ASSIMILATION OF SATELLITE DATA USING VARIATIONAL TECHNIQUES

J. Pailleux
Météo-France
Toulouse, France

Abstract

An important effort has been put on variational techniques in research on data assimilation, since the mid 80s. The motivations of this effort are described, one important motivation being the use of satellite data (and especially TOVS data). The principles and the tools relevant to the use of satellite data in a variational analysis are described, and some simple examples are given. Then the state of the art regarding use of TOVS data is summarized, and the main evolutions of satellite data and of their assimilation in Numerical Weather Prediction (NWP) are discussed.

1. GENERAL FORMULATION OF VARIATIONAL ASSIMILATION

1.1. Principle of 4D variational assimilation (4D-VAR)

4D-VAR is an algorithm which computes directly a model trajectory entirely consistent with all the observations on a time period (t_1,t_2), and entirely consistent with the model equations; see Courtier et al (1992), Pailleux (1992) and Thépaut et al (1992). This algorithm integrates in one single consistent process the usual steps of an operational data assimilation suite: analysis step, initialization step and forecast step. It allows a more optimal use of observations (and of any kind of information on the atmosphere) than intermittent schemes.

To perform a 4D assimilation on a time period (t_1,t_2), one needs to find the model trajectory which minimizes the fit to all the observations on the time period, as well as the fit to a background (first guess) at time t_1. This background X_b is the best estimate of the model state X at time t_1, prior to any collection of observations at and after time t_1. As in all the operational assimilation schemes, the background X_b represents a summary of all the information on X accumulated before t_1. It is generally provided by the most recent numerical forecast available for time t_1. As documented in several papers, see for example Thépaut and Courtier (1991) or Courtier (this volume), the 4D-VAR technique consists in minimizing the following cost function:

$$\mathbf{J(X) = J_o + J_b + J_c} \quad \mathbf{(1)} \qquad \text{where:}$$

- J_o is the distance to the observations; the natural distance is the quadratic form built on the inverse of matrix O (covariance matrix of observation errors) and on the vector of observation departures: HX-y (y: vector containing all the observed data; H is the **observation operator** which produces the equivalent of vector y from model state X);

- J_b is the distance to the background Xb; the natural distance is the quadratic form built on the inverse of matrix B (covariance matrix of forecast errors) and on the vector $X-X_b$ (departures of model state to the background);

- J_c is an optional penalty term which contains physical constraints to apply on the model state X.

1.2. Some historical aspects: use of satellite data and 3D-VAR

The previous 4D-VAR algorithm "degenerates" into a 3D variational algorithm (3D-VAR) if $t_1 = t_2$ (i.e. if the assimilation period has a zero length). In 3D-VAR the minimization of cost function (1) is still performed with respect to $X = X(t_1)$, the background J_b is the same as in 4D-VAR, but the observation cost function J_o is using only the observations y which are at time t_1 (or close to time t_1). Consequently the comparison of the model to the observations can be done without any integration of the forecast model.

A 3D-VAR scheme can be used like any instantaneous analysis such as Optimal Interpolation (OI) in an intermittent assimilation scheme. Then one loses the above mentioned advantage of 4D-VAR: to combine optimally the observations and the information contained in the model dynamics. However, with respect to a standard operational OI analysis, a 3D-VAR scheme has two other advantages:

- It does not require any data selection as it solves a global minimization problem which uses all the observations;
- It allows the use of a larger variety of observed parameters, as the observation operator H does not need to be linear, does not need to be inverted and does not need to be simple.

As explained in Pailleux (1988), the possibility of using cloud-cleared TOVS radiances directly by including the direct radiative transfer computation in the observation operator H was a major motivation for developing 3D-VAR schemes. This was seen as a major step toward more optimal use of TOVS information which is represented in a very crude way in retrieved profiles. Problems in using TOVS retrievals in an analysis scheme have been extensively studied in Gallimore and Johnson (1986), Andersson et al (1991) and Kelly et al (1991). Experiments on the use of TOVS radiances in 3D-VAR have already been documented in Andersson et al (1992): they show the power of a 3D-VAR scheme to assimilate radiances directly. More results on direct assimilation of TOVS radiances in a variational analysis are given in this volume: see McNally et al, Andersson et al, Thépaut et al.

1.3. <u>Computation and minimization of the cost function</u>

Each term of the cost function (1) can be computed by applying a chain of operators to the control variable X (ensemble of model variables). The minimization of the cost function requires also the computation of its gradient. The gradient of each term of the cost function can be computed by using the adjoint of each operator involved in the cost function computation, and by applying them in the reverse order. This "adjoint technique" is illustrated in the following simple example of observation cost function.

Let us assume that the vector X of model variables is to be analysed at time t_1 (3D-VAR) from one observation y (only!) of a linear combination of the first two components of X: x_1 and x_2. The single observed quantity y is also made at time t_1.

$$X = \begin{vmatrix} x_1 \\ x_2 \\ .. \\ x_N \end{vmatrix}$$ is the control variable.

One observes the model quantity $z = HX = hx_1 + (1-h)x_2$, h being a constant interpolation coefficient.

The cost function to minimize is: $\mathbf{J(X) = (z\text{-}y)^2}$ which is the quadratic distance between the model and the observation (it is not necessary to weigh the term by the observation error variance as the term is unique).

The gradient of J is easy to compute analytically with respect to z: $\mathbf{\partial J/ \partial z = 2(z\text{-}y)}$.

However the minimization algorithm requires the gradient of J with respect to the control variable X, not with respect to z. In this simple case the gradient with respect to X can be computed analytically:

$$Grad_X J = \begin{vmatrix} \partial J/ \partial z.\ \partial z/ \partial x_1 \\ \partial J/ \partial z.\ \partial z/ \partial x_2 \\ 0. \\ .. \\ 0. \end{vmatrix} = \begin{vmatrix} h\ \ \partial J/\partial z \\ (1\text{-}h)\ \ \partial J/\partial z \\ 0. \\ .. \\ 0. \end{vmatrix}$$

$$\mathbf{Grad_X J} = \begin{vmatrix} \mathbf{h} \\ \mathbf{1\text{-}h} \end{vmatrix} \mathbf{\partial J/ \partial z = H^t\ \partial J/ \partial z} \qquad \mathbf{(2)}$$

The observation operator can be expressed in a matrix form: H is a matrix with only one row and two columns (two coefficients), its transpose is a 1 column vector containing the same two coefficients.
The gradient with respect to X can also be computed using the "adjoint rule": if H is an operator computing z from X, H' the corresponding tangent linear operator, H^* the adjoint operator, then H^* computes the gradient with respect to X of any function J from its gradient with respect to z:

$$X \xrightarrow{\quad H \quad} z$$
$$X \xrightarrow{\quad H' \quad} z$$
$$\mathrm{Grad}_X J \xleftarrow{\quad H^* \quad} \mathrm{Grad}_z J$$

So by definition of the adjoint, one can write:

$$\mathbf{Grad_X J = H^* \; \partial J/\partial z} \qquad \mathbf{(3)}$$

Comparing (2) and (3), one can verify on this trivial example that the adjoint of a linear operator H expressed in a matrix form is the transpose matrix:

$H^* = H^t$ (In addition H = H' as H is linear).

Let us now assume that the same observation y is made at an observation time $t>t_1$, and that one still wants to use it to analyse X at time t_1 (4D-VAR). In order to compare X with the observation, one needs first to integrate a forecast model from t_1 to t (operator M), then one needs to apply the same observation operator H as before to X(t).

$z = H(X(t)) = H[M(X)]$

Then the cost function J can be computed in the same way as before (3D case), as well as its gradient with respect to z (trivial) and the gradient with respect to X(t) (application of the trivial adjoint of H). Applying also the adjoint rule to the forecast model M, one can compute the gradient with respect to $X = X(t_1)$:

$\mathrm{Grad}_X J = M^*[\mathrm{Grad}_{X(t)} J]$

One can verify the chain rule for the adjoint computation: to compute z from X, one applies first operator M, then operator H; the adjoint of the compound operator is obtained by applying the adjoint of the individual operators in the reverse order: H^* first, then M^*.
The chain rule illustrated on this simplified example is used systematically to compute the observation cost function of (1) and its gradient. This adjoint technique is especially powerful when one wants to include complex satellite data in the variational analysis: any observed quantity can be used provided it is linked to the model variables through a differentiable observation operator (to allow gradient computations), even if this link is very indirect. **One important point is that the observation operators do not need to be inverted: only their adjoint is required, not their inverse.** Examples of observation operators are given in the next section.

2. COMPUTATION OF THE Jo COST FUNCTION AND ITS GRADIENT

In the following examples, only the vertical part of the observation operators is envisaged: it depends on the type of the observation variable whereas the horizontal part is common to all observations (inverse spectral transforms if the forecast model is spectral, and horizontal interpolation to the observation point).

2.1. Example of a simple conventional observation operator: wind components from radiosondes (linear operator)

In order to compute the model equivalent of the observation, one has to interpolate each wind component in the vertical to go from the model levels to the observation levels. One can use for example a linear interpolation in p between model levels p_k and p_{k-1} to observation level p, as illustrated in fig. 1.

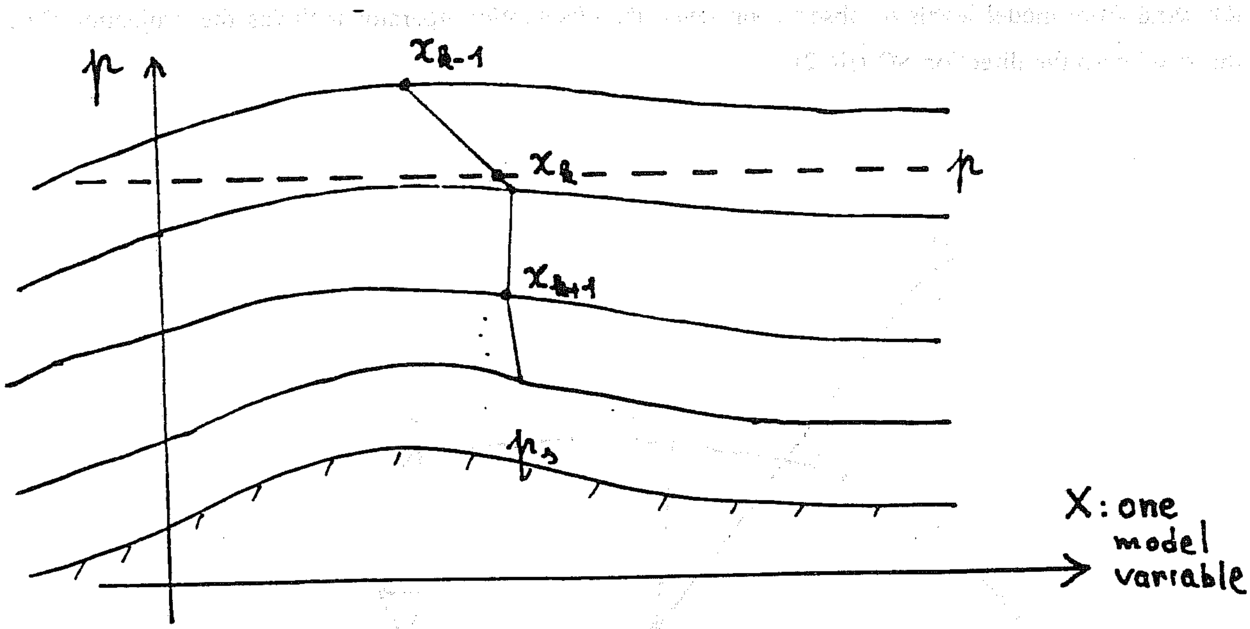

Fig.1: An example of vertical observation operator. Use of linear interpolation in p from model hybrid levels (dependent on p_s) to observation level p.

The two following points are worth noting in relation with the gradient computations:

- x_k being any wind component (u or v) at level k, the interpolated value at observation pressure p depends on the model level wind values x_{k-1}, x_k, but also on surface pressure p_s if the model has a sigma or hybrid coordinate system; consequently the surface pressure which drives the model geometry in the vertical will be affected by any upper air observation, regardless of its type;
- in the case of the simple vertical interpolation presented in fig.1, let us note that the observation operator is not a differentiable function of surface pressure: in the case when the observation pressure level is very close to say model level k, a small perturbation p_s on the surface pressure is sufficient to move model level k from one side of the observation level to the other. The interpolation is then suddenly performed from x_k and x_{k+1} rather than from x_{k-1} and x_k (non differentiable operator as illustrated by the angular points on fig.1).

2.2. <u>Two examples of simple satellite observation operators: wind component observed from a satellite lidar - total water vapour content from a microwave instrument (linear operators)</u>

A satellite lidar can observe the wind component at an observation point O in the atmosphere, along the line going from the satellite S to the observation point O. In addition to the vertical interpolation I_v of the wind from model levels to observation level, the observation operator includes the projection P of the wind onto the direction SO (fig.2).

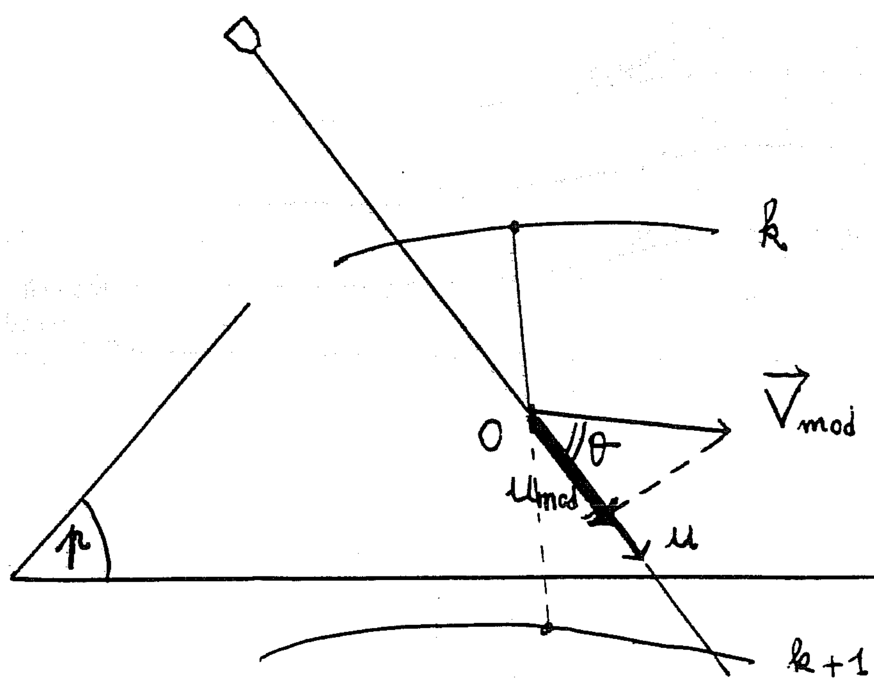

Fig.2: Example of the satellite wind lidar observation operator. The observed quantity is the wind component up along the line SO going from the satellite to the observation point.

Let us call V_{mod} the model wind vector interpolated in the vertical to the observation pressure level p, u_{mod} its projection onto the line SO, u the wind component observed by the lidar along SO.

To compute u_{mod}, one applies first the vertical interpolation I_v to the model variables, then the projection operator P. At this point one can compute the contribution of lidar observation u to the observation cost function:

$$J_o = ((u_{mod}-u)/s_0)^2$$

(s_0 is the standard deviation of the lidar observation error, necessary to weigh properly the lidar observations with respect to other sources of information entering the variational analysis).

The gradient of J_o with respect to u_{mod} is trivial to compute:

$$\partial J_o / \partial u_{mod} = 2(u_{mod}-u)/s_0^2$$

To compute the gradient with respect to model variables ,one follows the adjoint rule by applying first P^*, then I_v^*. The operator P is a projection involving angle ,it is linear;its adjoint involves trivial multiplications by $\cos(\theta)$ and $\sin(\theta)$.

Another simple linear satellite observation operator is the one involved in the use of an observed total water content of the air column (say for example the water content as observed by the current SSM/I instrument). The specific humidity q is a model variable, the vertical observation operator is its vertical integral, discretized in an appropriate way:

$$PWC_{mod} = \int_0^{ps} q.dp$$

PWC_{mod} has to be compared to the observation PWC in order to compute its contribution to the observation cost function. The gradient computations requires the adjoint of the discretized integral: simple multiplication by the weights involved in the discretized integral.

2.3.Example of a nonlinear satellite observation operator: wind speed as observed by an altimeter

Some meteorological satellites like ERS1 are equipped with altimeter instruments which allows the indirect observation of the surface wind over the oceans. The comparison of the forecast model to such an observation requires:

. The application of a boundary layer model B_l deriving the surface wind from the model variables;

. The computation of the model wind speed (square root operator).

Let us note that both the B_l and square root operators are nonlinear: this means that the gradient computations require the basic model state, the weights given to the different observations are then situation dependent (as these weights are driven by the gradient). This would not be possible in a classical linear algorithm used in most of the operational analysis schemes. Using the wind speed only (without any information on the wind direction) is also very easy to achieve in the context of a variational analysis, but it is not possible in an OI analysis because of the nonlinearity of the square root operator: OI has to go around this difficulty by creating pseudo observations from (e.g.) the observed wind speed and the first guess direction. Along the same lines, the variational analysis methodology allows the direct use of a wind direction observation only: this may be useful for using some cloud winds which are sometimes reliable in direction but not in speed.

The use of wind speed observations requires the adjoint of the two above operators. The adjoint of the square root operator is trivial, but the adjoint of the boundary layer operator may be very complex, depending on the physical parameterizations which have been included in the forecast model. B_l^* is needed not only for optimum use of altimeter data, but of all near surface observations: SYNOPs, SHIPs, buoys, scatterometer data.

2.4. Use of direct/adjoint radiative transfer code to assimilate TOVS radiances

The methodology illustrated on the previous examples can be applied to the TOVS radiances in a straightforward way. The standard radiance computation packages are radiative transfer observation operators which use as input vertical profiles of temperature and humidity (plus some surface parameters). In order to evaluate the distance J_o (say in a 3D-VAR problem) between the forecast model and TOVS radiance observations, one needs to include this radiance computation into the chain of operators H at each satellite observation point. One also needs to include the adjoint of the radiance package into the chain of adjoint operators H^*. Let us note that the inverse of the radiance computation is not needed (the minimization package will take care implicitly of the temperature/humidity retrievals using all the information available in 3D-VAR: model background, radiosondes...). Let us note also that the radiance observation operator is dependent on the profile of meteorological variables (i.e. nonlinear): this means that the appropriate amount of information present in the TOVS radiances is extracted from each profile in 3D-VAR. This is not possible in a linear analysis such as OI. More details about the TOVS observation operators can be found in Andersson et al (1992).

Such a variational analysis scheme is also attractive for using TOVS, because it allows the integration of retrieval (1D-VAR), 3D variational analysis (3D-VAR) and 4D variational assimilation (4D-VAR) in a single concept:

- If the vector X in expression (1) is a temperature and humidity profile (rather than all the 3D fields), the 3D variational scheme is degenerated into a 1D variational analysis, i.e. a retrieval scheme using a background information X_b in addition to the TOVS radiances.
- One can try to compare the vector X of model variables at time t_1 with data which are observed at any time $t>t_1$, by including the forecast model in the chain of operators (see 1.3 above). The 3D scheme is then generalised into a 4D variational scheme.

2.5. Computation of Jo for radiances: an algorithmic aspect

The computation of the cost function Jo for satellite radiances is less straightforward as for most of the other observation types, because the radiance errors are expected to be correlated in the horizontal. Consequently the matrix O is far from being diagonal, and a special numerical technique is needed. One practical way of solving the problem consists in using the following two remarks:

- The radiance cost function can still be split into contributions coming from different satellites (NOAA11, NOAA12,..) and from different TOVS types (clear, partly cloudy, cloudy).

-Assuming the observation error correlation can be split into the product of a vertical correlation (interchannel correlation) by a horizontal one (using separability assumption), the radiance cost function can be split into contributions coming from the different eigen-vectors of the interchannel correlation matrix of TOVS errors.

3. STATUS AND PERSPECTIVES

3.1. Current status of 1D/3D/4D - VAR

As mentioned before, 1D-VAR retrieval, 3D variational analysis and 4D variational assimilation can be presented as different versions of a single concept: minimizing a cost function representing the distance of the model to various observations and representing also other penalty terms. 1D, 3D and 4D can also be applications of a same code. For example the same direct/adjoint radiance computation package can be used for these three applications.

1D-VAR TOVS retrieval has computer memory requirements which are much lower than 3D and 4D-VAR because of the small size of the control variable which is reduced to one vertical profile. A 1D-VAR scheme has been used operationally at ECMWF since 1992 (see McNally et al (this volume)). The "interactive retrieval" scheme which became operational at NMC (Washington) is based on the same variational principle (see Derber in this volume). Although the use of a model background in the retrieval procedure provides retrieved profiles which are very accurate, one has to be careful when using these profiles in the full 3D analysis: the difficulty is to weigh them properly as they contain information from the model background in addition to information from the satellite radiances.

3D-VAR schemes are much more expensive than 1D-VAR in computer memory because of the minimization procedure performed directly in the 3D space. However their costs in computer time has the same order of magnitude as the current operational analysis scheme. A 3D-VAR scheme (called "Statistical Spectral Interpolation", see Derber (this volume)) has been operational at NMC (Washington) since 1991. A similar scheme should become operational at ECMWF in 1994. 4D-VAR schemes are at least as expensive in memory as 3D-VAR schemes, and they are much more expensive in computer time because of the integration of direct/adjoint models which has to be repeated several times. However realistic simplifications of 4D-VAR can be found, such as the "incremental approach" (see Thépaut et al (this volume)), and it is reasonable to assume that 4D-VAR can reach the operational stage near the mid-90s.

3.2. Some difficulties in using satellite data in NWP models

Satellite instruments can provide a lot of information on the atmosphere which is badly represented (or not represented) in NWP models: cloud information, cloud liquid water, parameters defining the physical properties of the ground... On the other end the assimilation procedure works well only if the model variables can be compared with a reasonable accuracy to the observations. The assimilation of the above mentioned parameters will not be completely achieved until these parameters are well represented in NWP models. The direct use of raw radiances is also a problem as most of the channels are contaminated by clouds. Consequently the variational approach previously described works well only when using cloud-cleared radiances. The direct use of TOVS raw radiances implies:

- either a good representation of clouds in the assimilating model;
- or the availability of cloud information from another source (for example another satellite instrument such as AVHRR).

In spite of these limitations, the main trend of these last few years has been to assimilate data which are closer and closer to the genuine observed quantity as it is measured by the instrument. This is especially true for satellite instruments which have been developed recently and provide information on the atmosphere which is less and less directly linked to NWP variables: one example is the scatterometer data. This implies that a large variety of observation operators H has to be developed (together with their adjoints to fit the variational approach). Some of these operators can be very complex, and with the development of future satellite instruments these operators are expected to become more and more complex.

Another specific aspect of satellite data is its big volume compared to conventional observations. In the ECMWF system, the TOVS clear radiances have to go through a prescreening before entering 3D-VAR: in addition to quality control tests, this prescreening (currently called "PRESAT") reduces the number of TOVS observation points by a ratio which is about 3. When one tries to use the TOVS raw radiances directly the data volume will be increased by roughly one order of magnitude. It will become even bigger when ATOVS replaces TOVS in the mid 90s. A preliminary processing of these data sets will be necessary to avoid having the satellite information swamping all the other observations analysis: the problem is to weigh the satellite information correctly with respect to the other sources of information, and to describe correctly (through appropriate covariance error matrices) the redundancy which is present in the satellite radiances: redundancy between different channels in the vertical, redundancy between different observation points in the horizontal.

In the longer term future, there are plans for developing satellite instruments (e.g.: interferometers) with a very high number of observed quantities. The total amount of satellite data could be further increased by two orders of magnitude, with the equivalent of several thousands of highly redundant channels in the vertical. This rapid increase of data volume may slow down or even stop the current tendency toward assimilating parameters which are closer and closer to the genuine observed quantity. Some techniques have to be found to compress the future satellite information into a form which:

- is still representative of what the instrument can observe in the atmosphere;
- is suitable to be presented to the variational analysis, i.e. leads to observation operators H which are tractable and to observation error covariance matrices which are easy enough to model.

4. CONCLUSION

Data assimilation work carried out since the mid 80s has demonstrated that the variational algorithm is a flexible and powerful methodology for using current and future satellite data. The approach consists in developing all the required observation operators to go from the model variable space to the space of observed quantities, in order to compare the forecast model with the observations. The idea is also, as far as possible, to avoid inverting these observation operators, as very often the satellite information on its own is insufficient to represent the forecast model variables correctly. Ultimately 4D variational assimilation should provide the optimal mix of the whole information available in a NWP model: satellite and conventional observations, information contained in the dynamics of the model...

The main current applications of the variational methodology in the area of satellite data (illustrated in this volume) are the use of TOVS radiances and the use of scatterometer data. In the future one will notice rapid developments of use of satellite data along the same methodology. However it is likely that significant extra work will be needed to assimilate the huge satellite data volumes which are announced for the year 2000 and beyond.

ACKNOWLEDGEMENTS

The concepts described and illustrated in the present paper are coming from the work and discussions held since the mid 80s in the ECMWF data assimilation and satellite sections. Special thanks are due to the current and past members of these two sections, including the consultants from outside Europe.

REFERENCES

Andersson,E.,A. Hollingsworth,G. Kelly,P. Lönnberg,J. Pailleux and Z. Zhang, 1991: Global observing system experiments on operational statistical retrievals of satellite soundings.
Mon. Wea. Rev., 118, No 8, 1851-1864

Andersson,E.,J. Pailleux,J-N. Thépaut,J.R. Eyre, A.P. McNally, G.A. Kelly and P. Courtier, 1992: Use of cloud-cleared radiances in three/four-dimensional variational data assimilation.
Proceedings of the ECMWF workshop on variational assimilation with special emphasis on three-dimensional aspects. Reading. 9-12 November 1992. Available from ECMWF. Also submitted to QJRMS.

Andersson,E., A. McNally, J.N. thépaut and J. Eyre, 1993: Assimilation of satellite data by 3D-VAR at ECMWF. This volume.

Courtier,P.,J.N. Thépaut and A. Hollingsworth, 1992: A strategy for operational implementation of 4D-VAR.
Proceedings of the ECMWF workshop on variational assimilation with special emphasis on three-dimensional aspects. Reading. 9-12 November 1992. Available from ECMWF.

Courtier,P., 1993: Introduction to Numerical Weather Prediction data assimilation methods. This volume.

Derber,J., 1993: Assimilation of TOVS data at NMC. This volume.

Gallimore,R.G. and D.R. Johnson, 1986: A case study of GWE satellite data impact on GLA assimilation analysis of two cyclones. Mon. Wea. Rev., 114, 2016-2032.

Kelly,G.A., E. Andersson, A. Hollingsworth, P. Lönnberg, J. Pailleux and Z. Zhang, 1991: Quality control of operational physical retrievals of satellite sounding data. Mon. Wea. Rev., 119, 1866-1880.

McNally A., G.A. Kelly, E. Andersson and J. Eyre, 1993: Assimilation of TOVS data by 1D-VAR at ECMWF. This volume.

Pailleux,J., 1988: Variational analysis: Use of observations - example of clear radiances. Proceedings of the ECMWF Seminar on Data Assimilation and the use of satellite data. Reading. 5-9 September 1988. Available from ECMWF.

Pailleux,J., 1992: Organisation of 3D variational analysis within the "IFS/ARPEGE" project. Future plans at Météo-France.
Proceedings of the ECMWF workshop on variational assimilation with special emphasis on three-dimensional aspects. Reading. 9-12 November 1992. Available from ECMWF.

Thépaut,J.N. and P. Courtier, 1991: Four-dimensional variational data assimilation using the adjoint of a multilevel primitive equation model. Q.J.R.Met. Soc., 117, 1225-1254.

Thépaut,J.N., P. Courtier and R.N. Hoffman, 1992: Use of dynamical information in four-dimensional variational assimilation.
Proceedings of the ECMWF workshop on variational assimilation with special emphasis on three-dimensional aspects. Reading. 9-12 November 1992. Available from ECMWF.

Thépaut,J.N., E. Andersson, P. Courtier and R. Hoffman, 1993: Assimilation of satellite data by 4D-VAR at ECMWF. This volume.