Tellus (2001), 53A, 425–446
Printed in UK. All rights reserved

TELLUS
ISSN 0280–6495

Three-dimensional variational data assimilation for a limited area model

Part I: General formulation and the background error constraint

By N. GUSTAFSSON[1,*], L. BERRE[1,2], S. HÖRNQUIST[1,3], X.-Y. HUANG[4], M. LINDSKOG[1], B. NAVASCUÉS[5], K. S. MOGENSEN[4] and S. THORSTEINSSON[1,6], [1]*Swedish Meteorological and Hydrological Institute, S-60176 Norrköping, Sweden;* [2]*Météo-France, Toulouse, France;* [3]*ABB Corporation, Copenhagen, Denmark;* [4]*Danish Meteorological Institute, Copenhagen, Denmark;* [5]*National Meteorological Institute of Spain, Madrid, Spain;* [6]*Icelandic Meteorological Office, Reykjavik, Iceland*

(Manuscript received 17 March 2000; in final form 6 February 2001)

ABSTRACT

A 3-dimensional variational data assimilation (3D-Var) scheme for the HIgh Resolution Limited Area Model (HIRLAM) forecasting system is described. The HIRLAM 3D-Var is based on the minimization of a cost function that consists of one term J_b, which measures the distance between the resulting analysis and a background field, in general a short-range forecast, and another term J_o, which measures the distance between the analysis and the observations. This paper is concerned with the general formulation of the HIRLAM 3D-Var and with J_b, while the companion paper by Lindskog and co-workers is concerned with the handling of observations, including the J_o term, and with validation of the 3D-Var through extended parallel assimilation and forecast experiments. The 3D-Var minimization requires a pre-conditioning that is achieved by a transformation of the minimization control variable. This change of variable is designed as an operator approximating an inverse square root of the forecast error covariance matrix in the model space. The main transformations are the subtraction of the geostrophic wind increment, the bi-Fourier transform, and the projection on vertical eigenvectors. The spectral bi-Fourier approach allows one to derive non-separable structure functions in a limited area model, in the form of vertically dependent horizontal spectra and scale-dependent vertical correlations. Statistics have been accumulated from differences between +24 h and +48 h HIRLAM forecasts valid at the same time. Results from single observation impact studies as well as results from assimilation cycles using operational observations are presented. It is shown that the HIRLAM 3D-Var produces assimilation increments in accordance with the applied analysis structure functions, that the fit of the analysis to the observations is in agreement with the assumed error statistics, and that assimilation increments are well balanced. It is also shown that the particular problems associated with the limited area formulation have been solved. These results, together with the results of the companion paper, indicate that the 3D-Var scheme performs significantly better than the statistical interpolation scheme.

1. Introduction

The present operational HIRLAM data assimilation system is based on 3-dimensional multivariate statistical interpolation (or OI, Optimum Interpolation) of observed deviations from a short range forecast (Lorenc, 1981) and includes also a non-linear normal mode initialization scheme (Machenhauer, 1977). This data assimilation is applied intermittently with a 3- or 6-h data assimilation cycle. On the basis of experience with this

* Corresponding author.
e-mail: nils.gustafsson@smhi.se

data assimilation system, the design of a new data assimilation system for HIRLAM was studied by Gustafsson et al. (1997). They suggested that a new data assimilation for HIRLAM should be based on variational techniques. The main arguments in favour of variational techniques are: (1) the possibilities of improving initial baroclinic structures through stronger utilization of the time dimension and the model equations during the data assimilation, and (2) the possibilities for a more rational use of observations that are non-linearly coupled to the forecast model variables. The research and development work for the HIRLAM variational data assimilation started in 1995. The target is 4D-Var, 4-dimensional variational data assimilation. The natural step towards 4D-Var is 3D-Var, 3-dimensional variational data assimilation. The aim of this paper is to give a description of the HIRLAM 3D-Var, in particular the treatment of the background field. The handling of observations and results from parallel assimilation and forecast experiments compared with the operational data assimilation are the subjects of a companion paper (Lindskog et al., 2001).

Elements of variational techniques have been used for data assimilation purposes throughout the history of numerical weather prediction (Sasaki, 1958). With the introduction of the adjoint model technique to make it possible to use the forecast model equations as strong constraints during the assimilation process (Le Dimet and Talagrand, 1986; Lewis and Derber, 1985), intensive efforts have been devoted to introducing 3D-Var and 4D-Var operationally. The first operational 3D-Var system (Parrish and Derber, 1992) was introduced at the National Centers for Environmental Prediction (NCEP). Recently, the European Centre for Medium-Range Weather Prediction (ECMWF) replaced their operational 3D-Var system (Courtier et al., 1998) with an operational 4D-Var version (Courtier et al., 1994; Rabier et al., 1998b). The NCEP and ECMWF variational data assimilation schemes are based on global forecast models. Global variational data assimilation schemes have also been implemented by the Canadian (Gauthier et al., 1999), French and UK (Lorenc et al., 2000) meteorological services. Regional variational data assimilation schemes have been developed at NCEP (Zupanski, 1993), at the UK Meteorological Office (Lorenc, 1997), at the Canadian meteorological service (Laroche et al., 1999), within the HIRLAM Project (Gustafsson and Huang, 1996; Huang et al., 1997; Berre, 1997; Gustafsson et al., 1998) and at Météo-France (Berre, 2000; Sadiki et al., 2000).

The small number of observations compared to the number of variables in the initial state of the forecast model is a central problem in meteorological data assimilation. To resolve this, we introduce a priori information in the form of a short-range forecast and statistical knowledge about the errors of this forecast, in addition to statistical knowledge about the observation errors. We may also utilize additional a priori information regarding, for example, balances between different forecast model variables. Lorenc (1986) introduced Bayes' theorem and the maximum likelihood principle for treating a priori information in the derivation of a variational data assimilation equation. In the case of Gaussian probability distribution functions for the background field errors, as well as for the observation errors, the variational data assimilation problem consists of finding the model state vector $\boldsymbol{x}$ that minimizes the following cost function J:

$$J = J_b + J_o = \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^b)^T \boldsymbol{B}^{-1}(\boldsymbol{x} - \boldsymbol{x}^b) + \tfrac{1}{2}(H\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{R}^{-1}(H\boldsymbol{x} - \boldsymbol{y}).$$

Here J_b measures the distance to a background model state $\boldsymbol{x}^b$ and J_o measures the distance to the vector $\boldsymbol{y}$ of the observations. The observation operator H transforms a model state into the observed quantities. $\boldsymbol{B}$ is the matrix containing the covariances of the background field errors, while $\boldsymbol{R}$ is a matrix containing the covariances of the errors in the observations.

The HIRLAM 3D-Var is given an incremental formulation (Courtier et al., 1994), such that the assimilation increment $\delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}^b$ can be applied at a different (lower) horizontal resolution than the full model state. The new features of the HIRLAM 3D-Var, as compared to 3D-Var schemes developed elsewhere, are the use of a bi-Fourier formulation for the background error constraint J_b and the balance constraint based on the geostrophic wind relation for the assimilation increments with temperature, logarithm of surface pressure, ageostrophic wind components and specific humidity as assimilation control variables. The bi-Fourier approach allows for a LAM

3D-Var without the data selection problems of statistical interpolation, for an appropriate preconditioning, for non-separable assimilation structure functions and for non-zero increments near the lateral boundaries.

The incremental formulation of HIRLAM 3D-Var is described in Section 2. A preconditioning for the HIRLAM 3D-Var is introduced in the form of a series of transforms that are applied to the analysis increments, and in such a way that they provide an assimilation control variable for which the background error cost function J_b has a simple expression. The change of variable is designed as an operator transforming the forecast error in the model space into a variable whose covariance matrix can be assumed to be an identity matrix. The transformation of the assimilation control variable is described in Section 3, while the derivation of forecast error statistics needed for this transformation is treated in Section 4. A validation of the HIRLAM 3D-Var is given by single observation impact experiments in Section 5 and by diagnostics of a single data assimilation cycle using operational observations in Section 6. Some concluding remarks are presented in Section 7.

2. Incremental formulation

HIRLAM variational data assimilation can be applied with the grid point or spectral HIRLAM forecast models. When applied with the spectral HIRLAM model, the spatial representation of the assimilation increment $\delta \boldsymbol{x}$ in spectral space is identical to the spatial representation of the model state variable $\boldsymbol{x}$, which is also in spectral space, provided the same spectral truncation is applied for the forecast model integration as for the variational data assimilation. When the grid point model is applied together with the variational data assimilation, however, the full resolution model background field is available in grid point space. The high resolution part of the spectrum of the model background field is mainly generated by the model itself, for example, due to orographic height differences, land–sea contrasts, and other processes related to the lower boundary condition for the forecast model. Observations will also include contributions from this high resolution part of the spectrum. The horizontal scales of the assimilation increments are prescribed by the autocorrelation spectra of the background field errors and these contain very small contributions in the high-resolution part of the spectrum. Thus the assimilation will not contribute significantly to change the model background field in this part of the spectrum. For this reason, and in order to avoid aliasing of small scale observed and/or modeled variations to larger scales, it is important to avoid any spatial truncation/filtering of the background field before comparing it with the observed values. Therefore the HIRLAM variational data assimilation may be applied in its *incremental* form, for example, when the background field is provided by the grid point model or by a spectral model with a truncation that is different from the truncation of the assimilation increments. The cost function for the incremental formulation is defined as follows:

$$J = J_b + J_o = \tfrac{1}{2}\delta \boldsymbol{x}'^{\mathrm{T}} \boldsymbol{B}'^{-1} \delta \boldsymbol{x}' + \tfrac{1}{2}(H\boldsymbol{x}^b + \boldsymbol{H}'\delta \boldsymbol{x}' - \boldsymbol{y})^{\mathrm{T}} \boldsymbol{R}^{-1}(H\boldsymbol{x}^b + \boldsymbol{H}'\delta \boldsymbol{x}' - \boldsymbol{y}).$$

The model background field $\boldsymbol{x}^b$ is provided at full model resolution, while the assimilation increment $\delta \boldsymbol{x}'$ is provided at another, in general lower, spatial resolution. The background error covariance matrix $\boldsymbol{B}'$ now has the squared dimension of the assimilation increment. We may formally introduce an operator $\boldsymbol{G}$ that transforms a full resolution model state $\boldsymbol{x}$ to the resolution of the assimilation increment $\boldsymbol{x}' = \boldsymbol{G}\boldsymbol{x}$. The generalized inverse $\boldsymbol{G}^{-1}$ transforms the low resolution assimilation increments to the full model resolution $\delta \boldsymbol{x} = \boldsymbol{G}^{-1}\delta \boldsymbol{x}'$. In the case of application of incremental variational data assimilation to the grid point HIRLAM model, with the full resolution background field $\boldsymbol{x}^b$ defined in grid point space, this generalized inverse consists in setting truncated spectral coefficients to zero and application of an inverse Fourier transform back to grid point space.

The observation operator H may, in general, be non-linear, and the non-linearity of H may, for example, be important for the assimilation of satellite radiance information. For the incremental HIRLAM 3D-Var, at least in its first version, we have chosen to use the full non-linear observation operator H for the background field $\boldsymbol{x}^b$, while a linearized observation operator $\boldsymbol{H}'$ is applied to the analysis increment $\delta \boldsymbol{x}' = \boldsymbol{x}' - (\boldsymbol{x}^b)'$. The linearization is carried out with reference to the

background field $\boldsymbol{x}^{\mathrm{b}}$. The observation operators include, for example, horizontal and vertical interpolations of model background and incremental fields. More details about the observation operators can be found in the companion paper (Lindskog et al., 2001).

Standard minimization software packages (Gilbert and Lemaréchal, 1989) are used to find the minimum of J with respect to the analysis increment $\delta\boldsymbol{x}'$. These software packages solve the minimization problem iteratively. For each iteration, calculation of the gradient of the cost function with respect to the model state increment vector $\delta\boldsymbol{x}'$ (the control vector) is required:

$$\nabla_{\delta x'} J = \boldsymbol{B}'^{-1}\delta\boldsymbol{x}' + \boldsymbol{H}'^{\mathrm{T}}\boldsymbol{R}^{-1}(H\boldsymbol{x}^{\mathrm{b}} + \boldsymbol{H}'\delta\boldsymbol{x}' - \boldsymbol{y}).$$

With regard to the notations in the rest of this paper and the companion paper (Lindskog et al., 2001), $\delta\boldsymbol{x}$ may either mean a full resolution spectral increment or a lower resolution spectral increment $\delta\boldsymbol{x}'$.

3. Formulation of the background error constraint for a limited area model

For HIRLAM, the model state increment vector $\delta\boldsymbol{x}$ includes all spectral components of the following model increment variables to be determined by the variational data assimilation:

$$\delta\boldsymbol{x} = \begin{pmatrix} \delta\boldsymbol{u} \\ \delta\boldsymbol{v} \\ \delta\boldsymbol{T} \\ \delta\boldsymbol{q} \\ \delta \ln \boldsymbol{p}_{\mathrm{s}} \end{pmatrix},$$

where $\delta\boldsymbol{u}$ is the increment vector of the wind component in the x-direction, $\delta\boldsymbol{v}$ the wind component increment in the y-direction, $\delta\boldsymbol{T}$ the temperature increment, $\delta\boldsymbol{q}$ the specific humidity increment and $\delta \ln \boldsymbol{p}_{\mathrm{s}}$ the increment of the logarithm of the surface pressure. The HIRLAM model state includes some further surface and soil variables, e.g., soil moisture and surface temperature. These are not included among the variables to be controlled by the variational data assimilation, mainly because the transforms to be applied to the control variables are not applicable to these surface variables, defined only in grid point space.

3.1. Preconditioning and change of variable

For a fast convergence of the minimization algorithm, a pre-conditioning of the minimization problem is needed. An ideal pre-conditioning is obtained if the Hessian matrix (i.e., the second order derivative of the cost function) is an identity matrix. A good approximation of this is to ensure that the Hessian of the background error contribution J_{b} to the cost function is equal to the identity matrix (Lorenc, 1988). This can be achieved by defining a change of variable $\boldsymbol{U}$ to be applied to the assimilation increment $\delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}^{\mathrm{b}}$ such that it transforms the forecast error ε in model space into ε', a variable whose covariance matrix is an identity matrix. This change of variable can be written as $\varepsilon' = \boldsymbol{U}\varepsilon$, or equivalently $\varepsilon = \boldsymbol{U}^{-1}\varepsilon'$. This implies an equality between the corresponding covariance matrices:

$$\boldsymbol{B} = \overline{(\varepsilon\varepsilon^{\mathrm{T}})} = \boldsymbol{U}^{-1}\overline{(\varepsilon'\varepsilon'^{\mathrm{T}})}\boldsymbol{U}^{-\mathrm{T}} = \boldsymbol{U}^{-1}\boldsymbol{U}^{-\mathrm{T}},$$

if $\overline{(\varepsilon'\varepsilon'^{\mathrm{T}})} = \mathbf{I}$ and $\overline{(\ )}$ denotes ensemble averaging. $\boldsymbol{U}$ appears to be a left-hand inverse square-root of the forecast error covariance matrix.

$$\boldsymbol{B}^{-1} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U},$$

and the expression for J_{b} with respect to $\chi = \boldsymbol{U}\delta\boldsymbol{x}$ becomes:

$$J_{\mathrm{b}} = \tfrac{1}{2}\delta\boldsymbol{x}^{\mathrm{T}}\boldsymbol{B}^{-1}\delta\boldsymbol{x}$$
$$= \tfrac{1}{2}\chi^{\mathrm{T}}\boldsymbol{U}^{-\mathrm{T}}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U})\boldsymbol{U}^{-1}\chi = \tfrac{1}{2}\chi^{\mathrm{T}}\chi.$$

We will now proceed to describe how $\boldsymbol{U}$ is designed. This amounts to describing what transforms are applied to convert the forecast error in physical model space to a variable, whose covariance matrix can be assumed to be an identity matrix. The main difficulties here are the de-coupling of the geostrophic balance between mass and wind field forecast errors and the de-correlation of horizontally and vertically correlated forecast errors.

3.2. Mass/wind balane

Forecast error fields are known to be in near geostrophic balance in the extratropics and assimilation increment fields should ideally also be in near geostrophic balance. One simple way to obtain de-correlated forecast errors with regard to this balance is therefore to transform the full wind forecast error to an ageostrophic forecast error by subtraction of the geostrophic wind error,

as determined from the mass field forecast error. The ageostrophic wind forecast errors are moreover assumed to be de-correlated with the temperature and surface pressure forecast errors.

Since the development of the adjoint model for the HIRLAM 4D-Var started from the spectral version of the model (Gustafsson, 1999), it was natural to utilize the corresponding tangent-linear geostrophic balance relation to de-correlate the wind and mass field forecast errors. The continuous geostrophic wind components (u_g and v_g) in the spectral HIRLAM are given as follows for the hybrid vertical coordinate system:

$$u_g = -\frac{1}{f}\left[\frac{R_d T_v}{h_y}\frac{\partial \ln p}{\partial y} + \frac{1}{h_y}\frac{\partial \Phi}{\partial y}\right],$$

$$v_g = \frac{1}{f}\left[\frac{R_d T_v}{h_x}\frac{\partial \ln p}{\partial x} + \frac{1}{h_x}\frac{\partial \Phi}{\partial x}\right],$$

where f denotes the Coriolis parameter, h_x and h_y map-factors in the x- and y-directions respectively, R_d the dry air gas constant, T_v virtual temperature, p pressure and Φ geopotential. The vertical discretization of this geostrophic wind calculation is carried out in accordance with Simmons and Burridge (1981). Tangent-linear and adjoint geostrophic expressions were derived directly from the vertically discretized non-linear expressions (Gustafsson and Huang, 1996). Horizontal derivatives are calculated by Fast Fourier Transforms (FFTs) from the corresponding derivatives in spectral space.

The introduction of the ageostrophic wind analysis increments as control variables was chosen for its simplicity and because of the possibility of avoiding the use of vorticity and divergence as control variables for the wind field. The mean wind is needed, in addition to vorticity and divergence, to represent the wind field on a limited area domain, and using these three variables as control variables was considered more complicated regarding the formulation of the geostrophic balance. On the other hand, use of ageostrophic wind components as control variables leads to serious problems with data assimilation in tropical areas. Neglecting the correlation between mass and ageostrophic wind is also likely to be suboptimal at low levels if one wants to represent, for example, frictional effects. A balance constraint for the HIRLAM 3D-Var based on vorticity and divergence, and including a representation of frictional effects, is being developed (Berre, 2000).

3.3. *Horizontal spectral transforms*

Forecast errors are known to be horizontally correlated in physical space. In global variational data assimilation schemes (Parrish and Derber, 1992; Courtier et al., 1998), the assimilation control variable is defined in spectral space (spherical harmonics). This is related to the possibility of considering that different spectral forecast error modes are statistically independent, given the assumption that spatial covariances are horizontally homogeneous in physical space (Boer, 1983).

A similar property exists in a spectral space corresponding to a limited area model (see, e.g., Berre (2000) for a bi-Fourier spectral representation). The application of spectral transforms to meteorological fields over a limited model integration area poses a particular problem, since the fields do not have a periodic variation over the area. One possibility would have been to assume zero-valued analysis increments on the lateral boundaries and to apply sine transforms (Lorenc, 1997). For the HIRLAM 3D-Var, we have followed Haugen and Machenhauer (1993) and used an area extension (Fig. 1) to obtain periodic variations in both horizontal dimensions. This makes it possible to have non-zero analysis increments on the lateral boundaries and to apply standard Fast Fourier Transforms (FFTs) for the transformation of the control vector.

Any 2-dimensional model error field $\varepsilon_m(i_x, i_y)$ at level m can be represented by the coefficients $\hat{\varepsilon}_{klm}$ of the following series of orthogonal functions,

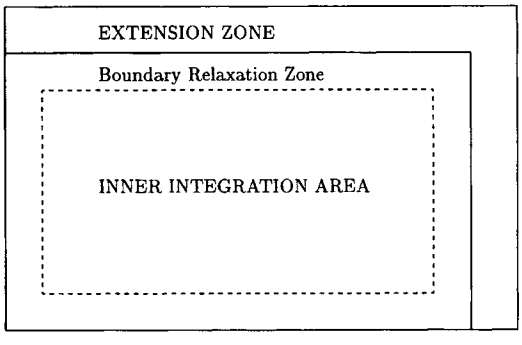

Fig. 1. Integration area of the spectral HIRLAM.

using the HIRLAM Fourier transform pair:

$$\varepsilon_m(i_x, i_y) = \frac{1}{\sqrt{NX\,NY}} \sum_{k=-K}^{K} \sum_{l=-L}^{L} \hat{\varepsilon}_{klm} \times \exp\left[+2\pi i \left(\frac{ki_x}{NX} + \frac{li_y}{NY} \right) \right],$$

$$\hat{\varepsilon}_{klm} = \frac{1}{\sqrt{NX\,NY}} \sum_{i_x=1}^{NX} \sum_{i_y=1}^{NY} \varepsilon_m(i_x, i_y) \times \exp\left[-2\pi i \left(\frac{ki_x}{NX} + \frac{li_y}{NY} \right) \right],$$

where NX and NY are the number of grid points of the zonal and meridional axis of the extended domain, respectively. K and L are the corresponding maximum wave numbers. An elliptical truncation defined by $(k/K)^2 + (l/L)^2 \leqslant 1$ is used in order to obtain an isotropic and homogeneous resolution over the extended area. This is omitted in the notations for the sake of simplicity. Note also that $\hat{\varepsilon}_{-k-lm} = \hat{\varepsilon}^*_{klm}$, where * is the complex conjugate operator, because ε represents a real variable. This implies that only coefficients with $k \geqslant 0$ need to be explicitly handled.

One should note that the assumption of horizontal homogeneity is known to be more valid for correlations than for covariances, so that it is useful to normalize the forecast errors in physical model space by their standard deviations before applying the spectral transforms. This gives the possibility of introducing spatial and temporal variations of these standard deviations. This has not been implemented yet in the HIRLAM 3D-Var, but it could be done in the future following, e.g., the ideas of Fisher and Courtier (1995).

After the normalization by the standard deviations in physical model space and the spectral transforms, each spectral coefficient of the forecast error field has a variance which, according to the homogeneity assumption, corresponds to the spectral density of the horizontal correlation. Normalizing with the square-root of this spectral variance provides spectral coefficients whose variance is equal to one.

3.4. *Vertical transforms*

After subtracting the geostrophic wind forecast error, application of the normalizations by the forecast error standard deviations and the spectral transforms, the residual correlations that need to be accounted for are those involving the same horizontal spectral modes at different vertical levels.

Projecting each bi-Fourier forecast error on the eigenvectors of the corresponding vertical correlation matrix provides variables whose covariance matrix is diagonal. This diagonal matrix contains the associated eigenvalues. Dividing by the square-roots of these eigenvalues allows one to obtain a variable with a covariance matrix that is equal to the identity matrix (given the assumptions of homogeneity and of de-correlation between mass and ageostrophic wind forecast errors).

The use of such horizontal and vertical transforms allows us to use non-separable assimilation structure functions on a limited area, since different vertical levels may be associated with different spectral density representations and different horizontal modes may be associated with different vertical representations (different vertical eigenvectors and eigenvalues).

3.5. *Inverse change of variable and its adjoint*

The complete transformation from the analysis increments in model space to the control variable can thus be written:

$$\chi = \boldsymbol{PVLFS}^{-1}\boldsymbol{AF}^{-1}\delta\boldsymbol{x} = \boldsymbol{U}\delta\boldsymbol{x} = \boldsymbol{U}(\boldsymbol{x} - \boldsymbol{x}^{\mathrm{b}}),$$

where $\delta\boldsymbol{x}$ is the assimilation increment, $\boldsymbol{F}$ is the Fourier transform to spectral space, $\boldsymbol{F}^{-1}$ is the inverse Fourier transform, $\boldsymbol{A}$ is the subtraction of the geostrophic wind increment from the full wind increment, $\boldsymbol{S}^{-1}$ is the normalization with the forecast error standard deviation, $\boldsymbol{L}$ is the normalization by the square-root of the spectral density of the horizontal forecast error correlation, $\boldsymbol{V}$ is the projection on the eigenvectors of the vertical forecast error correlation matrix and $\boldsymbol{P}$ is the normalization by the square root of the vertical eigenvalue.

It should be mentioned here that these transforms are never applied in their forward form during the variational data assimilation. At the start of the minimization process, all elements of the control vector χ are put equal to zero. Then, during each iteration, the inverse transform $\boldsymbol{U}^{-1}$ is applied to obtain the spectral assimilation incre-

ments in the model space:

$$\delta \boldsymbol{x} = \boldsymbol{U}^{-1}\chi = \boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{S}\boldsymbol{F}^{-1}\boldsymbol{L}^{-1}\boldsymbol{V}^{-1}\boldsymbol{P}^{-1}\chi.$$

Knowing that $\boldsymbol{B}^{-1} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$, the value of the cost function is evaluated as:

$$J = J_{\mathrm{b}} + J_{\mathrm{o}} = \tfrac{1}{2}(\chi)^{\mathrm{T}}(\chi) + \tfrac{1}{2}(H\boldsymbol{x}^{\mathrm{b}} + \boldsymbol{H}\boldsymbol{U}^{-1}\chi - \boldsymbol{y})^{\mathrm{T}}\boldsymbol{R}^{-1} \times (H\boldsymbol{x}^{\mathrm{b}} + \boldsymbol{H}\boldsymbol{U}^{-1}\chi - \boldsymbol{y}).$$

The adjoint of the inverse transform $\boldsymbol{U}^{-\mathrm{T}}$ is then applied to obtain the gradient of the cost function with respect to the control variable χ:

$$\nabla_{\chi} J = \nabla_{\chi} J_{\mathrm{b}} + \nabla_{\chi} J_{\mathrm{o}} = \chi + \boldsymbol{U}^{-\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1}(H\boldsymbol{x}^{\mathrm{b}} + \boldsymbol{H}\boldsymbol{U}^{-1}\chi - \boldsymbol{y}).$$

It may be remarked that $\boldsymbol{H}^{\mathrm{T}}$ and $\boldsymbol{U}^{-\mathrm{T}}$ include the application of the adjoint of the inverse Fourier transform (which is the same as a direct Fourier transform) to a grid point field of departures between the analysis and the observations. As there are no observations in the extension zone, it is natural to set zero values there before applying the adjoint of the inverse Fourier transform.

As the minimization uses control variables that are real-valued, there is a final transformation from the complex variable containing the spectral coefficients to the real variable containing the corresponding real and imaginary parts. Note that the inverse of this transformation is straightforward, while the adjoint inverse operation involves a factor 2 to account for the contribution from the complex conjugate of each spectral coefficient (except for $(k, l) = (0, 0)$). The gradient of J with respect to the final real control vector is thus twice the gradient of J with respect to the complex variable χ. This final transformation is omitted in the notations above.

4. Estimation of background error statistics

4.1. Data set

Statistics were accumulated on differences between SMHI +24 h and +48 h operational HIRLAM forecasts, that are valid at the same time. The operational SMHI HIRLAM is applied with a resolution of 0.4° × 0.4° in the horizontal and with 31 vertical levels. The data set covers the period 1 December 1997 through 28 February 1998. Variables which were studied are those that are used in the control vector of the HIRLAM 3D-Var (temperature (T), ageostrophic wind in the x-direction (u_{ag}) and in the y-direction (v_{ag}), specific humidity, logarithm of surface pressure), and full wind (u, v) in addition. As statistics for v_{ag}, v are quite similar respectively to those for u_{ag}, u, only the latter will be shown here.

These statistics of forecast differences were considered as an approximation to the forecast error covariances. This corresponds to "the NMC method" (Parrish and Derber, 1992; Rabier et al., 1998a). The theoretical justification for this approach is rather weak. An interesting alternative is to derive the forecast error covariances from ensemble assimilation experiments (Houtekamer et al., 1996) by perturbing, for example, observations and forecast model parameterizations. Another method could be to use the time-averaged covariances of an Extended Kalman Filter (EKF) in research mode to specify the static error covariances (Bouttier, 1996), although this is much more costly.

Standard deviations of the forecast differences are usually rescaled to match the amplitude of +6 h forecast errors (Parrish and Derber, 1992; Rabier et al., 1998a). A scaling factor equal to 0.6 was empirically chosen to obtain background error standard-deviations in agreement with the orders of magnitude of the observation errors and of the innovation vectors (which are departures between +6 h forecasts and observations).

4.2. Horizontal homogeneity and isotropy

Under the assumption that spatial covariances are horizontally homogeneous in grid point space, the bi-Fourier spectral modes are de-correlated unless they correspond to the same pair of horizontal wave numbers (k, l):

$$\overline{\hat{\varepsilon}_{klm}\hat{\varepsilon}^{*}_{k'l'n}} = \delta_k^{k'}\,\delta_l^{l'}\,\overline{\hat{\varepsilon}_{klm}\hat{\varepsilon}^{*}_{kln}},$$

where $\delta_k^{k'}$ is the Kronecker symbol and m and n are two model levels. As mentioned in Subsection 3.3 this property motivates the use of a spectral transform for the definition of the control variable. Moreover, in this case a Fourier transform relates covariances in the spectral and grid point spaces. This means that the calculation of a full matrix of two-dimensional (i.e., horizontal) covariances in grid point space is equivalent to the calculation of a diagonal matrix

containing spectral variances. Similarly, it is possible to represent implicitly a full matrix of 3-dimensional covariances in grid point space through the calculation of a block-diagonal spectral covariance matrix (with each block corresponding to a vertical covariance matrix for a given couple of wave numbers).

It thus appears more efficient to calculate covariances in spectral space than in grid point space if one intends to use the homogeneity assumption. In practice, the first step of the algorithm consists in applying a bi-Fourier transform to the forecast differences in grid point space. This step includes a preliminary area extension to obtain a bi-periodic field of forecast differences (see Subsection 3.3). A second step is to calculate spectral covariances.

Such a procedure amounts to applying the homogeneity assumption over the whole bi-periodic extended domain. To verify that this is reasonable, we have estimated horizontal correlations for temperature at a few levels with two different algorithms. A first estimate is obtained with the aforementioned method, which will be referred to as "the spectral algorithm": after calculating spectral variances at a given level, the corresponding horizontal correlation function in grid point space is provided by an inverse FFT and a normalization by the value at the origin. A second estimate is obtained by a direct calculation of the average correlation function for the integration area in grid point space. This will be referred to as "the grid point algorithm". An example is shown for temperature at level 18 (500 hPa) (Fig. 2): both algorithms give similar results. As the spectral algorithm is much less costly than the grid point algorithm, it was used in our study to estimate background error statistics.

Another simplification that has been used is the isotropy assumption. This is indeed a reasonable approximation for homogeneous covariances that are calculated over a long period, as was shown by Thiebaux et al. (1986). This is illustrated for temperatures near 500 hPa in Fig. 3. Wind autocorrelations, however, present noticeable anisotropic features, as will be shown in Subsection 4.4. Using such an isotropy assumption allows one to improve the sampling of the covariance estimations. In spectral space, this consists of calculating the average spectral covariance of wave number pairs (k, l) that are linked with the same total

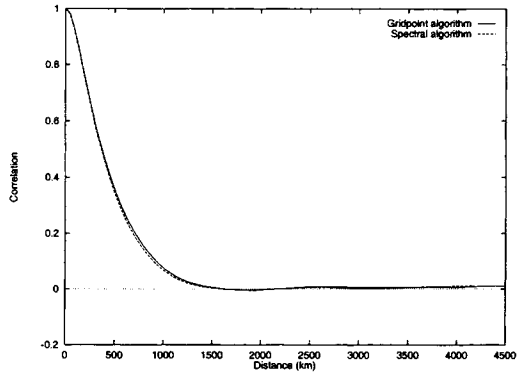

Fig. 2. Cross-sections in the y-direction of average auto-correlation functions of the temperature at level 18 (500 hPa), provided by grid point statistics in the integration area (solid line) and by spectral statistics (dashed line).

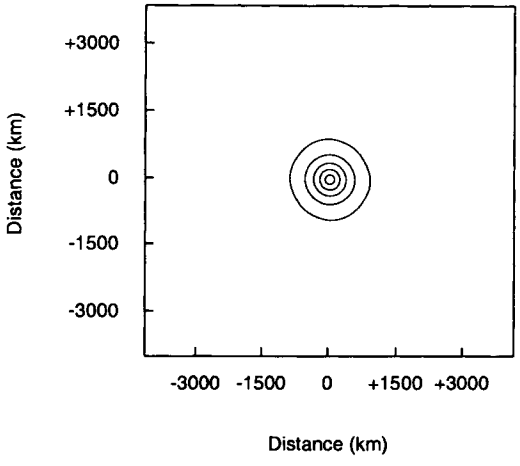

Fig. 3. Auto-correlation function of the temperature at level 18 (500 hPa). The first isoline near the centre is 0.9, and the isoline spacing is 0.2. The zero contour has been omitted.

wave number k^t:

$$B_{k^t}^{m,n} = \frac{1}{2\pi} \int_0^{2\pi} \overline{\hat{\varepsilon}_{pqm} \hat{\varepsilon}_{pqn}^*} \, \mathrm{d}\theta,$$

with (p, q) such that

$$N_s \sqrt{\left(\frac{p}{NX}\right)^2 + \left(\frac{q}{NY}\right)^2} = k^t,$$

$$\theta = \tan^{-1}\left(\frac{q/NY}{p/NX}\right).$$

The integral in the expression is calculated as a sum for each discrete k^t. N_s is an arbitrary scaling factor; in this study we used $N_s = NX$.

4.3. *Non-separability*

In the past, it was usual to assume that three-dimensional auto-correlation functions were separable. This meant that they could be written as the product of a horizontal correlation function that was independent of height, and a vertical correlation function which was supposed to be independent of horizontal coordinates. It is, however, relatively easy to formulate non-separable structure functions in spectral space, as has been shown with spherical harmonics in global models (Courtier et al., 1998; Rabier et al., 1998a) and with Bessel functions (Hollingsworth and Lönnberg, 1986). This is also possible in limited area models when using a bi-Fourier formulation (Berre, 2000).

The spectral covariance can be written as:

$$\overline{\hat{\varepsilon}_{klm}\hat{\varepsilon}^*_{kln}} = \sigma_m \sigma_n \sqrt{\gamma_m(k^t)\gamma_n(k^t)}\, r_{k^t}(m, n),$$

where

$$k^t = N_s \sqrt{\left(\frac{k}{NX}\right)^2 + \left(\frac{l}{NY}\right)^2}$$

and σ_m, σ_n are the horizontally averaged standard deviations of levels m, n:

$$\sigma_m = \sqrt{\sum_{k_0^t=0}^{K^t} B^{m,m}_{k_0^t} 2\pi k_0^t \frac{1}{N_s^2}},$$

γ_m, γ_n are the normalized bi-Fourier spectral densities of the variance at levels m, n (they are closely related to spectral coefficients of the horizontal correlations, under the homogeneity assumption):

$$\gamma_m(k^t) = \frac{B^{m,m}_{k^t}}{\sum_{k_0^t=0}^{K^t} B^{m,m}_{k_0^t} 2\pi k_0^t \frac{1}{N_s^2}}$$

$r_{k^t}(m, n)$ is the vertical correlation between levels m, n for what concerns the total wave number k^t:

$$r_{k^t}(m, n) = \frac{B^{m,n}_{k^t}}{\sqrt{B^{m,m}_{k^t} B^{n,n}_{k^t}}}$$

and K^t is the maximum total wave number.

The shape of the vertical profiles of average standard-deviations (Fig. 4) agrees well with those of previous studies such as in Rabier et al. (1998a). Note, for example, the maxima for the full wind near the tropopause around 300 hPa, for temperature near 200 hPa and for specific humidity around 850 hPa. A first non-separable feature is the vertical dependence of spectral densities (of horizontal correlation functions). Plotting the corresponding power spectra

$$P_m(k^t) = \frac{2\pi k^t}{N_s^2}\gamma_m(k^t)$$

of temperature at different levels (Fig. 5) indicates some shift of the energy towards small scales when approaching low levels. In grid point space, this

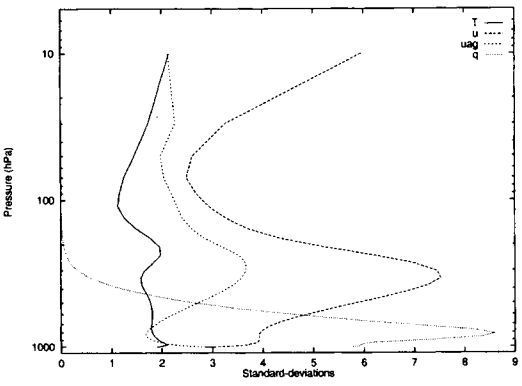

Fig. 4. Horizontally averaged standard deviations as a function of pressure (in hPa) for the component of the full wind in the x-direction (u (m/s)), the component of the ageostrophic wind in the x-direction (u_{ag} (m/s)), the temperature (T (K)), and the specific humidity (q (10^{-5} kg/kg)), J_b uses these values scaled by 0.6 (exept u, which is not in the control vector).

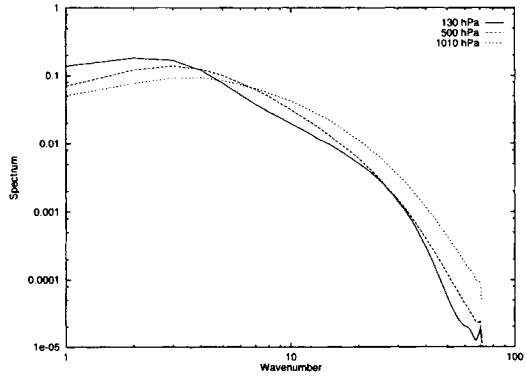

Fig. 5. Horizontal auto-correlation spectra as functions of the horizontal total wave number for the temperature at level 7 (130 hPa), level 18 (500 hPa) and level 31 (1010 hPa).

leads to sharper horizontal correlations with decreasing height (Fig. 6). A survey of the height dependence of horizontal auto-correlations can be obtained by plotting the vertical profiles of their length-scales. The length scale is here defined such that its squared value is equal to -2 times the inverse of the Laplacian of the correlation function at the origin (Daley 1991, page 110). Fig. 7 indicates that this vertical dependence is well marked for the temperature and the full wind, while values tend to be more uniform vertically for the specific humidity and the ageostrophic wind. (Values for specific humidity are not represented above 180 hPa, since statistics at these heights are not very reliable as there is little humidity.) The specific humidity and the ageostrophic wind have smaller horizontal scales than the temperature and the full wind. This is in agreement with the importance of mesoscale diabatic processes for the humidity field and with the lower degree of geostrophy in the smaller scales. The increase of the temperature length scale with height agrees well with previous studies such as Rabier et al. (1998a). Values tend to be smaller than those for global models, due to the enhanced resolution of the HIRLAM model.

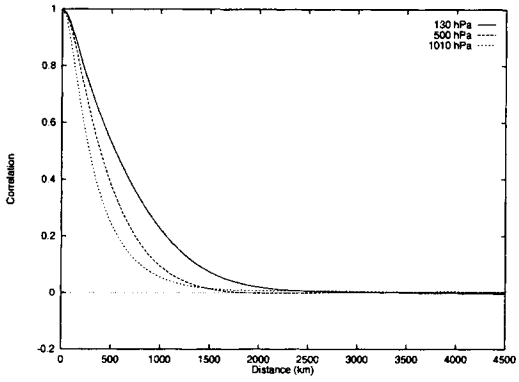

Fig. 6. Horizontal auto-correlation as functions of horizontal distance for the temperature at level 7 (130 hPa), level 18 (500 hPa) and level 31 (1010 hPa).

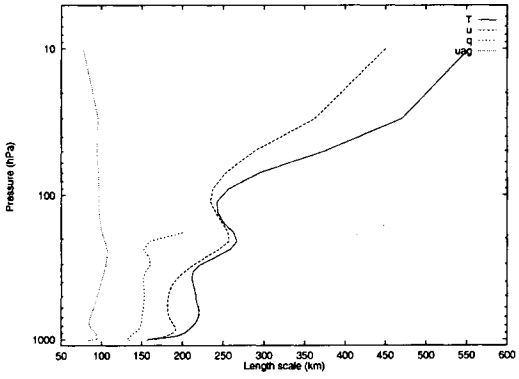

Fig. 7. Horizontal length scales (km) as a function of pressure (hPa) for the temperature (T), the component of the full wind in the x-direction (u), the specific humidity (q) and the component of the ageostrophic wind in the x-direction (u_{ag}).

Another aspect of non-separability is the dependence of vertical correlations on horizontal scale. For all variables, the vertical auto-correlations tend to become broader towards larger horizontal scales. This can be illustrated, for example, for the temperature at model level 18 (Fig. 8). This result is in accordance with what was found at ECMWF (see Fig. 7 in Derber and Bouttier (1999)). Some unrealistically large vertical correlations were found for the full as well as the ageostrophic wind at the smallest horizontal scales. These are likely to be due to un-realistic gravity wave oscillations in the forecast model integrations. For the implementation of the ageostrophic wind statistics in the 3D-Var, these spurious small scale features were corrected by a simple ad-hoc extrapolation of spectral statistical densities from larger horizontal scales to the smallest scales.

4.4. *Ageostrophic wind and full wind*

A comparison of statistics for the ageostrophic wind and for the full wind allows one to illustrate specific features of these two quantities.

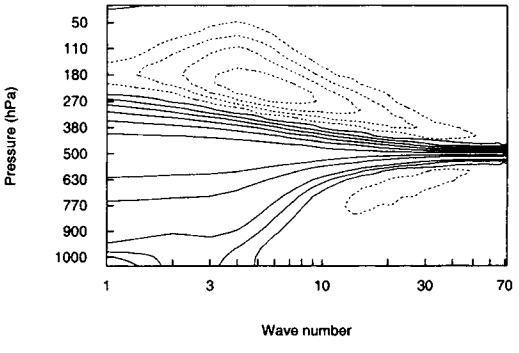

Fig. 8. Vertical correlations for the temperature at level 18 (500 hPa) as a function of horizontal wave number k^t and pressure (hPa). Positive (resp. negative) values are indicated by full (resp. dashed) isolines. The isoline spacing is 0.15, and the zero contour has been omitted.

The vertical profiles of their average standard deviations (Fig. 4) indicate a maximum at the jet-stream level in both cases. One can note moreover that the ageostrophic wind standard deviation tends to increase towards the lowest levels, whereas the full wind standard deviation decreases. These opposite evolutions are likely to be due to surface friction effects which make the full wind more ageostrophic and which reduce its speed.

The examination of horizontal auto-correlations at 500 hPa (panels (a) and (b) in Fig. 9) indicates that errors on the ageostrophic wind are in smaller scale than those on the full wind, as expected. Both ageostrophic and full winds appear to be more rotational than divergent since there are negative lobes north and south of the origin (Daley, 1985).

Similar patterns can be found when plotting horizontal auto-correlations at a level close to 1000 hPa, but these patterns appear to be rotated (panels (c) and (d) in Fig. 9). The rotation is anti-clockwise for the ageostrophic wind, while it is clockwise for the full wind. These different rotations reflect respectively a positive cross-correlation between the ageostrophic vorticity and the ageostrophic divergence, and a negative cross-correlation between vorticity and divergence for the full wind (Daley, 1985). This can be interpreted when considering the expected mass/wind coupling at these low levels. Surface friction makes the full and ageostrophic winds converge towards low geopotentials. As the full wind has a positive vorticity in regions of low geopotentials, this contributes to the aforementioned negative correlation between rotational and divergent parts of the full wind. Inversely, as the ageostrophic vorticity tends to be negative in these regions of low geopotentials, it is positively correlated with the ageostrophic divergence.

One may represent these anisotropic features of the ageostrophic wind auto-correlations in the HIRLAM 3D-Var. On the other hand, to be consistent, one should also represent the corresponding cross-correlations with temperature and surface pressure. This is under investigation following the ideas of Derber and Bouttier (1999).

5. Single observation influence experiments

An efficient way to check the variational data assimilation software is to carry out single observation influence experiments. Such influence experiments have been recommended by the WMO Working Group on Numerical Experimentation (WGNE) for a basic comparison of different variational data assimilation schemes.

We have carried out experiments where we have inserted a single observed 500 hPa temperature value, corresponding to an observation increment of -3 K, or a single observed 500 hPa wind vector, corresponding to a westerly wind component increment of 5 ms^{-1}. We have inserted the isolated observation in the middle of the integration area for a clean study of the assimilation structure functions as given by the derived influence of the observation. We have also inserted the isolated observations in the vicinity of the lateral boundaries to study the effect of different lateral boundary formulations as well as to examine whether there are any effects caused by the assumption of bi-periodic assimilation increments.

All experiments have been carried out at a horizontal resolution of $0.4° \times 0.4°$ with 31 vertical levels and with the non-separable analysis structure functions derived by the NMC method from the differences between SMHI $+48$ h and $+24$ h operational HIRLAM forecasts valid at the same time. The (inner) integration area consisted of 202×178 grid points, while the extended integration area used for representation of, for example, analysis increments in spectral space consisted of 240×216 grid points. This means that the width of the extension zone is approximately 1500 km, which should be sufficient to avoid any influence from, for example, observations close to the east boundary on analysis increments in grid points close to the west lateral boundary. For small model domains, however, the extension zone will probably have to be relatively larger as compared with the inner integration area.

5.1. *Observation in the middle of the integration area*

The horizontal effect on the temperature and wind analysis increments of a single 500 hPa temperature observation at 60°N 30°W is illustrated for model level 18 (approximately at 500 hPa over sea) in Fig. 10. The vertical impact on the temperature increments is illustrated by a vertical cross-section in the west–east direction in Fig. 11. We may notice that the temperature increments are

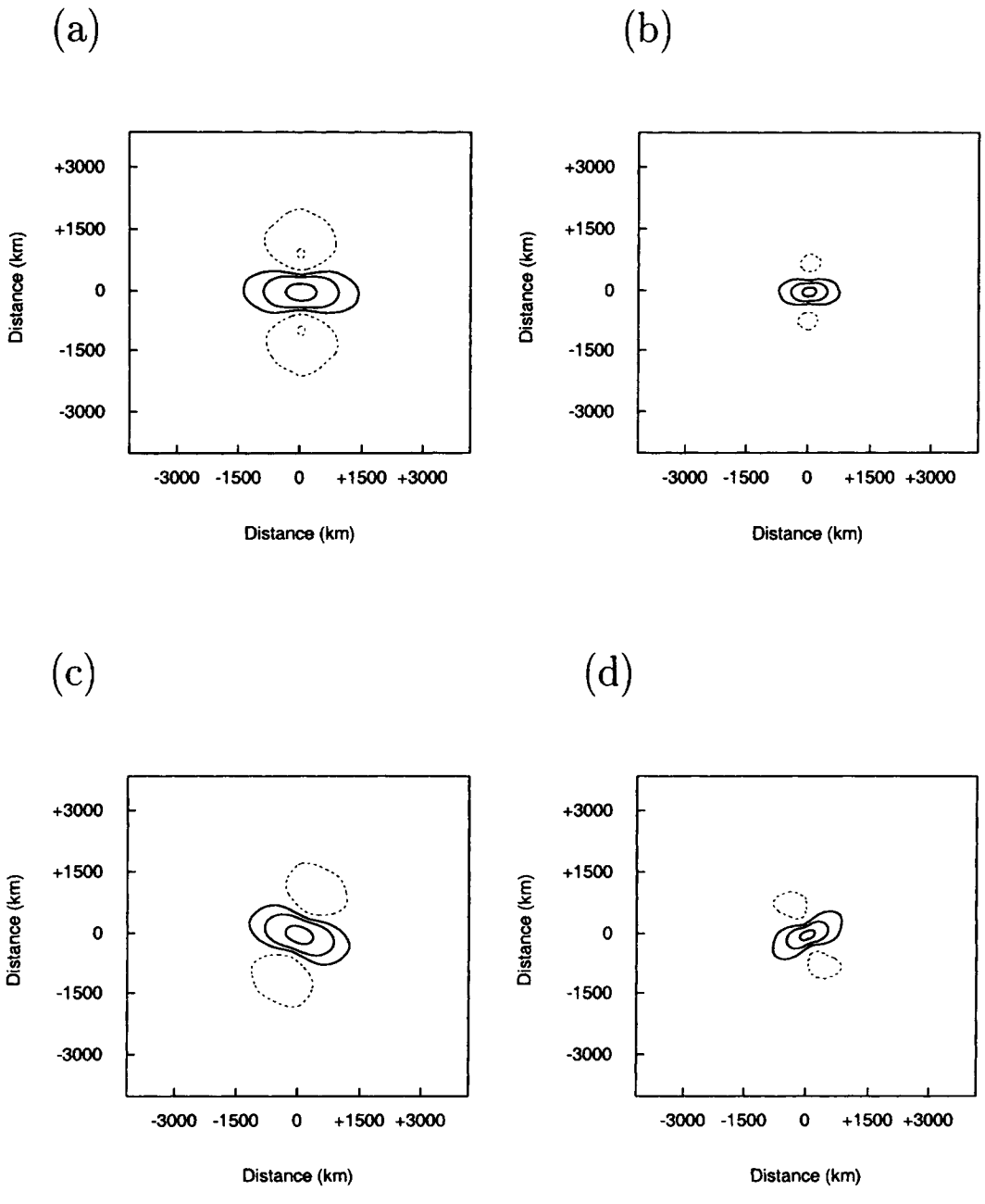

Fig. 9. Horizontal auto-correlations of the component in the x-direction for the full wind at levels 18 (500 hPa) and 31 (1000 hPa) [panels (a) and (c) respectively], and for the ageostrophic wind at levels 18 (500 hPa) and 31 (1000 hPa) [panels (b) and (d) respectively]. Positive (resp. negative) values are indicated by full (resp. dashed) isolines. From the point in the middle of the domain, the first two isolines are 0.5 and 0.15, and then the isoline spacing is 0.1 for lower values; the zero contour has been omitted.

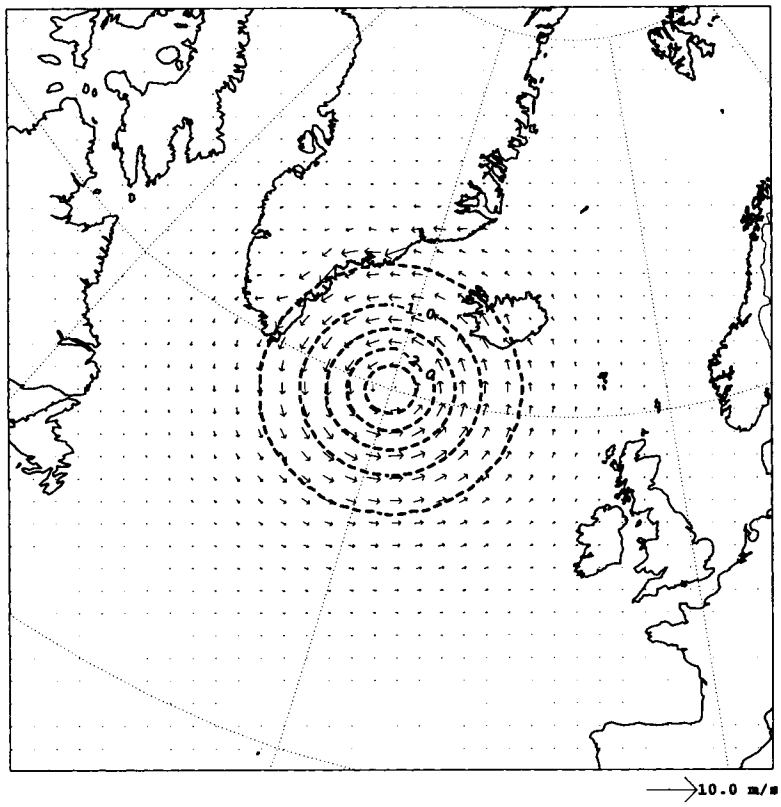

Fig. 10. Impact of a single 500 hPa temperature observation increment of −3 K on temperature and wind analysis increments of model level 18. Isoline spacing is 0.5 K. Dashed lines indicate negative increment values. Wind arrows have been plotted at every 3rd grid point.

isotropic in the horizontal with a maximum impact of about −2.8 K. The inserted temperature observation was assumed to have a very small observation error with a standard deviation of 0.2 K, which explains the strong impact of this single observation. The vertical cross-section illustrates the decrease in correlation with vertical distance between the observation level and the level of the analysis. The change in sign of the vertical correlations above the tropopause is also clearly illustrated: with a cooling of the troposphere we will have a lower and warmer tropopause. The geostrophic relation built into the structure functions is also illustrated in Fig. 10. With cooler mid-tropospheric temperatures, we will have lower mid-tropospheric pressure level heights and a corresponding cyclonic assimilation increment in the wind field.

The model level 21 wind and model level 18 temperature analysis increments for a single westerly component wind observation increment of 5 ms^{-1} at 500 hPa is illustrated in Fig. 12. We can notice that, due to the assumed larger error variances of the geostrophic part of the flow, the increment flow is dominated by a non-divergent (rotational) component with an elliptical influence area for the observed westerly wind component, a cyclonic flow to the north of the observation position, and an anti-cyclonic flow to the south of the observation position. The corresponding temperature increments on model level 18 are in close geostrophic balance with the cold temperature increments to the north of the wind observation and the warm temperature increments to the south of the wind observation.

5.2. *Sensitivity to the lateral boundary formulation*

Three options for handling of assimilation increments near the lateral boundaries have been tested

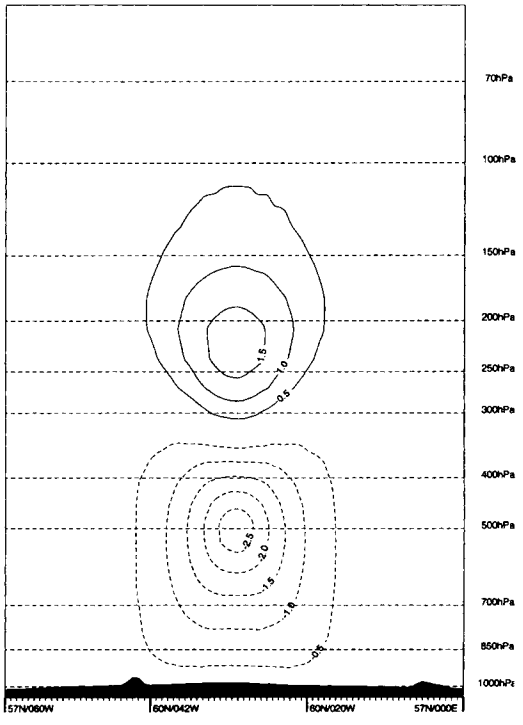

Fig. 11. Vertical cross-section in the west–east direction of temperature analysis increments from a single 500 hPa temperature observation increment of −3 K. Isoline spacing is 0.5 K.

in the HIRLAM 3D-Var: (1) Force the assimilation increments to be zero along the lateral boundaries, (2) Force the assimilation increments to be zero along the lateral boundaries with a gradual increase of the amplitude of the increments towards the interior of the assimilation area and (3) No forcing of the assimilation increments to be zero along the lateral boundaries. The first option, a simple forcing of the assimilation increments to be zero along the lateral boundaries turned out to give very un-realistic assimilation increments close to the lateral boundaries, and will not be described in the following.

The effect of inserting a 500 hPa temperature observation close to the eastern lateral boundary at 50°N 60°E is illustrated for model level 18 temperature and wind increments in Fig. 13. We let the analysis increments gradually decrease to zero towards the lateral boundaries in this experiment. This is achieved by multiplying the background error standard deviations with a cosine-shaped scaling factor with a value 0.01 at the lateral boundary and with a value 1.0 approximately 300 km from the lateral boundary inside the assimilation area. This selection of the scaling factor is somewhat arbritary, it was chosen mainly to be consistent with the boundary relaxation scheme that is used in the spectral HIRLAM model (Gustafsson, 1999). As can be seen from Fig. 13, the effect of this procedure is a strong gradient in the temperature analysis increments approximately 150 km from the lateral boundary, where the horizontal derivative of the cosine-shaped scaling factor has its maximum. We can also notice that the wind increments are in near geostrophic balance. With regard to the effects of the assumption that analysis increments have a bi-periodic variation over the extended integration area, very minor effects could possibly be seen in the wind field increments at the western boundary (not shown in the figure).

With the area extension technique, it is no longer necessary to assume that the analysis increments are zero on the lateral boundaries. The control of the lateral boundaries during 3D-Var, and particularly during 4D-Var, is an open unsolved problem, however. It may be argued that the lateral boundaries should be controlled through the data assimilation in the model that provides the lateral boundary conditions. On the other hand, the timing in the production of these lateral boundary conditions may have the effect that available lateral boundaries originate from an earlier, less accurate, forecast run.

With particular reference to 3D-Var applications, one may also argue that the boundary relaxation procedure during the ensuing forecast model integration will throw away increments of lateral boundary analysis in any case. One should then notice, however, that the degree of this effect depends on the particular selection of boundary values for the initial time of the forecast model integration. If the analysis itself is used as the lateral boundary condition for the initial time, the analysis increments along the lateral boundaries are utilized during the time period between the first two boundary data sets and may be advected into the inner integration area and thus also affect the forecast. If the lateral boundary values for the initial time come from another model, the effect of the analysis increments along the lateral boundaries will of course be negligible.

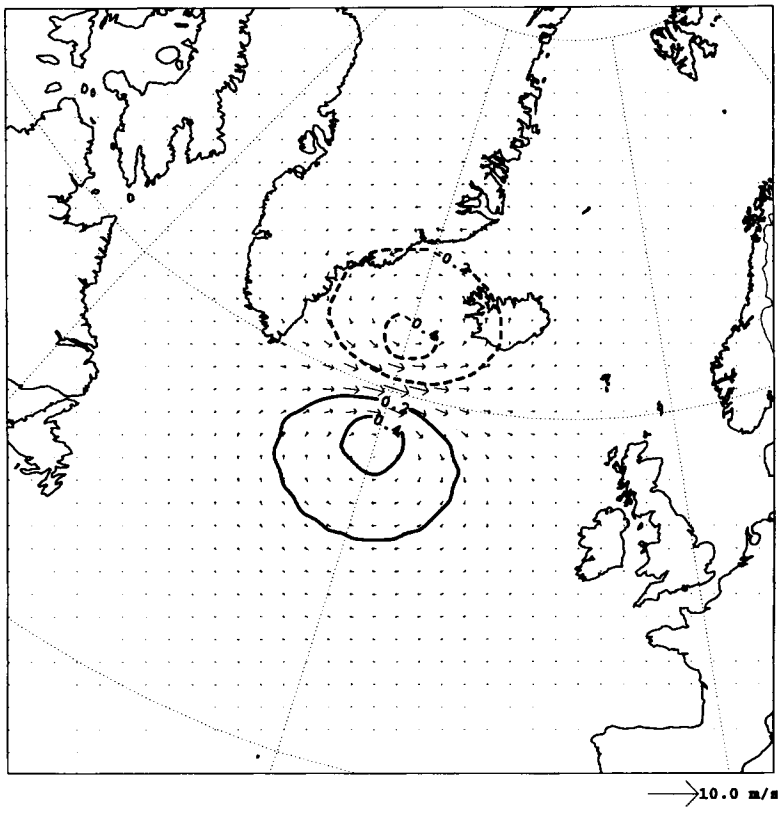

Fig. 12. Impact of a single 500 hPa westerly wind observation increment of 5 ms^{-1} on temperature analysis increments of model level 21 (approximately at 600 hPa) and wind analysis increments of model level 18 (approximately at 500 hPa). Wind arrows have been plotted at every 3rd grid point and temperature isoline spacing is 0.2 K.

If one considers 4D-Var applications as well, it has already been demonstrated by Gustafsson et al. (1998) that lateral boundary conditions may be controlled and improved by down-stream observed information well inside the lateral boundaries. This is achieved through the (up-stream) advection of cost function gradient information by the adjoint model. Since, in such a case, one is able to improve the lateral boundaries by advection of observed information, it is also natural to try to improve the lateral boundaries directly through observations and the analysis structure functions. We will now describe a sensitivity experiment where we remove the condition on the increments of the lateral boundary analyses to be zero-valued.

The effect of inserting a single 500 hPa temperature observation close to the eastern lateral boundary, again at 50°N 60°E but now with the analysis increments along the lateral boundaries being analyzed, is illustrated for temperature and wind increments of model level 18 in Fig. 14. The inserted observed value as well as the standard deviation of the observation error were exactly the same as in the previous experiment. It is obvious from these results that, at least for this simple case of a single observation, the calculation of analysis increments in the grid points along the lateral boundaries does not pose any particular difficulty. Neither does the assumption that analysis increments have a bi-periodic variation over the extended integration area introduce any increased negative effects in the case when the lateral boundary grid points are being analyzed. It is important to point out, however, that this result was made possible through the use of an

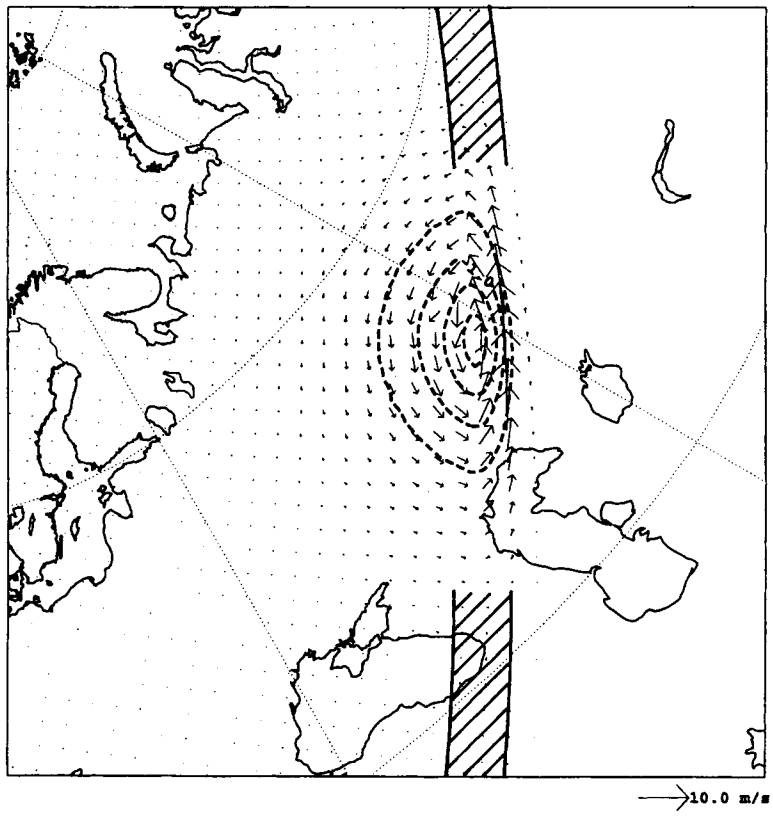

Fig. 13. Impact of a single 500 hPa temperature observation increment of −3 K on temperature and wind analysis increments of model level 18. Increments have been forced to zero along the lateral boundaries. Isoline spacing is 0.5 K. Dashed lines indicate negative increment values. Wind arrows have been plotted at every 3rd grid point. The width of the boundary relaxation zone has been indicated by shading.

extension zone to obtain bi-periodic variations and through the use of full Fourier transforms. With sine transforms it will of course not be possible to analyze the lateral boundary grid points. One should also be careful in the definition of the extension zone: With smaller analysis areas and increased horizontal resolutions, the extension zone must be selected so as to be relatively larger than in the current case.

6. Diagnostics of a single data assimilation cycle

The performance of the HIRLAM 3D-Var scheme has been evaluated through parallel assimilation and forecast experiments comparing 3D-Var results with results from the operational data assimilation based on statistical interpolation (Lorenc, 1981). Most of the results of these comparison experiments are presented in the companion paper (Lindskog et al., 2001). We will here only present some examples of diagnostics directly related to the performance of the background error constraint, the main subject of the present paper. The spatial structures of the assimilation increments were dealt with in the previous section and we will restrict the presentation to aspects of (a) Convergence of the minimization, (b) Fit of assimilated fields to observations and (c) Model balance and initialization effects.

The parallel data assimilation and forecast experiments were carried out for one 30-day winter period and one 30-day summer/early autumn

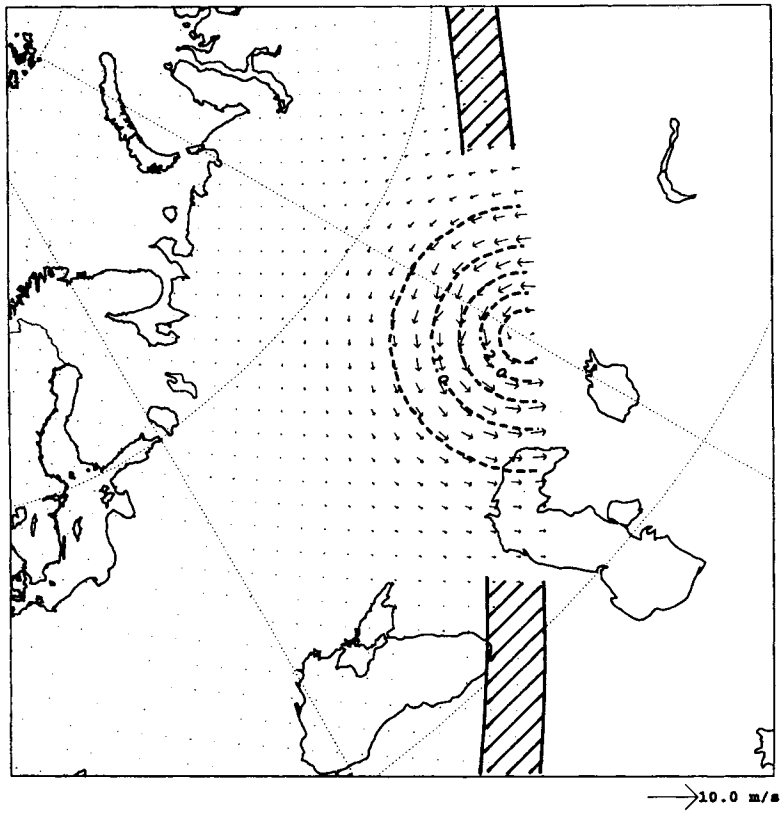

Fig. 14. Impact of a single 500 hPa temperature observation increment of −3 K on temperature and wind analysis increments of model level 18. Increments are analyzed along the lateral boundaries. Isoline spacing is 0.5 K. Dashed lines indicate negative increment values. Wind arrows have been plotted at every 3rd grid point. The width of the boundary relaxation zone has been indicated by shading.

period. The winter period was 10 February–9 March 1998. The experiments were done for the following combinations: (1) Statistical interpolation with the HIRLAM grid point model, (2) 3D-Var with the HIRLAM grid point model and (3) 3D-Var with the HIRLAM spectral model. The model geometry was as described for the single observation influence experiments: horizontal resolution of $0.4° \times 0.4°$ with 31 vertical levels and with 202×178 grid points on the operational SMHI area. We will present a selection of diagnostic results for a randomly chosen assimilation cycle, 5 March 1998 12 UTC, belonging to the winter season experimentation period. Since the diagnostic results for the spectral and the grid point models were almost identical, we have chosen to show results from the grid point model experiment only. It can be argued that the grid point model provides a more severe test of the performance of the incremental HIRLAM 3D-Var formulation, since the background error constraint, and in particular the balance constraint, has been designed on the basis of the spectral model and its geometry.

6.1. *Convergence of the minimization*

The convergence of the minimization process to determine 3D-Var assimilation increments for the selected assimilation cycle is illustrated in Fig. 15. The cost function value for each evaluation of this value (and its gradient) during the minimization is presented. The convergence criterion for this illustration was selected such that the norm of the

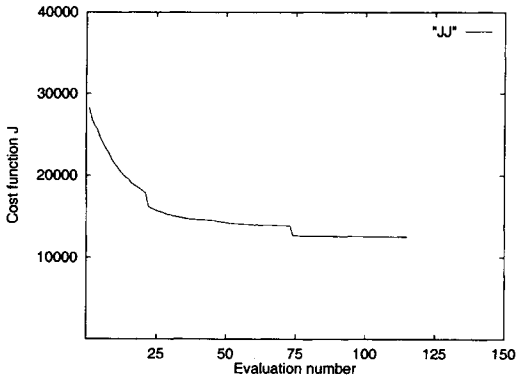

Fig. 15. Cost function value as a function of the cost function evaluation number during 3D-Var minimization for the assimilation for 5 March 1998 12 UTC with the grid point version of HIRLAM.

gradient of the cost function should decrease by a factor 100. With this criterion for the minimization package M1QN3 (Gilbert and Lemaréchal, 1989), the cost function evaluation continued until 115 function evaluations. At this stage in the minimization process, the cost function had decreased by a factor 2. It can be noticed, however, that very little reduction of the cost function value is achieved after approximately 80 cost function evaluations. This is typical for almost all assimilation cycles carried out during the parallel assimilation experiments. It indicates that the maximum number of cost function evaluations may be reduced to approximately 80 in order to reduce computing time for operational applications of the 3D-Var scheme. The large cost function reduction at cost function evaluation number 20 can be explained by the activation of the variational quality control at this specific point. The further large reduction around cost function evaluation number 75 can be explained by the rejection at this point of observations assigned small weights by the variational quality control. The variational quality control is described and the performance of the variational quality control is further examined in the companion paper (Lindskog et al., 2001).

6.2. *Observation fit statistics*

One way to evaluate the performance of the 3D-Var assimilation is to compare statistics of the deviations between observed values and analyzed values with the corresponding statistics of deviations between observed values and background field values. One example of such statistics is given in Fig. 16 for radiosonde observations utilized during the 3D-Var analysis cycle for 5 March 1998 12 UTC with the grid point model +6 h forecast utilized as the analysis background field. Only observations accepted by the variational quality control are included in the statistics. First of all, we can notice that the number of observed values is of the order of 200–800 in layers of 50–150 hPa thickness, while the number of radiosonde soundings in the HIRLAM area is of the order of 100–150. This occurs because all data from significant levels are used in the HIRLAM 3D-Var. This is in contrast to the HIRLAM statistical interpolation analysis scheme, for which only standard pressure level data are utilized. We can also notice that the observed values are significantly closer to the analyzed values than to the background values. This indicates that the 3D-Var analysis has been drawn reasonably well for the observations. Ideally the observed values in the rms measure should deviate from the analyzed values according to the assumed standard deviations of the observation errors. Assumed standard deviations for the observation errors are, for example, approximately 2 m/s for wind components and 0.8 K for temperature in the lower troposphere increasing to 3–4 m/s and 1.1 K, respectively, in the upper troposphere and lower stratosphere. These values agree well with the rms values in Fig. 16.

6.3. *Balance and initialization increments*

Any intermittent data assimilation can be taken as consisting of three types of increments added to the model state variables for every assimilation cycle: (1) the analysis increment, (2) the initialization increment and (3) the forecast increment reflecting the atmospheric change over the time period of the assimilation cycle. For an ideal data assimilation system (Daley, 1991, page 28) the forecast increments should have the largest amplitudes, followed by the analysis increments, while the initialization increments ideally should be very small, taking care of, for example, non-linear deviations (Machenhauer, 1977) from simpler balance relations included in the assimilation background error constraint.

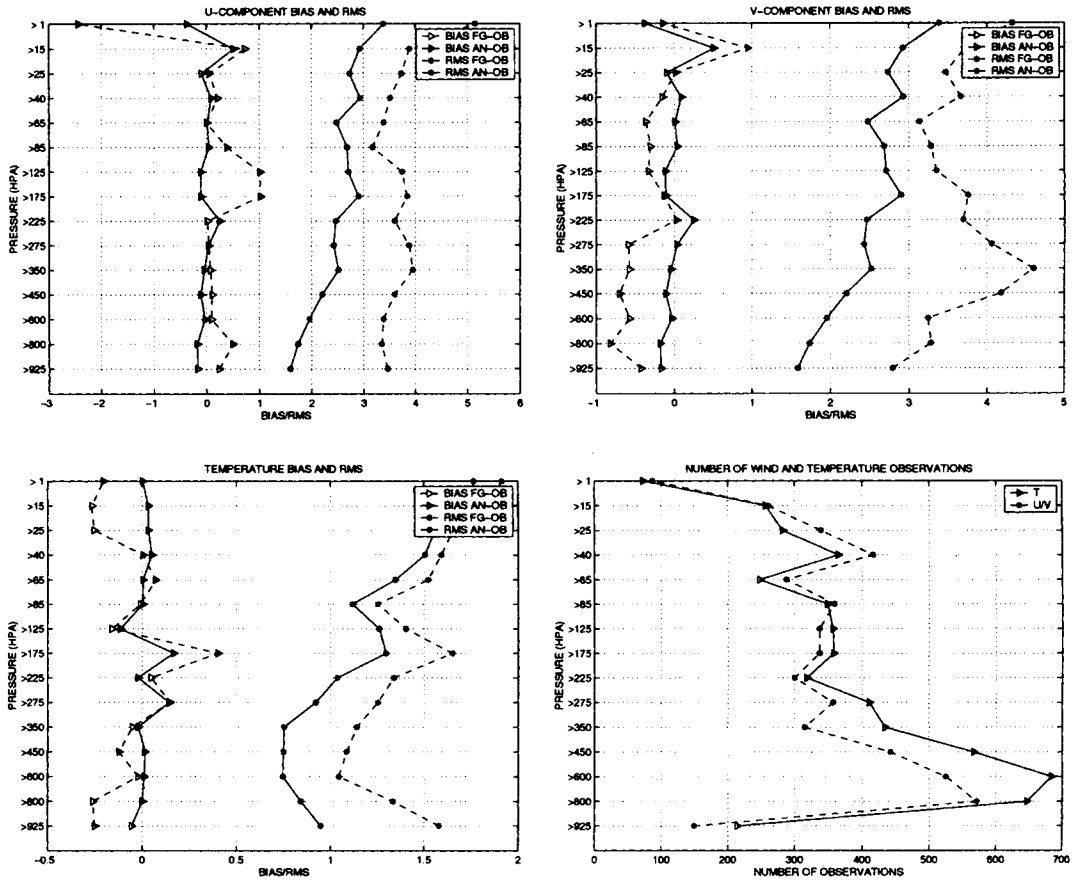

Fig. 16. Statistics of observed minus background (dashed lines) and observed minus analysis values (full lines) for u-wind component (top left), v-wind component (top right) and temperature (bottom left). Bias (marked with $\triangleright$) and root mean square (marked with *). Number of temperature observations (bottom right, full line) and number of wind observations (bottom right, dashed line). 3D-Var analysis for 5 March 1998 12 UTC with a +6 h grid point HIRLAM background field.

To illustrate that the HIRLAM 3D-Var assimilation compared to the operational assimilation based on statistical interpolation means a step forward toward the ideal data assimilation, surface pressure initialization increments for both assimilation schemes with the grid point model and for the assimilation cycle 5 March 1998 12 UTC are presented in Fig. 17. It is clear that the initialization increments are significantly smaller for the 3D-Var assimilation than for the statistical interpolation assimilation, in agreement with similar results for global models (Parrish and Derber, 1992; Gauthier et al., 1999). Non-linear normal mode initialization (Machenhauer, 1977) is applied in both assimilation schemes. It may be added that the global usage of data inherent in 3D-Var, as compared to the box data selection scheme (Lorenc, 1981) in the statistical interpolation, is likely to be the main contributing factor to the improved model balance of the 3D-Var assimilation (Cohn et al., 1998).

Another way to illustrate the improved model balance achieved through the HIRLAM 3D-Var is to study diagnostic quantities related to the level of high frequency oscillations (gravity wave noise) in forecast runs without any initialization. One such diagnostic quantity, related to the level of activity of the external gravity waves, is the

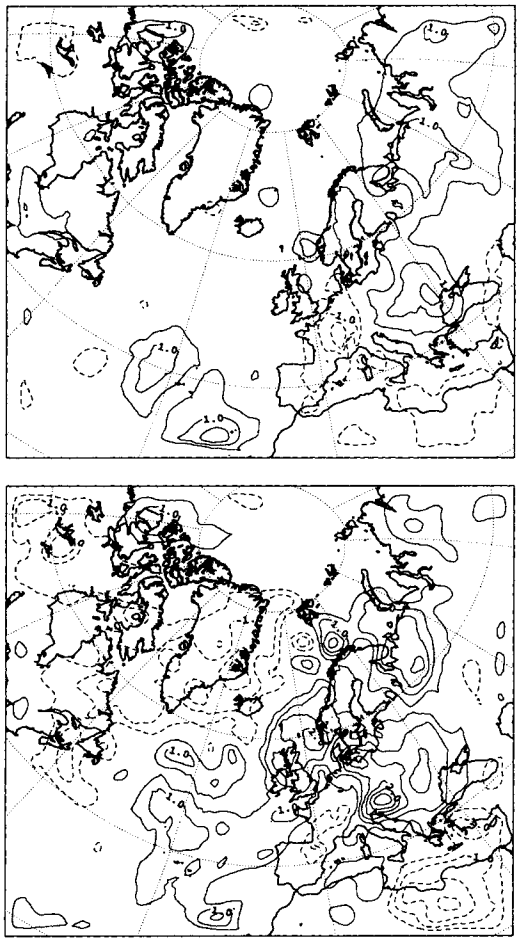

Fig. 17. Initialization increments for surface pressure, 5 March 1998 12 UTC. 3D-Var (top) and statistical interpolation (bottom). Isoline spacing is 0.5 hPa and dashed lines indicate negative increment values. The isoline for zero increments has been omitted.

horizontal average of the absolute surface pressure tendency during the first hours of model integration. This quantity has been plotted for forecasts based on statistical interpolation and 3D-Var assimilation, both with and without initialization, in Fig. 18. It is obvious that the level of noise in the forecast based on 3D-Var without any initialization is significantly lower than for the corresponding forecast based on statistical interpolation. It is also clear that the level of noise is significantly reduced by the aid of the initialization for both assimilation schemes.

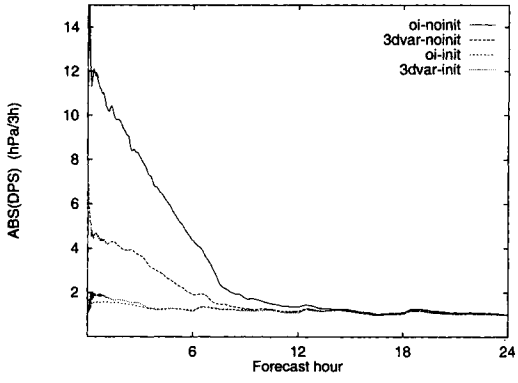

Fig. 18. Model area average of the absolute surface pressure tendency as a function of forecast length with the grid point HIRLAM model. Initial time is 5 March 1998 12 UTC. Statistical interpolation without initialization (curve marked oi-noinit) and with initialization (oi-init); 3D-Var without initialization (3dvar-noinit) and with initialization (3dvar-init).

7. Concluding remarks

The first version of a 3-dimensional variational data assimilation (3D-Var) for HIRLAM has been developed. It has been given an incremental formulation to permit application with the spectral as well as the grid point HIRLAM models. A change of variable is introduced to ensure a good preconditioning of the minimization. It is designed as an operator transforming the forecast error in the model space into a variable whose covariance matrix is an identity matrix. Such an operator is an inverse square root of the forecast error covariance matrix. Applying it to the analysis increment provides a control variable for which the expression of J_b is simple. The main transformations are the subtraction of the geostrophic wind increments, the spectral bi-Fourier transforms and the projection onto the eigenvectors of the vertical forecast error correlation matrices.

The geostrophic part of the wind component increments are calculated from the mass field increments via a tangent-linear geostrophic balance derived from the spectal model equations. Analysis structure functions for the 3D-Var were derived by the NMC method from the time history of operational SMHI HIRLAM forecasts. The first version of these structure functions was based on the assumption of homogeneity and isotropy with regard to horizontal and vertical forecast

error covariances of temperature, surface pressure, humidity and ageostrophic wind components. The bi-Fourier formulation of the structure functions allows representation of non-separable features in a limited area model, namely the height dependence of horizontal correlations and the scale dependence of vertical correlations.

The performance of the HIRLAM 3D-Var has been demonstrated by single observation impact experiments and by diagnostics of a single data assimilation cycle using operational observation data. It has been shown that assimilation increments have spatial structures in accordance with the structure functions, that assimilation increments are well balanced, and that the fit between utilized observations and assimilation fields is good. A more elaborate performance evaluation, including parallel assimilation and forecast experiments to compare with the operational HIRLAM data assimilation based on statistical interpolation, is the subject of a companion paper (Lindskog et al., 2001). The results of these parallel experiments indicate that the HIRLAM 3D-Var outperforms the reference HIRLAM data assimilation based on statistical interpolation. Major factors contributing to the superiority of the 3D-Var are believed to be the global usage of data, as compared to the box data selection in statistical interpolation, the non-separable structure functions, the variational quality control and the use of all significant level data in multi-level observation reports.

The main emphasis in this paper has been on the 3D-Var assimilation of wind and (dry) mass variables. The moisture assimilation certainly needs some further consideration, for example introduction of temperature-dependent background error standard deviations for specific humidity, or a change to relative humidity as the moisture assimilation control variable.

The further development of the HIRLAM variational data assimilation includes analysis structure functions based on a generalized statistical balance, including also the moisture field (Berre, 2000), utilization of GPS (Global Positioning System) atmospheric delay data, satellite radiance data, as well as radar radial wind vector data and an incremental 4-dimensional variational data assimilation (4D-Var). The first trials with HIRLAM 4D-Var have already been carried out successfully.

8. Acknowledgements

There are many persons who have helped us in the development of the HIRLAM 3-dimensional variational data assimilation. Without any particular order of preference, we would especially like to thank Erik Andersson, Philippe Courtier, Declan Horgan, Erland Källén, Peter Lönnberg, Jean Pailleux, Olivier Talagrand, Jean-Noël Thépaut, Per Undén and Drasko Vasiljević. We also thank the anonymous reviewers for useful comments. The development of the HIRLAM 3D-Var was partly carried out with support from the European Commission's programme Environment and Climate (Project NEWBALTIC II; Contract No. ENV4-CT97-0626) and with support for assimilation of ground-based GPS data from the Swedish National Space Board (Contract DNr 153/98 SSF).

REFERENCES

Berre, L. 1997. Non-separable structure functions for the HIRLAM 3DVAR. *HIRLAM Technical report* **30**, November 1997, 40 pp.

Berre, L. 2000. Estimation of synoptic and meso scale forecast error covariances in a limited area model. *Mon. Wea. Rev.* **128**, 644–667.

Boer, G. J. 1983. Homogeneous and isotropic turbulence on the sphere. *J. Atm. Sci.* **40**, 154–163.

Bouttier, F. 1996. Application of Kalman filtering to numerical weather prediction. *Proceedings 1996 ECMWF Seminar on Data assimilation and Workshop on Non-linear aspects of data assimilation.* ECMWF, Reading, UK, pp. 61–90.

Cohn, S. E., Da Silva, A., Guo, J., Sienkiewicz, M. and Lamich, D. 1998. Assessing the effects of data selection with the DAO physical-space statistical analysis system. *Mon. Wea. Rev.* **126**, 2913–2926.

Courtier, P., Thépaut, J.-N. and Hollingsworth, A. 1994. A strategy for operational implementation of 4D-Var using an incremental approach. *Q. J. Roy. Meteorol. Soc.* **120**, 1367–1388.

Courtier, P., Andersson, E., Heckley, W., Pailleux, J., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F. and Fisher, M. 1998. The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part I: Formulation. *Q. J. Roy. Meteorol. Soc.* **124**, 1783–1808.

Daley, R. 1985. The analysis of synoptic scale divergence

by a statistical interpolation procedure. *Mon. Wea. Rev.* **113**, 1066–1079.

Daley, R. 1991. *Atmospheric data analysis.* Cambridge University Press, Cambridge, UK, 460 pp.

Derber, J. and Bouttier, F. 1999. A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus* **51A**, 195–221.

Fisher, M. and Courtier, P. 1995. Estimating the covariance matrix of analysis and forecast errors in variational data assimilation. ECMWF Res. Dep. Technical Memorandum, No. 220.

Gauthier, P., Charette, C., Fillion, L., Koclas, P. and Laroche, S. 1999. Implementation of a 3D variational data assimilation system at the Canadian Meteorological Centre. Part I: the global analysis. *Atmosphere-Ocean.* **37**, 103–156.

Gilbert, J. C. and Lemaréchal, C. 1989. Some numerical experiments with variable storage quasi-Newton algorithms. *Math.* Prog. **B25**, 407–435.

Gustafsson, N. 1999. The numerical scheme and lateral boundary conditions for the spectral HIRLAM and its adjoint. *ECMWF Seminar Proceedings. Recent developments in numerical methods for atmospheric modelling.* ECMWF, Reading, UK, 7–11 September 1998, pp. 335–363.

Gustafsson, N. and Huang, X.-Y. 1996. Sensitivity experiments with the spectral HIRLAM and its adjoint. *Tellus* **48A**, 501–517.

Gustafsson, N., Lönnberg, P. and Pailleux, J. 1997. Data assimilation for high resolution limited area models. *J. Meteorol. Soc. Japan* **75**, 367–382.

Gustafsson, N., Källén, E. and Thorsteinsson, S. 1998. Sensitivity of forecast errors to initial and lateral boundary conditions. *Tellus* **50A**, 167–185.

Haugen, J.-E. and Machenhauer, B. 1993. A spectral limited-area model formulation with time-dependent boundary conditions applied to the shallow-water equations. *Mon. Wea. Rev.* **121**, 2618–2630.

Hollingsworth, A. and Lönnberg, P. 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: the wind field. *Tellus* **38A**, 111–136.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H. and Mitchell, H. L. 1996. A system simulation approach to ensemble prediction. *Mon. Wea. Rev.* **124**, 1225–1242.

Huang, X.-Y., Gustafsson, N. and Källén, E. 1997. Using an adjoint model to improve an optimum interpolation based data assimilation system. *Tellus* **49A**, 161–176.

Laroche, S., Gauthier, P., St-James, J. and Morneau J. 1999. Implementation of a 3D variational data assimilation system at the Canadian Meteorological Centre, Part II: the regional analysis. *Atmosphere-Ocean* **37**, 281–307.

Le Dimet, F. X. and Talagrand, O. 1986. Variational algorithms for analysis and assimilation of meteorological observations. Theoretical aspects. *Tellus* **38A**, 97–110.

Lewis, J. M. and Derber, J. C. 1985. The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus* **37A**, 309–327.

Lindskog, M., Gustafsson, N., Navascués, B., Mogensen, K. S., Huang, X.-Y., Yang, X., Andræ, U., Berre, L., Thorsteinsson, S. and Rantakokko, J. 2001. Three-dimensional variational data assimilation for a limited area model. Part II: observation handling and assimilation experiments. *Tellus* **53A**, this issue.

Lorenc, A. 1981. A global three-dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.* **109**, 701–721.

Lorenc, A. 1986. Analysis methods for numerical weather prediction. *Q. J. R. Meteor. Soc.* **112**, 1177–1194.

Lorenc, A. 1988. Optimal nonlinear objective analysis. *Q. J. Roy. Meteor. Soc.* **114**, 205–240.

Lorenc, A. 1997. Development of an operational variational assimilation scheme. *J. Meteorol. Soc. Japan* **75**, 339–346.

Lorenc, A. C., Ballard, S. P., Bell, R. S., Ingleby, N. B., Andrews, P. L. F., Barker, D. M., Bray, J. R., Clayton, A. M., Dalby, T., Li, D., Payne, T. J. and Saunders, F. W. 2000. The Met. Office Global 3-Dimensional Variational Data Assimilation Scheme. *Q. J. Roy. Meteor. Soc.*, **26**, 2991–3012.

Machenhauer, B. 1977. On the dynamics of gravity oscillations in a shallow water model, with application to normal mode initialization. *Beitr. Phys. Atmos.* **50**, 253–271.

Parrish, D. F. and Derber, J. C. 1992. The National Meteorological Centre's spectral statistical interpolation analysis system. *Mon. Wea. Rev.* **120**, 1747–1763.

Rabier, F., McNally, A., Andersson, E., Courtier, P., Undén, P., Eyre, J., Hollingsworth, A. and Bouttier, F. 1998a. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). Part II: structure functions. *Q. J. Roy. Meteorol. Soc.* **124**, 1809–1830.

Rabier, F., Thépaut, J.-N. and Courtier, P. 1998b. Extended assimilation and forecast experiments with a four-dimensional variational assimilation system. *Q. J. Roy. Meteorol. Soc.* **124**, 1861–1888.

Sadiki, W., Fischer, C. and Geleyn, J.-F. 2000. Mesoscale background error covariances: recent results obtained with the limited-area model ALADIN over Morocco. *Mon. Wea. Rev.* **128**, 3927–3935.

Sasaki, Y. 1958. An objective analysis based on the variational method. *J. Meteorol. Soc. Japan* **36**, 77–88.

Simmons, A. J. and Burridge, D. M. 1981. An energy and angular momentum conserving vertical finite-difference scheme and hybrid vertical coordinates. *Mon. Wea. Rev.* **109**, 758–766.

Thiébaux, H. J., Mitchell, H. L. and Shantz, D. W. 1986. Horizontal structure of hemispheric forecast error correlations for geopotential and temperature. *Mon. Wea. Rev.* **114**, 1048–1066.

Zupanski, M. 1993. Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Mon. Wea. Rev.* **121**, 2396–2408.