Variational assimilation of conventional meteorological observations with a multilevel primitive-equation model

By JEAN-NOËL THÉPAUT*, DRASKO VASILJEVIC and PHILIPPE COURTIER

European Centre for Medium-range Weather Forecasts, Reading

and

JEAN PAILLEUX
Météo-France, Toulouse, France

(Received 19 May 1992; revised 19 August 1992)

SUMMARY

The paper describes a four-dimensional variational assimilation using a multilevel global primitive-equation spectral model. The experiments consist in minimizing the distance between the model solution and conventional observations spread over 24 hours.

A first set of experiments was performed at low resolution (spectral truncation T21). The model solution converges to fit the observations satisfactorily over the whole assimilation period. The meteorological quality of the analyses, in terms of the large-scale part of the flow, is comparable with the ECMWF operational analysis. Confirming early results, the control of gravity waves in the solution can be easily handled using a combination of a penalty term and a normal-mode initialization scheme within the forward–backward process of the variational assimilation. Experiments excluding data over an area with a strong baroclinic development show that the information contained in the dynamics of the model is used successfully in the analysis over this area. In addition, the dynamics are able to infer implicitly flow-dependent structure functions.

Higher-resolution experiments (spectral truncation T42) were performed and compared with optimum interpolation (OI) analyses performed under the same conditions. Although the lack of a first-guess term in the cost function leads to small-scale noise generation if the minimization is pursued, the quality of the variational assimilation is comparable with OI, and the error growth of the following 24-hour forecast is smaller when performed from the variational analysis. Moreover, experiments excluding wind data from aircraft show a clear advantage for the variational approach against OI in using, in a consistent way, the information coming from the model and the observations.

1. INTRODUCTION

The problem of four-dimensional assimilation of observations in meteorology can be stated as: how to find the best initial conditions of a numerical weather forecast using all the available information distributed over a given time period—namely the observations and the dynamics of the model. Moreover, the advent of remote-sensing data, which will provide a large quantity of new sources of information, makes it crucial to find a method which easily handles asynoptic data properly and whose link with the model variables is not necessarily trivial.

Several techniques of assimilation have been used so far (see the extensive review given by Ghil and Malanotte-Rizzoli (1991)). In particular, the optimum interpolation (OI) scheme has been broadly used amongst most operational weather centres (see Lorenc (1981) and Shaw et al. (1987) for a thorough description of the method and its implementation at the European Centre for Medium-range Weather Forecasts (ECMWF)). Several weaknesses inherent in the method and its practical implementation are now identified. For instance, the OI analysis extracts information poorly from observations nonlinearly related to the model variables (see Andersson et al. 1991). The data selection algorithm is also generally a source of noise in the final analysis (Parrish and Derber 1992).

The variational approach circumvents some of the practical OI weaknesses, since it allows the analysis to use all the observations at every model grid point, and can easily

* Corresponding author.
handle a non-trivial link between the observations and the model state. In its four-dimensional version, the method consistently uses the information coming from the observations and the dynamics of the model. Over a limited period of time, it produces the same result as the full extended Kalman-filter approach (see Thépaut and Courtier (1991) hereafter referred to as TC91, and Daley (1991)) and for a much lower cost.

These advantages led to a considerable effort being undertaken at the ECMWF and Météo-France to develop a global variational assimilation scheme in its three-dimensional (3D-VAR) and four-dimensional (4D-VAR) versions (Courtier et al. 1991). It may be noted that a 3D-VAR analysis has already been successfully implemented operationally at the National Meteorological Center (USA) (Derber et al. 1991), giving consistently better analyses and forecasts when compared with their classical OI system.

Following the early encouraging 4D-VAR studies performed along the lines developed by Lewis and Derber (1985) and LéDimet and Talagrand (1986) and with simple models (Courtier and Talagrand 1987; Derber 1987; Courtier and Talagrand 1990), experiments have been recently carried out with more sophisticated models, namely primitive-equation models. These recent 4D-VAR experiments have proved to be successful (TC91; Rabier and Courtier 1992; Navon et al. 1991) in identical-twin-experiment context. The ability of 4D-VAR to extract information on temporal evolution from the dynamics of the numerical model has been particularly addressed in an idealized case in Rabier and Courtier (1992).

Despite the high cost of the method and remaining problems (such as, for instance, the introduction of non-differentiable processes in the model), the theoretical advantages of 4D-VAR compared with the current operational data-assimilation systems make it a good candidate for a possible future operational assimilation scheme.

It then becomes crucial to study the potential of the method in a more realistic context. As a natural continuation of the previous works, this paper investigates the feasibility of a ‘truer-size’ 4D-VAR, taking advantage of having a primitive-equation model and of using the real conventional observations available on the Global Telecommunication System (GTS) that are currently used in most weather centres.

Firstly, a brief description of the 4D-VAR method and of all the ingredients involved in the assimilation process is presented in section 2. A first set of experiments performed at low resolution (spectral model truncated at wave number 21) is then presented in section 3. Problems such as the speed of convergence, the control of gravity waves and the impact of sparsely distributed data are particularly addressed. The second set of experiments was performed at higher resolution (truncation of the model at wave number 42). The good meteorological quality of the assimilations and the subsequent forecasts is demonstrated in section 4 through comparison with OI analyses performed at the same resolution and in similar conditions. General conclusions and remaining problems are discussed in section 5.

2. THE METHOD

As stated above, the principle of variational assimilation is very simple: the goal is to find a model solution which is as close as possible (in a least-square sense) to the observations available over a time period \([t_0, t_1]\). The model trajectory is constrained to fit observations by adjusting its initial conditions.

Let us define \(x_0\), the model state at time \(t_0\). We then introduce a scalar function, called a cost function, measuring the discrepancy between the model solution integrated from \(x_0\) and the observations available over \([t_0, t_1]\).
A natural expression for the measure of the discrepancy is given by:

\[
\hat{\mathcal{J}}_o(x_0) = \frac{1}{2} \sum_{i=0}^{n} (H_i(x_i) - y_i)^T O_i^{-1}(H_i(x_i) - y_i)
\]  

(1)

where \(H_i\) represents the observation operator relating the observation vector \(y_i\) to the model state variable \(x_i\) at time \(t_i\). Superscripts T and \(-1\) denote matrix transpose and inverse respectively. \(O_i\) represents the observation-error covariance matrix, which takes into account errors in the operator \(H_i\), representativeness errors and instrumental errors (Lorenc 1986).

Let us now introduce the atmospheric model \(\mathcal{M}\) leading to the time evolution of \(x\):

\[
x_{i+1} = \mathcal{M}(t_{i+1}, t_i)x_i
\] 

(2)

Following the notations adopted in TC91, we introduce the tangent-linear model, \(\mathcal{R}\), and its adjoint:

\[
\delta x_{i+1} = \mathcal{R}(t_{i+1}, t_i)\delta x_i
\] 

(3)

\[
\delta x_{i+1}' = \mathcal{R}^*(t_{i+1}, t_i)\delta x_i'
\] 

(4)

where \(\delta x\) is a perturbation in the vicinity of a trajectory \(x\) and \(\delta x'\) is called the adjoint variable (the superscript * denotes the adjoint). We will then try to minimize \(\hat{\mathcal{J}}_o\) in the model space at time \(t_0\). The minimization algorithm requires the gradient of the cost function with respect to \(x_0\).

To a perturbation \(\delta x_0\) of \(x_0\) corresponds a perturbation \(\delta \hat{\mathcal{J}}_o\) of the cost function \(\hat{\mathcal{J}}_o\) which characterizes the gradient of \(\nabla_{x_0} \hat{\mathcal{J}}_o\) by:

\[
\delta \hat{\mathcal{J}}_o = \langle \nabla_{x_0} \hat{\mathcal{J}}_o, \delta x_0 \rangle
\] 

(5)

where \(\langle , \rangle\) is an inner product defined in the model space.

The use of (4), to which we add a forcing term as described in Talagrand and Courtier (1987), leads to the so-called adjoint equation:

\[
\delta x_{i+1}' = \mathcal{R}^*(t_{i+1}, t_i)\delta x_{i+1}' + H_i^*O_i^{-1}(H_i(x_i) - y_i)
\] 

(6)

where \(H^*\) represents the adjoint of the tangent-linear operator of \(H\). The forcing term corresponds indeed to the contribution of the particular observation \(y_i\) to the gradient of \(\hat{\mathcal{J}}_o\) with respect to \(x_i\).

The backward integration of (6) starting from \(H_n^*O_n^{-1}(H_n(x_n) - y_n)\) and pursued to time \(t_0\) gives the value of the gradient of the cost function with respect to \(x_0\):

\[
\delta x_0' = \nabla_{x_0} \hat{\mathcal{J}}_o.
\] 

(7)

The observations are not necessarily the only source of information that we want to be taken into account to define the analysed model state. For instance Courtier and Talagrand (1990) have shown that a reasonable mass–wind balance of the analysis could not be obtained if the model solution was only constrained to fit the observations.

It is then possible to introduce additional terms in the cost function to force the model solution to obey a priori known physical properties (for instance the geostrophic equilibrium). Let us call the term containing these additional weak constraints \(\hat{\mathcal{J}}_c\) (following the terminology defined in Sasaki (1970)). The gradient contribution \(\nabla_{x_0} \hat{\mathcal{J}}_c\) will then be required for our minimization problem.

In the experiments presented below, only the observation term \(\hat{\mathcal{J}}_o\) and the constraint term \(\hat{\mathcal{J}}_c\) have been used. In particular, we have not considered the introduction of a first-
guess term taking into account a priori knowledge of the model state $x_0$. The consequence of excluding this term will be discussed in section 4 and in the conclusion. The total cost function to minimize then becomes:

$$\tilde{J} = \tilde{x}_0 + \tilde{J}_c. \quad (8)$$

(a) The model and its adjoint

The primitive-equation model used in the following experiments is the IFS/ARPEGE code (cycle 8) which is described in Courtier et al. (1991). The vertical coordinate is the hybrid pressure/sigma coordinate (Simmons and Burridge 1981). The model uses the spectral representation of the field $x$ (vorticity $\xi$, divergence $D$, temperature $T$, specific humidity $q$ and logarithm of surface pressure ln$\sigma$) based on spherical harmonics with a triangular truncation at degree 21 or 42, depending on the experiments. The number of vertical levels is 19.

No physical parametrization was used in the model, but a horizontal diffusion close to the one described in TC91 was introduced. A refinement of the horizontal diffusion for temperature was added so that it approximate diffusion on pressure levels to avoid unrealistic warming of mountain tops. The option of a collocation grid with reduction of the number of points in the vicinity of the poles (Hortal and Simmons 1991) was used in the T42 experiments to save computing time. The numerical validation of the tangent-linear and adjoint models was performed following the tests described in TC91.

(b) The minimization algorithm

The minimization problem involved in these 4D-VAR experiments can be considered as 'large scale' (Gilbert 1991). Indeed, the number of degrees of freedom in the control variable is either 37268 for experiments performed at truncation T21 or 142373 for experiments at truncation T42. An appropriate robust and efficient descent algorithm is then required.

The following experiments were performed with a variable-storage quasi-Newton-type algorithm provided by the Institut National de Recherche en Informatique et en Automatique (INRIA), France. A description of the algorithm and the performance of the code are given in Gilbert and Lemaréchal (1989). Essentially, the method uses the available in-core memory provided by the user to update an approximation of the inverse Hessian matrix of the cost function. In the experiments presented here, storage was provided to allow ten updates of the Hessian matrix, this number being appropriate for obtaining a good efficiency of the algorithm (Gilbert, personal communication).

Once the memory is used up, the quasi-Newton matrix (approximation of the inverse Hessian) keeps being modified during the minimization process by dropping information coming from the oldest gradient, and inserting information coming from the more recently computed gradient.

The norm $\| \cdot \|$ in the minimization space has been chosen as follows:

$$\| x \|^2 = \frac{1}{2} \int_\xi \left[ \nabla \Delta^{-1} \xi \cdot \nabla \Delta^{-1} \xi + \nabla \Delta^{-1} D \cdot \nabla \Delta^{-1} D + R_s T_s (\ln \pi)^2 + \right.$$

$$\left. \frac{C_p}{T_r} T^2 + \frac{L_v}{q_r} q^2 \right] d\Sigma \left( \frac{\partial p_r}{\partial \eta} \right) d\eta \quad (9)$$

that is a quadratic invariant of the primitive equations linearized in the vicinity of a state of rest defined by a uniform surface pressure $\pi$, and a uniform temperature $T_r$ (Talagrand 1981). $R_s$ is the gas constant for the dry air, $C_p$ the specific heat at constant pressure, $p$
the pressure and $\eta$ the vertical hybrid coordinate. This energy form for the inner product is very close to the one described in TC91. An additional term has been introduced for the humidity component, taking into account the effects of latent heat in the Bernoulli equation (Gill 1982, p. 83). The latent heat of vaporization of water, $L_v$, is $2.5 \times 10^6$ J kg$^{-1}$, the reference specific humidity $q_r$ has been empirically taken equal to $10^{-5}$ kg kg$^{-1}$.

In the low-resolution experiments, we stopped the minimization process after 60 iterations, or once the norm of the gradient had been reduced by a factor of 100. Unless stated otherwise, the minimization process was interrupted after 40 iterations for higher-resolution experiments.

It may be noted that one iteration of the minimization algorithm basically consists in determining the descent direction and the optimal step-size (automatically computed inside the algorithm) which will minimize the cost function along that direction. The determination of the optimal step-size along the descent direction (called the line-search process) may require several computations of the cost function and of its gradient. This operation (function–gradient evaluation) will be referred to as a ‘simulation’, following the mathematical language introduced in Gilbert and Lemaréchal (1989). It is indeed the number of simulations involved during the minimization that determines the cost of the method.

(c) The $\tilde{\zeta}$ term

The observations used in the following low-resolution experiments consist of all the conventional observations (which passed the ECMWF operational OI quality control) for the 24-hour period from 12 UTC 9 February 1989 to 12 UTC 10 February, except for SATEMs (satellite temperature/humidity soundings), 10 m wind and 2 m temperature

<table>
<thead>
<tr>
<th>Time slot</th>
<th>SYNOP</th>
<th>AIREP</th>
<th>SATOB</th>
<th>DRIBU</th>
<th>TEMP</th>
<th>PILOT</th>
<th>PAOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 h</td>
<td>2927</td>
<td>130</td>
<td>399</td>
<td>22</td>
<td>416</td>
<td>25</td>
<td>156</td>
</tr>
<tr>
<td>13 h</td>
<td>4</td>
<td>155</td>
<td>0</td>
<td>23</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14 h</td>
<td>1</td>
<td>143</td>
<td>0</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15 h</td>
<td>181</td>
<td>144</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16 h</td>
<td>16</td>
<td>126</td>
<td>312</td>
<td>4</td>
<td>1</td>
<td>41</td>
<td>0</td>
</tr>
<tr>
<td>17 h</td>
<td>63</td>
<td>127</td>
<td>485</td>
<td>2</td>
<td>10</td>
<td>75</td>
<td>0</td>
</tr>
<tr>
<td>18 h</td>
<td>3434</td>
<td>94</td>
<td>422</td>
<td>46</td>
<td>115</td>
<td>88</td>
<td>0</td>
</tr>
<tr>
<td>19 h</td>
<td>27</td>
<td>56</td>
<td>0</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20 h</td>
<td>18</td>
<td>78</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21 h</td>
<td>126</td>
<td>87</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22 h</td>
<td>3</td>
<td>100</td>
<td>364</td>
<td>9</td>
<td>38</td>
<td>56</td>
<td>0</td>
</tr>
<tr>
<td>23 h</td>
<td>93</td>
<td>105</td>
<td>466</td>
<td>18</td>
<td>177</td>
<td>72</td>
<td>0</td>
</tr>
<tr>
<td>24 h</td>
<td>3075</td>
<td>106</td>
<td>480</td>
<td>35</td>
<td>448</td>
<td>42</td>
<td>145</td>
</tr>
<tr>
<td>01 h</td>
<td>25</td>
<td>138</td>
<td>0</td>
<td>40</td>
<td>9</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>02 h</td>
<td>3</td>
<td>141</td>
<td>0</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>03 h</td>
<td>211</td>
<td>175</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>04 h</td>
<td>3</td>
<td>195</td>
<td>0</td>
<td>13</td>
<td>1</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>05 h</td>
<td>90</td>
<td>165</td>
<td>501</td>
<td>28</td>
<td>11</td>
<td>112</td>
<td>0</td>
</tr>
<tr>
<td>06 h</td>
<td>3382</td>
<td>156</td>
<td>478</td>
<td>32</td>
<td>91</td>
<td>99</td>
<td>0</td>
</tr>
<tr>
<td>07 h</td>
<td>30</td>
<td>118</td>
<td>0</td>
<td>31</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>08 h</td>
<td>3</td>
<td>88</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>09 h</td>
<td>188</td>
<td>98</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>10 h</td>
<td>15</td>
<td>108</td>
<td>298</td>
<td>5</td>
<td>15</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>11 h</td>
<td>77</td>
<td>118</td>
<td>522</td>
<td>28</td>
<td>214</td>
<td>71</td>
<td>0</td>
</tr>
<tr>
<td>12 h</td>
<td>3328</td>
<td>130</td>
<td>490</td>
<td>49</td>
<td>433</td>
<td>27</td>
<td>146</td>
</tr>
</tbody>
</table>

Observations are packed in one-hour time slots centred on each hour.
measurements from SYNOP-SHIPs (synoptic land and sea surface data) and DRIBUs (measurements from drifting buoys). More precisely, all measurements of surface pressure or geopotential from SYNOP-SHIPs, DRIBUs and PAOBS (Australian pseudo observations), geopotential heights, relative humidity and wind components from TEMPS, wind components from PILOTs (both radiosonde measurements), AIREPs (aircraft wind data) and SATOBs (wind measurements derived from geostationary satellites) were used.

These observations are performed mainly at synoptic hours, but some of them are available at intermediate times. Table 1 summarizes the temporal distribution of the different observation types during the assimilation period. Figure 1 shows the geographical distribution of SYNOP-SHIPs, AIREPs, TEMPS and DRIBUs (main sources of information) during this period and for the northern and southern hemispheres,
pointing out the dramatic difference of data coverage between the two hemispheres and between land and ocean.

Indeed, the 4D-VAR approach allows the use of observations at the appropriate time, but for practical reasons the observations used in the experiments described below were packed by time slots of one hour (centred on each hour). It should be noted that the quality control of the observations has been performed within the ECMWF operational assimilation at a truncation of T106. We can then first expect a mismatch with a T21 model, especially in the vicinity of mountains, since the surface pressure observations were not corrected for height differences between T21 and T106 orography. Moreover, a temporal redundancy check on data like DRIBUs or AIREPs is done operationally so that only the report closest to the analysis time is retained (within a temporal window of 6 hours). Only 15 standard levels from the radiosonde observations are used in the
analysis for practical reasons. The results of a 4D-VAR experiment should be improved by a relaxation of these simplifications.

In Eq. (1) we can see that an observation operator $H$ is required for each observed quantity. The common part of all these observation operators consists in inverse spectral transforms to go from spectral to grid-point representation, bilinear horizontal interpolation (performed on the hybrid coordinate surfaces) from the grid points to the observation points, and vertical interpolation from the model levels to the observation levels.

Similarly to the operational OI, observation errors are assumed uncorrelated by observation types. Moreover, we assume a zero correlation in the horizontal between errors in different observations of the same type. The cost function can then be split into single cost functions corresponding to each observation, and the $O$ matrix reduces to the appropriate error variance. An exception is made for geopotential heights from radiosondes which are assumed to be correlated in the vertical. The error standard deviations used in the assimilation are displayed in Table 2.

### TABLE 2. OBSERVATION-ERROR STANDARD DEVIATIONS USED IN $\chi_v$ FOR EACH TYPE OF OBSERVATION

<table>
<thead>
<tr>
<th>Pressure (hPa)</th>
<th>TEMP cat. 1 height (m)</th>
<th>TEMP cat. 2 height (m)</th>
<th>TEMP cat. 3 height (m)</th>
<th>PILOT, TEMP wind (m s$^{-1}$)</th>
<th>AIREEP wind (m s$^{-1}$)</th>
<th>SATOB wind (m s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>4.3</td>
<td>6.5</td>
<td>8.6</td>
<td>2.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>850</td>
<td>4.4</td>
<td>6.6</td>
<td>8.8</td>
<td>2.4</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>700</td>
<td>5.2</td>
<td>7.8</td>
<td>10.4</td>
<td>2.5</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>500</td>
<td>8.4</td>
<td>12.6</td>
<td>16.8</td>
<td>3.4</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>400</td>
<td>9.8</td>
<td>14.7</td>
<td>19.6</td>
<td>3.6</td>
<td>3.5</td>
<td>6.0</td>
</tr>
<tr>
<td>300</td>
<td>10.7</td>
<td>16.1</td>
<td>21.4</td>
<td>3.8</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>250</td>
<td>11.8</td>
<td>17.7</td>
<td>23.6</td>
<td>3.2</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>200</td>
<td>13.2</td>
<td>19.8</td>
<td>26.4</td>
<td>3.2</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>150</td>
<td>15.2</td>
<td>22.8</td>
<td>30.4</td>
<td>3.2</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>100</td>
<td>18.1</td>
<td>27.2</td>
<td>36.2</td>
<td>2.2</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>70</td>
<td>19.5</td>
<td>29.3</td>
<td>39.0</td>
<td>2.0</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>50</td>
<td>22.5</td>
<td>33.8</td>
<td>45.0</td>
<td>2.0</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>30</td>
<td>25.0</td>
<td>37.5</td>
<td>50.0</td>
<td>2.0</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>20</td>
<td>32.0</td>
<td>48.0</td>
<td>64.0</td>
<td>2.5</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>10</td>
<td>40.0</td>
<td>60.0</td>
<td>80.0</td>
<td>3.0</td>
<td>4.0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SHIP height (m)</th>
<th>PAOB height (m)</th>
<th>SYNOP (land) height (m)</th>
<th>DRIBU height (m)</th>
<th>SYNOP-SHIP wind (m s$^{-1}$)</th>
<th>DRIBU wind (m s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.0</td>
<td>32.0</td>
<td>7.0</td>
<td>14.0</td>
<td>3.6</td>
<td>5.4</td>
</tr>
</tbody>
</table>

For the upper-air observations, the standard deviations are given for the standard 15 pressure levels.
- cat. 1: North America (80°N–30°N, 50°W–170°W)
- cat. 2: All except cat. 1 and cat. 3.

(d) The $\chi_v$ term

As stated above, one can seek as the best analysis a solution close to the observations and to the so-called slow manifold (Leith 1980). As shown in Courtier and Talagrand
(1990), one way to achieve this is to add to the cost function a penalty term on the
tendency of the gravity modes $\partial G/\partial t$ of the analysed state. We then define $\mathcal{C}_c$ as follows:

$$\mathcal{C}_c = \alpha_c \left| \frac{\partial G}{\partial t} \right|^2$$

(10)

with $\| \cdot \|$ defined by (9).

The coefficient $\alpha_c$ has been chosen empirically as $10^7 \text{s}^4 \text{kg}^{-1}$.

The calculations of $\mathcal{C}_c$ and of its gradient require one time-step integration of the
model to determine $G$, the computation of the norm, and one backward time-step
integration of the adjoint model.

We will see in section 3(c) that another way to control the gravity waves in the
solution can be performed through the use of a normal-mode initialization and of its
adjoint within the forward–backward integration process (see TC91).

3. NUMERICAL EXPERIMENTS

The 24-hour 4D-VAR assimilations have been performed mainly at low resolution
(truncation T21) and their quality is evaluated by comparing them with the operational
OI analysis.

Figures 2(a) and (b) represent the northern hemisphere 500 hPa geopotential fields
(top panels) and mean-sea-level pressure fields (bottom panels) for, respectively, 12 UTC
9 February 1989 and 12 UTC 10 February 1989. These fields come from the T21 truncated
representation of the T106 ECMWF operational analysis.

The meteorological situation is characterized by a block centred over central Europe.
The surface low immediately to the west of the block was filling quickly between 9 and
10 February over the eastern part of the Atlantic Ocean, while cyclogenesis was taking
place in the west Atlantic Ocean (low moved eastward from 60°W to 40°W while
deepeining by 30 hPa). In the Pacific Ocean, one can see that at 500 hPa a ridge moved
eastward from 165°E to 180°E while intensifying. On the surface map, the corresponding
anticyclone (1026 hPa) centred at 170°E, 35°N moved east to be replaced 24 hours later
by a frontal zone. Temperature fields at 500 hPa and 850 hPa (not shown here) indicate
a fairly strong baroclinic signature. The analysed field in Fig. 2(b) will be considered as
the truth to evaluate the quality of the 4D-VAR assimilations. Figure 3, which represents
the same field as Fig. 2(b) but for the southern hemisphere, will allow us to measure the
quality of 4D-VAR over this area. We will, for instance, focus more particularly on the
500 hPa ridge and its associated surface anticyclone located south-east to east of the New
Zealand area where a fair amount of data are present. Another ridge located at 115°W
is of interest because it is located in an area almost completely void of data.

(a) Reference experiment

As a first basic experiment we performed a T21 4D-VAR assimilation aimed at
finding a solution close to the observations and to the slow manifold. The terms $\mathcal{C}_c$ and
$\mathcal{C}_d$ described above were used. In practice the initial point of the minimization can come
from a previous estimate of the atmosphere at the beginning of the assimilation period.
Moreover, the minimum of the cost function should not depend on the initial point
of the minimization, at least as long as the tangent-linear hypothesis is valid both for the
observation operators and the model. At T21 the dimension of the control variable is
37268 and the set of observations described above provides 120000 pieces of information.
The $\mathcal{C}_c$ constraint applied on the first five vertical modes provides 4840 pieces of
Figure 2. Meteorological situation in the northern hemisphere for (a) 12 UTC 9 February 1989 and (b) 12 UTC 10 February 1989. Top panels: 500 hPa geopotential field, bottom panels: mean-sea-level pressure.
Figure 3. As Fig. 2(b) but for the southern hemisphere.
information. The problem is then globally over-determined, but with possible local under-determination in data-void areas.

It is interesting, in the absence of a first-guess term, to study the behaviour of 4D-VAR, in particular in areas with little data. The initial point for the minimization was chosen far from the presumed solution (we took the ECMWF initialized analysis for 12 UTC 6 February 1989). The subsequent 24-hour forecasts displayed in Figs. 4 and 5 are to be compared with the analyses displayed respectively in Fig. 2(b) and Fig. 3. Note the huge differences in several respects between the fields. As far as the northern hemisphere is concerned, the ridge located at 170°E in Fig. 2(b) is replaced by a trough, and the circulation in the Atlantic Ocean at 50°N is zonal. For the southern hemisphere, large differences can be noticed in the two areas of interest mentioned above. In particular, a trough is situated at 170°W with an associated surface low of 965 hPa centred at 60°S.

The minimization experiment was stopped after 57 simulations, the norm of the gradient having been reduced by a factor of 100. Seventy per cent of the decrease of the cost function was obtained in about 20 simulations, at which point the observation cost function was reduced by almost one order of magnitude. We also noticed that after an overshoot at the beginning of the minimization the $\partial_\nu$ term remained close to zero. Figure 6 represents the contribution of individual times to the time integral which defines the total observation cost function, before and after minimization. Indeed, since the vertical coordinate is logarithmic and the two curves are nearly parallel, one can see that the fitting to observations seems rather uniform over the assimilation period. Moreover, we noticed a decrease of the model–observation difference for all observation types. However, the model–observation difference after minimization was slightly, but systematically, larger than the specified observation-error standard deviations (by 2 to 3 m s$^{-1}$ for the wind in the northern hemisphere). This could be due to incorrectly specified observational errors, since no inflation has been applied on the observation-error standard deviations to account for a greater representativeness error at truncation T21. Several other factors may also be important, such as the use of an imperfect assimilating model or the lack of convergence after 57 simulations.

Between the beginning and the end of the minimization the global root-mean-square (r.m.s.) difference between the model solution at the end of the assimilation period and the ECMWF operational analysis decreased from 14.2 to 8.2 m s$^{-1}$ for the $u$-component of the field, from 16.0 to 7.2 m s$^{-1}$ for the $v$-component of the field, from 3.8 K to 2.0 K for the temperature field and from 6.3 g kg$^{-1}$ to 5.5 g kg$^{-1}$ for the specific-humidity field. These figures remain large compared with the fit of the ECMWF operational analysis to the observations (typically 2.0 m s$^{-1}$ for the zonal wind as described by Lönneberg (1988)).

Figures 7 and 8, to be compared with Fig. 2(b) and Fig. 3, represent the final model solution at the end of the period of assimilation for, respectively, the northern and southern hemispheres. In the northern hemisphere all the large structures have been reconstructed in the mid latitudes. In particular, the frontal system in the Pacific Ocean has been reasonably analysed, although with too weak an intensity. The Atlantic depression (40°W, 60°N), which was not present in the initial solution, is also now present and properly located.

In the southern hemisphere (Fig. 8) the meteorological structures have been reconstructed and located satisfactorily, despite the overall lack of data. In particular the anticyclone located at 170°W, 45°S is present in the final solution. The associated ridge at 500 hPa is, however, much too weak. In the area void of data between 100°W and 120°W we can notice on the 500 hPa and surface map the poor performance of the assimilation in reconstructing the ridge.
Figure 4. Forecast from the initial point of the minimization (12 UTC 6 February 1989 analysis). Top-panel: 500 hPa geopotential field over the northern hemisphere, bottom-panel: mean-sea-level pressure field over the northern hemisphere.
Figure 5. As Fig. 4 but for the southern hemisphere.
(b) Impact of the lack of data

When comparing Fig. 7 and Fig. 8 with, respectively, Fig. 2(b) and Fig. 3 one clearly sees that even if 4D-VAR produces fields similar to the ones produced by the operational OI in the northern hemisphere, there are noticeable differences in the southern hemisphere. Bearing in mind that over this area the operational OI assimilation was performed at higher resolution, using satellite data and accurate first-guess information provided by a sophisticated model, one can still reasonably consider the OI analysis as the ‘truth’. The differences between the two fields then reveal that the poorer 4D-VAR performance is due to the general lack of conventional observations in the southern hemisphere.

We then studied in a simple framework the following problem: to what extent in a four-dimensional context can the dynamics of the model compensate for the deficiency of data coverage? The Pacific area, where a fairly strong baroclinic structure is developing, was chosen for the study of this problem.

A first experiment was carried out on the basis of the reference field, but excluding all the observations (360 SYNOPs, 34 TEMPS, 458 AIREPs and 152 SATOBs) from the area 160°E–180°W, 20°N–60°N. At 45°N it corresponds to a zonal distance of roughly 1800 km; therefore with a typical mean flow velocity of 50 m s⁻¹, an advected particle has enough time to cross this area.

The result of the assimilation (after 56 simulations) is shown in Fig. 9 which represents the 500 hPa geopotential and temperature fields (top panel) and the sea-level pressure field (bottom panel) over the area of interest and for the final model solution.

Of note is that the perturbed system, absent from the initial solution, has been retrieved with a good accuracy at the surface. The 500 hPa geopotential and temperature fields clearly indicate a baroclinic signature in the analysed field. The dynamics of the model has therefore been able to infer information on the vertical structure of the atmosphere in the masked area. However, the systems are slightly less intense. The anticyclone is 3 hPa weaker than for the assimilation using all the observations, the depression 1 hPa weaker.
Figure 7. As Fig. 4 but the forecast is from the final point of the minimization (57 simulations).
Figure 8. As Fig. 7 but for the southern hemisphere.
Figure 9. Final model solution at the end of the assimilation period. The assimilation has been performed excluding all the observations from the area 160°E–180°E, 20°N–60°N. Top-panel: 500 hPa geopotential and temperature field, bottom-panel: mean-sea-level pressure field.
When increasing the masked area to 150°E–170°W, 10°N–60°N, corresponding to the exclusion of 415 SYNOPs, 44 TEMPS, 638 AIREPs and 410 SATOBS, Fig. 10 shows that, to a certain extent, the frontal signature at the surface is still present in the final solution, as is the 500 hPa temperature ridge. However, both the anticyclone and the depression are much less intense, as we can see from the surface pressure gradient between 160°E and 180°E.

The spread of the difference between the increased masked-area field and the reference one is interesting. Figure 11 represents the geopotential differences at 500 and 250 hPa between the two experiments.

The negative maximum difference located at 45°N, 180°W occurs at the maximum of the geopotential ridge, whereas the positive patterns occur in low-height areas (see Fig. 10), showing the difficulties that the model has in intensifying structures in the total absence of observations. More interesting is the downstream spread of the error patterns, showing that information coming from the upstream observations has been advected by the flow in the immediate adjacent area, whereas the quality of the analysis eastward of the masked area is contaminated by the upstream lack of data. It is worth mentioning that the error pattern has an upstream component at time \( t_0 \) (not shown), stressing that the adjustment of the model solution to the observations is temporally global, leading to a downstream advection into the future as well as an upstream advection into the past. This example of advection of information by the dynamics confirms the results found by Courtier and Talagrand (1987) with a barotropic vorticity-equation model. Moreover, we can see from Fig. 11 that the differences between error maxima at 500 hPa and 250 hPa are out of phase. However, with this experiment it is difficult to know if this phase difference comes from the fact that about the same number of observations were present at these two levels, giving a complete picture of the perturbed system, or if it comes from the propagation of information in the vertical through more complex dynamical processes.

To understand the mechanism of vertical propagation of information, a 4D-VAR experiment was run by only excluding AIREPs from the above mentioned area (95% of them being confined between 200 and 250 hPa, none below 350 hPa). For comparison with the experiment where all the observations were used, the minimization algorithm was forced to perform up to 100 simulations for both assimilations to make sure that differences did not reflect convergence problems. Figure 12 represents the geopotential differences at 250 and 500 hPa between the two experiments. Although the geopotential difference is much larger at 250 hPa, wind observations at 250 hPa have a clear impact on the geopotential field at 500 hPa.

Figure 13 shows a cross-section of the atmosphere from 60°N–150°E to 10°N–170°W, a direction roughly orthogonal to the position of the frontal structure. The background field (dashed thin lines) is the vorticity field for the reference experiment, showing an alternation of positive, negative and again positive regions. One can notice a vertical tilt of the vorticity field characteristic of a baroclinic perturbation. The thick full lines represent the vorticity-difference field between the experiment where AIREPs were excluded and the one where all the observations were used.

The error maximum occurs roughly where the AIREP data are missing. It is interesting that the error is vertically spread down to the ground and the vertical-difference pattern gets the tilted structure of the original perturbation. In particular the negative pattern with an extremum located at 177°E, 25°N and at 250 hPa spreads vertically along the vorticity front itself and leads to a second extremum at 170°W, 10°N near the ground. In fact this difference field is a measure of how the information contained in the AIREPs has been transported in their vicinity through the dynamics of the model.
Figure 10. As Fig. 9 but the assimilation has been performed excluding all the observations from the area 150°E−170°W, 10°N−60°N.
Figure 11. Geopotential differences at (top) 500 hPa and (bottom) 250 hPa between the reference 4D-VAR experiment and the experiment where all the observations have been excluded from the area 150°E–170°W, 10°N–60°N. Differences are shown for 12 UTC 10 Feb. 1989.
Figure 12. Geopotential differences at (top) 500 hPa and (bottom) 250 hPa between the reference 4D-VAR experiment and the experiment where the AIRePs have been excluded from the area 150°E-170°W, 10°N-60°N. Differences are shown for 12 UTC 10 Feb. 1989 after 100 simulations of the minimization algorithm.
Figure 13. Cross-section of vorticity differences (full lines) between the reference experiment and the one without AIR-EPs. The background field (dashed lines) is the vorticity field produced by the reference experiment. Contour intervals are $5 \times 10^{-6}$ s$^{-1}$ for the background field and $10^{-6}$ s$^{-1}$ for the field of differences.

It gives an insight into the structure functions implicitly used in 4D-VAR.

The vertical spread of the vorticity difference displayed in Fig. 13, intrinsically linked to the four-dimensional nature of the assimilation, is very promising because it shows that 4D-VAR is able to specify flow-dependent structure functions. As a consequence, in this particular case of baroclinic instability, the vorticity-difference extrema show that the model dynamics were able to move a front thanks to wind observations at 250 hPa. In addition, a similar shape for the geopotential-error pattern was found (not shown here). This mass–wind coupling shows the multi-dimensionality of the variational approach.

(c) Control of gravity waves

As said in section 2, the introduction of a $\tilde{\zeta}_c$ term defined by Eq. (10) constrains the model solution to remain close to the so-called slow manifold.

Bearing in mind that the initial point of the minimization is far from the expected solution, no major problem of convergence was encountered during the minimization process, although two trials in the line-search per iteration were needed on several occasions, and 57 simulations were needed altogether, to reduce the norm of the gradient of the cost function by 100. As a comparison only 50 simulations were needed when the weak constraint term $\tilde{\zeta}_w$ was not added to the observation term $\tilde{\zeta}_o$. A fair measure of the degree of balance of the field is given by the energy implied by the gravity-mode tendencies of the analysed field. We have also checked that the amount of gravity waves in the model solution was controlled during the minimization descent.

Another way of constraining the solution to lie on the slow manifold is to use a nonlinear normal-mode initialization (NNMI) scheme before each temporal integration of the model, and its adjoint after each adjoint model integration. The use of NNMI in idealized cases has proved to be successful to hasten convergence and to control to some extent the amount of gravity waves in the solution (TC91). However, Courtier and Talagrand (1990) and TC91 found that during the minimization process spurious gravity waves were generated in proportion to the degree of invertibility of the NNMI process. A combination of $\tilde{\zeta}_c$ and NNMI has been tried in a shallow-water context (Courtier and Talagrand 1990) and it turned out to have the advantages of both approaches.
Three additional experiments were conducted in the same conditions with only the following differences: the first one being an assimilation without any constraint on the gravity modes, the second one containing NNMI within the integration process, and the third one using the combination of \( \zeta_c \) and NNMI. The normal-mode initialization scheme was a mixed implicit–explicit scheme (Temperton 1988, 1989). The first five vertical modes were initialized and two iterations of the NNMI process were performed.

The results led to conclusions similar to those discussed by Courtier and Talagrand (1990) and by TC91, which can be summarized as follows:

(i) The total absence of constraint leads to an increase of the energy in the gravity components of the final solution.

(ii) The use of a \( \zeta_c \) term controls the gravity waves in the solution but can make the convergence slower. (However, we did not experience dramatic convergence problems as in Courtier and Talagrand (1990).)

(iii) The use of NNMI avoids the overshoot problem encountered in the ‘\( \zeta_c \) only’ experiment. It also hastens the convergence, but gravity waves can be excited if the minimization is pursued owing to the invertibility of the NNMI process.

(iv) The combination of NNMI and a \( \zeta_c \) term seems beneficial, NNMI speeding up the convergence and \( \zeta_c \) ensuring a balanced solution.

### 4. High-resolution experiments

In the previous section experiments were performed at truncation T21, the ECMWF operational analyses being taken as the truth to evaluate the quality of 4D-VAR. Another way of evaluating 4D-VAR is to compare the performance of the method with another assimilation system in conditions as similar as possible, in particular in terms of the utilized observations, error statistics and the resolution of the model. The coarsest resolution at which it was achievable to run an OI assimilation within the ECMWF experimental framework easily was T42. Moreover, the meteorological evaluation of 4D-VAR makes more sense at this resolution than at T21.

An OI T42 assimilation was run from 12 UTC 9 February 1989 to 12 UTC 10 February 1989 in the following conditions. Usually, observations are taken in a 6-hour temporal window around the nominal time of each analysis. To ensure that exactly the same set of observations was used in the OI as in the 4D-VAR, an exception was made so that only observations available less than 30 minutes before the first analysis and after the last analysis were used.

Another modification of the OI assimilation consisted of using the ECMWF operational model in its adiabatic version (both for the initialization and the 6-hour forecast part). It is worth stressing that although the model codes used for the OI and 4D-VAR are different, the parallel runs revealed that the two models produce very similar forecasts up to 4 days. This modification of the OI assimilation changed the quality control of the observations, so that the set of observations used in the T42 experiments is slightly different from those displayed in Table 1. However, in terms of quantity of observations, the relative temporal distribution over the 24-hour period was not changed significantly.

(a) General behaviour of 4D-VAR performed at T42

A 4D-VAR experiment was run at T42 using this new set of observations. As far as the control of gravity waves is concerned, a combination of \( \zeta_c \) and NNMI (two iterations) was applied. It is worth stressing that although the number of pieces of information coming from the observations is roughly the same as in the previous experiments, the dimension of the control variable is now 142373, which means that the problem is,
globally, slightly under-determined, despite the use of NNMI and the presence of $\tilde{\gamma}_c$ in the cost function.

The initial point for the minimization was taken equal to a 6-hour forecast available for 12 UTC 9 February 1989. The same field is used as the first-guess in the T42 OI assimilation. The minimization was run up to 40 simulations.

Figure 14 represents the variation of the cost function with the number of simulations. We note that the amount of energy in the gravity tendency, estimated with the $\tilde{\gamma}_c$ curve, remains low during the descent process, although small spikes are noticeable on the $\tilde{\gamma}_c$ curve during the minimization process. These spikes correspond in fact to two trials during the line-search within some iterations during the minimization process. Another important fact is that after 40 simulations the slope of the observational cost function indicates that convergence has not been reached. In fact the decrease of the cost function does not necessarily mean that the analysis is still being improved, but can correspond to the generation of small-scale noise adjusting to the observations. This phenomenon, described for instance in Courtier and Talagrand (1987) for variational experiments performed with a vorticity-equation model, is illustrated here in Fig. 15, which displays the mean squared difference of the vorticity field between the 4D-VAR solution and the ECMWF operational T106 analysis (truncated at T42) taken here as the truth. The distance is shown as a function of total wave number. The four curves correspond to the difference after 0, 10, 20 and 40 simulations of the minimization process. The solid line corresponds to the difference between a 24-hour forecast performed from the initial point of the minimization and the corresponding operational OI analysis. After 10 simulations the distance to the OI analysis has decreased over all the spectrum. The distance after

![Figure 14. Variation of the cost function with the number of simulations (T42 4D-VAR experiment): $\tilde{\gamma}_s$ (full line), $\tilde{\gamma}_c$ (dotted line).](image-url)
20 simulations still decreases for the large wave numbers, but the impact tends to be neutral for the small scales (from total wave number 25). After 40 simulations the distance is roughly similar to what it was after the first 10 simulations for the large-scale part of the field (up to wave number 12). It then becomes consistently larger for smaller scales to end up with values larger than the original distance for total wave numbers greater than about 30.

The last descent steps seem to consist essentially in generating small-scale noise to fit the available observations better. This phenomenon was even more spectacular on the meteorological maps at the beginning of the period of assimilation (not shown here). Indeed the presence of horizontal diffusion in the model tends to reduce the small-scales’ differences after 24 hours (the e-folding time of the horizontal diffusion in the model is typically 4 hours for the smallest scales). This result is a clear indication that at this model resolution, and with the observations currently used over 24 hours, a first-guess term is needed in the cost function to prevent the minimization from generating artificially noisy features.

Comparison of the 4D-VAR with the ECMWF operational analyses shows that the smallest r.m.s. differences were obtained for 20 simulations, to reach typical values at 500 hPa of 5 m s$^{-1}$ for the wind field and 1.3 K for the temperature field; these are still relatively large when compared with the fit of the T42 OI analysis to observations (3.5 m s$^{-1}$ for the wind).

(b) Comparison between 4D-VAR and OI analysis

Quality of the forecasts. In numerical weather prediction the best analysis can be defined as the analysis which will lead to the best forecast, provided a quality criterion
has been chosen. A 5-day adiabatic forecast was run from both the T42 OI and the 4D-VAR analyses. The forecasts were then verified every 12 hours against the ECMWF operational analyses.

Figure 16 represents the growth of the temperature (top panels) and kinetic energy (bottom panels) r.m.s. differences between the model solution and the verifying operational analyses at, respectively, 150 hPa (Fig. 16(a)) and 500 hPa (Fig. 16(b)) during a 5-day forecast performed from the 4D-VAR assimilation (thick line) and the T42 OI assimilation (dashed line).

At the analysis time the 150 hPa r.m.s. difference is larger for 4D-VAR than for OI, both for the temperature and kinetic energy fields. At 500 hPa the r.m.s. differences are very similar for OI and 4D-VAR. More generally, the verification of the analyses against the radiosondes (not shown here) led to the following conclusion: OI analysis

![Graphs showing temperature and kinetic energy differences over time](image)

Figure 16. Temporal variation over a 5-day forecast performed from 4D-VAR (full line) and OI (dashed line) of the root-mean-square error (verified against the ECMWF operational analyses) of the 150 hPa (a) and 500 hPa (b) temperature (top panels) and kinetic energy (bottom panels) fields.
was fitting the observations significantly better than 4D-VAR at the end of the assimilation period. A first explanation for the relatively poor performance of 4D-VAR compared with OI is that both methods did not use the same amount of information about the atmosphere. Indeed, the first-guess information was only used in the OI assimilation.

A second reason is that, although 4D-VAR is supposed to produce a good description of the atmospheric flow over the entire assimilation period, we know that the model-generated error is not taken into account in the variational process. A description (although crude) of this error exists in the OI formalism. Although it is difficult to evaluate accurately this kind of error over 24 hours, this definitely penalizes 4D-VAR.

However, when looking at the temporal evolution of the r.m.s. differences, the error growth of the subsequent forecast turns out to be smaller overall when integrated from 4D-VAR assimilation. For the 500 hPa temperature field the slopes are very similar

Figure 16. Continued.
during the 5-day period; and for the kinetic-energy field the error growth tends to become significantly larger for the forecast performed from OI after 3 days.

This forecast behaviour is confirmed at the 150 hPa level (a level roughly corresponding to the position of jet streams). The error growth of the forecast for the temperature field is slightly smaller when it is performed from 4D-VAR. Moreover, a clear advantage for 4D-VAR is noticeable after 3 days on the kinetic-energy curve. After 5 days the 150 hPa kinetic-energy r.m.s. error is 10% smaller in the 4D-VAR case. We have noticed that the improvement concerned essentially the rotational part of the wind (not shown).

This result is a strong indication that the forecast has benefited from the fact that the analysis produced by 4D-VAR was fully consistent with the dynamics of the model at the end of the assimilation period.

Transfer of dynamical information. One of the main conclusions drawn in section 3(b) was the ability of 4D-VAR to extract information from the dynamics and implicitly to specify flow-dependent structure functions. Since the ECMWF operational OI assimilation is four-dimensional it seemed interesting to compare the efficiency of the two approaches in that respect.

For this purpose, experiments were run in the same framework as in section 3(b). Two sets of two parallel assimilations were performed, the first set consisting of the T42 4D-VAR and OI assimilations just described, the second set consisting of the same assimilations but excluding AIREP data over the area 150°E–170°W, 10°N–60°N. Because of the higher resolution of the model the difference patterns between experiments with and without AIREPs were more localized than for the T21 experiments.

As in section 3(b), we studied the vertical spread of the differences between experiments with and without AIREPs, in a direction roughly orthogonal to the original perturbed system. Figure 17 represents a cross-section of the vorticity difference (thick full curves) between 4D-VAR experiments with and without AIREP data. The position of the system being slightly different with the T42 assimilation, the section is drawn along the direction 170°E–180°E, 20°N–60°N which is now the region of interest (alternation of a ridge and a trough at 250 hPa). The dashed curves show the vorticity field of the solution coming from 4D-VAR with AIREPs. The distribution of the vorticity field shows an alternation of positive and negative features, with again a north-eastward vertical tilt characteristic of our perturbed system. It is notable that there is good overall agreement between the vertical spread of positive vorticity differences (between 45°N and 30°N) and the original vertical tilt of the vorticity field. Even a secondary extremum located at 30°N, 177.5°E can be seen near the ground. The flow-dependent structure of the difference patterns confirms the results obtained in section 3(b).

Figure 18 represents the same cross-section but for the OI experiments. Although some differences from the previous section are noticeable in the background vorticity field (dashed lines) whose structure is more noisy, the basic features are similar. In particular a positive vorticity structure at 250 hPa is located at around 175°E, 40°N and spreading vertically down to the ground. The vorticity fronts also present a slightly tilted vertical structure.

The field of vorticity differences between the OI with and without AIREPs shows a much larger extremum than for 4D-VAR experiments in the area where data are missing. It is interesting to look at the vertical spread of the vorticity differences. Remembering that the observations are removed only at one level, the difference fields provide an image of the vertical-structure functions used in the assimilation. In the case of OI the vertical spread of the difference is purely barotropic, and the shape of the vertical spread looks relatively independent from the dynamics of the background.
Figure 17. Cross-section of vorticity differences (thick lines) between the T42 4D-VAR reference experiment and the one without AIREPs. The background field (dashed lines) is the vorticity field produced by the T42 4D-VAR reference experiment. Contour intervals are $10^{-4}$s$^{-1}$ for the background field and $4 \times 10^{-4}$s$^{-1}$ for the field of differences.

Figure 18. Same as Fig. 17 but for the T42 OI experiments.

meteorological situation. This propagation is in good agreement with the barotropic-structure functions specified in the OI, consisting of empirical exponential functions in the vertical. The cross-sections of geopotential differences (not shown here) show a similar signal.

On this example, 4D-VAR shows a use of the dynamical information present in the model equations that is much more consistent than for the OI four-dimensional assimilation.

5. Discussion and Conclusions

The results presented in this article confirm earlier results obtained with either simplified numerical models or idealized observations: four-dimensional variational
assimilation using the adjoint technique is numerically feasible. The assimilations performed at T21 produced analyses whose quality, in terms of reconstruction of meteorological features, looks reasonable with respect to what can be expected at this resolution. The control of gravity waves in the model solution was easily handled by the use of a $\mathcal{N}_c$ term.

The experiments performed at T42 when compared with the classical OI reveal several interesting points. Firstly, a $\mathcal{N}_g$ term which constrains the solution to fit an a priori knowledge of the atmosphere, provided that this a priori knowledge is sufficiently accurate, is certainly needed to avoid small-scale noise in data-sparse areas.

This a priori knowledge of the model state $x_g$ can be, for instance, a 6-hour previous forecast or even a previous 4D-VAR assimilation. An additional term $\mathcal{N}_g = \frac{1}{2}(x_0 - x_g)^T P^{-1}(x_0 - x_g)$, where here $P$ represents the background-error covariance matrix, would then be added in the cost function.

The introduction of this term allows, in particular, the transportation in time of information coming from previous observations. The development of a sophisticated $\mathcal{N}_g$ is ongoing at the ECMWF within the variational analysis project and will soon be tested in the 4D-VAR context.

The relatively poor fit of the 4D-VAR analysis to the observations (compared with OI) can be explained by another weakness of the experiments; namely, the assumption that the model is perfect (although we did not have a direct measure of its consequence in our experiment). However, results obtained by Derber (1989) and Courtier and Talagrand (1990) show that there is scope to relax this assumption.

The experiments have shown several advantages in favour of 4D-VAR. Several authors (see, for example, Gustafsson (1985) and Daley (1992)) have pointed out the significant local variability of the forecast errors. One well-known weakness of OI concerns its poor specification of the anisotropy, tilt and flow dependence of the forecast-error statistics. One of the main advantages of 4D-VAR compared with OI demonstrated in this paper is the natural specification of flow-dependent structure functions in 4D-VAR.

Another strong indication of the quality of the four-dimensional variational approach is given by the error growth of the forecast ensuing from 4D-VAR, which is significantly smaller than when it is performed from an OI assimilation (roughly 10% for the 500 hPa and 150 hPa kinetic-energy r.m.s. error after 5 days). However, this is only one case study and several similar studies will have to be performed on different meteorological situations to confirm this promising result.

One crucial problem inherent in 4D-VAR is obviously its computer cost. Indeed, if the adjoint integration of the model costs roughly twice a forward integration of the model, one computation of the gradient of the cost function is equivalent to a 3-day forecast for a 24-hour assimilation period. Under these conditions the convergence has to be efficient.

A possibility could be the use of the formalism developed in Gauthier (1992) and Rabier and Courtier (1992) to approximate the diagonal of the Hessian of the cost function by a limited number of gradient computations. Thépaut and Moll (1990) have shown that, in a small-scale minimization problem at least, the provision of the diagonal of the Hessian to the minimization algorithm significantly speeded up the convergence. This idea is to be tested in a 4D-VAR context. Hopefully this will give us an insight about the cost/benefit ratio that we can expect from 4D-VAR.
ACKNOWLEDGEMENTS

This work was done with the collaboration and the support of the data-assimilation team at the ECMWF which is gratefully acknowledged. The authors are indebted to M. Hamrud and O. Talagrand for their support and helpful comments. Dr J. C. Gilbert (INRIA) kindly provided the new minimization package used in the variational experiments. The study was partially supported by the Groupement de Recherche ‘Méthodes variationnelles en Météorologie et Océanographie’ funded by Centre National de la Recherche Scientifique and Météo-France.

REFERENCES


Lönnberg, P. 1988 Developments in the ECMWF analysis system. Pp. 75–119 in the Proceedings of the ECMWF seminar on data assimilation and the use of satellite data, 5–9 September


