ECMWF Feature article

from Newsletter Number 123 – Spring 2010

METEOROLOGY

The new Ensemble of Data Assimilations

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

doi:10.21957/vzbtkbf8

This article appeared in the Meteorology section of ECMWF Newsletter No. 123 – Spring 2010, pp. 17–21.

The new Ensemble of Data Assimilations

Lars Isaksen, Jan Haseler, Roberto Buizza, Martin Leutbecher

An Ensemble of Data Assimilations (EDA) system will be introduced at ECMWF with cycle 36r2 of the Integrated Forecasting System (IFS). The EDA system consists of an ensemble of ten independent lower-resolution 4D-Var assimilations that differ by perturbing observations, sea-surface temperature fields and model physics. The computing cost is significant, similar to running the deterministic analysis suite.

The main justification for implementing the EDA is that it quantifies analysis uncertainty.

- It is the first system implemented at ECMWF that provides estimates of analysis uncertainty. A properly designed EDA will complement the data assimilation system with important information about the quality of the deterministic analysis.
- It can be used to estimate flow-dependent background errors in the deterministic 4D-Var assimilation system; this will potentially improve the medium-range forecast. Flow-dependent background errors will be introduced in the second phase, in the autumn of 2010.
- It can improve the representation of initial uncertainties for the Ensemble Prediction System (EPS). When the EDA is introduced in IFS cycle 36r2, EDA-based perturbations will replace evolved singular vectors (SVs) to generate the EPS initial conditions. This change will improve the EPS skill, especially over the tropics.

The application of EDA in EPS will be described in an accompanying article by *Buizza et al.* in this issue of the *ECMWF Newsletter* (pages 22 to 28).

The EDA is expected to become an important part of the ECMWF data assimilation system, with the introduction of a hybrid 4D-Var/EDA system. Also, in the coming years, the EPS and EDA are expected to be further integrated to the benefit of both systems.

Rationale for developing an EDA system

The EDA is based on a perturbed lower-resolution version of the operational analysis system. If the perturbations of observations and model physics are realistic, the EDA will provide good estimates of analysis uncertainty. Because four-dimensional data assimilation, similar to that used for the operational analysis, is an integral part of EDA it has the potential to provide very valuable information about analysis uncertainty of the operational assimilation system. This is difficult to obtain by other means. The system will also provide short-range forecast error uncertainty. Many applications would benefit from accurate estimates of the uncertainty in analysis and short-range forecast errors. This could provide guidance for the quality of ECMWF's short-range forecasts. The EDA can also be used to improve the data assimilation system and the EPS.

In data assimilation, one of the crucial aspects is the estimation of the background error variances. To a large degree these are static in the operational 4D-Var system. This is unrealistic, especially for extreme events, where the background error variances can be underestimated significantly.

The EDA system is able to produce flow-dependent estimates of analysis uncertainty and background error uncertainty based on the ensemble spread, measured as the standard deviation of the difference between independent short-range background forecasts (*Fisher*, 2003; *Tan et al.*, 2007). This information gives an estimate of the error-of-the-day and can also be used for the estimation of seasonally varying background errors. Recent research at ECMWF has shown a beneficial impact from using the EDA errors of the day in the operational high-resolution 4D-Var. This is expected to be implemented in the second half of 2010.

Currently, covariance statistics of background error are generated from an offline EDA that is run over a period of one month. Only rarely are the statistics updated, and a single set of statistics is used for all seasons. The availability of an operational EDA will allow more frequent updates of background error statistics, with the possibility of accounting for seasonal variation of error covariances.

 doi:10.21957/vzbtkbf8

Characteristics of the EDA system

4D-Var at ECMWF is based on the incremental approach to minimising a cost function. The first minimisation (inner loop) takes place at low resolution to produce preliminary low-resolution analysis increments with a simplified representation of linearized model physics. The subsequent loops are at higher resolution with a more advanced linearized model physics applied. The comparison of observations against model fields takes place at a high resolution with all the non-linear aspects included (outer loop). This incremental approach provides considerable flexibility in the use of computer resources.

Isaksen et al. (2007) describe the design of the EDA in detail, based on analyses run with a T255 outer loop and T95/T159 inner loops and 91 vertical levels. In the EDA, for each observation, perturbations are defined by randomly sampling a Gaussian distribution with zero mean and standard deviation equal to the estimate of the observation error standard deviation. For cloud-track wind observations, perturbations are horizontally correlated. Sea-surface temperature fields are also perturbed, with correlated patterns as currently used in the Seasonal Forecasting System. At the first assimilation cycle, the randomly-perturbed observations are the only source of difference between the perturbed analyses, while for the subsequent cycles differences will evolve in the first-guess fields and contribute to the analyses spread. Model error is simulated by stochastically perturbing the model tendencies using same method applied in the EPS – the method used is described later in the section on '*The operational EDA configuration*'.

Isaksen et al. (2007) demonstrated the EDA system's ability to produce flow-dependent spread and deliver promising results for some extreme meteorological events. The EDA produces a realistic horizontal distribution of analysis error and background error, with small values over the data rich areas of the USA, Europe and Australia.

Figure 1 shows a snapshot of background error estimates for 850 hPa zonal wind on a day in October 2006 for the northern hemisphere extra-tropical region from the operational randomization method (*Fisher & Courtier*, 1995) and the standard deviation of a ten member EDA. The EDA values have been scaled to get more realistic global average amplitude of variances. It is clear that the day-to-day background error estimates for wind have very different structures and amplitudes for the operational method (Figure 1a) and the EDA (Figure 1b). The operational method takes some account of flow curvature, but primarily samples the static background error variances, taking account of the observation coverage.

It is seen from Figure 1 that the EDA method results in more flow-dependent variability of the background error variances. The largest values are seen east of Japan where an extra-tropical low is developing. The EDA method really captures the dynamically active regions, like extra-tropical lows and troughs; an ability that to a large extent is lacking for the operational method.

The impact of EDA resolution and ensemble size has been investigated for the case used in Figure 1. Some results are now described that focus on the region near Japan that is dynamically very active. For the low-resolution assimilation system with one T95 inner loop (Figure 2c) there is a clear flow dependence, but this is not the case when the operational randomization method is used (Figure 2a).

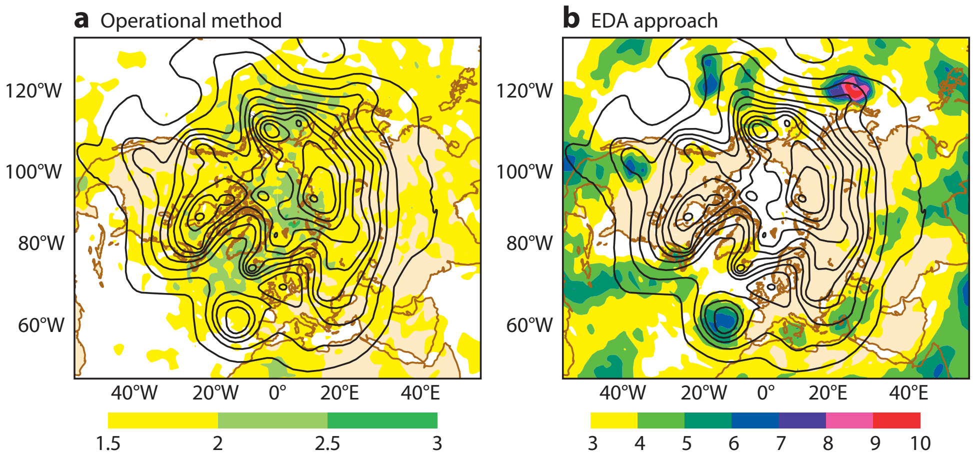

Figure 1 Background error standard deviation estimates for 850 hPa zonal wind at 00 UTC on 16 October 2006 using (a) the operational randomization method of *Fisher & Courtier*, 1995 (maximum value 2.97 ms^{-1}) and (b) the EDA approach based on 2 times the standard deviation of 10 ensemble members using T255 outer loop and T95/T159 inner loop (maximum value 9.99 ms^{-1}). Both panels have 500 hPa geopotential height field overlaid (8 dm contouring). The factor of two makes the average global EDA spread more realistic.

Also, the low-resolution assimilation system (Figure 2c) delivers less focussed, but still similar results compared to a higher-resolution system with two (T95 and T159) inner loops (Figure 2b). Both use a T255 outer-loop resolution.

Comparing Figures 2c and 2d, one can see the impact of increasing the number of ensemble members from 10 to 50. The patterns are very similar, but the result of the 50-member ensemble is smoother with fewer spikes and also reduces regions with very low variance. The 50 members are basically giving a statistically better sampling of the forecast errors, but they also describe flow-dependent features at a higher resolution. The results shown in Figure 2 suggest that computer resources may be better spent on more members with a simpler low-resolution version of the 4D-Var system. But more research investigating case studies of extreme events will have to be performed before the final conclusions can be drawn on this subject.

The capturing of flow dependence for all three ensemble-based versions shown in Figure 2 is clearly visible. The smoothing property of using 50 members is also marked. Indeed, the amplitudes and structures are surprisingly similar for the low- and higher-resolution analysis system. This may well be due to the fact that all systems for these experiments used a T255 outer-loop resolution. It is at this stage and resolution that the observations are perturbed. The uncertainty information is also propagated in time with the same T255 model resolution.

The EDA is most beneficial for extreme events. As an example, consider the Category-3 Hurricane Emily on 20 July 2005 just before it made landfall in Mexico. Figure 3a shows the precipitation measured by the local weather radar. The EDA spread (i.e., the standard deviation of the ten, in this case T399 members) for zonal wind at approximately 850 hPa is given in Figure 3b. Typical standard deviations in the region would be 2–3 ms^{-1}, but the flow-dependent estimates from the EDA system yield standard deviations up to 13 ms^{-1}. Note that the spread is concentrated in the vicinity of Hurricane Emily, identified by the mean sea level pressure contours.

The ability of the EDA system to identify regions of large background error associated with extreme events has the potential to significantly improve quality control decisions and give higher weight to observations used in the analysis from such regions.

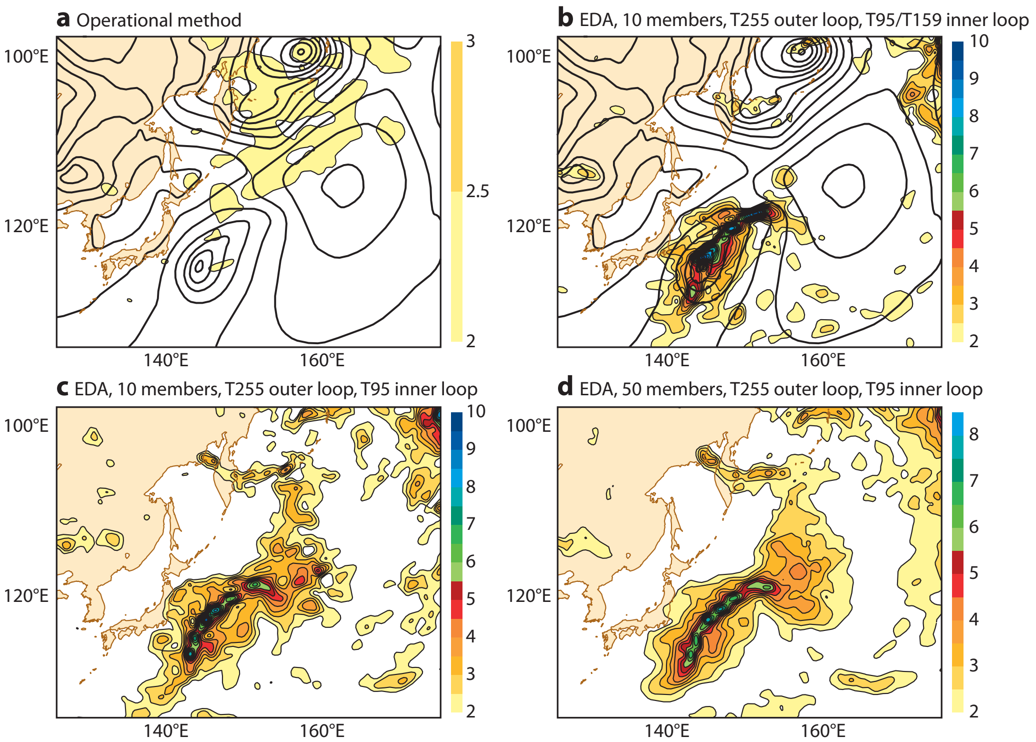

Figure 2 850 hPa zonal wind background error estimates valid at 00 UTC on 16 October 2006 for a baroclinic area near Japan. (a) The operational cycling randomization method (global max. 2.8 ms^{-1}). (b) The 10-member T255 outer loop with T95/T159 inner loop (global max. 13.8 ms^{-1}). (c) The 10-member T255 outer loop with T95 inner loop (global max. 13.9 ms^{-1}). (d) The 50-member T255 outer loop with T95 inner loop (global max. 12.0 ms^{-1}). Panels (b), (c) and (d) all use 2 times standard deviation of the ensemble members. The factor of two makes the average global EDA spread more realistic.

doi:10.21957/vzbtkbf8

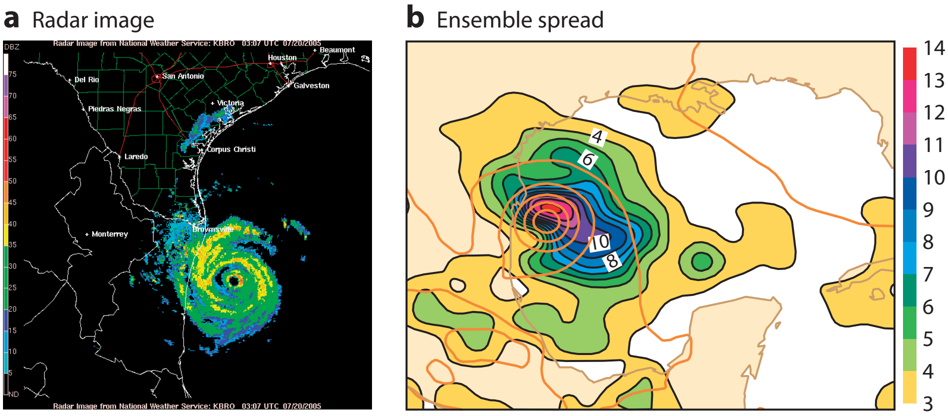

Figure 3 Atlantic hurricane Emily on 20 July 2005 near the coast of Mexico. (a) Radar image from National Weather Service. (b) Ensemble spread for zonal wind at 850 hPa near Hurricane Emily; the maximum value is 13.4 ms^{-1}. The mean sea level pressure contours are overlaid.

Deciding the operational configuration

It is important to note that the EDA system is designed to serve the needs of applications in the data assimilation system and the needs of the EPS. This has major implications for the design of the Ensemble of Data Assimilations that will become operational.

To be able to use the system to calculate background error statistics and estimate flow-dependent background errors, the EDA system must be designed to be sufficiently similar to the operational 4D-Var assimilation system. So it requires the same number of vertical levels (91 at present). The horizontal resolution cannot be significantly lower than the operational 4D-Var resolution in order to get a flow-dependent impact on extreme events and analysis uncertainty estimates that describe the operational 4D-Var system. Also, for reasons of consistency, a 4D-Var configuration in the EDA is preferable to a 3D-Var configuration.

When EDA products are used to calculate initial uncertainties for the EPS it will require vertical interpolation from 91 to 62 levels and horizontal interpolation from T399 to T639. It has been decided not to force the EPS and the EDA resolutions to be the same. This leaves more flexibility in the future system design of EPS and EDA. If the resolution of the EDA becomes an issue for the EPS the situation needs to be revisited. However, the computer codes have been designed to cater for any resolution change, as this may be required for research experiments or future operational configurations.

The performance and computing costs of a number of potentially suitable low-resolution configurations were investigated to choose an assimilation system that was close in skill to the operational system, but still significantly cheaper. The investigations described above, performed at T255, showed that a 10 member EDA system gave similar ensemble spread patterns to a 50-member EDA system, though with more noise. Because noise-filtering methods are available it was decided to choose 10 members for the initial operational implementation.

It was also seen that a higher-resolution outer loop was the main contributor to increased, more detailed ensemble spread and more accurate analysis uncertainty estimates. For the same inner-loop resolution, it was found that an increase in outer-loop resolution from T255 to T399 also improves the forecast scores significantly. The investigations showed that a T95 inner loop is not capable of representing tropical cyclones and other extreme events accurately. On the other hand, the use of a T255 inner loop added significant extra computing costs without a significant gain in EDA variance estimates.

It turned out to be more beneficial to increase the outer-loop resolution combined with use of a moderate inner-loop resolution. It should be kept in mind that because the EDA system only produces 15-hour forecasts (not 10-day forecasts) the cost of increasing the outer loop and forecast resolution from T255 to T399 is relatively small.

The operational EDA configuration

Based on these investigations it was decided to use the following configuration for the operational EDA system.

- The EDA system is run at T399L91 resolution with a control (unperturbed) analysis and 10 perturbed analyses.
- The analyses are 12-hour 4D-Var, with two minimizations, first at T95, then at T159 with advanced linearized physics.
- The EDA is run twice daily, with the midnight analyses using observations from 2101 UTC to 0900 UTC and the midday analyses using observations from 0901 UTC to 1500 UTC.
- The observations used are those which have already been extracted for the operational high-resolution delayed cut-off 12-hour 4D-Var analysis, so the EDA analyses can run as soon as these observations become available.
- Unperturbed observations are used for the control analysis, while the other members use observations which have been modified by a random perturbation which is proportional to the observation error.
- For Atmospheric Motion Vector (AMV) observations, the perturbations are horizontally correlated.
- Input sea surface temperature fields are perturbed using the same method as for the seasonal forecasts.

In the EDA it is important to represent model error to account for the fact that the forecast model is not perfect. To simulate the impact of model uncertainty, the stochastically perturbed parametrization tendency (SPPT) scheme is used; this perturbs the total parametrized tendency of physical processes. Positive results have also been obtained with the stochastic backscatter (SPBS) scheme that is based on the idea of backscatter of kinetic energy from unresolved scales (see *Palmer et al.*, 2009 for a review of ECMWF work on stochastic parametrization schemes). To date, however, the use of the SPPT scheme alone gave the best performance (work is in progress to assess the impact of also introducing a backscatter scheme).

On average, if the EDA spread is measured in terms of the 700 (850) hPa temperature standard deviation, the SPPT scheme increases the global average by 19% (23%), and if the EDA spread is measured in terms of the 700 (850) hPa the kinetic energy standard deviation is increased by 33% (39%).

It is interesting to see the geographical distribution of model error impact on short-range forecast uncertainty implied by the stochastic methods. Figure 4 shows the zonally averaged EDA spread for temperature (Figure 4a) and zonal wind (Figure 4b) valid on 14 October 2008 when SPPT is used. The impact of SPPT on the spread can be assessed by calculating the ratio of EDA spread from an experiment with SPPT applied compared to EDA spread from an experiment without model error parametrization. It is clear that the increase in spread due to SPPT is significant for both temperature (Figure 4c) and zonal wind (Figure 4d) throughout the atmosphere. The largest SPPT impact for temperature is at the top of the planetary boundary layer, especially in the stratocumulus regions. For wind the largest impact is near 700 hPa in the tropics. The SPBS only perturbs the wind field directly, so the increased spread in temperature (Figure 4e) is small. The increased wind spread for the SPBS scheme (Figure 4f) is mainly located in the planetary boundary layer, where the convection is most active. It is clear that the SPPT scheme provides more widespread perturbations than the SPBS. The larger level of spread looks reasonable and gives the best improvement of the EPS system.

Figure 4 confirms that the SPBS and SPPT methods complement each other, but the new tendency stochastic physics is the more effective of the two methods. Further testing will be performed with both methods applied. This is linked to the medium term goal of unifying the model error representations in the EPS and EDA.

doi:10.21957/vzbtkbf8

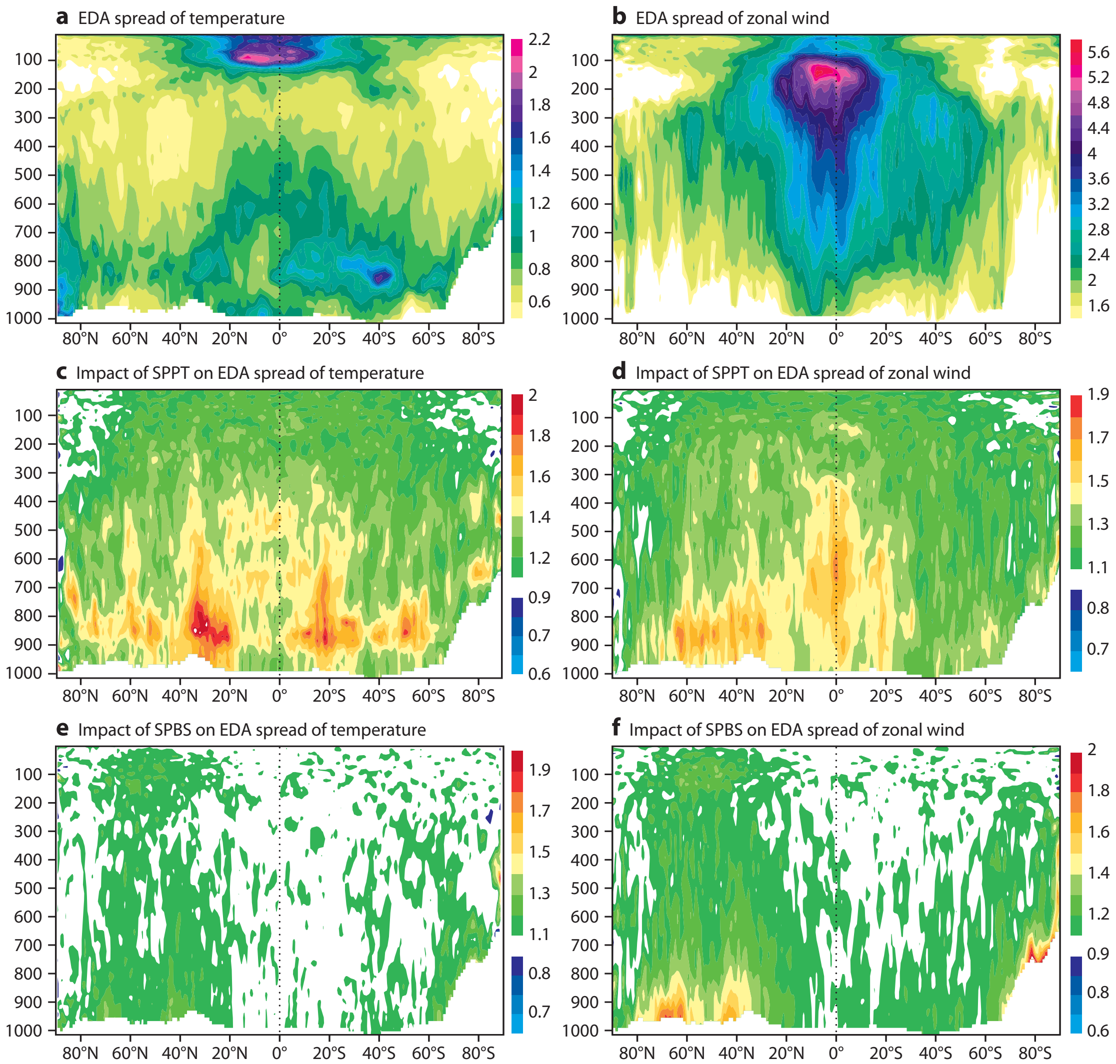

Figure 4 Impact of model error representation on EDA spread. Zonal mean values for (a) temperature and (b) zonal wind on 14 October 2008 using the operational SPPT configuration. The ratio of EDA spread for SPPT applied compared to EDA spread without model error parametrization for (c) temperature and (d) zonal wind. The same ratios for (e) temperature and (f) zonal wind when SPBS is used. As expected, the ratios are almost always greater than one.

Future developments

A well-designed EDA system will enable improved estimates of analysis uncertainty. This will be a potentially valuable output from the Centre's data assimilation system.

For the EDA, the short-term improvements will take account of horizontal correlations for radiance observation errors, use OSTIA instead of NOAA/NESDIS sea-surface temperatures (as already is the case in the operational 4D-Var), and improve the perturbations applied to the surface observations and parameters. Further investigations will be performed to assess the benefit of more ensemble members, various model error representations and different assimilation configurations. Finally, a significant effort is ongoing to develop and implement the use of flow-dependent background error in the deterministic 4D-Var system.

Further Reading

Buizza, **R.**, **M. Leutbecher**, **L. Isaksen** & **J. Haseler**, 2010: The use of EDA perturbations in the EPS. *ECMWF Newsletter No. 123*, 22–28.

Fisher, **M.**, 2003: Background error covariance modelling. In Proc. *ECMWF Seminar on Recent developments in data assimilation for atmosphere and ocean,* 8–12 September 2003, 45–63. http://www.ecmwf.int/publications/library/do/references/list/17111

Fisher, **M.** & **P. Courtier**, 1995: Estimating the covariance matrices of analysis and forecast error in variational data assimilation, *ECMWF Tech. Memo. No. 220.* http://www.ecmwf.int/publications/library/do/references/list/14

Isaksen, **L.**, **M. Fisher** & **J. Berner**, 2007: Use of analysis ensembles in estimating flow-dependent background error variance. In *Proc. ECMWF Workshop on Flow-dependent Aspects of Data Assimilation*, 11–13 June 2007, 65–86. http://www.ecmwf.int/publications/library/do/references/list/14092007

Palmer, **T.N.**, **R. Buizza**, **F. Doblas-Reyes**, **T. Jung**, **M. Leutbecher**, **G.J. Shutts**, **M. Steinheimer** & **A. Weisheimer**, 2009: Stochastic parametrization and model uncertainty. *ECMWF Tech. Memo. No. 598.* http://www.ecmwf.int/publications/library/do/references/list/14

Tan, **D.G.H.**, **E. Andersson**, **M. Fisher** & **L. Isaksen**, 2007: Observing system impact assessment using a data assimilation ensemble technique: application to the ADM-Aeolus wind profiling mission. *Q. J. R. Meteorol. Soc.*, **133,** 381–390.

doi:10.21957/vzbtkbf8