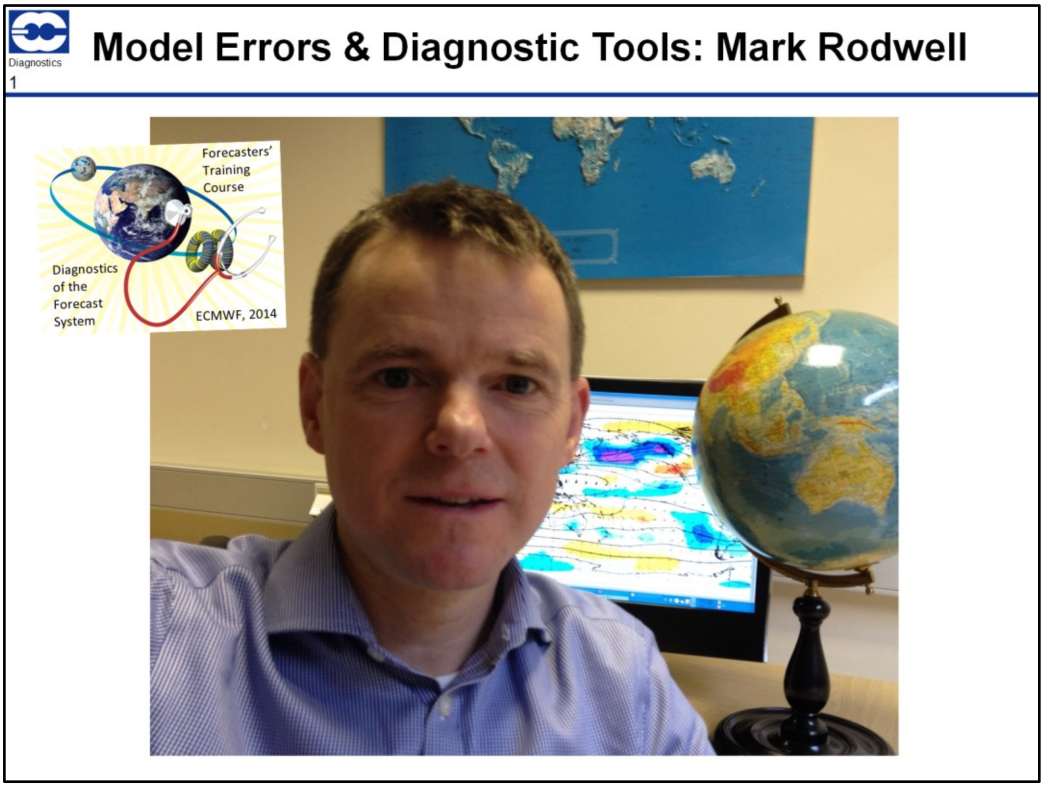


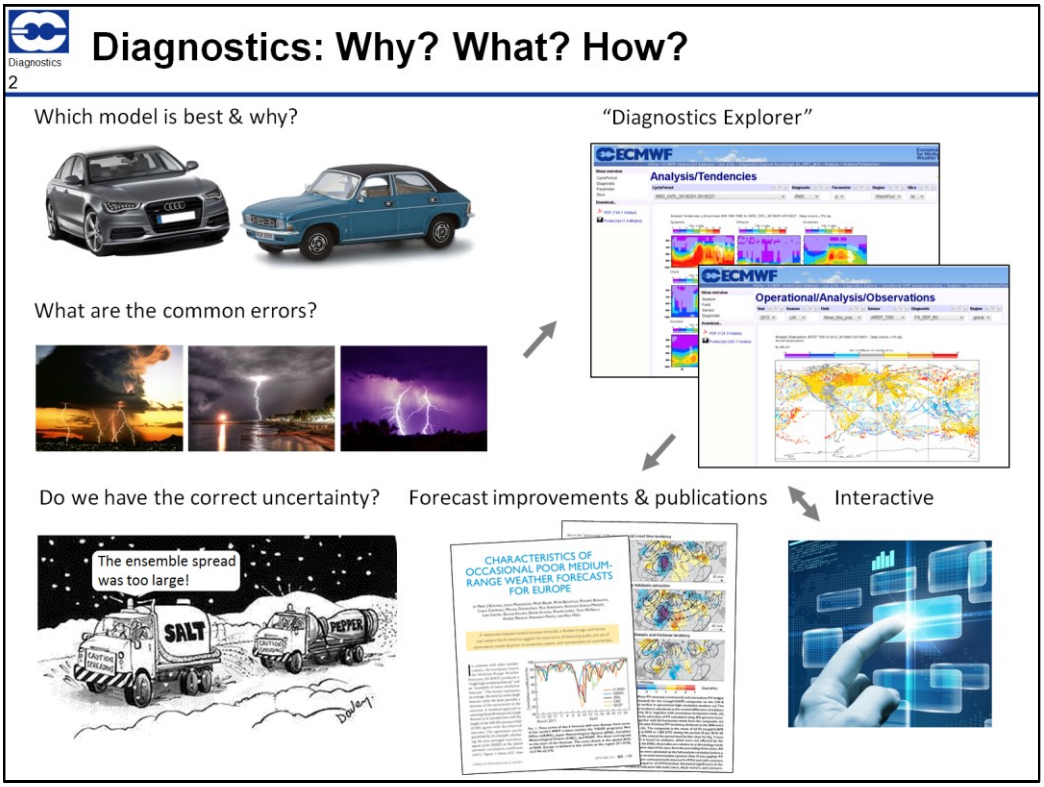
Diagnostics
1

Model Errors & Diagnostic Tools: Mark Rodwell

Forecasters' Training Course
Diagnostics of the Forecast System
ECMWF, 2014

Hello! I am Mark Rodwell. I work on Diagnostics of the Integrated Forecast System (IFS).

Diagnostics at ECMWF is about looking for weaknesses in the IFS, trying to identify their causes, working with developers, and documenting the resulting changes in performance. Here, with the help of a few examples, I will discuss the development of some of our diagnostic tools, and how they are helping us identify residual deficiencies.



Diagnostics: Why? What? How?

Diagnostics
2

Which model is best & why?

What are the common errors?

Do we have the correct uncertainty?

"Diagnostics Explorer"

Forecast improvements & publications

Interactive

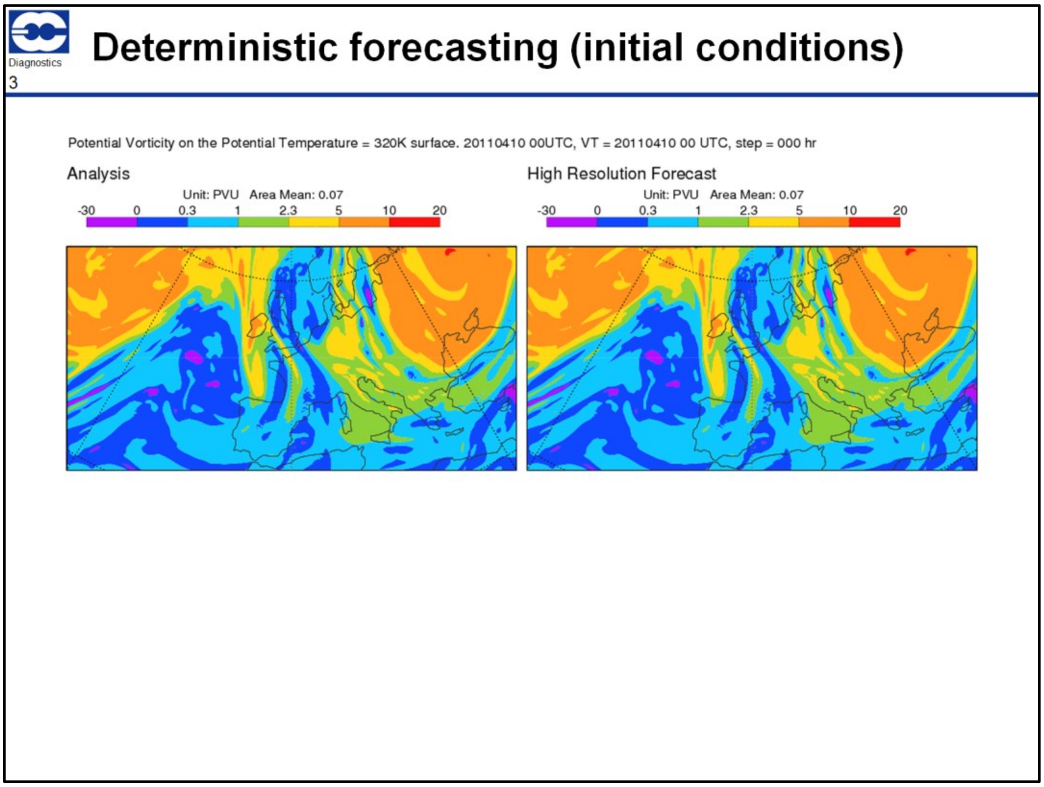
ECMWF
Analysis/Tendencies
ECMWF
Operational/Analysis/Observations
The ensemble spread was too large!
SALT
PEPPER
CAUTION SPREADING
CAUTION SPREADING
Dooley
CHARACTERISTICS OF OCCASIONAL POOR MEDIUM-RANGE WEATHER FORECASTS FOR EUROPE

The scope of Diagnostics is broad-ranging. Some think of diagnostics as primarily associated with case-studies of individual events that happened (a severe storm or the medieval warm period may be two examples) and perhaps were poorly forecast (or hindcast). Individual case-studies can be a useful approach for understanding the evolution of the event and for identifying gross errors, but have limitations if not backed-up by other evidence. At the heart of these limitations is the question "was it a model error, or was the event highly unpredictable?". By "unpredictable", I mean that it was a feature for which chaotic uncertainties grew quickly. Wrongly assuming it is all model error (in the weather or climate context) opens-up the possibility of making detrimental changes to our models. Others consider diagnostics to be about metrics of the model climate (the Lorenz energy cycle for example). This approach side-steps the issue of predictability, but at the possible expense of making it difficult to identify the root-cause of any deficiency. Nevertheless, it can be useful for monitoring and comparing weather or climate models. One could argue, however, that the two central aspects of weather and climate forecasting are precisely the quantification of predictability and the identification of the root-causes for errors! Central to my research have been attempts to enable Diagnostics to provide useful information on these two aspects.

The choice between cars in the slide is fairly clear (although I'm sure the Austin Maxi was ground-breaking in its day!). These days, it is perhaps more difficult to choose between cars without reading the specifications very closely. The same is true for weather and climate models. Our diagnostics need to be ever more precise if we are to judge unambiguously whether any given update to the forecasting system is a real improvement. This might involve looking at many cases (compositing on many storm events for example) but, because of limitations of computing time, also requires that our diagnostics distil as much information from the available test cases as possible, and are not sensitive to sampling uncertainties. Making sure that diagnostics are calculated from un-interpolated data, targeted carefully at a given hypothesis, and include statistical significance testing are examples and good practice.

These days, we are not just interested in developing one good model. Increasingly, users are demanding (and researchers are recognising the importance of) the estimation of uncertainty in all aspects of our forecast system. We need to be sure that the predicted uncertainty accurately reflects the true uncertainty of the situation. The cartoon strip represents an example of where the ensemble spread may have been too large! The correct tuning, or calibration, of ensemble spread also requires consideration of many cases and further development of diagnostics is needed.

The "Diagnostics Explorer" has been developed at ECMWF with the aim of providing all researchers the ability to make precise diagnostics of their experiments as easily as possible. The diagnostics include "model-space" and "observation-space" tools. The hope is that this will help speed-up the development process.

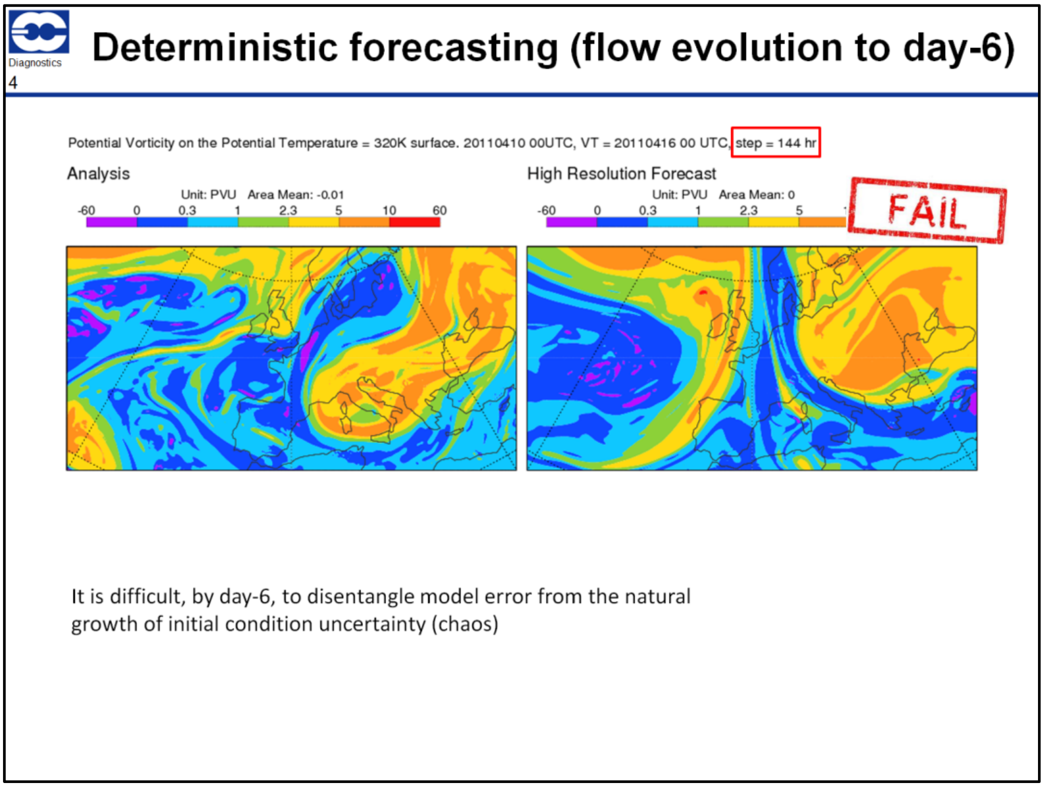


Diagnostics
3

Deterministic forecasting (initial conditions)

Potential Vorticity on the Potential Temperature = 320K surface. 20110410 00UTC, VT = 20110410 00 UTC, step = 000 hr

Here is an example of the traditional deterministic forecasting approach. I am showing Potential Vorticity (PV) on the 320K isentropic surface. The orange colours highlight high PV air, primarily from the polar stratosphere. The blue colours highlight tropospheric air from lower latitudes. The green band in-between highlights the tropopause. PV is materially conserved in adiabatic, frictionless situations. Hence PV highlights nicely the atmospheric flow and the physical process (convection etc.) that are modifying it. In the next slide, we see an animation of the subsequent analysed and forecast flow.



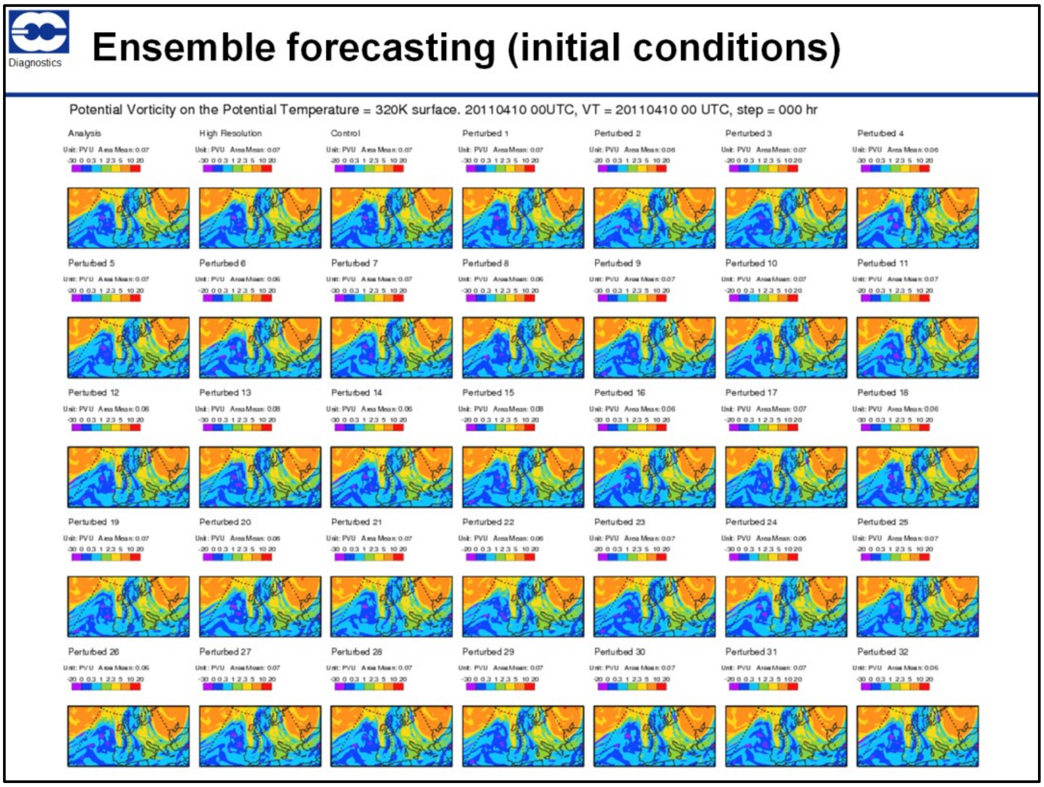
Diagnostics
4

Deterministic forecasting (flow evolution to day-6)

Potential Vorticity on the Potential Temperature = 320K surface. 20110410 00UTC, VT = 20110416 00 UTC, step = 144 hr

It is difficult, by day-6, to disentangle model error from the natural growth of initial condition uncertainty (chaos)

Over the first 6 days of the forecast, we see how the flow evolves. We see examples of anticyclonic wave-breaking that, in the analysis, lead to a blocked-type flow over Europe at day-6. The single high-resolution forecast is particularly poor by day-6, but what led to this forecast failure? Was it a model error? Was it that the initial conditions were unusually poor? Was it that the flow was more un-predictable than usual? We cannot begin to disentangle these issues at day-6 without moving to the ensemble (probabilistic) forecast system.

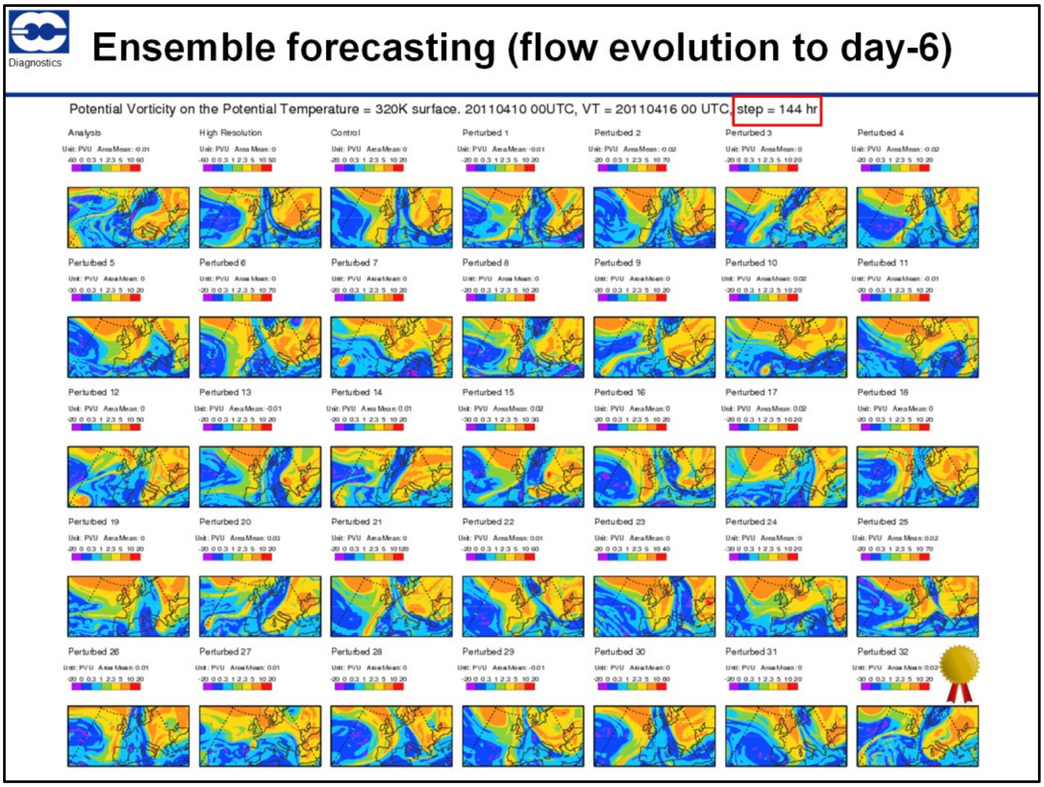


Diagnostics

Ensemble forecasting (initial conditions)

Potential Vorticity on the Potential Temperature = 320K surface. 20110410 00UTC, VT = 20110410 00 UTC, step = 000 hr

Here we see the same initial conditions, but I have also included the ensemble forecast system. For clarity, I am only showing the first 32 of the 50 ensemble members (labelled ‘Perturbed’). Looking carefully, it is possible to see slight differences in the initial conditions of these ensemble members. These differences represent our uncertainty in the observations and also errors that are likely to grow most quickly in the early stages of the forecast. I have also included the ‘control’ forecast for the ensemble. The initial conditions for the control forecast match more closely those of the high-resolution forecast, but are at the lower-resolution of the ensemble (although this is not easy to see differences here).



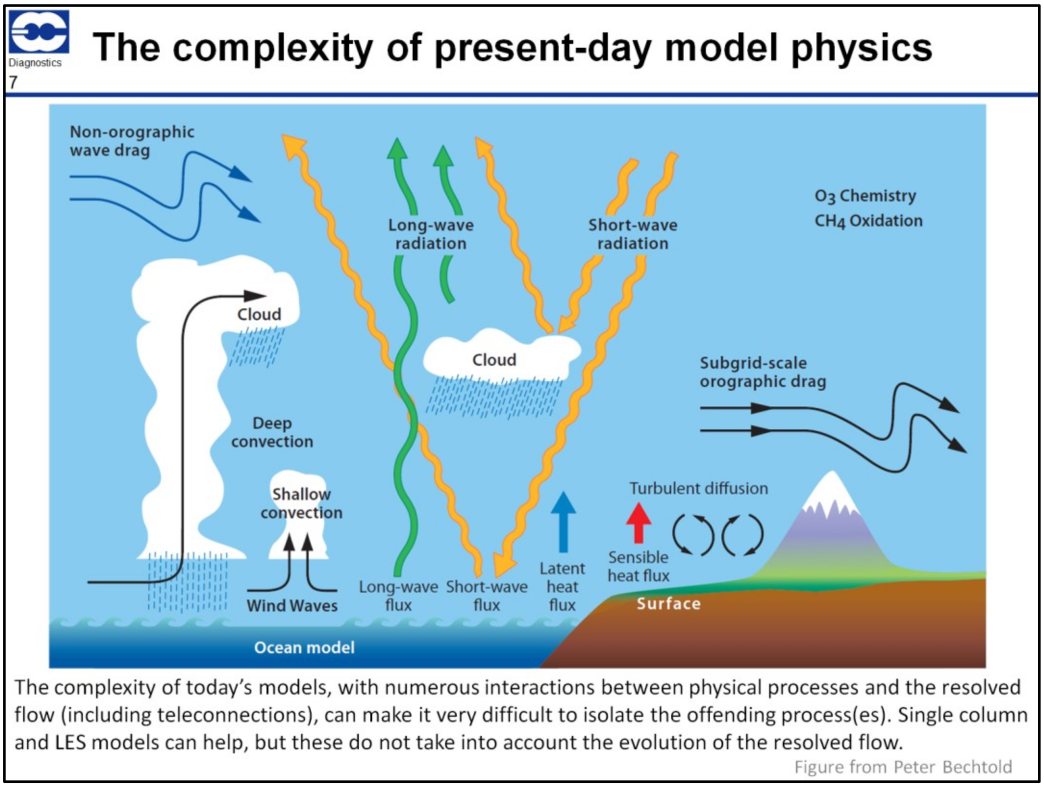
Diagnostics

Ensemble forecasting (flow evolution to day-6)

Potential Vorticity on the Potential Temperature = 320K surface. 20110410 00UTC, VT = 20110416 00 UTC, step = 144 hr

Again we see how the forecasts evolve over the first 6 days. In addition to the differences in initial conditions, the ensemble forecasts also include ‘stochastic physics’ (stochastically-perturbed physical tendencies – SPPT) which acts to represent sub-grid-scale uncertainty (or non-determinism) for a given resolved flow field.

At day-6, there is a lot of ‘spread’ in the ensemble – reflecting our uncertainty in the outcome. But do any members capture the real outcome? One member (‘Pertubed 32’) matches quite well the analysis. It is clearly possible that the issue is purely associated with low inherent predictability. In which case, any improvement would require reducing uncertainties in the initial conditions. Later in this talk, I highlight an aspect of the flow (a meso-scale convective system over North America) that acts within the forecasts to amplify the initial uncertainties. Hence it appears there is still scope for improving the forecast at day-6 through improvements to the modelling (or parametrization) of such meso-scale systems.



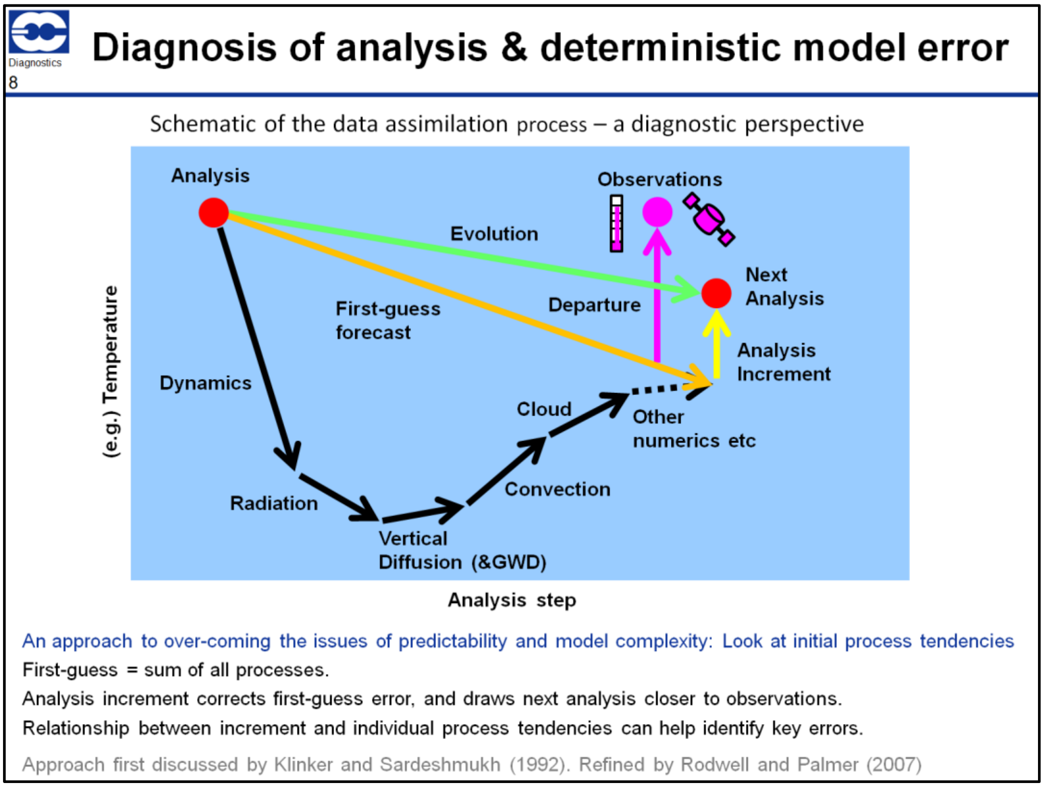
The complexity of present-day model physics

The complexity of today's models, with numerous interactions between physical processes and the resolved flow (including teleconnections), can make it very difficult to isolate the offending process(es). Single column and LES models can help, but these do not take into account the evolution of the resolved flow.

Figure from Peter Bechtold

In addition to the issue of predictability, there is another issue of model complexity.

Much model development is driven by a `bottom-up' desire to improve the representation of underlying physical processes. Such development is undertaken at ECMWF by researchers with responsibility for a given physical process. They use, for example, single-column models driven by fixed boundary conditions. However, we also need a `top-down' approach to diagnosing the impact of such changes in the full forecast system, and to identify residual issues. This top-down task is one of the functions of Diagnostics. With the increasing complexity (and accuracy) of present-day models, which include increasing numbers of physical and micro-physical processes, with considerable scope for interaction between themselves and with the resolved flow, our task is always challenging!



Diagnostics
8

Diagnosis of analysis & deterministic model error

Schematic of the data assimilation process – a diagnostic perspective

An approach to over-coming the issues of predictability and model complexity: Look at initial process tendencies

First-guess = sum of all processes.

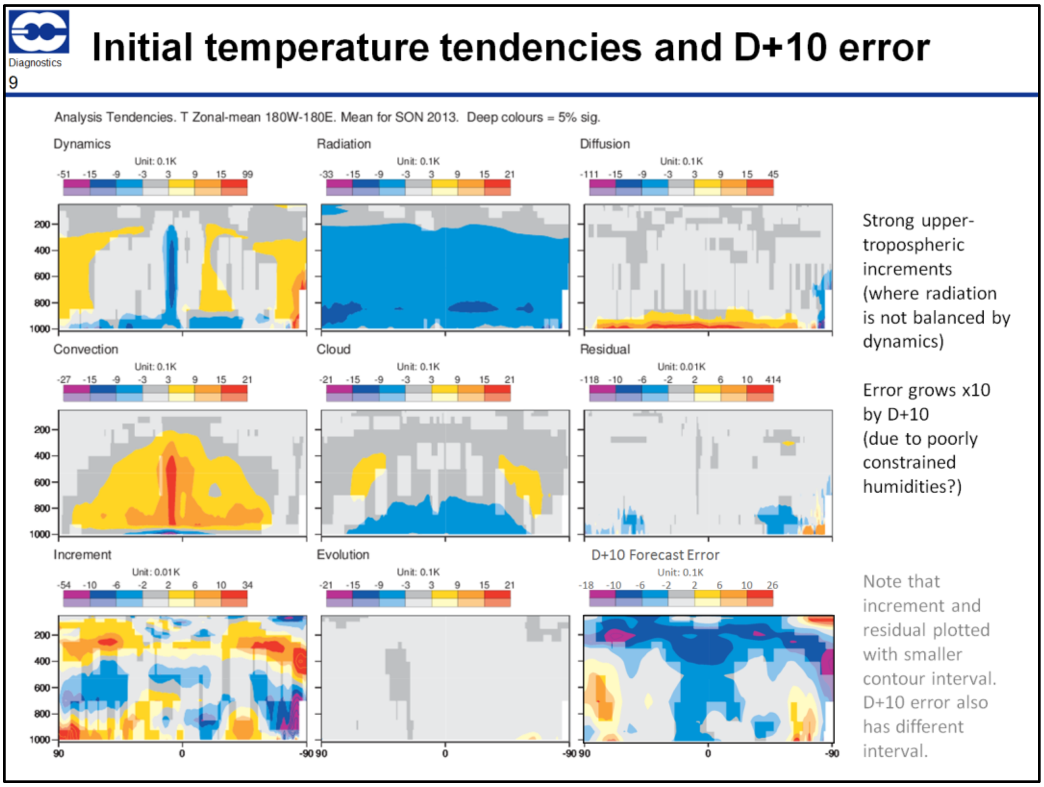
Analysis increment corrects first-guess error, and draws next analysis closer to observations.

Relationship between increment and individual process tendencies can help identify key errors.

Approach first discussed by Klinker and Sardeshmukh (1992). Refined by Rodwell and Palmer (2007)

One approach to over-coming the issues of predictability and model complexity in our quest for top-down model assessment is to look very early-on in the forecast. Indeed, to look at the forecast model within the data assimilation system itself. With every data assimilation cycle, the model is effectively assessed against millions of new observations. I will discuss such an approach within this talk.

The data assimilation system acts to draw the analysis away from the first-guess (prior forecast) and closer to the observations in a way that is consistent with estimated observation and model errors. The difference between the final analysis and the first-guess is known as the 'analysis increment'. This can be viewed as a correction to the first-guess forecast. What I term the 'analysed evolution' of the flow is the difference between successive analyses. Note that the first-guess forecast is simply the sum of the tendencies from the dynamics and the physical processes (and any other numerics) represented within the model. In the schematic, the impact of each process has been accumulated over all model timesteps within the data assimilation window - so that we can see the accumulated effect of each process. If the forecast model is perfect and the observations unbiased then, when averaged over many data assimilation cycles, the mean analysis increment should be zero (or at least very small). In such a situation, the contributions from all the processes should be almost in balance (since the analysed evolution is also small when averaged over many cycles). If the mean analysis increment is not zero, then this indicates that the model processes are not correctly in balance and that some aspect(s) of the model contain errors. (This assumes that the observations are unbiased, which is a reasonable assumption since the 'variational bias correction scheme' acts to remove large-scale systematic departures from the first-guess). How might such an imbalance arise? The concept of 'radiative-convective' equilibrium embodies the idea that radiative processes act to destabilise the atmosphere (heat the surface and cool the mid-to-upper troposphere) and the convection induced by this destabilisation acts to restore balance by cooling the surface and heating the mid-to-upper troposphere. With this idealised concept in mind, either a convection scheme that is too weak (given the observed temperature and humidity profiles) or a radiation scheme that is too strong (given the observed conditions – humidity *etc.*; as embodied by the analysis) would lead to a systematic initial *net* cooling of the mid-troposphere. Assuming there are relevant observations present, this cooling would be corrected with positive analysis increments. With the mean initial tendencies (or analysis increments), we therefore have a diagnostic that can quantify *local* model physics error before significant interactions have taken place with (and via) the resolved dynamics.



Diagnostics

9

Initial temperature tendencies and D+10 error

Strong upper-tropospheric increments (where radiation is not balanced by dynamics)

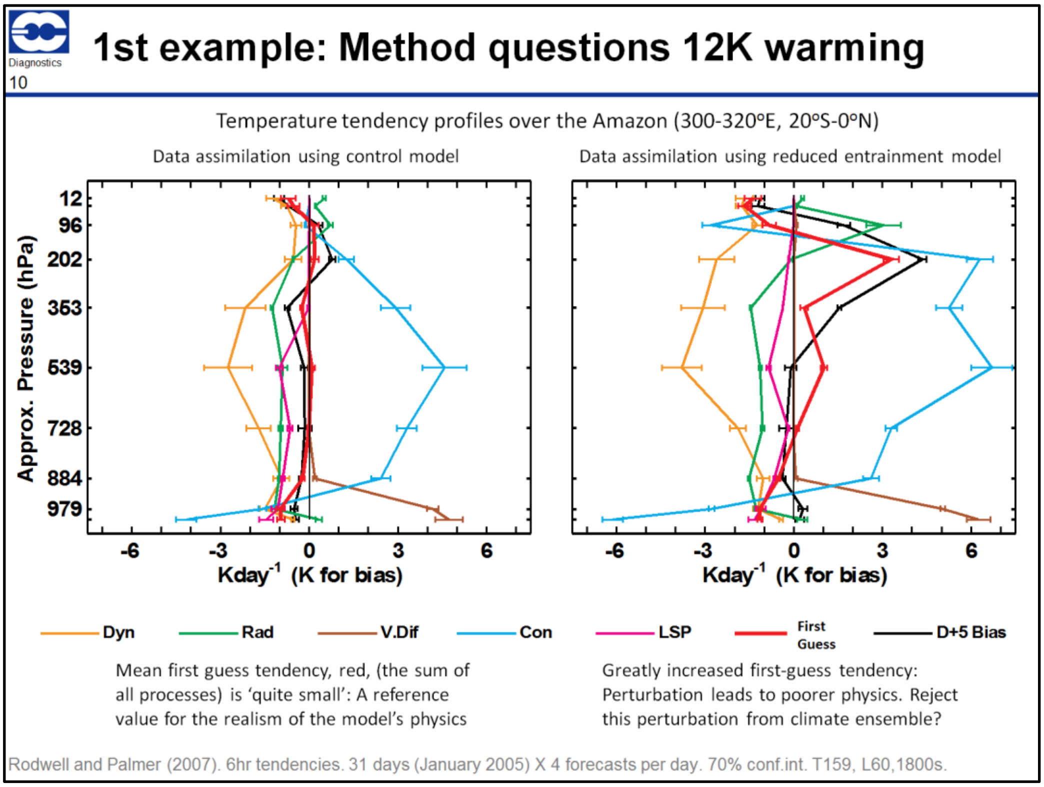
Error grows x10 by D+10 (due to poorly constrained humidities?)

Note that increment and residual plotted with smaller contour interval. D+10 error also has different interval.

A long-standing issue for the ECMWF model has been a cold bias near the (primarily extratropical) tropopause. The analysis increments (bottom left) show that the analysis is around 0.1K warmer on average than the model at a lead-time of 12 hours. The mean forecast error at day 10 (bottom right) is around 1K. This ten-fold growth confirms that the root-cause is likely to be model error, rather than observation error. Consideration of the process tendencies shows that there are two strong contributors to the temperature budget in the extratropical upper troposphere: dynamics and radiation. What appears to be happening is that the radiative cooling extends slightly higher than the dynamical warming.

Why might this be the case? Separate experiments (Leroy and Rodwell; 2014) highlight an issue with humidity at this level. There does not appear to be a lot of upper-tropospheric humidity information in the observations, and consequently humidities are able to drift towards the model's own attractor (which is difficult to improve for the same reason). The radiation then responds to these humidity biases and leads to erroneous cooling.

Note that the 'residual' in the budget is plotted with the same contour interval as the analysis increments. The is important because the conclusions drawn require that there are no other hidden (numerical) reasons for the budget imbalance.



1st example: Method questions 12K warming

Mean first guess tendency, red, (the sum of all processes) is 'quite small': A reference value for the realism of the model's physics

Greatly increased first-guess tendency: Perturbation leads to poorer physics. Reject this perturbation from climate ensemble?

Rodwell and Palmer (2007). 6hr tendencies. 31 days (January 2005) X 4 forecasts per day. 70% conf.int. T159, L60,1800s.

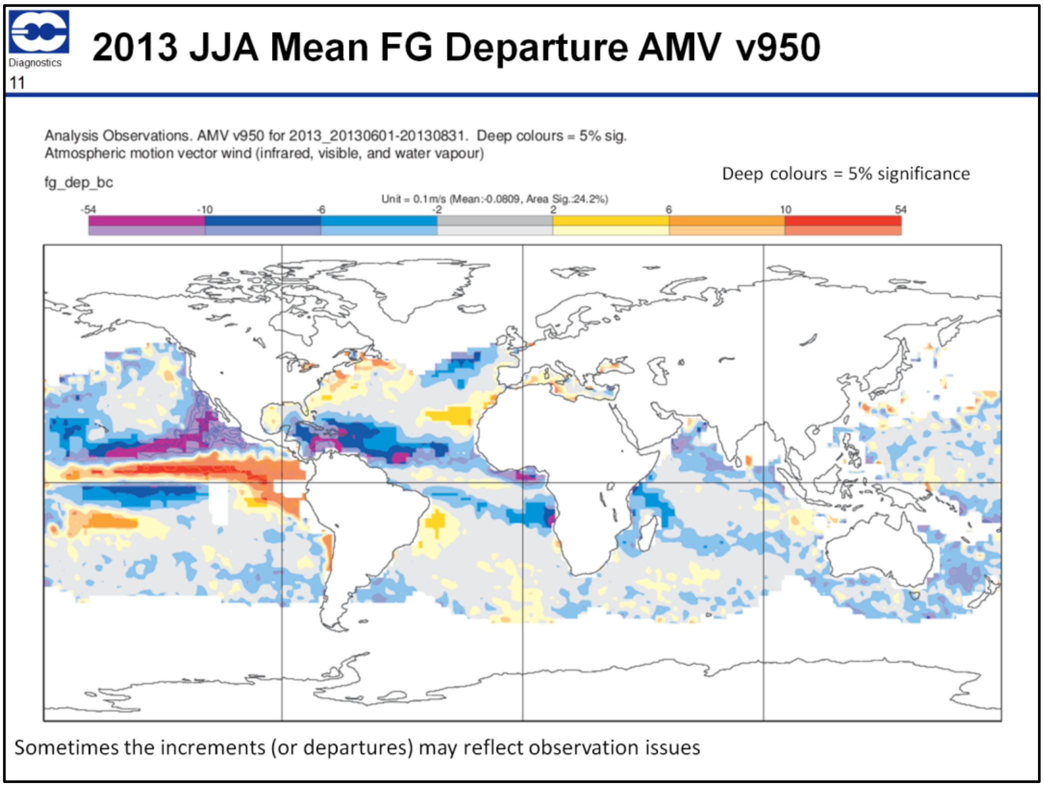
This slide shows our first use of this approach.

The left figure shows the initial tendencies associated with the dominant physical processes (and the dynamics) within the ECMWF model for the Amazon region (based on a model version that was operational a few years ago). It can be seen that convective (Con) heating is balanced by dynamic (Dyn) cooling due to ascent and also by radiative (Rad) cooling. The first-guess tendency is simply the sum the individual process tendencies. It can be seen that this total initial tendency is 'small' and thus the model physics is 'reasonable'.

Stainforth et al. (2005) found that the perturbation that led to the largest global warming (up to 12K) in their ensemble was a reduction in the turbulent entrainment coefficient. When this perturbation is applied to the ECMWF model and a new data assimilation / forecast cycle experiment made, it is found (see right figure) that the initial tendencies are not so well in balance.

Rodwell and Palmer (2007) argued that the larger total initial tendency implies that this perturbation is very unrealistic (unphysical) and can thus be down-weighted or even rejected within the perturbed ensemble. In this particular case, the uncertainty in climate change would be substantially reduced. The initial tendency approach is much less computationally expensive than running a coupled model for hundreds of years. This provides justification for pursuing a seamless approach to weather and climate forecasting.

We can begin to understand physically what is happening in the model with reduced entrainment: with less entrainment, less buoyancy is detrained from a convective plume, which thus rises higher and heats more. Initially, this increased heating is not balanced by increased dynamical cooling since the large-scale dynamics are better constrained by the observations, and respond more slowly. Later in the forecasts, the processes within this perturbed model must also come into balance (when the atmospheric state approaches the climate attractor of the perturbed model). Interesting the Amazonian precipitation climate of the perturbed model is less than that of the control model – highlighting the issues of interpretation after processes have had time to interact: a key reason for looking at the shortest relevant timescales.



Diagnostics
11

2013 JJA Mean FG Departure AMV v950

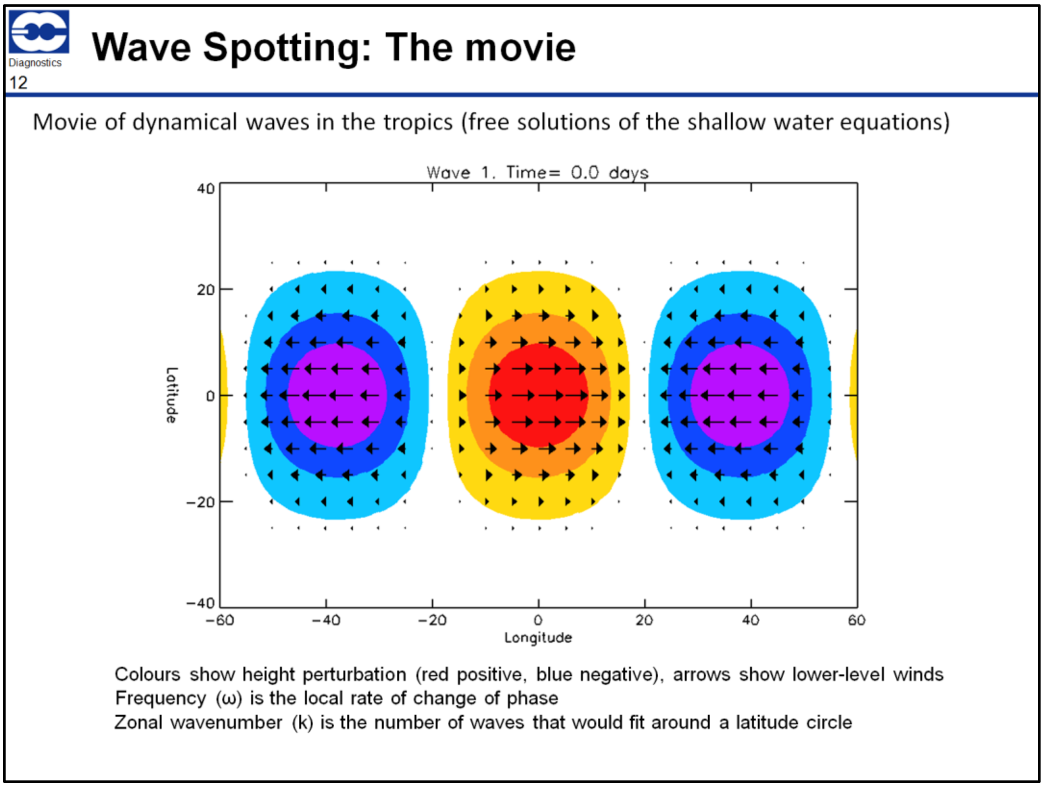
Analysis Observations. AMV v950 for 2013_20130601-20130831. Deep colours = 5% sig.
Atmospheric motion vector wind (infrared, visible, and water vapour)

fg_dep_bc

Deep colours = 5% significance

Sometimes the increments (or departures) may reflect observation issues

It is not always obvious from examination of the data assimilation system that the model is at fault. Here is an example of the differences between the observed meridional wind speed, as observed using satellite-derived atmospheric motion vectors (AMVs). The various geostationary satellites are obvious from there circular foot-prints. Is the meteorology of the eastern tropical Pacific that different from the meteorology of the other ocean basins to account for the much larger departures, or is there an issue with the satellite observations in this region? This is one question we are investigating at the moment.



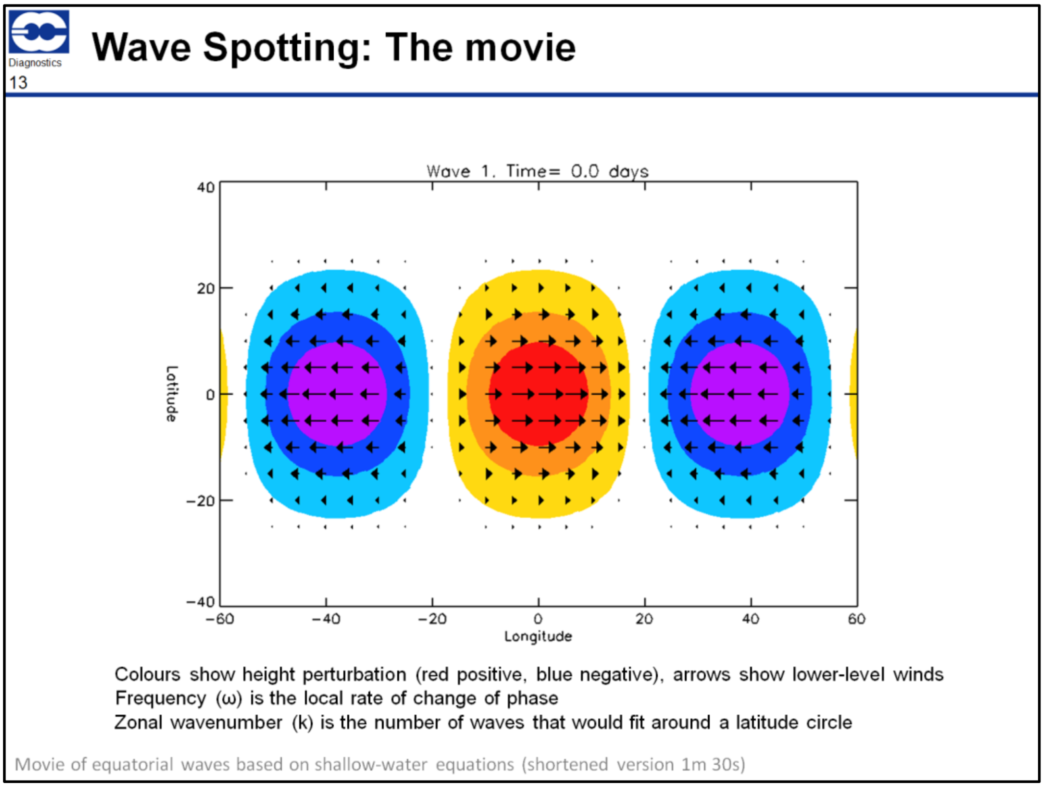
Diagnostics
12

Wave Spotting: The movie

Movie of dynamical waves in the tropics (free solutions of the shallow water equations)

Colours show height perturbation (red positive, blue negative), arrows show lower-level winds
Frequency (ω) is the local rate of change of phase
Zonal wavenumber (k) is the number of waves that would fit around a latitude circle

In addition to reducing mean errors, we also wish to better represent variability in the flow. This is (the first frame of) an animation that highlights the horizontal structure and phase-speed of the various equatorial waves (assuming a stationary background flow). The animation highlights the fact that local anomalies (or errors) can affect other regions through the actions of waves and teleconnections (quasi-stationary waves). This animation also highlights the fact that variations in the atmosphere can have specific spatio-temporal characteristics. We can use this fact to target diagnostics towards particular features.

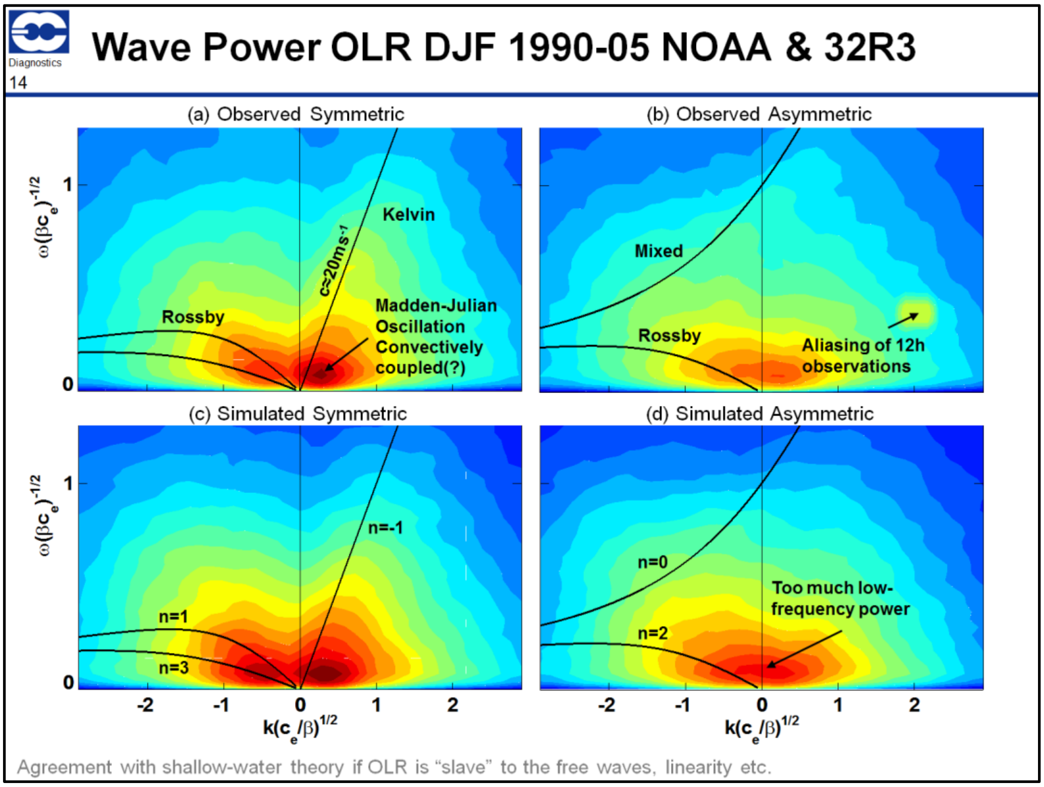


Wave Spotting: The movie

Colours show height perturbation (red positive, blue negative), arrows show lower-level winds
Frequency (ω) is the local rate of change of phase
Zonal wavenumber (k) is the number of waves that would fit around a latitude circle

Movie of equatorial waves based on shallow-water equations (shortened version 1m 30s)

In addition to reducing mean errors, we also wish to better represent variability in the flow. This is an animation that highlights the horizontal structure and phase-speed of the various equatorial waves (assuming a stationary background flow). The animation highlights the fact that local anomalies (or errors) can affect other regions through the actions of waves and teleconnections (quasi-stationary waves). This animation also highlights the fact that variations in the atmosphere can have specific spatio-temporal characteristics. We can use this fact to target diagnostics towards particular features.

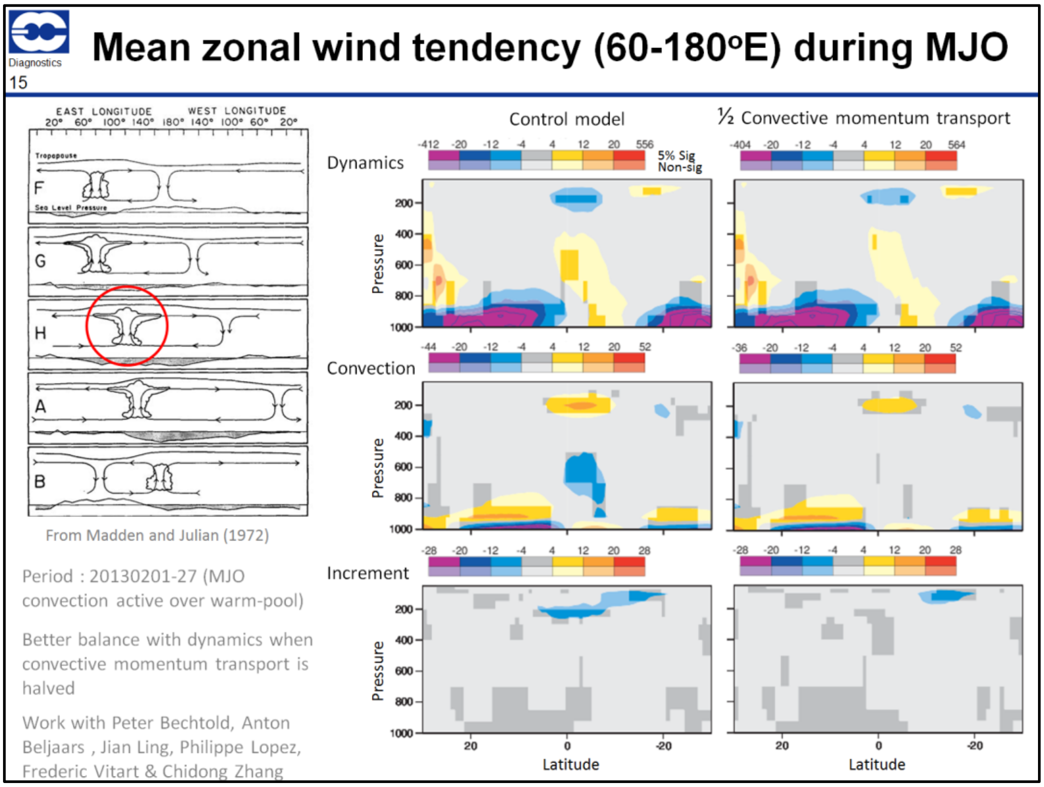


Diagnostics
14

Wave Power OLR DJF 1990-05 NOAA & 32R3

Agreement with shallow-water theory if OLR is “slave” to the free waves, linearity etc.

Using out-going long-wave radiation (for example) it is possible to isolate the wave power for each wave-number and frequency. This has been done in panel (a) for the observed waves that are symmetric about the equator, and over-laid with the theoretical dispersion diagram for equatorial waves. The agreement is quite amazing (to me at least!). The higher wave power can be seen for both the Rossby and Kelvin waves. The observations are not frequent enough to allow us to calculate the power in the gravity waves. The region of high power at small positive wave-number and small frequency relates to the Madden-Julian Oscillation (MJO). The MJO is not on any of the curves and this demonstrates that the MJO is not adequately represented by the Shallow water model. A likely reason could be that the MJO involves coupling with the physics (convection, radiation etc). A reasonable agreement can also be seen between the theoretical symmetric waves and those found in the ECMWF model cycle 32R3, here run at resolution T_L159L91 (panel c). The ECMWF model cycle 32R3 had a major change to its convection. The entrainment of moisture into a convective plume used to be partly due to turbulent processes and partly related to the large-scale convergence. At cycle 32R3, the explicit connection to the large-scale convergence was removed. This had many beneficial impacts on forecast scores and on synoptic activity. Unfortunately, it also strongly increased low-frequency, planetary activity. This over-estimation of low-frequency activity is clearly visible for symmetric and asymmetric waves (compare panels (c) and (d) to panels (a) and (b), respectively). The entrainment change may well have had a major impact on wave-convective coupling. It remains to be seen if subsequent model cycles build on this change and improve our ability to simulate the Madden-Julian Oscillation. For the asymmetric waves, the mixed Rossby-gravity waves are apparent in the observed OLR data (panel b). The model (panel d) tends to capture the asymmetric Rossby waves but doesn’t appear to capture the distinct set of mixed Rossby-gravity waves seen in the observations. (Note that the large power near non-dimensionalised wavenumber 2 is spurious and due to satellite sampling that involves 14 passes around the length of the equator).



Diagnostics
15

Mean zonal wind tendency (60-180ºE) during MJO

From Madden and Julian (1972)

Period : 20130201-27 (MJO convection active over warm-pool)

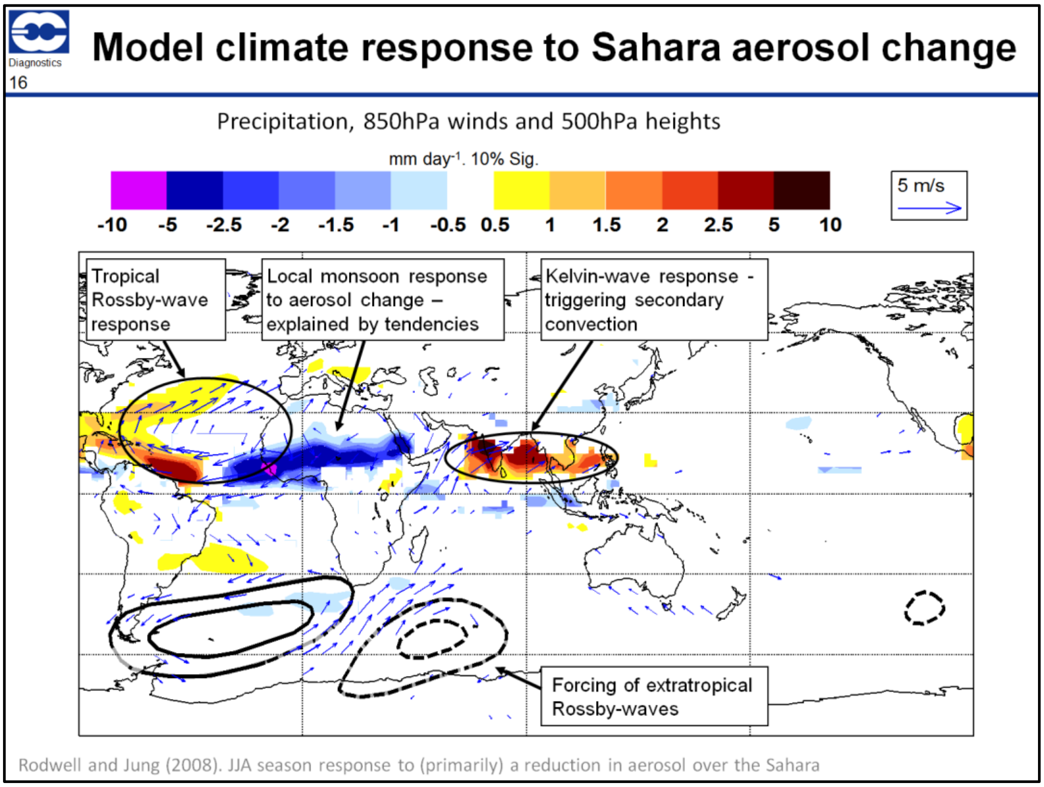
Better balance with dynamics when convective momentum transport is halved

Work with Peter Bechtold, Anton Beljaars , Jian Ling, Philippe Lopez, Frederic Vitart & Chidong Zhang

The panel on the left is part of a schematic showing the eastward propagation of the Madden and Julian Oscillation MJO) around the equator. The MJO is not a solution of the shallow water equations (that were shown in the wave-spotting movie), partly because it involves an interaction between the dynamical flow and the convection. It is these interactions that we need focus-on diagnostically.

We have investigated the MJO with the initial tendency approach too. The temperature budget is a bit tricky to interpret as several processes are important. Here I am showing (in the central column of panels) the zonal wind budget, that is dominated by the dynamics and convection terms. This is the budget averaged between 60 and 180ºE when the convection is active over the Warm Pool and western equatorial Pacific. In the upper tropical troposphere, the dynamics are accelerating the (prevailing easterly) winds and the convection is acting to decelerate these winds. The physical process for this deceleration is the turbulent mixing with lower-tropospheric air, which has a smaller easterly component. The increments suggest that the convective deceleration is twice as strong as it should be (assuming the resolved dynamics are blameless!). In a sensitivity study, we halved the convective momentum transport globally and the resulting tendency budget (when data assimilation is performed with this perturbed model) is shown in the right-hand panels. The result was quite pleasing – the upper-tropospheric balance was improved so that the mean increments effectively disappear. Since the convective momentum transport must integrate vertically to zero, one worry was that the reduction is convective momentum transport would also reduce the easterly convective acceleration in the mid-troposphere and lead to a new imbalance with the dynamics. What was also pleasing to see was that the mid-tropospheric dynamical westerly forcing appears to weaken (as an indirect response) so that the mid-tropospheric balance is maintained. This example demonstrates that the initial tendency approach can motivate, as well as assess, model improvements.

We have yet to assess whether the better upper-tropospheric momentum budget allows the MJO to propagate more freely through the Warm-Pool region. This is a particular problem for the model at present and one that, once solved, could lead to improved extended-range predictions in the extratropics as well as the tropics.



Diagnostics 16

Model climate response to Sahara aerosol change

Precipitation, 850hPa winds and 500hPa heights

Tropical Rossby-wave response

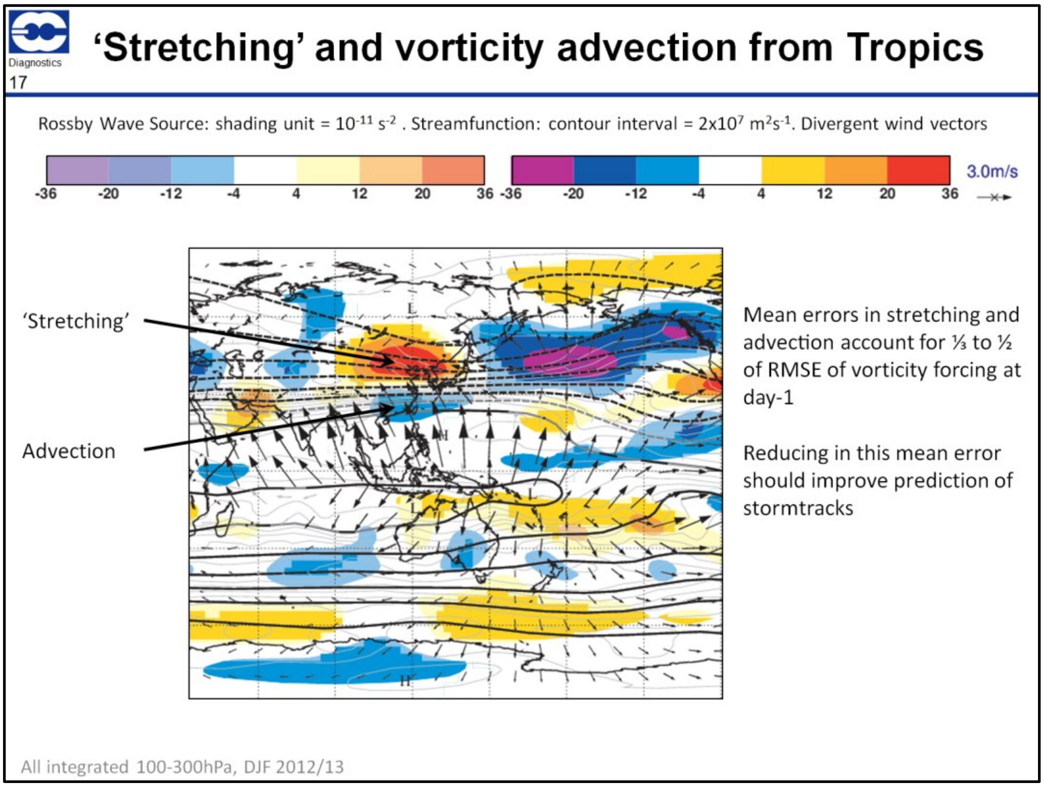
Local monsoon response to aerosol change – explained by tendencies

Kelvin-wave response - triggering secondary convection

Forcing of extratropical Rossby-waves

Rodwell and Jung (2008). JJA season response to (primarily) a reduction in aerosol over the Sahara

This figure shows the model climate changes in precipitation, low-level winds and 500 hPa geopotential heights associated with a change in model aerosol climatology – primarily over the Sahara region (Rodwell and Jung; 2008). The main effect was a dramatic (and welcome) decrease in strength of the North African monsoon. The physical basis for this improvement was understood through the use of initial (and medium-range) tendency budgets. From what I have discussed about equatorial waves, we are able to explain the change in low-level circulation over the subtropical north Atlantic as a clear Rossby-wave response to the reduced monsoon heating over northern Africa (the circulation anomaly is the opposite to that seen for the `Gill response' (Gill; 1980 – see also Matsuno; 1966) because we have a cooling anomaly rather than his heating anomaly). In addition, increased precipitation over the Indian Ocean in the June-August season (JJA) can be explained as being triggered by the upwelling Kelvin wave response to the reduced heating over tropical north Africa. As with the North African monsoon response itself, feedbacks with the convection tend to strongly enhance the precipitation response over the Indian Ocean. Notice the additional response in the southern extratropics. I discuss briefly our diagnostics of tropical-to-extratropical forcing in the next slide.



Diagnostics
17

‘Stretching’ and vorticity advection from Tropics

Rossby Wave Source: shading unit = 10^{-11} s^{-2} . Streamfunction: contour interval = 2×10^{7} $m^{2}s^{-1}$. Divergent wind vectors

‘Stretching’

Advection

Mean errors in stretching and advection account for ⅓ to ½ of RMSE of vorticity forcing at day-1

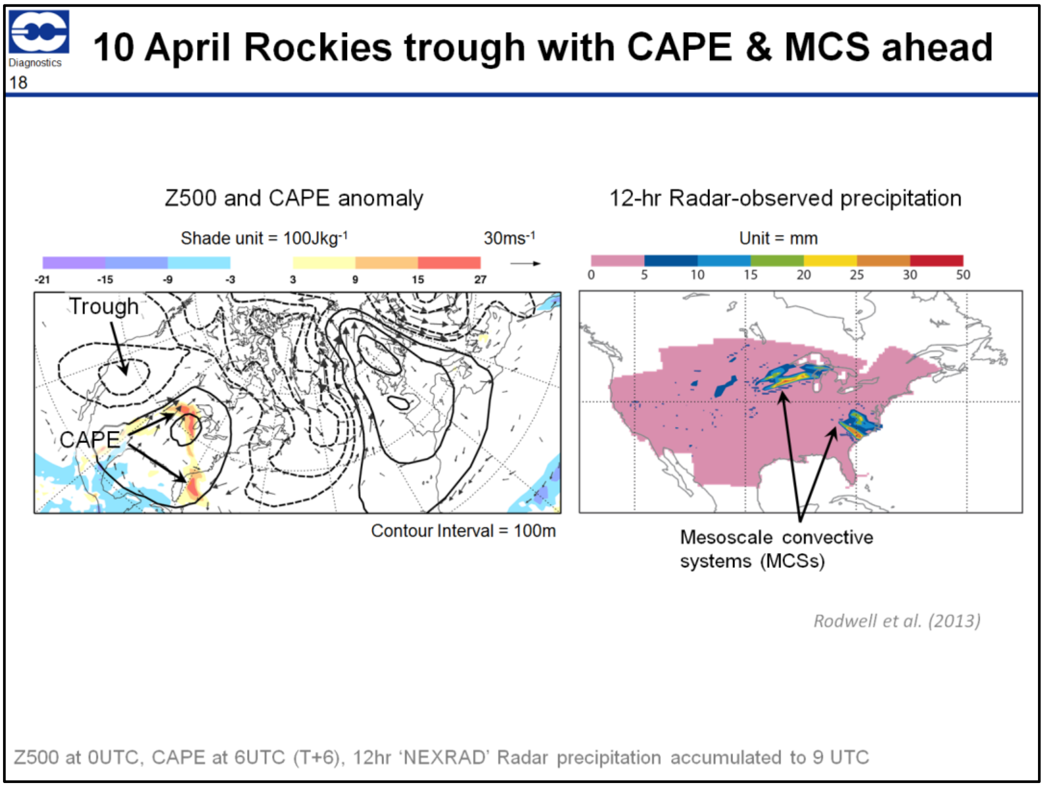
Reducing in this mean error should improve prediction of stormtracks

All integrated 100-300hPa, DJF 2012/13

Sardeshmukh and Hoskins (1988) discussed the ‘Rossby wave source’ and how tropical convective anomalies could lead to the excitation of extratropical Rossby waves. This plot shows, for DJF 2012/13, shows mean upper-tropospheric divergent winds (integrated between 100 and 300hPa), mean streamfunction (thick contours) and the Rossby wave source (shaded). The streamfunction is associated with the rotational flow, which reaches speeds of ~$50ms^{-1}$ (*e.g.* over Tibet). The Rossby wave source comprises two elements. (1) A ‘stretching term’ - which is visible (in red) on the down-slopes of Tibet - whereby horizontal convergence associated with stretching leads to the spin-up of vorticity (the ‘ballerina effect’). (2) The blue shading over southern China represents the advection of low values of planetary vorticity by the divergent outflow from tropical convection. Such vorticity forcing results in Rossby waves along the Jet stream. In some cases (such as for the aerosol change on the previous slide) the resulting extratropical response involves a strong stationary component (when the Rossby wave has a scale that allows it to propagate upstream at the same speed as it is advected downstream by the Jet). Regardless of whether the extratropical response is stationary or not, it is clearly worth evaluating our ability to predict the Rossby wave source.

Mean errors in the Rossby wave source presently account for ⅓ to ½ of the day-1 root-mean-square errors in this region (and indeed around the other large mountain ranges of the Earth too). Although it maybe unclear by how much we can reduce the random component of the Rossby wave source error, it does seem feasible that we could tackle this sizeable mean component. Doing so should help us better predict the evolution of the world’s stormtracks, such as that over the North Pacific.

For more discussion of this diagnostic approach, see Rodwell and Jung (2008).



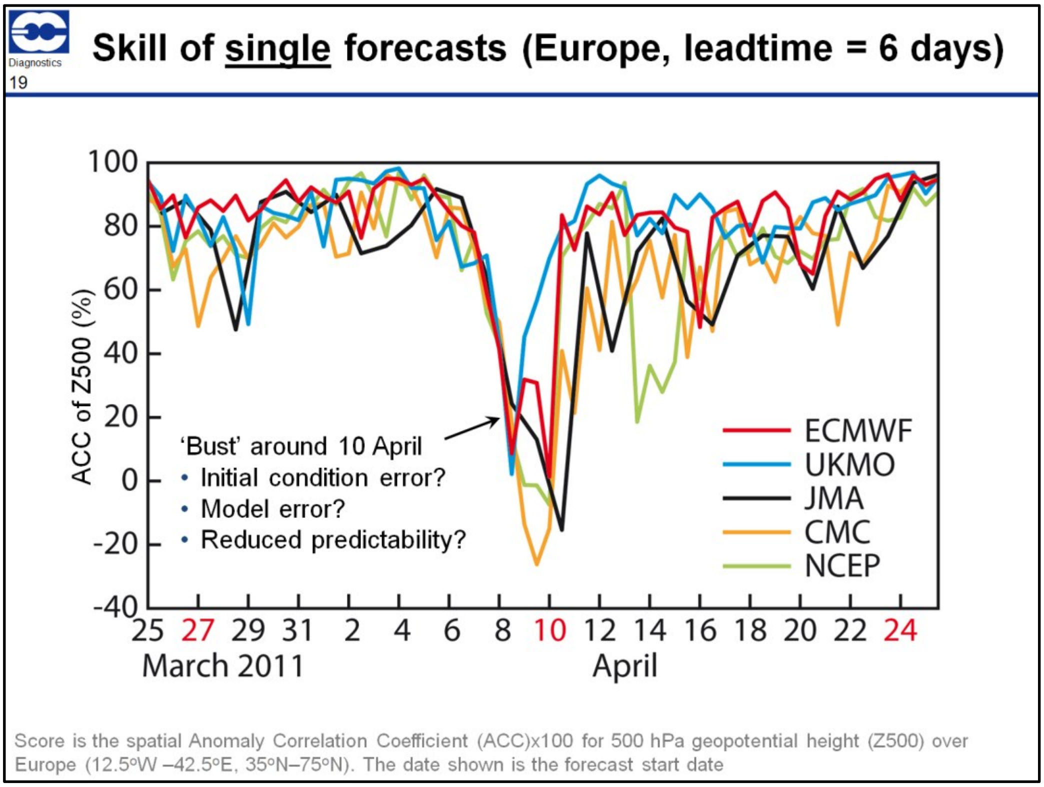
Diagnostics
18

10 April Rockies trough with CAPE & MCS ahead

Rodwell et al. (2013)

Z500 at 0UTC, CAPE at 6UTC (T+6), 12hr 'NEXRAD' Radar precipitation accumulated to 9 UTC

Rossby waves crossing the North Pacific, and baroclinic developments, lead to variations in the flow over North America. One particular flow-regime over North America has been shown to lead to poorer subsequent forecasts over the North Atlantic and Europe (Rodwell et al.; 2013). An example, for 10 April 2011, is shown here. The key features are a trough over the Rockies and, associated with warm moisture southerly flow on its leading edge, high levels of Convective Available Potential Energy (CAPE). This CAPE can give rise to Mesoscale Convective Systems (MCSs; shown in the right panel) over northern North America which can disrupt the upper-level Jet Stream. This is actually the case I discussed at the beginning – when the high resolution forecast for Europe was particularly bad at day-6


Diagnostics
19

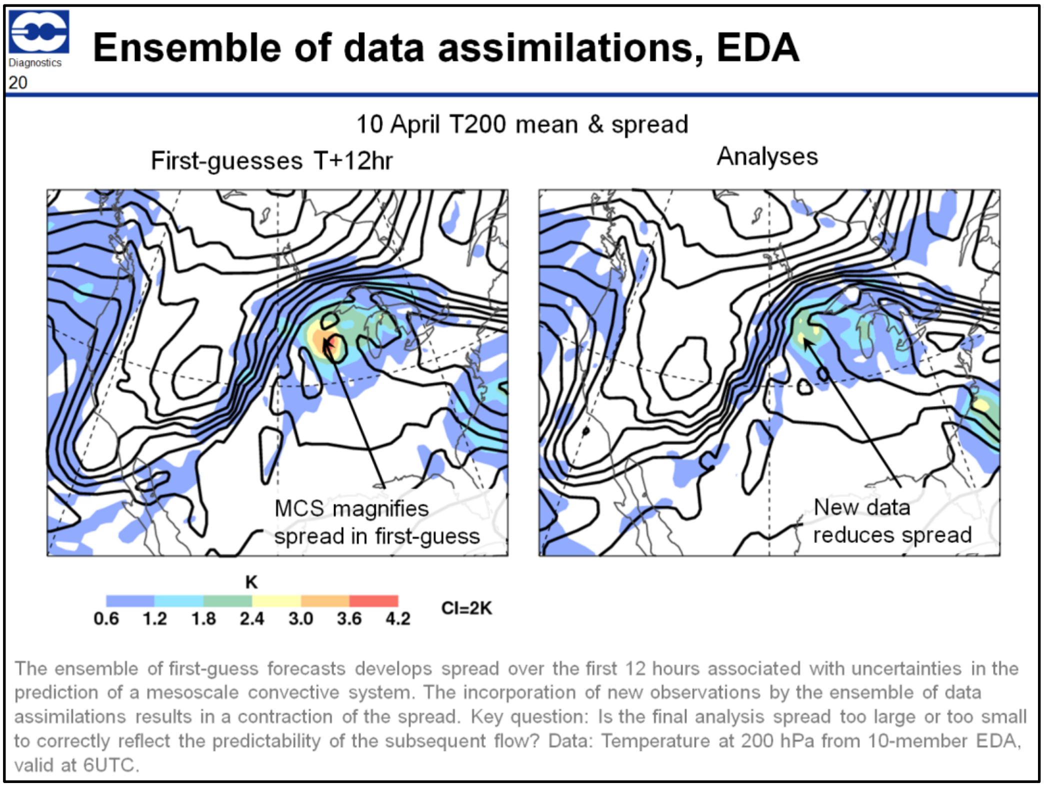
Skill of single forecasts (Europe, leadtime = 6 days)

‘Bust’ around 10 April
- Initial condition error?
- Model error?
- Reduced predictability?

Score is the spatial Anomaly Correlation Coefficient (ACC)x100 for 500 hPa geopotential height (Z500) over Europe (12.5°W –42.5°E, 35°N–75°N). The date shown is the forecast start date

This figure shows the time series of spatial anomaly correlation coefficient (ACC) for day-6 forecasts of Z500 over Europe from several of the world’s forecast centres. In general, scores fluctuate around the 80% level, but around 10 April (corresponding the 10 April case shown on the previous slide) a ‘bust’ occurs. On this particular occasion all centres suffered, with the UK Met. Office (UKMO) recovering earliest.

The fact that all centres suffered suggests that much of this bust must be associated with a drop in predictability (and not purely due to model error). Hence we need to start think in probabilistic terms and investigating our ensemble prediction systems. Looking for “best ensemble members” can give an indication of the likely regions where initial condition uncertainty is most important. In this case, it was over central North America. The next slide shows the development of this uncertainty over the subsequent 12 hours.



Diagnostics
20

Ensemble of data assimilations, EDA

10 April T200 mean & spread

First-guesses T+12hr

Analyses

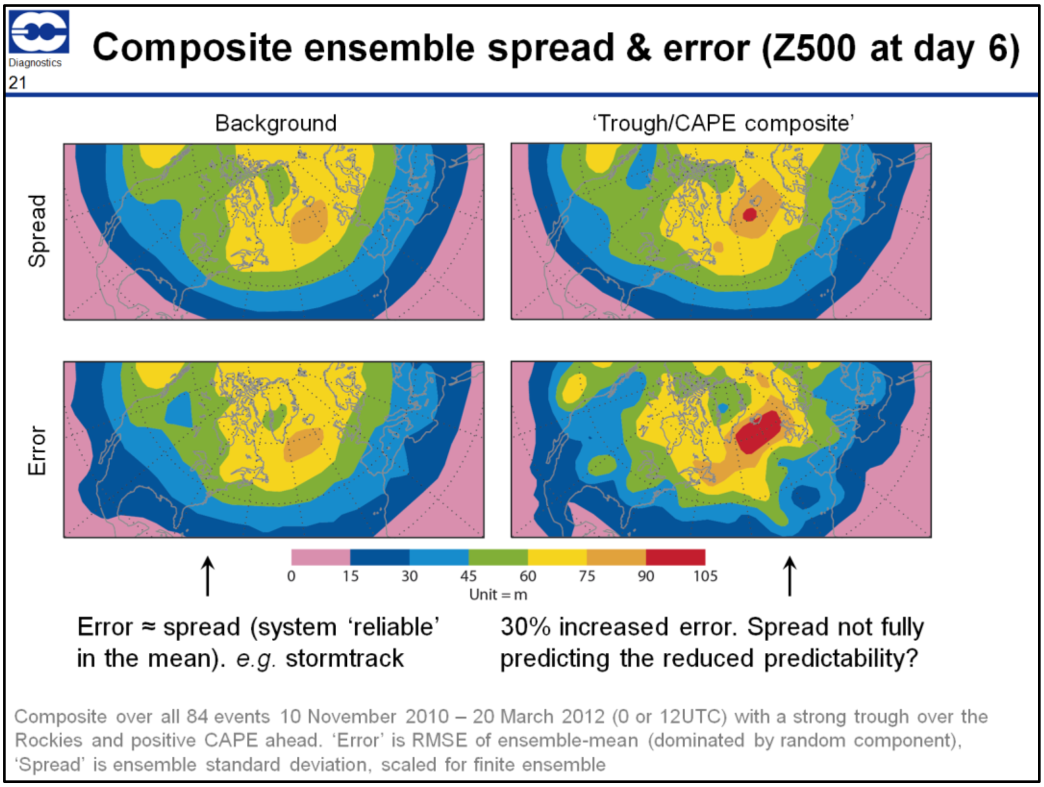
MCS magnifies spread in first-guess

New data reduces spread

The ensemble of first-guess forecasts develops spread over the first 12 hours associated with uncertainties in the prediction of a mesoscale convective system. The incorporation of new observations by the ensemble of data assimilations results in a contraction of the spread. Key question: Is the final analysis spread too large or too small to correctly reflect the predictability of the subsequent flow? Data: Temperature at 200 hPa from 10-member EDA, valid at 6UTC.

In addition to its ‘best estimate’ high-resolution analysis, ECMWF runs an Ensemble of Data Assimilations (EDA) which assimilates observations that are perturbed to reflect observation error. At the time this analysis was conducted the EDA involved 10 members. This slide shows (shaded) the EDA standard deviation in 200 hPa temperatures for the 10 April case. The first-guess forecasts (left) show strong growth of uncertainty associated with the development of the mesoscale convective system. When new observations are then assimilated (right), we see that the spread of the final set of analyses is reduced.

To first-order, the MSC is clearly acting to magnify uncertainty. Not only does the MCS disrupt the Jet Stream – and thus affect subsequent European weather - but there is also considerable uncertainty in how it disrupts the Jet Stream – and thus it enhances uncertainty in the subsequent European weather. A key question for Diagnostics is not about the 0th-order disruption, nor about the 1st-order enhancement of uncertainty, but rather the 2nd-order question of whether the uncertainty is enhanced by the correct amount.



Diagnostics
21

Composite ensemble spread & error (Z500 at day 6)

Error ≈ spread (system 'reliable' in the mean). *e.g.* stormtrack

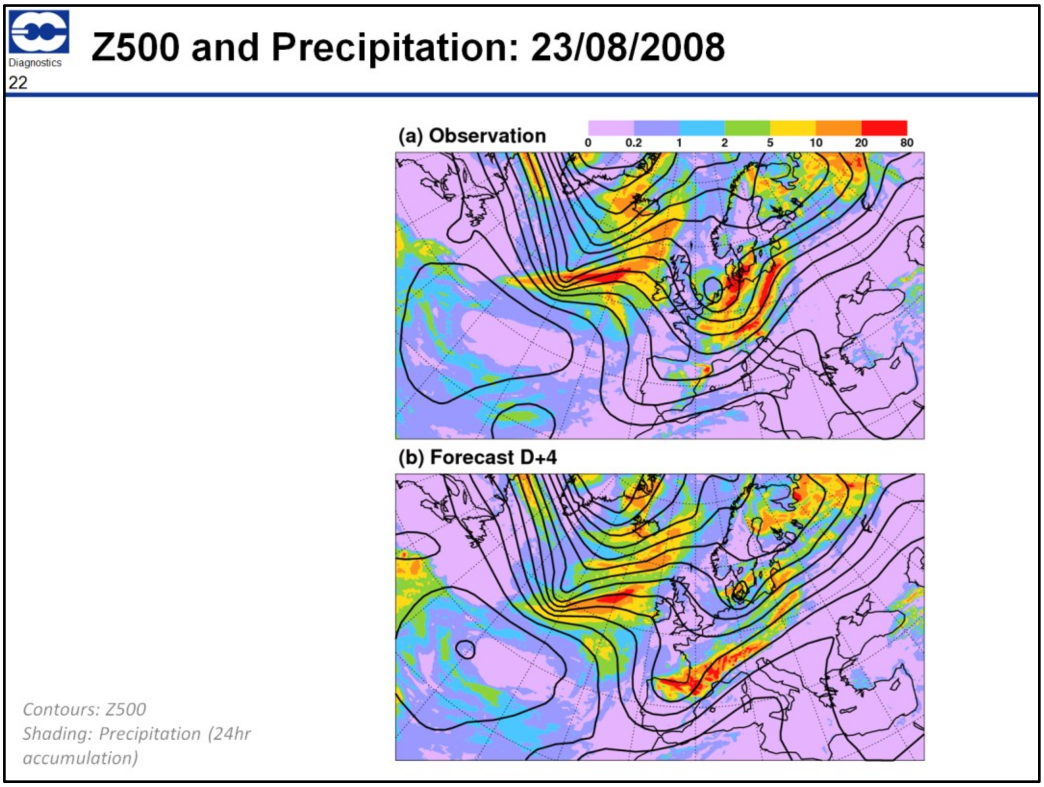
30% increased error. Spread not fully predicting the reduced predictability?

Composite over all 84 events 10 November 2010 – 20 March 2012 (0 or 12UTC) with a strong trough over the Rockies and positive CAPE ahead. 'Error' is RMSE of ensemble-mean (dominated by random component), 'Spread' is ensemble standard deviation, scaled for finite ensemble

This figure shows mean day-6 spread and ensemble-mean error based on 84 trough/CAPE events (right) and for the background (left). It can be seen that the trough/CAPE situation does indeed lead to increased spread and error. Experimentation has shown that both the trough and the CAPE are necessary aspects.

Note that 'spread' is actually the ensemble standard deviation - scaled slightly so that, for long-term averages, it would be equal to the root-mean-square-error of the ensemble-mean for a reliable system. Notice that the spread and error are in good agreement for the background. However, general agreement between mean spread and mean error is not sufficient for an ensemble forecast system to be useful. The system needs to be able to predict variations in error about this mean value. If anything, the spread for the trough/CAPE situation is too small compared to the error. However, the sample size of 84 may still be too small to be completely sure about this.

Given a sufficient sample size, we would need to know whether the first-guess forecasts (e.g. on the previous slide) realistically represent the uncertainty prior to the assimilation of new observations, and whether the assimilation correctly reduces this spread. Key aspects would include the estimates of observation and background error, and assumptions made within the tangent-linear version of the model used within the assimilation process (does linear physics produce too much spread in fast-maturing non-linear MCSs for example?). We would also need to consider how well we represent sub-grid-scale physics uncertainty (`stochastic physics') within the ensemble forecasts? This investigation has provided pointers to how we might assess flow-dependent uncertainty in the future, but Diagnostics has not got much further yet.



Diagnostics
22

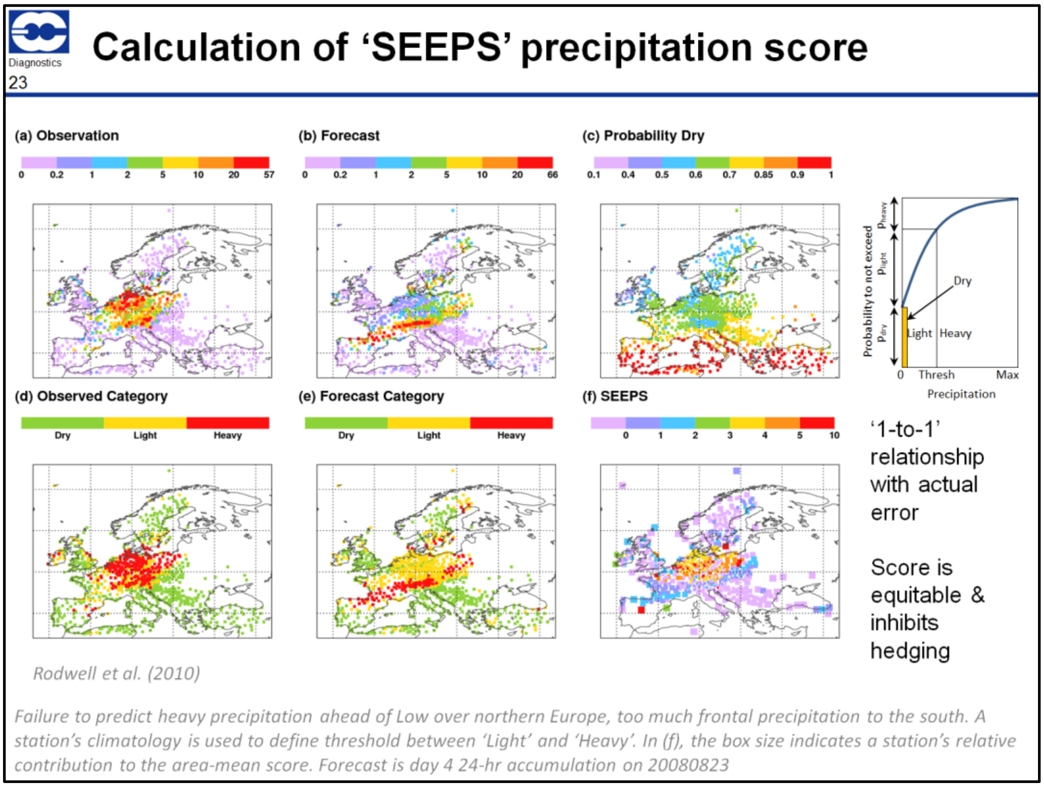
Z500 and Precipitation: 23/08/2008

(a) Observation

(b) Forecast D+4

Contours: Z500
Shading: Precipitation (24hr accumulation)

There is a need to focus more diagnostic attention on the prediction (and predictability) or weather features such as precipitation. Here we see a situation where the 500 hPa height field is reasonably well predicted at a lead-time of 4 days but the precipitation forecast is very poor over Europe. (The ‘observed’ precipitation here is actually a 1-day forecast). Weather parameters are much more problematic to verify owing to their small-scale variability and highly skewed distribution. Note, for example, the non-linear colour scale used here. However, such verification (and monitoring of trends in scores) is essential because it targets more directly the model physics (including cloud micro-physics) and should accelerate improvements in the prediction of the weather features that our users are interested in.



Diagnostics 23

Calculation of 'SEEPS' precipitation score

'1-to-1' relationship with actual error

Score is equitable & inhibits hedging

Rodwell et al. (2010)

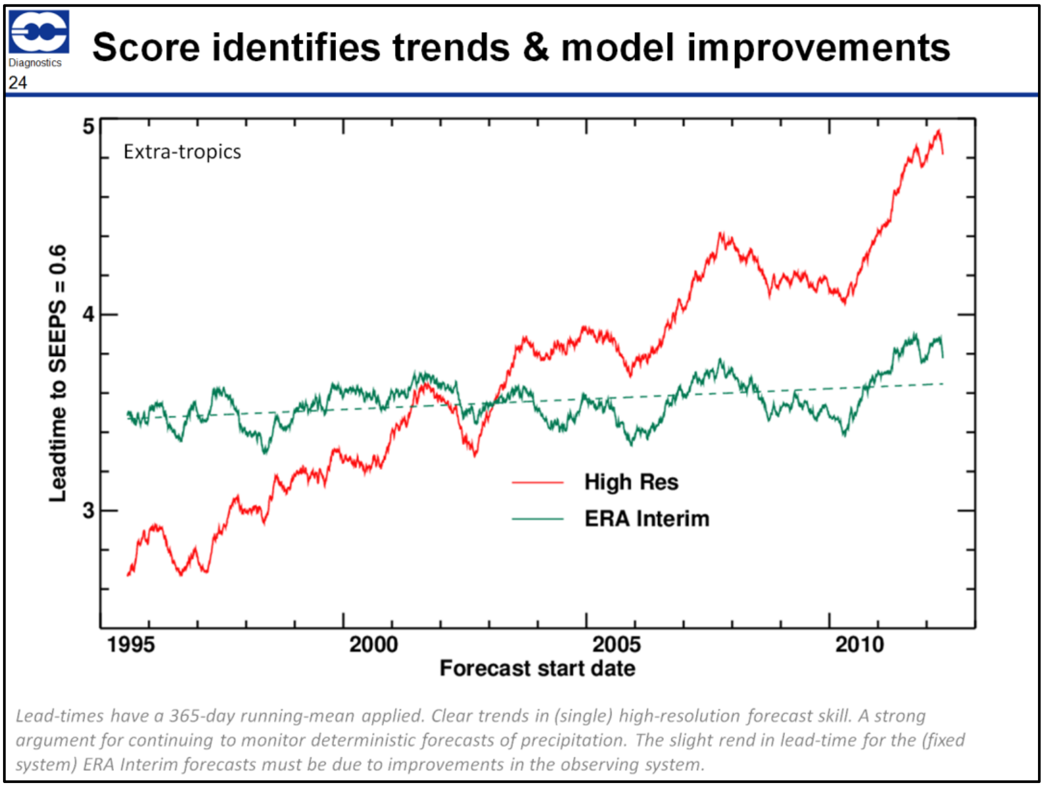
Failure to predict heavy precipitation ahead of Low over northern Europe, too much frontal precipitation to the south. A station's climatology is used to define threshold between 'Light' and 'Heavy'. In (f), the box size indicates a station's relative contribution to the area-mean score. Forecast is day 4 24-hr accumulation on 20080823

This slide demonstrates how the 3-category 'SEEPS' score can assess the forecast performance for precipitation for the same case as the previous slide. Panel (a) shows the observed 24hr-accumulated precipitation (this time it is station observations) and (b) shows the corresponding day-4 forecast.

Panel (c) is based on a long-term station climatology. It shows the climatological probability that a day in August will be dry (and demonstrates why northern Europeans head south for their summer vacation!). We use this climatology to define three precipitation categories. The schematic graph, which shows an typical cumulative distribution curve, demonstrates how this is done. The probability of dry weather p_{dry} is at the intercept with the y-axis (for compatibility with reporting practices, we actually take the intercept at a precipitation value of 0.2mm). We define 'light' and 'heavy' precipitation by dividing the remaining probability into two (we choose the ratio 2:1 so that light precipitation occurs twice as often as heavy precipitation: $p_{light} = 2p_{heavy}$). Drawing a horizontal line as shown (at $p_{dry} + p_{light}$) to the cumulative distribution and then down to the x-axis, we fine the precipitation threshold that divides light and heavy precipitation categories. This threshold is dependent on location and month of the year, and ensures that the 3-categories are meaningful whatever the local climate. Panel (d) and (e) show the observed and forecast categories.

Using a 3x3 scoring matrix, SEEPS then attaches an error to each (observation, forecast) category pair (panel f). The scoring matrix is designed so that SEEPS possess useful attributes such as being 'equitable'. The expected (*i.e.* long-term mean) error for a completely unskilful forecast system is 1, while the error is 0 for a perfect forecast system. SEEPS is essentially (1 minus) the average of two Peirce skill scores, one assessing the dry/light boundary and one assessing the light/heavy boundary. Notice that SEEPS penalises the failure to predict precipitation ahead of a low over northern Europe and the over-prediction of frontal convection to the south. Since 'heavy precipitation' is quite rare at this time of year (it occurs <10% of days, as deduced from panel c), the under-prediction to the north is penalised more heavily than the over-prediction to the south.

Note that the boxes in (f) are drawn with differing sizes. Larger squares reflect lower spatial density of observations. We use such an approach to ensure that area-averaged scores are not dominated by data-rich regions.



Diagnostics

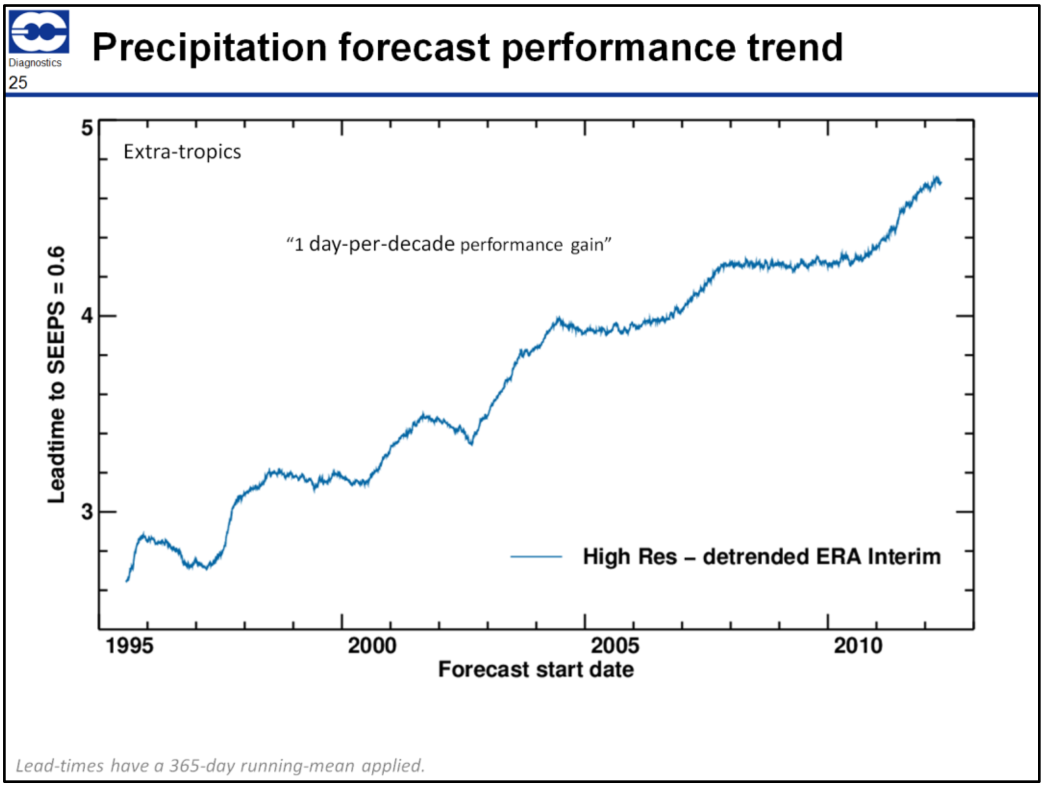
24

Score identifies trends & model improvements

Lead-times have a 365-day running-mean applied. Clear trends in (single) high-resolution forecast skill. A strong argument for continuing to monitor deterministic forecasts of precipitation. The slight rend in lead-time for the (fixed system) ERA Interim forecasts must be due to improvements in the observing system.

The attributes of the SEEPS score enable area-averages to be calculated (and meaningful). This graph shows the lead-time for SEEPS to reach 0.6 (the longer the lead-time, the better the forecast system) for the extra-tropics (North and South). The red curve shows the lead-time for ECMWF's operational 'high-resolution' forecast. The upward trend over the years indicates continual improvement. This is due to improvements in the observing system, the data assimilation system, the forecast model and increasing resolution over the years. Such improvement is a strong argument for the continued monitoring of 'deterministic' scores (even though we know that the prediction of precipitation is very much a probabilistic exercise). Note that there are also inter-annual variations along the red curve – what is the cause of these variations?

The green curve shows the lead-time for forecasts made with a fixed system (the forecast system used within the ECMWF re-analysis project, 'ERA Interim'). There is a slight upward trend here that must be due to changes in the observing system. However, the main feature of note is that interannual variations are very similar to those of the operational forecast. This suggests that a large part of the interannual variation of the operational scores is simply 'regime-related' (precipitation is easier to predict in some years than others due to changes in the frequencies of different weather regimes *etc*.). The indication is that it might be worth using the ERA Interim scores to remove this regime-related variability (see next slide).

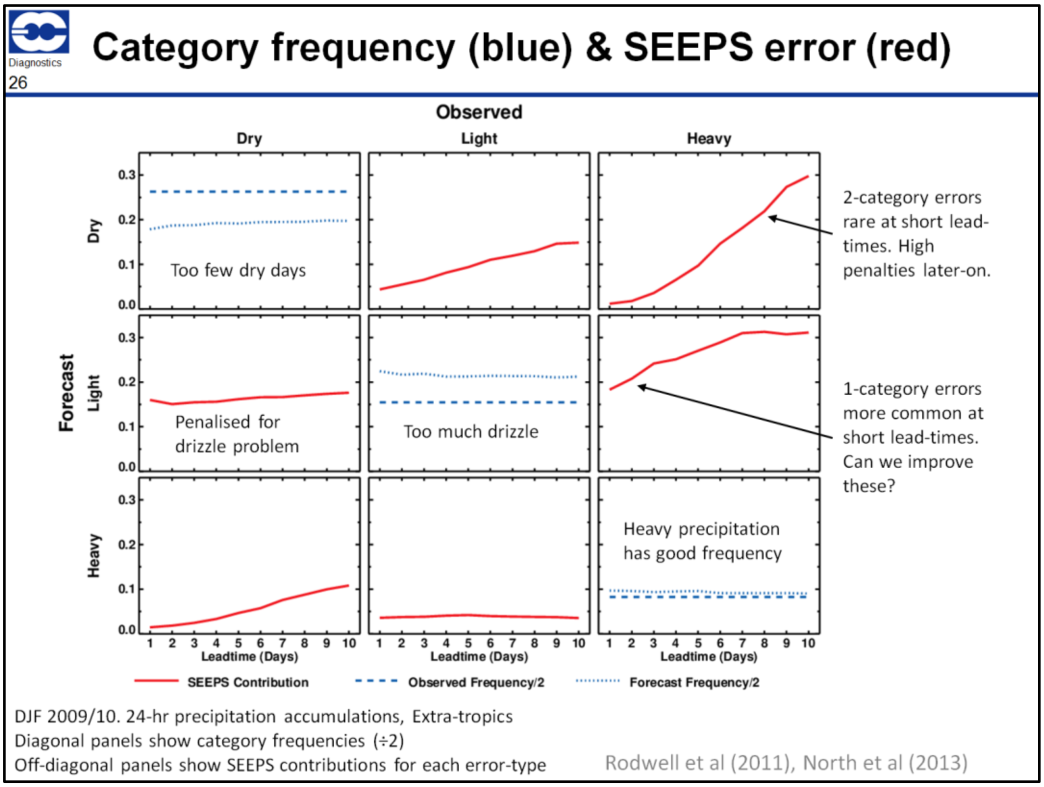


Diagnostics
25

Precipitation forecast performance trend

Lead-times have a 365-day running-mean applied.

When we subtract the detrended ERA-Interim lead-times from the operational high-resolution forecast lead-times (to remove the regime-related variability, but preserve the trend associated with the observing system), we obtain a clearer quantification of forecast system improvements. The trend is about 1 day lead-time gain per decade (this is actually the target set by ECMWF's governing body).



Category frequency (blue) & SEEPS error (red)

Too few dry days

Penalised for drizzle problem

Too much drizzle

Heavy precipitation has good frequency

2-category errors rare at short lead-times. High penalties later-on.

1-category errors more common at short lead-times. Can we improve these?

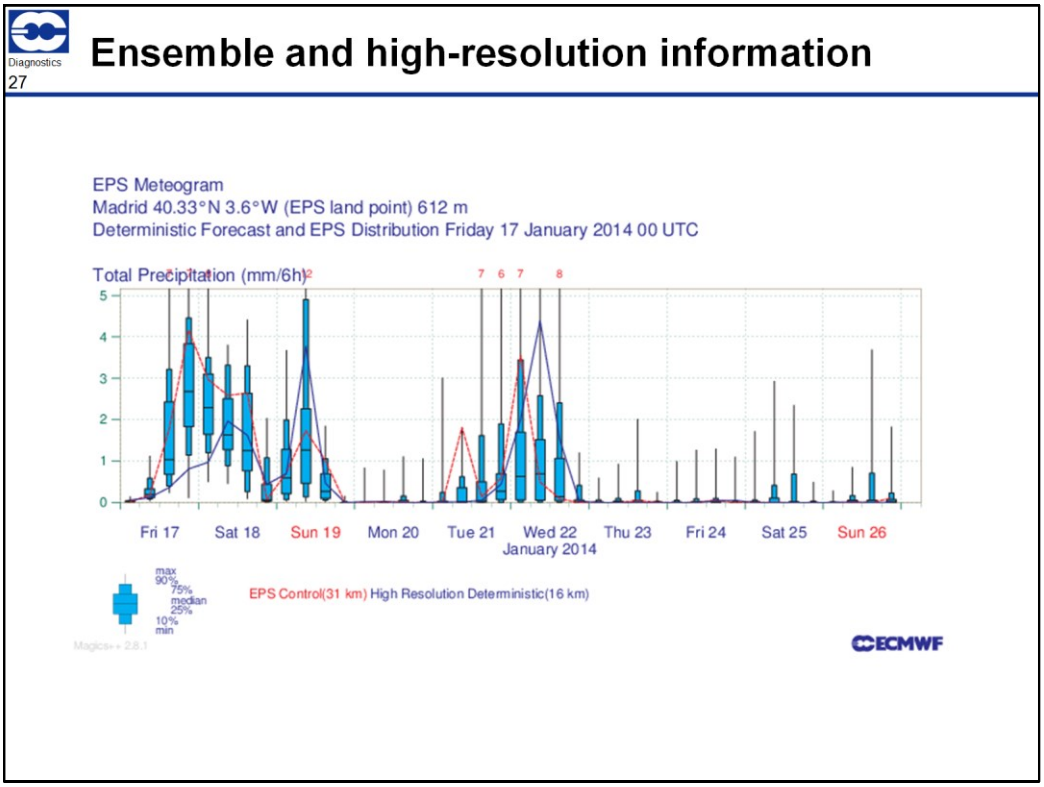
DJF 2009/10. 24-hr precipitation accumulations, Extra-tropics
Diagonal panels show category frequencies (÷2)
Off-diagonal panels show SEEPS contributions for each error-type

Rodwell et al (2011), North et al (2013)

The main plot here shows two aspects of the forecast system based on ~1000 European stations during December-February 2009/10. The red curves show the contributions to SEEPS from particular categorical errors (when predicting 24hr accumulations). For example, the top, middle graph shows the contribution to the European-averaged SEEPS from occasions where dry weather was forecast but light precipitation was observed. The graph shows this contribution as a function of forecast lead-time.

On the diagonal (where there is zero contribution to SEEPS because the forecast category is correct), the blue curves show the frequencies of each category for the observations (dashed) and high-resolution deterministic forecast (dotted). As discussed above, the observed frequency of light precipitation is twice that of heavy predication (to within variations due to the fact that the climatology was deduced from independent data). These frequency graphs show a clear under-prediction of dry weather and an over-prediction of light precipitation. The frequency of prediction of heavy precipitation appears to be remarkably good.

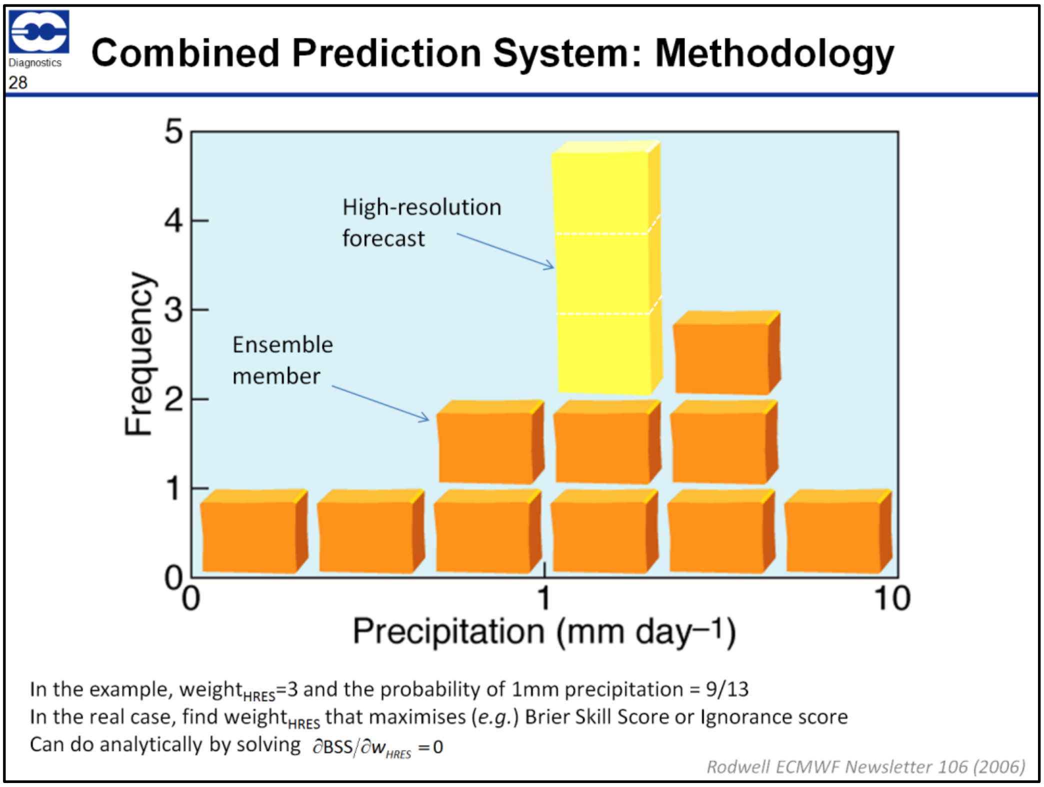
The contributions to SEEPS from particular categorical errors highlight the issue of over-prediction of ‘drizzle’ from the very start of the forecast. At longer lead-times, failures to predict heavy precipitation become more important. A paper by North et al. (2013) discusses the verification of 6-hr accumulations, and demonstrates that the drizzle issue is strongest at mid-day. The difficulty to predict heavy precipitation is worst for summertime evening convection. This decomposition of the SEEPS scores should help in the prioritisation of aspects to improve within the forecast model.



Diagnostics
27

Ensemble and high-resolution information

As has been noted, precipitation forecasting is perhaps best done in a probabilistic way (although not all users would agree with this!). Here we see a precipitation ‘Meteogram’ for Madrid. The blue boxes highlight the distribution of forecast precipitation from ECMWF’s 50-member ensemble. Also plotted is the single high-resolution forecast (dark blue curve) and the single control forecast (consistent with the resolution of the ensemble, show by the red dashed curve). All three forecast ‘products’ are predicting precipitation in the first three days, but notice that the high-resolution forecast predicts quite a bit less than most of the ensemble members. Sometimes, the high – resolution forecast can be completely outside the range of the ensemble. This should, of course, happen once in about 25 forecasts on average for statistical reasons. However, perhaps the high-resolution forecast is providing additional information to the forecaster. Perhaps, for example, it is more accurate (at short lead-times) than a single ensemble member because it is run at higher resolution, and started from unperturbed initial conditions. One simple approach to estimating the addition information provided by the high-resolution system, and thus enabling the forecaster to make sense of the full set of forecasts, is discussed in the next slide.



Diagnostics
28

Combined Prediction System: Methodology

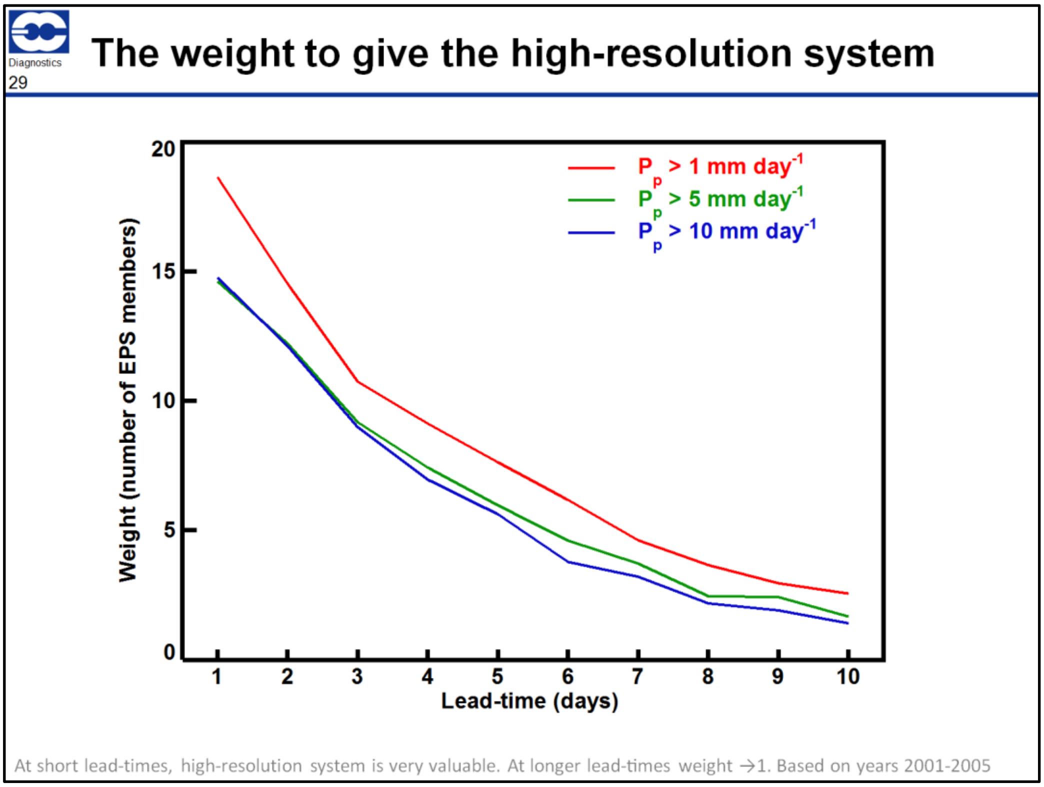
In the example, weight$_{HRES}$=3 and the probability of 1mm precipitation = 9/13
In the real case, find weight$_{HRES}$ that maximises (*e.g.*) Brier Skill Score or Ignorance score
Can do analytically by solving $\partial BSS/\partial w_{HRES} = 0$

Rodwell ECMWF Newsletter 106 (2006)

If would be useful if we could quantify the information that he ensemble and high-resolution systems bring to the forecaster on the bench. One intuitive approach (Rodwell 2006: Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better, *ECMWF Newsletter*, **106**, 17-23) is discussed here.

The figure shows how we could combine the ensemble prediction system and the high-resolution deterministic forecast into a single probabilistic forecast. Each ensemble member (shown by the orange squares) is given a weight of 1. If the high-resolution forecast (shown by the yellow rectangle) brings more information to the system than any individual ensemble member (by virtue of its higher resolution or more likely initial conditions), then this should be rewarded with a weight greater than 1. The schematic shows an ensemble of size 10 and gives a weight to the high-resolution forecast equivalent to 3 ensemble members. Hence the probability for the event that precipitation is greater than 1 mm is 9/13.

The aim in reality is not to assume a weight of 3, but to calculate the optimal weight to give to the high-resolution forecast. The optimal weight could be dependent on the user's application, but here we keep to the physical problem by maximising the Brier Skill Score for a given precipitation threshold. This is done for each year in a cross-validated way to ensure that we do not artificially inflate the score. The verification data is European gauge observations.



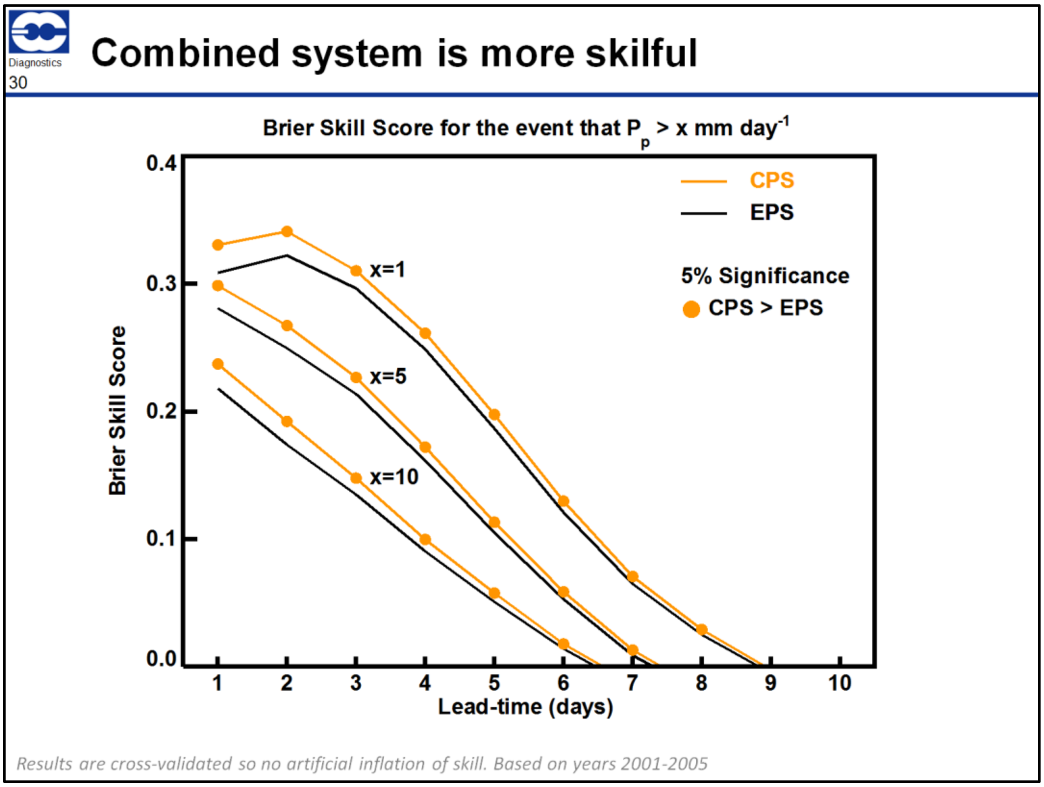
Diagnostics
29

The weight to give the high-resolution system

At short lead-times, high-resolution system is very valuable. At longer lead-times weight →1. Based on years 2001-2005

Here we see the mean optimal weights that should be applied to the deterministic forecast. The weights are shown as a function of lead-time and for three precipitation thresholds. In general, the weight is rather insensitive to the threshold. At a lead-time of 1 day, the deterministic forecast provides the same information as about 15-18 ensemble members. As the lead-time increases, the optimal weight decreases so that by day-10, the deterministic forecast is equivalent to perhaps 2 ensemble members.

This diagnostic is useful to the forecaster on the bench (and we can produce combined probabilistic products), but it could also be useful to help us decide how much computer resources to give to each forecast system. At the time of this work, the high resolution system was run at T511 while the ensemble was run at T255. This meant that the high resolution forecast required about 18 times as much computing power as one ensemble member. Hence, we can conclude that the deterministic system is valuable early in the forecast, but other strategies may be more optimal. For example making only short forecasts with the high resolution system, and devoting the saved computing resources to the ensemble may be one option. Of course, this diagnostic is only one source of information - there are many other (sometimes competing) considerations for how we configure our forecast system.



Diagnostics

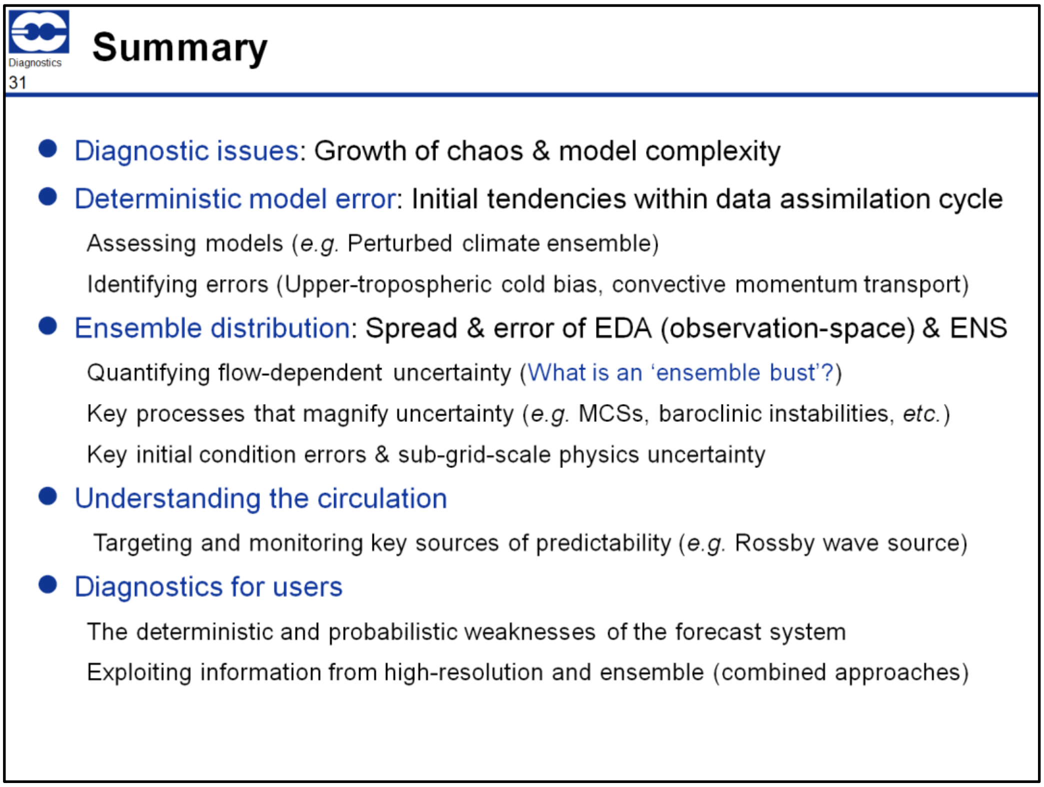
30

Combined system is more skilful

Brier Skill Score for the event that $P_p > x$ mm day^{-1}

Results are cross-validated so no artificial inflation of skill. Based on years 2001-2005

Finally, here is the comparison of the scores obtained by combining the high-resolution deterministic forecast with the ensemble. Again, results are cross-validated to ensure the combined scores are not artificially inflated. For each precipitation threshold (x), the Brier Skill Score is improved and this improvement is statistically significantly at the 5% level, as indicated by the orange circles.



Diagnostics
31

Summary

- **Diagnostic issues**: Growth of chaos & model complexity
- **Deterministic model error**: Initial tendencies within data assimilation cycle

  Assessing models (*e.g.* Perturbed climate ensemble)

  Identifying errors (Upper-tropospheric cold bias, convective momentum transport)
- **Ensemble distribution**: Spread & error of EDA (observation-space) & ENS

  Quantifying flow-dependent uncertainty (What is an 'ensemble bust'?)

  Key processes that magnify uncertainty (*e.g.* MCSs, baroclinic instabilities, *etc.*)

  Key initial condition errors & sub-grid-scale physics uncertainty
- **Understanding the circulation**

  Targeting and monitoring key sources of predictability (*e.g.* Rossby wave source)
- **Diagnostics for users**

  The deterministic and probabilistic weaknesses of the forecast system

  Exploiting information from high-resolution and ensemble (combined approaches)

I have discussed two fundamental issues for Diagnostics: the growth of chaos, and the increasing complexity of models. One approach to over-coming these issues is to examine the forecast very early on - during the data assimilation system in fact. By doing so, we minimise the time over-which chaos has had the opportunity to rear its head, and over-which model processes have had a chance to interact with each other. We also synergistically harness the great amount of work done by the data community to quality-control, estimate errors in, and 'forward-model', of as many observations as possible. (A forward model 'interpolates' the model fields to the observation so that a comparison can be made). I mentioned the assessment of the perturbations to a climate model and also the identification of root-causes for errors such as the upper-tropospheric cold bias and convective/dynamical coupling in the MJO.

However, we do not simply wish to over-come the issue of chaos. There is a new world (diagnostically at least) of uncertainty which we must embrace. The forecast bust example highlights our continued reliance on the deterministic forecast system, when we should really be thinking probabilistically. Do we predict the correct amount of spread for this and other initial flow regimes? What is the sampling uncertainty in the predicted spread and eventual error? What are the key process that magnify uncertainty and do we adequately represent these (MCS for example)?

An understanding of the circulation is also important if we are to be able to target diagnostics and monitoring at key sources of predictability. What are the most important aspects to focus on in order to make the biggest gains in performance? I mentioned the Rossby wave source as one key diagnostic. It is a means by which the memory within the tropical oceans can be harnessed to (potentially) improve extratropical forecast performance.

Users require diagnostics as well as scores. For example, they would like access to what we know about the key deterministic and probabilistic weaknesses of our system. The combined prediction system highlights a way for the user to optimally exploit the information from the high-resolution and ensemble systems. It can also provide information on optimal forecast system configuration.