960.

Haimberger, L., B. Ahrens, F. Hamelbeck, and M. Hantel, 2001: Impact of time sampling on atmospheric energy budget residuals. *Meteorol. Atmos. Phys.*, **77**, 167–184.

Hasler, J., 1982: An investigation of the impact at middle and high latitudes of tropical forecast errors. ECMWF Technical Report 31, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK. 42 pp.

Healy, S. B., 2008: Assimilation of GPS radio occultation measurements at ECMWF. In: *Proceedings of the GRAS SAF Workshop on applications of GPS radio occultation measurements.*, pp. 99–109. ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK.

Held, I. M. and M. J. Suarez, 1994: A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models. *Bull. Amer. Meteor. Soc.*, **73**, 1825–1830.

Hewson, T. D., 2002: A comparison of cyclone spectra in forecasts from the operational (G1) and new dynamics trial (NT) versions of the unified model - Aug to Dec 2001. *Met Office Forecasting Research Technical Report*, (376).

Hewson, T. D., 2009a: Diminutive frontal waves - A link between fronts and cyclones. *J. Atmos. Sci.*, **66**, 116–132.

Hewson, T. D., 2009b: Tracking fronts and extra-tropical cyclones. ECMWF Newsletter 121, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

Hewson, T. D. and H. A. Titley, 2010: Objective identification, typing and tracking of the complete life-cycles of cyclonic features at high spatial resolution. *Meteor. Appl.*. In press.

Isaksen, L., J. Hasler, R. Buizza, and M. Leutbecher, 2010: The new ensemble of data assimilations. ECMWF Newsletter 123, ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK.

Jung, T., G. Balsamo, P. Bechtold, A. Beljaars, M. Köhler, M. Miller, J.-J. Morcrette, A. Orr, M. J. Rodwell, and A. M. Tompkins, 2010a: The ECMWF model climate: Recent progress through improved physical parametrizations. *Quart. J. Roy. Meteor. Soc.*. In press.

Jung, T. and M. Leutbecher, 2008: Scale-dependent verification of ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **134**, 973–984.

Jung, T., M. J. Miller, and T. N. Palmer, 2010b: Diagnosing the origin of extended-range forecast error. *Mon. Wea. Rev.*, **138**, 2434–2446.

Jung, T., T. N. Palmer, M. J. Rodwell, and S. Serrar, 2008: Diagnosing forecast error using relaxation experiments. ECMWF Newsletter 116, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

Jung, T., T. N. Palmer, M. J. Rodwell, and S. Serrar, 2010c: Understanding the anomalously cold European winter 2005/06 using relaxation experiments. *Mon. Wea. Rev.*, **138**, 3157–3174.

Jung, T. and M. J. Rodwell, 2010: Diagnosing remote origins of forecast error and circulation anomalies using relaxation experiments. In: *ECMWF Seminar Proceeding on "Diagnosis of Forecasting and Data Assimilation Systems, 7–9 September 2009"*. ECMWF, Shinfield Park, Reading RG2 9AX, UK.

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.

Krzeminski, B., N. Bormann, M. A. P. Kelly, G. and, and P. Bauer, 2009: Revision of the HIRS cloud detection. Technical Report 19, EUMETSAT/ECMWF Fellowship Programme, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

Kuang, Z., P. N. Blossey, and C. S. Bretherton, 2005: A new approach for 3d cloud-resolving simulations of large-scale atmospheric circulation. *Geophys. Res. Lett.*, **32**(L02809), 1–4.

Leutbecher, M., 2009: Diagnosis of ensemble forecasting systems. In: *Diagnosis of Forecasting and Data Assimilatin Systems*, ECMWF Seminar, pp. 235–266. ECMWF, Reading, UK.

Leutbecher, M. and T. N. Palmer, 2008: Ensemble forecasting. *J. Comp. Phys.*, **227**, 3515–3539.

Lopez, P., 2010: Linearized physics: Progress and issues. In: *Workshop on Assimilating Satellite Observations of Clouds and Precipitation into NWP models*. ECMWF/JCSDA, Reading, UK.

Lopez, P. and E. Moreau, 2005: A convection scheme for data assimilation: Description and initial tests. *Quart. J. Roy. Meteor. Soc.*, **131**, 409–436.

Lord, S., T. Zapotocny, and J. Jung, 2004: Observing system experiments with NCEP's global forecast system. In: *Third WMO Workshop on the Impact of Various Observing Systems on Numerical Weather Prediction, World Meteorological Organization., Alpbach, Austria*.

Lu, Q., W. Bell, P. Bauer, N. Bormann, and C. Peubey, 2010: An initial evaluation of FY-3A satellite data. *ECMWF Technical Memorandum*, p. available at http://www.ecmwf.int/publications/.

Mogensen, K. S., M. A. Balmaseda, A. Weaver, M. Martin, and A. Vidard, 2009: NEMOVAR: A variational data assimilation system for the new NEMO ocean model. ECMWF Newsletter 120, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

Morcrette, J.-J., H. W. Barker, J. N. S. Cole, M. J. Cole, M. J. Iacono, and R. Pincus, 2008: Impact of a new radiation package, McRad, in the ECMWF integrated forecasting system. *Mon. Wea. Rev.*, **136**, 4773–4798.

Morneau, J., S. Pellerin, S. Laroche, and M. Tanquay, 2006: Estimation of adjoint sensitivity gradients in observation space using the dual (PSAS) formulation of the environment canada operational 4D-Var. In: *Proceedings, 2nd THORPEX Intl Science Symp, Landshut, Germany*, pp. 4–8.

Neale, R. B. and B. J. Hoskins, 2000: A standard test for AGCMs and their physical parameterizations. I: The proposal. *Atmos. Sci. Letters*, **1**, 101–107.

Orr, A., P. Bechtold, J. Scinocca, M. Ern, and M. Janisková, 2010: Improved middle atmosphere climate and analysis in the ECMWF forecasting system through a non-orographic gravity wave parametrization. *J. Climate*. *(accepted)*.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. *ECMWF Technical Memorandum*, **598**, available at http://www.ecmwf.int/publications/.

Prusa, J. M., P. K. Smolarkiewicz, and A. A. Wyszogrodzki, 2008: EULAG, a computational model for multiscale flows. *Comput. Fluids*, **37**, 1193–1207.

Purser, R. J. and H. L. Huang, 1993: Estimating effective data density in a satellite retrieval or an objective analysis. *J. Appl. Meteor.*, **32**, 1092–1107.

Radnoti, G., P. Bauer, and A. McNally, 2010: ECMWF study on the impact of future developments of the space-based observing system on nwp. In: *EUMETSAT project report, in preparation*.

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. Jung, 2010: Diagnostics at ECMWF. In: *ECMWF Seminar Proceeding on "Diagnosis of Forecasting and Data Assimilation Systems, 7–9 September 2009"*. ECMWF, Shinfield Park, Reading RG2 9AX, UK.

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, T. D. Hewson, and T. Haiden, 2010: A new equitable score suitable for verifying precipitation in NWP. *Quart. J. Roy. Meteor. Soc.*, **136**, 1344–1363.

Saetra, O., H. Hersbach, J. R. Bidlot, and D. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.*, **132**, 1487–1501.

Schumacher, R. S. and C. A. Davis, 2010: Ensemble-based Forecast Uncertainty Analysis of Diverse Heavy Rainfall Events. *Wea. Forecasting*.

Takaya, Y., F. Vitart, G. Balsamo, M. Balmaseda, M. Leutbecher, and F. Molteni, 2010: Implementation of an ocean mixed layer model in the ifs. Technical Report 622, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK. Available at http://www.ecmwf.int/publications/.

Talagrand, O. and G. Candille, 2009: Verification of ensemble systems. In: *Diagnostics of data assimilation system performance*, Workshop. ECMWF, Reading, UK.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. In: *Proc. of Workshop on Predictability*, pp. 1–25. ECMWF, Reading, UK.

Tompkins, A. M. and M. Janisková, 2004: A cloud scheme for data assimilation: Description and initial tests. *Quart. J. Roy. Meteor. Soc.*, **130**, 2495–2517.

Trémolet, Y., 2007a: Model-error estimation in 4d-var. *Quart. J. Roy. Meteor. Soc.*, **133**, 1267–1280.

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

Tukey, J. W., 1972: Data analysis, computation and mathematics. *Quart. Appl. Math.*, **30**, 51–65.

Velleman, P. F. and R. E. Welsch, 1981: Efficient computing of regression diagnostics. *American Statistician*, **35**(4), 234–242.

Vitart, F. and F. Molteni, 2010: Simulation of the MJO and its teleconnections in the ECMWF forecast system. *Quart. J. Roy. Meteor. Soc.*, **136**, 842–855.

Vitart, F., S. Woolnough, M. Balmaseda, and A. Tompkins, 2007: Monthly forecast of the Madden-Julian Oscillation using a CGCM. *Mon. Wea. Rev.*, **135**, 2700–2715.

Wada, A., 2009: Idealized numerical experiments associated with the intensity and rapid intensification of a stationary tropical-cyclone-like vortex and its relation to initial sea-surface temperature and vortex-induced sea-surface cooling. *J. Geophys. Res.*, **114**, doi:10.1029/2009JD011993.

Wahba, G., D. R. Johnson, F. Gao, and J. Gong, 1995: Adaptive tuning of numerical weather prediction models: Randomized GCV in three and four dimensional data assimilation. *Mon. Wea. Rev.*, **125**.

Wedi, N. P., 2010: Diagnostics of model numerical cores: a model hierarchy. Proc. ECMWF Seminar on Diagnosis of Forecasting and Data Assimilation Systems, pp. 191–203. Eur. Cent. For Medium-Range Weather Forecasts, Reading, UK.

Wedi, N. P. and P. K. Smolarkiewicz, 2006: Direct numerical simulation of the Plumb-McEwan laboratory analog of the QBO. *J. Atmos. Sci.*, **63**(12), 3226–3252.

Wedi, N. P. and P. K. Smolarkiewicz, 2009: A framework for testing global nonhydrostatic models. *Quart. J. Roy. Meteor. Soc.*, **135**, 469–484.

Wedi, N. P., K. Yessad, and A. Untch, 2009: The nonhydrostatic global IFS/ARPEGE: model formulation and testing. Technical Report 594, Eur. Cent. For Medium-Range Weather Forecasts, Reading, UK.

Weisheimer, A., F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Arribas, M. Déqué, N. Keenlyside, M. MacVean, A. Navarra, and P. Rogel, 2009: ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions – Skill and progress beyond DEMETER in forecasting tropical pacific SSTs. *Geophys. Res. Lett.*, **36**, doi:10.1029/2009GL040896.

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

Willett, M. R., P. Bechtold, J. C. Petch, S. F. Milton, D. L. Williamson, and S. J. Woolnough, 2008: Modelling suppressed and active convection. Comparisons between three global atmospheric models. *Quart. J. Roy. Meteor. Soc.*, **134**, 1881–1896.

Zhu, P., C. S. Bretherton, M. Köhler, A. Cheng, A. Chlond, P. Geng, Q. amd Austin, J. C. Golaz, G. Lenderink, A. Lock, and B. Stevens, 2005: Intercomparison and interpretation of single-column model simulations of a nocturnal stratocumulus-topped marine boundary layer. *Mon. Wea. Rev.*, **133**, 2741–2758.

Zhu, Y. and R. Gelaro, 2008: Observation sensitivity calculations using the adjoint of the Gridpoint Statistical Interpolation (GSI) analysis system. *Mon. Wea. Rev.*, **136**, 335–351.