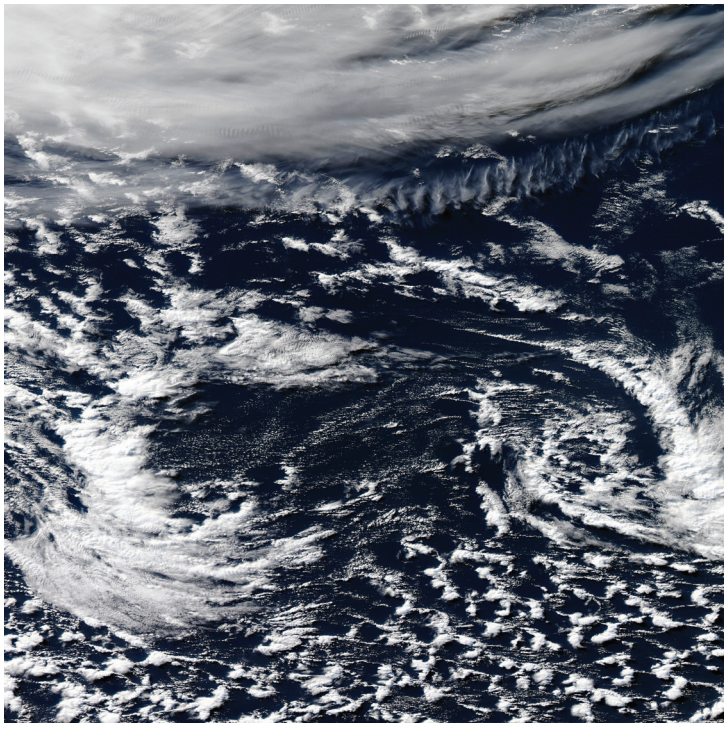
ECMWF Feature article

from Newsletter Number 146 – Winter 2015/16

METEOROLOGY

Using ensemble data assimilation to diagnose flow-dependent forecast reliability



NASA Worldview

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

doi:10.21957/xaxfw4c6

This article appeared in the Meteorology section of ECMWF Newsletter No. 146 – Winter 2015/16, pp. 29-34.

Using ensemble data assimilation to diagnose flow-dependent forecast reliability

Mark J Rodwell

Weather forecasting is fundamentally a probabilistic task due to the growth of unavoidable initial-state uncertainty. Moreover, the growth rates of these uncertainties can depend on the atmospheric flow so that predictability may vary from day to day. The established approach to representing uncertainty in probabilistic forecasting is to make an ensemble of forecasts, each starting from a slightly different initial state and including a different realisation of model uncertainty. A key question is how to assess the ensemble's ability to represent the flow-dependent growth of uncertainty.

Results suggest that such assessments are not easy to make at the medium range due to complications associated with error propagation and non-linear interactions. Using a specially developed ensemble reliability budget, appropriate for shorter-range assessments within the data assimilation window, these issues can be minimised and flow-dependent deficiencies in representing uncertainty can be identified. An analysis of the reliability budget can also help identify the causes of deficiencies in representing uncertainty. Results are illustrated for a flow situation where mesoscale convection is likely to occur over North America and which often results in reduced predictive skill for Europe several days later.

Forecast reliability

Figure 1 is a schematic representation of an ensemble forecast with ensemble members (blue curves) diverging in their prediction of two weather parameters (represented by the x- and y-axes) with increasing lead time t (coming out of the plane). Note that if we could produce an infinite number of ensemble members, they would describe the probability distributions depicted by the blue ellipses. A key question for numerical weather prediction is what constitutes a good ensemble forecast. Clearly users would like the ensemble distribution to be as narrow (or 'sharp') as possible in order to reduce uncertainty. However, this only makes sense if the eventual truth (black curve) lies within the ensemble distribution. More precisely, we require that the truth can be considered as another sampling of the ensemble distribution, and when this is true in general, the ensemble is said to be 'reliable'.

If we assume that forecast bias is negligible or accounted for, the next aspect of the ensemble distribution to assess in terms of reliability is its variance. The standard approach is to compare the mean ensemble variance (averaged over a set of forecast start dates) with the mean squared error of the ensemble mean (Error^2):

$$\text{Error}^2 = \text{EnsVar} + \text{Residual} \quad (1)$$

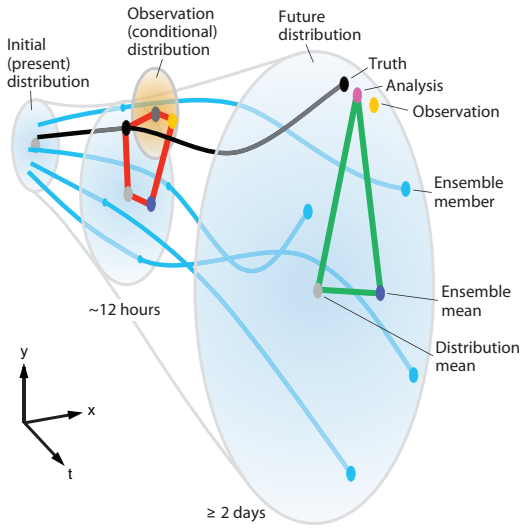


Figure 1 Schematic diagram depicting an ensemble forecast. The green and red polygons form the basis of the spread-error and 'EDA reliability budget' diagnoses of reliability, respectively. These are explained in Boxes A and B.

doi:10.21957/xaxfw4c6

This equation is valid at reasonably long lead times (perhaps greater than 2 days, as depicted by the right-hand distribution in Figure 1) when forecast error is sufficiently large for analysis error to be neglected in the calculation of Error^2. EnsVar in (1) is the mean sample variance of the ensemble (scaled to take account of the finite size of the ensemble), and the Residual represents any ensemble variance deficit associated with deficiencies in reliability, together with sampling uncertainties (due to the limited number of forecast start dates available). For a reliable system (and assuming no analysis error) the expected value of the residual is zero. The green triangle in Figure 1 depicts the theoretical basis for this 'spread-error relationship' and, for the interested reader, its derivation is discussed further in Box A.

Derivation of the spread-error relationship **A**

Figure 1 is a schematic representation of an ensemble forecast. The blue ellipses represent the initial and forecast distributions that one might obtain with an infinite ensemble size. For a finite ensemble size, m, the ensemble mean (dark blue dot) will not generally lie at the distribution mean (grey dot). For lead times ≥ 2 days, the analysis (pink dot) is often considered an adequate approximation for the truth (black dot).

The green triangle then shows how the (squared) error of the ensemble mean (right-hand side of the triangle) can be decomposed into the sum of the independent (squared) deviations of the truth and the ensemble mean from the distribution mean (the other two sides of the triangle). Assuming zero bias, the expected squared truth deviation can be written as the variance of the ensemble distribution plus a residual that indicates any systematic deficiency in this variance.

Because the ensemble members are independent, the expected squared ensemble-mean deviation can be written as $1/m$ times the ensemble distribution variance. The inclusion of this ensemble-mean variance, together with the desire for unbiased estimation of these terms using the available data, leads to the EnsVar term including an $\frac{m+1}{m-1}$ scaling factor in (1).

Using data from the ECMWF operational ensemble (ENS), Figure 2a shows northern hemisphere annual means of ensemble 'spread' and root mean square error (RMSE) of the ensemble mean (i.e. the square roots of EnsVar and Error^2) for 500 hPa height for the years 1996, 2005 and 2014. The reduction in RMSE at all lead times together with the better match with spread indicates substantial improvements in both sharpness and reliability over the years. Notice in particular the more realistic 'exponential' shape of the spread and error curves for 2014 over the first 5 or 6 days, which are much flatter at short ranges. These improvements have been achieved through many incremental changes to the forecasting system. These include the introduction of the Ensemble of Data Assimilations (EDA) and the development of the 'stochastic physics' parametrization, which represents, amongst other things, the upscale cascade of uncertainty from subgrid scales Improvements in the observation network and in the modelling of observation errors have also been important.

Figure 2b shows time series of the spread and RMSE for six-day forecasts for Europe from five of the world's leading operational forecasting centres. There cannot be such a good match between spread and RMSE on a day-to-day basis. Notice, however, the agreement between centres in terms of the variation in spread. This agreement suggests flow-dependent fluctuations in underlying predictability. The main reason for making ensemble forecasts is to be able to represent these flow-dependent variations in uncertainty and predictability. For a fully reliable ensemble forecast system, the ensuing flow-dependent probabilities for a given event will match the outcome frequencies when binned and averaged over a sufficiently large sample – as displayed in 'reliability diagrams'. Such correspondence is important to users because it allows them to make optimal decisions based on their own cost/loss models. Future steps towards such a reliable ensemble system are likely to come from more detailed flow-dependent diagnosis of model and observation error.

doi: 10.21957/xaxfw4c6

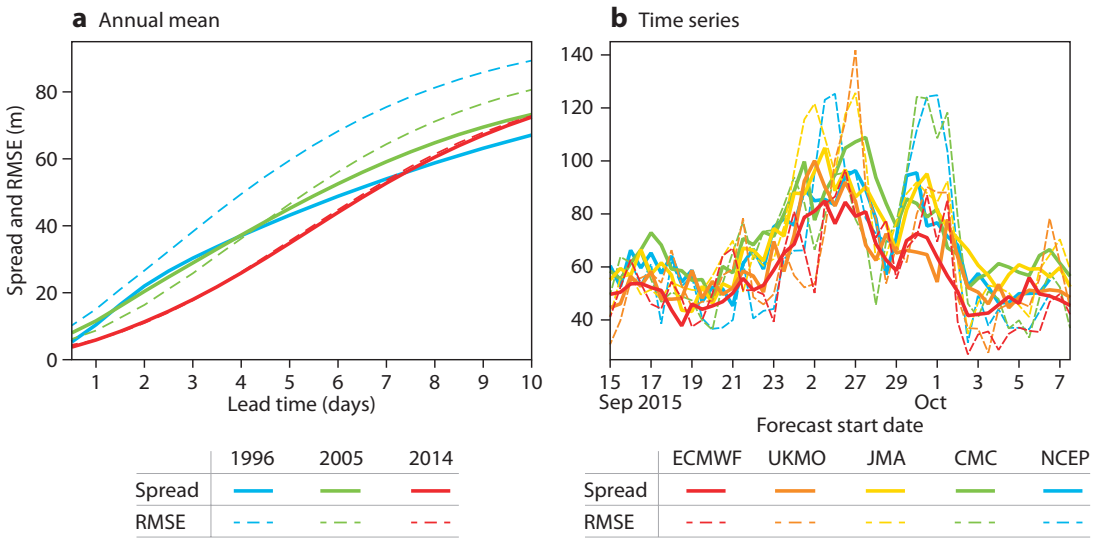


Figure 2 (a) Northern hemisphere annual means of ensemble spread and root mean square error (RMSE) of the ensemble mean for 500 hPa heights, and (b) time series of D+6 spread and RMSE for Europe from five of the world's leading operational forecasting centres: ECMWF; UKMO (UK Met Office); JMA (Japan Meteorological Agency); CMC (Canadian Meteorological Centre); NCEP (US National Centers for Environmental Prediction). Data retrieved from the TIGGE archive and interpolated to a 2.5° regular grid, with RMSE calculated relative to each centre's own analysis. The northern hemisphere includes latitudes north of 30°N, and Europe is defined here as 12.5°W–42.5°E, 35°N–75°N.

Difficulties in diagnosing flow-dependent reliability

Rodwell et al. (2012) highlighted reduced predictive skill for six-day forecasts for Europe in initial flow situations where a trough exists over the North American Rockies (e.g. as part of a Rossby wave), together with high Convective Available Potential Energy (CAPE) ahead of it. Such a situation is conducive to the formation of strong mesoscale convective systems (MCSs), which can interact with the jet stream. The uncertainty in intensity, location and timing of these MCS features (and possibly larger-scale dynamical instabilities coupled to orography) are thought to be important for the subsequent decrease in downstream skill.

The reliability of ensemble forecasts initiated from a particular flow regime, such as the trough/CAPE pattern, can also be assessed using (1). Figure 3 shows, for geopotential at 200 hPa (the jet stream level), the terms in (1) for a composite of the 54 cases during the period 19 November 2013 – 12 May 2015 (when cycle 40r1 of ECMWF's Integrated Forecasting System was operational) for which the initial conditions closely matched the trough/CAPE pattern, using the same method as in *Rodwell et al*. (2012). Large mean squared errors (Error^2) at forecast day 1 (D+1) can be seen over the Great Lakes region of North America (Figure 3a), associated with MCS activity. These errors are reasonably well predicted, on average, by the ensemble variance (EnsVar, Figure 3b). As the lead time increases, this error/spread signal is seen to propagate east across the North Atlantic (D+3 in Figure 3d,e and D+5 in Figure 3g,h). Notice in particular the large Error^2 at D+5 close to Western Europe (Figure 3g). This error is more than 15% greater than for the 'non-trough/CAPE' composite (consisting of the remaining about 1,000 forecasts, not shown), but the variance is actually slightly smaller than in the non-trough/CAPE composite (not shown). The right-hand panels in Figure 3 show the Residual (Error^2 minus EnsVar). The positive residual near Western Europe at D+5 (Figure 3i) suggests that the ensemble variance might be too small, but this is not statistically significant at the 5% level (saturated colours indicate statistical significance). The statistically significant negative residuals seen elsewhere in Figure 3 are generally not specific to the trough/CAPE situation.

The conclusion here is that there is broad agreement over the medium range between spread and error in this flow-specific situation, but it is difficult to identify the causes of residual differences. The crossing of the blue ensemble trajectories in Figure 1 represents the fact that errors are growing within a non-linear regime, interacting and dispersing through the action of teleconnections and waves in general. It is these effects that make it inherently difficult to assess flow-dependent ensemble reliability in the medium range, and even harder to identify the causes of any lack of reliability. By the same argument, it is also these effects that make it difficult to use off-line calibration techniques to improve flow-dependent reliability. Such improvements need to be made within the model itself and require more precise diagnostic tools.

doi:10.21957/xaxfw4c6

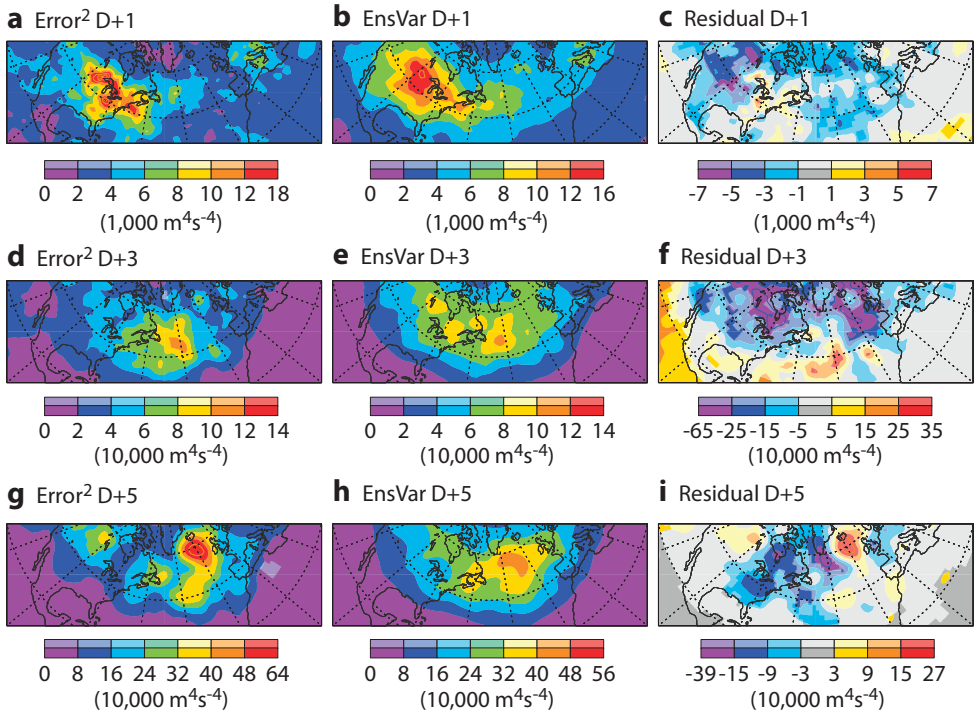


Figure 3 Terms in the spread-error relationship (1) for 200 hPa geopotential based on the 'trough/CAPE' composite for (a)–(c) D+1; (d)–(f) D+3; and (g)–(i) D+5. Statistical significance at the 5% level (from a null hypothesis of zero) is indicated by saturated colours and is deduced using a Student's t-test applied to the set of 54 forecasts.

The EDA reliability budget

In order to facilitate a more local and flow-dependent assessment of reliability, it proves useful to consider much shorter lead times (about 12 hours, as depicted by the central distribution in Figure 1), when errors are growing within a more linear regime and have not had so much time to interact or to disperse geographically. In anticipation of the development of a more seamless EDA/ENS system in the future, it makes sense to apply this test to the background forecasts of the EDA. These background forecasts include the most relevant part of the stochastic physics parametrization – the stochastic perturbation of physical tendencies (SPPT) scheme – and avoid potential complications associated with the re-centring and singular vector perturbations presently used in the initialisation of the ENS. At these short lead times, uncertainty in our knowledge of the truth cannot be neglected. We could incorporate an EDA analysis variance term into our reliability test to take account of this aspect, but we choose to work in observation space since the modelling of observation error (depicted by the orange ellipse in Figure 1) represents a more foundational aspect of the data assimilation process. Because observation errors cannot be neglected, we talk about 'departures' from observations rather than 'errors' from the truth. The new 'EDA reliability budget', the basis of which is depicted graphically by the red pentagon in Figure 1 and which is discussed further in Box B, is a decomposition of the mean squared departures of the form.

$$\text{Depar}^2 = \text{Bias}^2 + \text{EnsVar} + \text{ObsUnc}^2 + \text{Residual} \quad (2)$$

As before, EnsVar is the scaled sample variance, but now of the EDA background forecasts. ObsUnc^2 is the sample variance of the observation errors as modelled within the assimilation system. Bias^2 is the square of the estimated remaining bias of the model relative to the observations after the application of observation bias correction methods. Note that bias is sometimes neglected in the traditional spread-error assessment, but this risks the possibility that ensemble variance is erroneously inflated to achieve agreement with the squared error. The Residual quantifies the ensemble variance deficit (plus any deficiencies in the modelling of observation error). Non-zero values in either the Bias^2 term or the Residual are indicative of reliability deficiencies. Further details of the development of this budget are given in *Rodwell et al.* (2015).

To obtain a better understanding of the meaning of the terms in (2), and of the equation's utility in the assessment of flow-dependent ensemble reliability, it is useful to consider some examples. There are two key questions to answer: Is the EDA reliability budget able to identify statistically significant reliability deficiencies when it is analysed for a particular flow regime? If so, then what are these deficiencies?

Assessment of flow-dependent reliability

The EDA reliability budget will be computed for the trough/CAPE composite. First, however, it is useful to consider EDA reliability in a less flow-dependent context by computing the budget for the non-trough/CAPE composite. Figure 4 shows, for the non-trough/CAPE composite, the terms in (2) together with the observation density, for 200 hPa zonal wind speed, based on aircraft measurements and their corresponding values in the EDA background. Aircraft observations are numerous over central North America at this cruising altitude (Figure 4f) and, indeed, they are particularly influential in the data assimilation system. Observation uncertainty (Figure 4d) is computed from the independent observation errors assigned within the EDA. When averaged onto a 2°x2° grid, the observation uncertainty is naturally smallest where the observation density is largest. The EDA reliability budget decomposes the squared departure term (Depar^2; Figure 4a) into contributions from the bias (Bias^2; Figure 4b), ensemble variance (EnsVar; Figure 4c), observation uncertainty (ObsUnc^2; Figure 4d) and a Residual term (Figure 4e). The spatial structure of the Depar^2 term in the non-trough/CAPE composite largely follows that of the observation uncertainty. There is also a contribution from the ensemble variance, and possibly a more uniform offset associated with the bias. Figure 4e shows, however, that there is a small but statistically significant residual term, which suggests some general deficiency in reliability. This is not investigated further here because we are interested in flow-dependent reliability.

Derivation of the EDA reliability budget

B

At short lead times, uncertainty in our observational knowledge of the truth cannot be ignored. The red pentagon in Figure 1 shows how the spread-error relationship can be extended to include bias and random observation error. The departure of the ensemble mean from the observation is now written in terms of the remaining four sides of the pentagon. Assuming constant bias, these four terms are either independent or uncorrelated, so the expected square of the decomposition again simply involves the squares of the individual differences. Estimation with the available data then leads to the 'EDA reliability budget' (2).

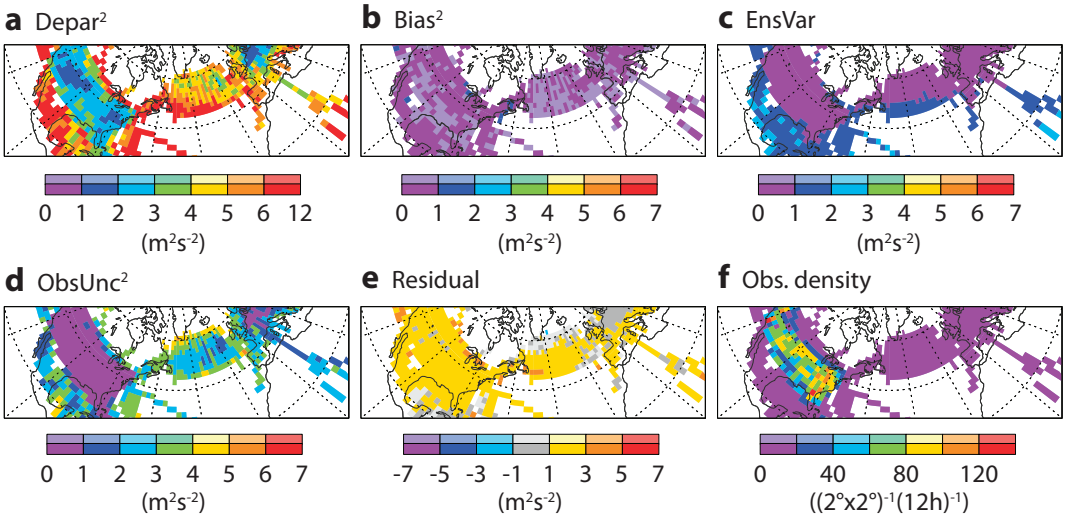


Figure 4 Panels (a) to (e) show the terms in the 'EDA reliability budget' (2) applied to 200 hPa zonal winds for the non-trough/CAPE composite. The data used are aircraft observations between 185 and 215 hPa that are actively assimilated, and the corresponding winds interpolated from the EDA background forecasts. Panel (f) shows the density of aircraft observations assimilated within the EDA control. To reduce noise, an average of at least one observation per 2°x2° grid box per 12-hour analysis cycle is required for the budget to be plotted. Values significantly different from zero at the 5% level are shown with saturated colours.

doi:10.21957/xaxfw4c6

Figure 5 shows the EDA reliability budget terms for the trough/CAPE composite. Comparison of Figure 5 with Figure 4 shows increased departures around the Great Lakes in the trough/CAPE composite. Larger departures are to be expected because of the strong (and less predictable) convection liable to be taking place. The increased ensemble variance indicates more forecast uncertainty in this region. Notice, however, that this increase does not fully account for the increased departures, and consequently the Residual term (note the different shading interval) increases markedly too in the region associated with MCS activity and has roughly twice the magnitude of the ensemble variance. One possibility is that the ObsUnc^2 term does not increase sufficiently in these convective situations. However, aircraft wind observations are thought to be quite accurate (they are assimilated without bias correction) and they are probably dense enough in this region (≥60 per 2°x2° grid box per 12 hours, and with little change in density) to rule out an increase in representativeness errors for the upper-tropospheric wind field. Hence it is likely that the background variance is strongly deficient in these trough/CAPE (and MCS) regimes. From a diagnostic development point of view, the key result here is that the EDA reliability budget is able to identify flow-dependent deficiencies in reliability. Next we consider what this budget can tell us about the root causes of reliability deficiencies.

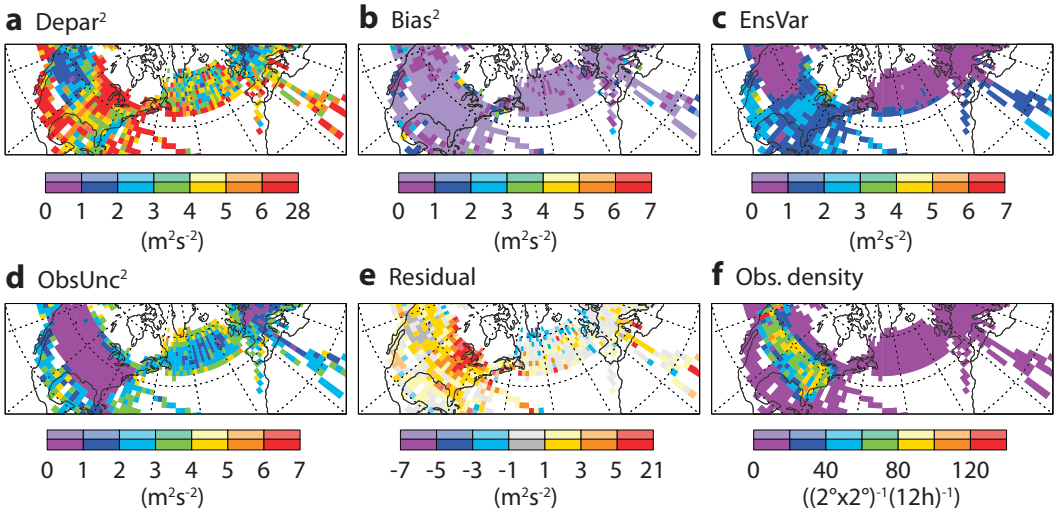


Figure 5 As Figure 4, but for the trough/CAPE composite.

Causes of flow-dependent reliability problems

The ensemble variance deficit following trough/CAPE situations could be due to deficiencies in stochastic physics – either in magnitude or formulation. Note, for example, that the 'Stochastic Kinetic Energy Backscatter' (SKEB) scheme is not included in the EDA background. The relatively low resolution of the background (T399, corresponding to a horizontal resolution of about 50 km) may also necessitate stronger stochastic physics in these mesoscale convection situations. Such a conclusion about deficiencies in stochastic physics is consistent with the results of *Rodwell et al.* (2015), who showed that the EDA reliability budget (applied to mid-tropospheric temperatures observed by AMSU-A satellite microwave channel 5) was able to highlight regional deficiencies in variance that were sensitive to changes in stochastic physics. For example, positive residuals in convective regions were deteriorated by turning off the stochastic physics. At the same time, negative residuals within subtropical anticyclone regions (where the mid-tropospheric meteorology is largely characterised by time-mean descent and clear-sky radiative cooling) were improved by turning off the stochastic physics.

There are, however, other possible causes for the residual seen in the MCS example. For example, 'analysis tendency and increment' diagnostics (*Rodwell & Palmer*, 2003) suggest that, for the trough/CAPE situation, the net physical heating within MCS events is placed too low in the atmosphere (not shown). One can hypothesise that this may weaken the interaction in the momentum field between the jet stream and t he MCS outflow, which could be another reason for the apparent lack of ensemble variance. Hence one can imagine situations where reliability deficiencies are associated with flow-dependent systematic error.

Although ruled out in the example shown, the EDA reliability budget is also sensitive to observation error assignment. This can be considered useful because a good modelling of observation error is also important for the reliable initialisation of ensemble forecasts. There can be situations where the residual term is most

clearly associated with the assignment of observation error variances, for example in regions where residuals of different signs are found for different observation types and are thus less likely to be associated with ensemble variance errors. While these observation error variances are already estimated by a variety of means, such as 'Desroziers statistics', *Rodwell et al.* (2015) demonstrated, for surface pressure observations over the oceans from ships and buoys, that the EDA reliability budget may also be useful in this regard. It is likely that there will be situations where deficiencies in both stochastic physics and the assignment of observation error are important. In such situations, additional information might be required to resolve the ambiguity.

In some situations, the bias term in (2) can also highlight errors in the ensemble distribution. For example, the EDA reliability budget for the AMSU-A satellite microwave channel 5 indicated significant biases off the west coast of South America, possibly associated with undetected shallow cloud in the observations. While (observation) bias is legitimately assumed to have been accounted for in the data assimilation process, it is clearly worth including this term in (2) from a diagnostic perspective.

Summary and outlook

A necessary condition for 'reliable' ensemble forecasts is that ensemble variance accurately represents the flow-dependent growth of initial and model uncertainties. A key question is how to diagnose deficiencies in this ensemble variance. For a particular initial flow known to be associated with reduced predictability, the traditional spread-error relationship was not able to identify the causes of medium-range variance deficiencies. This is partly because, by the medium range, errors are growing within a non-linear regime, interacting and dispersing. To minimise such issues, it is useful to look at the shorter timescales associated with the background forecasts of the Ensemble of Data Assimilations (EDA).

At these shorter lead times, as with data assimilation itself, uncertainties in our observational knowledge of the truth must be accounted for. This leads to the derivation of an 'EDA reliability budget' (*Rodwell et al.*, 2015) that decomposes mean squared departures (of the background ensemble mean relative to the observations) into squared-bias, ensemble-variance and observation-uncertainty terms, together with a residual that closes the budget. Results show that the residual (and bias) terms of the EDA reliability budget can identify regional and flow-dependent deficiencies in reliability. Although ambiguities may still arise, it is hoped that the EDA reliability budget will be a useful tool to diagnose our modelling of the non-linear fluid-dynamical/physical system (including, in future, the development of stochastically formulated physical parametrizations), our modelling of the observation operators (radiative transfer etc.), and our modelling of observation error.

For reduced forecast uncertainty, it is also important to increase the sharpness of probabilistic forecasts through reduced initial uncertainty. Subject to the ensemble being reliable, sharpness is chiefly addressed through improved observational information and data assimilation methods and can therefore be largely addressed separately from the modelling aspects associated with reliability.

ECMWF's proposed new strategy foresees the development of a more seamless EDA/ENS system, as depicted schematically in Figure 1. The use of a consistent model (same physics, stochastic physics and resolution) and consistent initialisation of the EDA background and ENS will mean that EDA reliability results like those presented here should facilitate improvements in medium-range reliability too. As lead time increases through the medium range and beyond, slower processes such as those associated with the land surface, ocean and sea ice also become important in the forecast. By examining reliability at all lead times, and representing uncertainties close to their sources, it may be possible to ensure that errors associated with these slower processes are also well represented.

 doi:10.21957/xaxfw4c6

Further reading

Desroziers, G., L. Berre, B. Chapnik & **P. Poli,** 2005: Diagnosis of observation background and analysis-error statistics in observation space. *Quart. J. R. Meteorol. Soc.*, **131,** 3385–3396, doi: 10.1256/qj.05.108.

Isaksen, L., J. Hasler, R. Buizza & **M. Leutbecher,** 2010: The new ensemble of data assimilations. *ECMWF Newsletter* **No. 123,** 17–21.

Leutbecher, M. & T.N. Palmer, 2008: Ensemble forecasting. *J. Comp. Phys.*, **227,** 3515–3539, doi: 10.1016/j.jcp.2007.02.014.

Rodwell, M.J., L. Magnusson, P. Bauer, P. Bechtold, C. Cardinali, M. Diamantakis, E. Källén, D. Klocke, P. Lopez, T. McNally, A. Persson, F. Prates & **N. Wedi,** 2012: Characteristics of occasional poor medium-range forecasts for Europe. *ECMWF Newsletter* **No. 131,** 11–15.

Rodwell, M.J., S.T.K. Lang, N.B. Ingleby, N. Bormann, E. Hólm, F. Rabier, D.S. Richardson & **M. Yamaguchi,** 2015: Reliability in ensemble data assimilation, *Quart. J. R. Meteorol. Soc.*, doi: 10.1002/qj.2663.

Rodwell, M.J. & **T.N. Palmer,** 2003: Using numerical weather prediction to assess climate models. *Quart. J. R. Meteorol. Soc.*, **133,** 129–146, doi: 10.1002/qj.23.

doi: 10.21957/xaxfw4c6