ECMWF Feature article

from Newsletter Number 125 – Autumn 2010

METEOROLOGY

Weak constraint 4D-Var

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

doi:10.21957/21fybybt

This article appeared in the Meteorology section of ECMWF Newsletter No. 125 – Autumn 2010, pp. 12–16.

Weak constraint 4D-Var

Yannick Trémolet, Mike Fisher

The fundamental purpose of 4D-Var (as implemented by ECMWF in 1997) is to correct a short-range forecast based on observations available since the last assimilation time. In this system, the correction is calculated in a four-dimensional domain: the three spatial dimensions and the time dimension. The atmospheric state over this domain is entirely determined by the state at the beginning of the assimilation window through the use of the forecast model. So, although 4D-Var finds the solution over a four-dimensional domain, it does so by adjusting the three-dimensional initial condition of the forecast (known as the control variable). This is equivalent to making the assumption that the forecast model is perfect over the length of the assimilation window. The model is said to be imposed as a strong constraint in the 4D-Var optimization problem.

Since 4D-Var became operational in 1997, many aspects of the data assimilation system have improved, and the amplitudes of many types of errors have reduced. The assumption that the model is perfect, or that model error is small enough relative to other errors in the system to be ignored, has become questionable. This is compounded by the fact that longer assimilation windows are desirable. Over long ranges, model error becomes larger and should be accounted for in the data assimilation process.

Relaxation of the perfect model assumption requires a modification of the 4D-Var algorithm. The resulting method is known as weak constraint 4D-Var. The remainder of this article will describe how weak constraint 4D-Var has been implemented in cycle 35r3 (8 September 2009) of ECMWF's Integrated Forecasting System (IFS). Also, directions for future research will be outlined.

Accounting for an imperfect model

The data assimilation process is a statistical problem where the best estimate of the state of the atmosphere is sought, given knowledge about the state and the error characteristics of the various sources of information. In weak constraint 4D-Var, the model is considered in the same way as the other sources of information, including taking into account that there is a degree of uncertainty about the information from the model.

There are several practical approaches to account for model imperfection in 4D-Var and to estimate model error (*Trémolet*, 2006). The first approach that has been implemented at ECMWF comprises adding a forcing term to the model which in principle would compensate for model error within each time step. The 4D-Var control variable is then augmented by these forcing terms and a term is added to the cost function to penalize model error according to its statistical characteristics. 4D-Var then determines what the optimal forcing terms should be, given the prescribed model error statistics and all other available information.

The forcing term at each time-step has in principle the same number of components as the model state. Thus the size of the control variable is multiplied by the number of time steps compared to the strong constraint 4D-Var control variable. This is unaffordable from a computing point of view but, even more importantly, there is not enough information available to determine so many parameters or to estimate their error characteristics, such as space and time correlations or flow dependence. Some simplifications are necessary to solve the weak constraint 4D-Var problem.

In our initial implementation of weak constraint 4D-Var, the main simplification is to assume that model error is constant in time over the length of one assimilation 'window' (currently 12 hours at ECMWF). With this assumption, the size of the control variable is doubled with respect to strong constraint 4D-Var, which is manageable on today's supercomputers. The model error covariance matrix becomes a three-dimensional matrix of the same dimension as the background error covariance matrix.

Model error statistics

Covariance statistics for both background and model errors are generated from large sets of samples. In both cases, since we do not know the true state of the atmosphere, we cannot explicitly generate the required samples of errors, and must instead turn to proxy quantities whose error statistics are similar to those of the actual errors. In the case of background errors, the proxy errors are generated as differences between forecasts from an ensemble of data assimilations. However, it is far from clear how to generate proxies for model error. In the current implementation of weak constraint 4D-Var, we use samples of differences between model tendencies from an ensemble of forecasts. The idea is that if each state from

 doi:10.21957/21fybybt

an ensemble is a possible representation of the actual atmospheric state, then the differences between the various tendencies generated by the model for different ensemble members give an indication of the likely size of the model uncertainty (see *Trémolet*, 2007 for more details).

The model error covariance matrix uses a spectral representation, as previously used for the background error. Since the covariances are computed from instantaneous quantities, rather than from short-forecast integrations, the correlation length scales for all variables are shorter in both the horizontal and the vertical. In this implementation, the covariances in one variable (temperature, say) are assumed to be uncorrelated with errors in other variables (e.g. vorticity). In reality, the errors affecting the model variables are not independent. So far attempts at accounting for multivariate relationships in model error have not been successful but this is an area for further research in the future.

More generally, very little is known about model error or its statistics. Whereas the estimation and modelling of the background error statistics has been an active area of research for many years, very little research has taken place to estimate or represent model error statistics in data assimilation. Moreover, the model error covariance matrix is an order of magnitude more complex than the background error covariance matrix due to the additional time dimension it involves.

Despite the relative lack of research into model error, a few approaches can already be identified. For example, ECMWF's Ensemble Prediction System (EPS) includes stochastic terms to represent model error. Such terms are mostly designed to make the ensemble prediction spread match the forecast error at medium range. Data assimilation is primarily a short-range problem. Nevertheless, valuable information and experience could be extracted from developments with the EPS. A practical step towards using the schemes developed in the EPS could be achieved by recognising that the method used by the stochastic backscatter scheme to convert the white noise output of the random number generator into a representation of spatially and temporally correlated model error, can be considered as defining the square-root of the model error covariance matrix. Use of this approach would allow a representation of model error in 4D-Var that is consistent with the approach adopted in the EPS.

In practice, observations are the only independent source of information available to estimate the actual model error. Ideally, estimations of model error statistics should use this information. Two directions of research can already be considered.

- Explore observation space consistency diagnostics, such as the ones proposed by *Desroziers et al.* (2005). Since strong constraint 4D-Var does not account for model error, any imperfect model should introduce inconsistencies with respect to the assumptions being made. The difficulty is then to extract useful model error information out of the internal signs of inconsistency.
- Extend work carried out in the context of lagged Kalman smoothers. It has been shown in that context that the difference in fit to the observations for analysis windows of different lengths can be attributed to model error. Again, the difficulty is to extract useful model error information from this signal.

Model error statistics are important for all data assimilation methods, including ensemble Kalman filters (where it is currently treated in a very crude way). Yet, model error is one of the least understood quantities in data assimilation. This is surprising, given the importance of the model in the process. Better estimation of model error statistics will be a major topic of research in data assimilation in the coming years.

Systematic model error

In the first implementation of weak constraint 4D-Var, model error is assumed to be independent from one assimilation cycle to the next. Although there are systematic errors that vary on longer timescales (seasonal for example), this is a safe approach that prevents positive feedback effects between model error on one hand and initial condition or observation bias correction on the other hand.

Experimentation is currently underway to account for model error on the longer timescales, such as temperature biases in the stratosphere in polar regions which typically vary on a seasonal timescale. This type of error is the largest in amplitude and can seriously affect the assimilation of high peaking channels for many satellite radiance observations. Since these biases vary on timescales that are slower than the length of the assimilation cycle, it is important to retain information from one cycle to the next. This is achieved by re-writing the model error penalty term in the 4D-Var cost function as a term that penalizes the variation in model error from one cycle to the next rather than the total model error. With this setup, however, there is no constraint to prevent model error from growing progressively over a large number of assimilation cycles. This could potentially have a major impact on the assimilation system.

Impact of the weak constraint 4D-Var

Weak constraint 4D-Var corrects model errors by adding a forcing term to each of the model's prognostic equations, in order to make the model consistent with the available observations. As it is known that errors in the ECMWF system are on average larger in the stratosphere than at lower levels, we have taken a cautious approach and initially restricted the model error term to apply only above 10 hPa (with a transition zone down to 40 hPa).

Figure 1 shows the monthly-mean model error forcing for temperature for July 2008. This indicates that to match the observations in the stratosphere and mesosphere there is the need for a systematic warming at polar latitudes and near the top of the model, with cooling at mid-latitudes in the upper stratosphere and mesosphere. Figure 2 shows the associated mean temperature analysis increments for strong and weak constraint 4D-Var. Note that the mean analysis increment is significantly reduced in weak constraint 4D-Var, which has correctly identified that the need for systematic corrections is due to errors in the model. Although the difference is relatively modest, Figure 2 also shows that oscillations in the increments over the North Pole have been removed and that they are reduced in amplitude over the South Pole. These oscillations are believed to be caused by model errors and their reduction should facilitate assimilation of satellite observations sensitive to temperature at these levels. All the results presented here were obtained with IFS cycle 35r2 at the resolution of T255 with 91 levels.

Figure 3a shows the average analysis and first-guess departures for radiances from AMSU-A channel 13 and the mean temperature analysis increment using weak constraint 4D-Var. Also shown is the model error forcing at the model level where this data is the most sensitive. The corresponding information when the model error is cycled and allowed to grow over time is shown in Figure 3b. In this case, the average observation first-guess departure is centred around the zero line which is not so when there is no cycling (red curves). This shows that the short-term forecast is improved by the model error cycling and model error information is retained and useful from one cycle to the next. The seasonal variation in observation bias correction (black curves) is also slightly reduced which goes in the right direction since this variation is due to seasonal variations in the model and not in the observations. The mean analysis increment (green curves) is also closer to the zero line where it should be in an unbiased system. (Note that it is not fully centred around the zero line. Most likely this is due to the fact that AMSU-A channel 14, which also has some sensitivity at that level, is not bias corrected.)

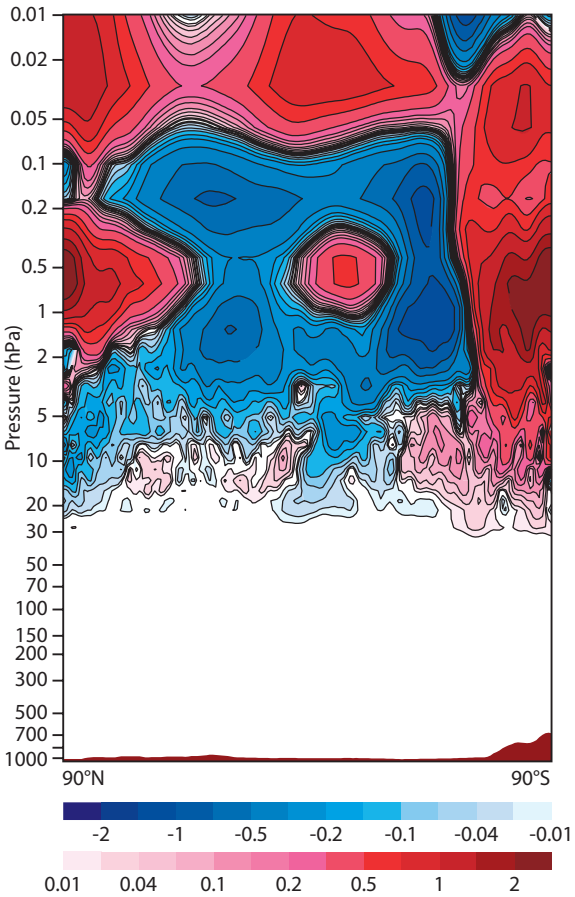

Figure 1 Monthly-mean temperature model error forcing (K/12h) in the stratosphere estimated by weak constraint 4D-Var for July 2008 (IFS Cy35r2, T255).

doi:10.21957/21fybybt

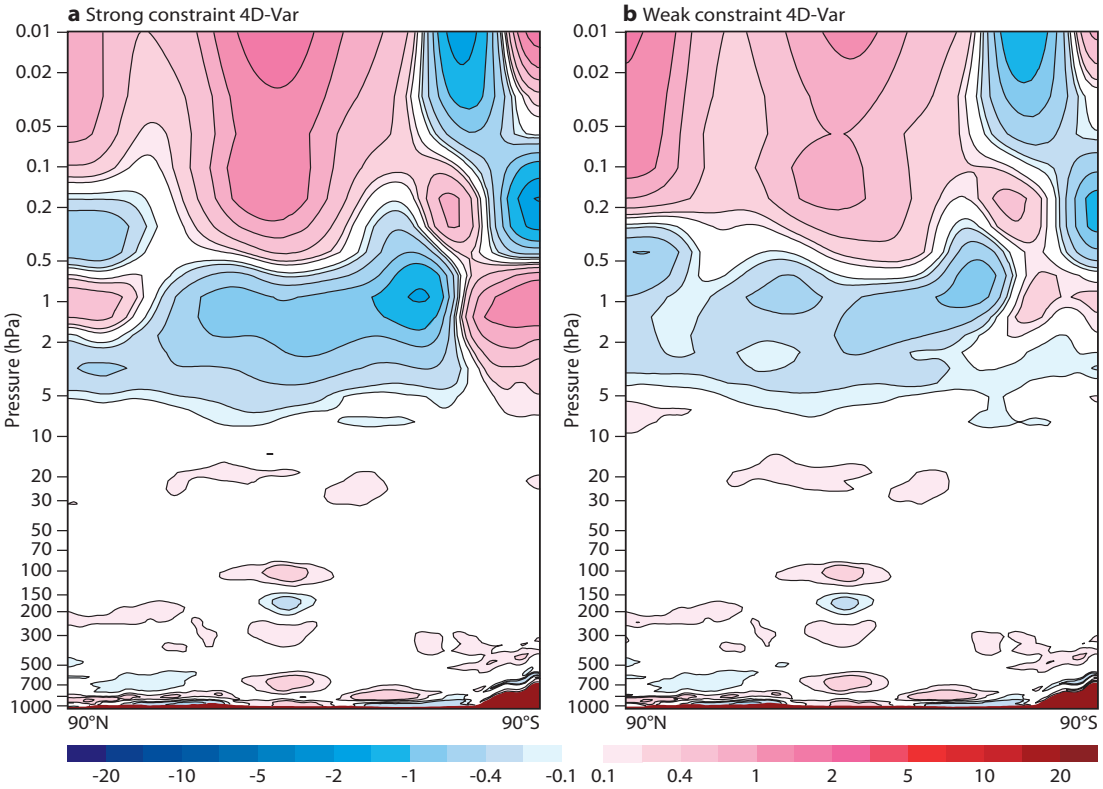

Figure 2 Monthly-mean temperature analysis increments (K) for July 2008 in (a) strong constraint 4D-Var and (b) weak constraint 4D-Var with model error applied in the stratosphere.

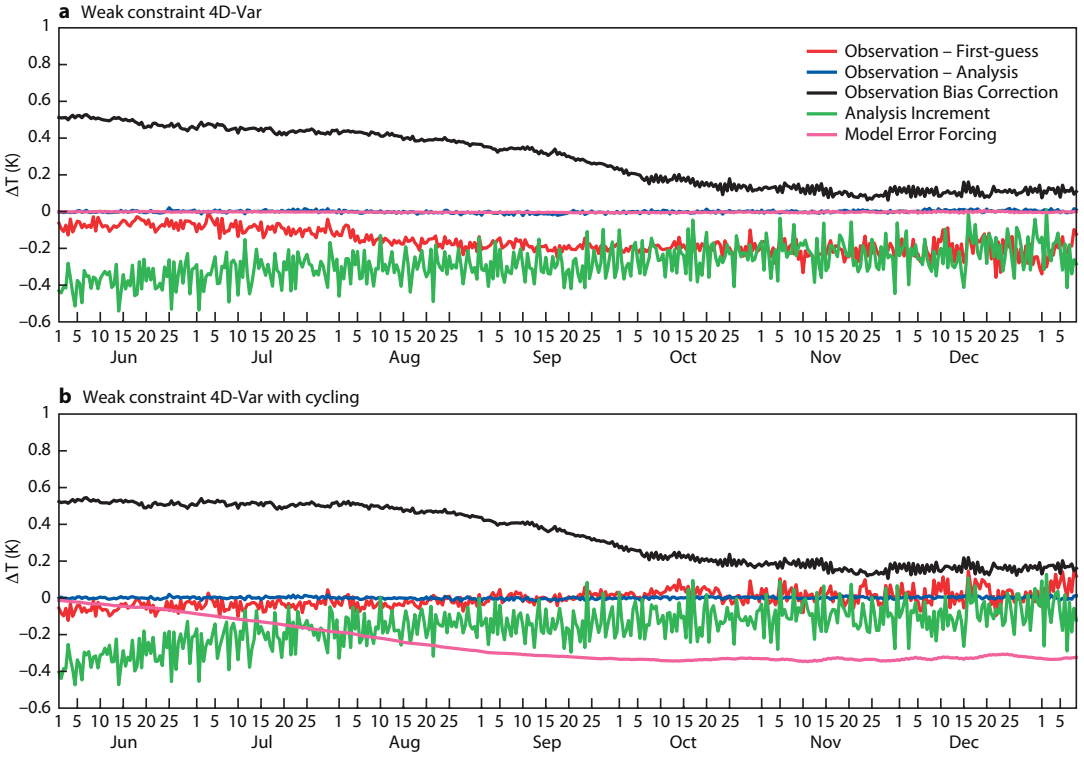

Figure 3 (a) Average analysis and first-guess departure of radiances from AMSU-A channel 13 for July 2008 for the northern hemisphere using weak constraint 4D-Var. Also shown is the mean temperature analysis increment and model error forcing at model level 14 where this channel is the most sensitive. (b) As (a) but the model error is cycled. See text for more details.

These results give an example of the complex interactions between the model error, the initial condition increment and the observation bias correction in 4D-Var. Overall, model error as estimated in these experiments varies on a slow timescale (the order of a few months on Figure 3) and should probably also vary on shorter timescales. However, interactions with the other parts of the control variable must be examined carefully.

4D-Var uses the covariance matrices of background, observation and model errors to partition the analysis error between the various terms of the cost function. Incorrect specification of any one of these covariance matrices can result in one source of error being misinterpreted as another type of error (e.g. observation error may be misinterpreted as model error). This highlights the necessity for a proper estimation of all the covariance matrices. The large number of available degrees of freedom in the model error makes it particularly important to correctly specify the model error covariances so as to avoid absorbing all the information contained in the observations into the wrong component.

Towards a longer assimilation window

Fisher et al. (2005) have shown that weak constraint 4D-Var with a long assimilation window is equivalent to a full rank Kalman smoother. This can be seen as theoretical justification for a move towards longer assimilation windows in 4D-Var. On a more pragmatic level, it seems obvious that a simultaneous analysis of all relevant observations ought to lead to a better analysis than an artificial splitting of observations into batches of length 12 hours, to be analysed independently. The practical implementation of this idea, however, requires that 4D-Var take into account the variation of model error within the assimilation window. For analysis windows longer than 12 hours, it is no longer sufficient to assume that model error remains constant throughout the window.

Because of the limitations imposed by the incremental 4D-Var algorithm, a long analysis window cannot be achieved with a formulation involving a model error forcing term. (The different-resolution models used in the inner and outer loops of 4D-Var react very differently to the forcing, and diverge significantly over the analysis window, preventing the convergence of the incremental algorithm.) Consequently, the weak constraint 4D-Var cost function has to be formulated directly as a function of the four-dimensional state over the length of the assimilation window. (In practice, for computational reasons, and because of the limited amount of available information, the state variable at regular sub-intervals over the assimilation window would be used.) For each time when the state variable is available, a model error term in the cost function is applied to minimize the gap between the state obtained by integrating the model from the previous time when the control variable is defined to the current time. We illustrate this schematically in Figure 4.

This approach has the significant advantage that the state at the start of each sub-interval is known at the start of each iteration of the minimisation. Evaluation of the cost function requires integrations of the model and its adjoint to be performed for each sub-interval, and these model integrations can be performed in parallel. This brings an additional dimension for parallelism in 4D-Var. However, the optimization problem that results from this formulation of 4D-Var has different properties than the (by now well-understood) strong constraint 4D-Var problem. Research into minimisation algorithms and preconditioning methods will be required to develop efficient minimisation strategies.

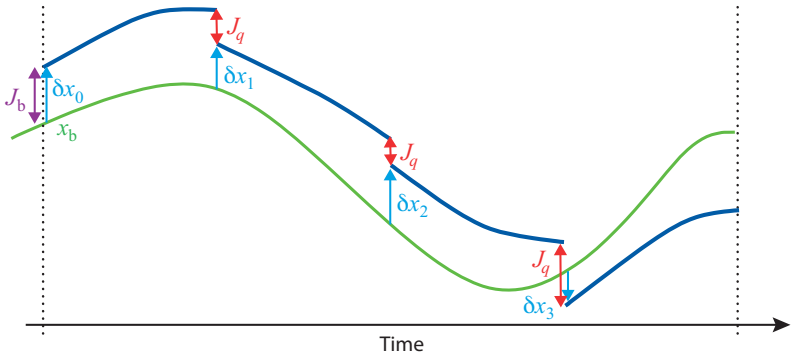

Figure 4 Long window 4D-Var. The schematic diagram shows a single analysis window. The green line represents a model integration started from the background state, which serves as the first guess for the analysis. The control variable for the analysis consists of the departures from the first guess at the start of a set of sub-intervals. Weak constraint 4D-Var adjusts these departures to simultaneously minimise the discrepancy between the analysis and the observations and background state, while also minimising the jumps between the sub-intervals.

 doi:10.21957/21fybybt

Further developments

The continued reduction in sources of analysis error since the introduction of 4D-Var means that it is no longer possible to ignore the model itself as a source of error. Accounting for this source of error requires the explicit inclusion of a model error term in the 4D-Var cost function, and a representation of the covariance matrix of model error.

Gathering information about the statistics of model error is difficult, and will be the topic of much future research. Our initial attempts have concentrated on approximating the model error covariance matrix by a covariance matrix of model tendencies. This is unlikely to be an accurate approximation. Nevertheless, it has proved sufficient to allow some significant model errors in the stratosphere to be accounted for in the analysis. Weak constraint 4D-Var was introduced into the ECMWF operational system with the implementation of IFS cycle 35r3 on 8 September 2009.

Improvements to the representation of systematic model error are planned through a modification of the model error penalty term that will allow information about model error to be retained from one analysis cycle to the next.

Although systematic model error is probably the largest component of model error, it will be necessary to account also for the time-varying component. It is likely that this component will have significantly different spatial structure than systematic error, requiring careful construction of the associated error covariance matrix. It is also highly likely that model error is correlated in time, and it will be necessary to take this correlation into account.

We believe that longer assimilation windows have the potential to significantly improve the quality of the 4D-Var analyses. However, to achieve longer windows, we require good statistical models of model error. Longer windows also require changes to the methods used to minimise the cost function. This presents a challenge, requiring the development of new approaches to minimisation and preconditioning. However, it also presents an opportunity to significantly improve the parallel efficiency and scalability of 4D-Var, by allowing parallel model integrations during the evaluations of the cost function. This increase in scalability will be important if 4D-Var is to remain practical on future computer architectures.

Further reading

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

Fisher, **M.**, **M. Leutbecher** & **G. Kelly**, 2005: On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.*, **131**, 3235–3246.

Trémolet, **Y.**, 2006: Accounting for an imperfect model in 4D-Var. *Q. J. R. Meteorol. Soc.*, **132**, 2483–2504.

Trémolet, **Y.**, 2007: Model error estimation in 4D-Var. *Q. J. R. Meteorol. Soc.*, **133**, 1267–1280.

doi:10.21957/21fybybt