IFS DOCUMENTATION

PART II: DATA ASSIMILATION (CY23R4)

Edited by Peter W. White

(Text written and updated by members of the ECMWF Research Department)

Table of contents

Chapter 1 'Incremental formulation of 3D/4D variational assimilation—an overview'

Chapter 2 '3D variational assimilation'

Chapter 3 '4D variational assimilation'

Chapter 4 'Background term'

Chapter 5 'Conventional observational constraints'

Chapter 6 'Satellite observational constraints'

Chapter 7 'Background, analysis and forecast errors'

Chapter 8 'Gravity-wave control'

Chapter 9 'Data partitioning (OBSORT)'

Chapter 10 'Observation screening'

Chapter 11 'Analysis of snow'

Chapter 12 'Land surface analysis'

Chapter 13 'Sea surface temperature and sea-ice analysis'

Chapter 14 'Reduced-rank Kalman filter'

REFERENCES

Part II: DATA ASSIMILATION

CHAPTER 1 Incremental formulation of 3D/4D variational assimilation—an overview

Table of contents

1.1 Introduction

1.2 Incremental Formulation

1.3 Practical implementation

1.3.1 Data flow

1.3.2 Formation of high-resolution analysis

1.3.3 Humidity and ozone

1.4 Preconditioning and control variable

1.5 Minimization

1.1 INTRODUCTION

This documentation on 3D and 4D–Var is meant to serve as a scientific guide to the 3D/4D–Var codes, a part of the IFS. The documentation is divided into eleven chapters. This, the first chapter deals with the scientific formulation, the practical implementation of the incremental method, and it includes some comments on minimization and preconditioning. The code structure and the computational details of the 3D/4D-Var cost-functions and their gradients are explained in Chapter 2 '3D variational assimilation' . There is a separate chapter on subjects specific to 4D-Var (Chapter 3 '4D variational assimilation'). Thereafter follows a description of the background term (Chapter 4 'Background term') and two chapters respectively on observation operators for conventional data (Chapter 5 'Conventional observational constraints') and satellite data (Chapter 6 'Satellite observational constraints'). Chapter 7 'Background, analysis and forecast errors' deals with the computation of background and analysis errors and Chapter 8 'Gravity-wave control' is on initialization. The modules for observation sorting and screening are described in Chapter 9 'Data partitioning (OBSORT)' and Chapter 10 'Observation screening' . Chapter 11 outlines the snow analysis, Chapter 12 describes the Soil analysis, Chapter 13 describes the sea surface temperature and sea-ice analysis and the final chapter Chapter 14 provides details of the reduced Kalman filter.

An extensive scientific description of 3D/4D-Var has been published in QJRMS, in ECMWF workshop proceedings and Technical Memoranda over the years. The incremental formulation was introduced by *Courtier et al.* (1994). The ECMWF implementation of 3D-Var was published in a three-part paper by *Courtier et al.* (1998), *Rabier et al.* (1998) and *Andersson et al.* (1998). The observation operators for conventional data can be found in *Vasiljevic et al.* (1992). The methods for assimilation of TOVS radiance data and ERS scatterometer data were developed by *Andersson et al.* (1994) and *Stoffelen and Anderson* (1997), respectively. The pre-operational experimentation with 4D-Var has been documented in three papers by *Rabier et a*l. (1998), *Mahfouf and Rabier* (1998) and *Klinker et al.* (1999).

3D-Var was implemented in ECMWF operations on 30 January 1996. The three-part paper mentioned above chiefly presented the scheme as it was at that point in time. There have been very significant developments of the system

during its time in operations. The first upgrade took place in connection with the move from a CRAY C90 system to a distributed memory Fujitsu VPP700 machine. The observation handling and data screening modules were replaced with new codes, see Chapter 9 'Data partitioning (OBSORT)' and Chapter 10 'Observation screening', respectively, and the paper by *Järvinen and Undén* (1997). Variational quality control of observations (*Andersson and Järvinen*, 1999, and Section 2.6) and a new algorithm for computing estimates of analysis and background errors (*Fisher and Courtier* 1995, and Chapter 7 'Background, analysis and forecast errors') were introduced.

In May 1997 there was a complete revision of the background term, see *Derber and Bouttier* (1999) and Chapter 4 'Background term'. The old background term, which was described in *Courtier et al.* (1998), is not covered by this documentation as it is now considered obsolete. Later that year (25 November 1997) 6-hour 4D-Var was introduced operationally, at resolution T213L31, with two iterations of the outer loop: the first with 50 iterations (simplified physics) and the second with 20 iterations (with tangent-linear physics). In April 1998 the resolution was changed to T_L319 and in June 1998 we revised the radiosonde/pilot usage (significant levels, temperature instead of geopotential) and we started using time-sequnces of data (*Järvinen et al.* 1999), so-called 4D-screening. Finally, the data assimilation scheme was extended higher into the atmosphere on 10 March 1999, when the $T_L319L50$ model was introduced, which in turn enabled the introduction in May 1999 of ATOVS radiance data (McNally *et al.* 1999). In October 1999 the vertical resolution of the boundary layer was enhanced taking the number of model levels to a total of 60. In summer 2000 the 4D-Var period was extended from 6 to 12 hours, whereas the ERA configuration was built as an FGAT (first guess at the appropriate time) of 3D-Var with a period of 6 hours. At the time of writing it is planned to increase the horizontal resolution of 4D-Var to $T_L511L60$, with inner loop resolution enhanced from T63L60 to $T_L159L60$ using the linearized semi-Lagrangian scheme.

1.2 Incremental Formulation

3D/4D–Var attempt to minimize an objective function J consisting of three terms:

$$J = J_b + J_o + J_c \tag{1.1}$$

measuring, respectively, the discrepancy with the background (a short-range forecast started from the previous analysis), J_b, with the observations, J_o and with the slow character of the atmosphere, J_c. The J_c-term controls the amplitude of fast waves in the analysis and is described in Chapter 8 'Gravity-wave control'. It is omitted from the subsequent derivations in this section.

In its incremental formulation (*Courtier et al.* 1994), we write

$$J(\delta\mathbf{x}) = \frac{1}{2}\delta\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\delta\mathbf{x} + \frac{1}{2}(\mathbf{H}\delta\mathbf{x} - \mathbf{d})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{H}\delta\mathbf{x} - \mathbf{d}) \tag{1.2}$$

$\delta\mathbf{x}$ is the increment and at the minimum the resulting analysis increment $\delta\mathbf{x}^a$ is added to the background $\mathbf{x}^b$ in order to provide the analysis $\mathbf{x}^a$:

$$\mathbf{x}^a = \mathbf{x}^b + \delta\mathbf{x}^a \tag{1.3}$$

$\mathbf{B}$ is the covariance matrix of background error while $\mathbf{d}$ is the innovation vector,

$$\mathbf{d} = \mathbf{y}^o - H\mathbf{x}^b \tag{1.4}$$

where $\mathbf{y}^o$ is the observation vector. $\mathbf{H}$ is a suitable low-resolution linear approximation of the observation operator

H in the vicinity of $\mathbf{x}^b$, and $\mathbf{R}$ is the covariance matrix of observation errors. Alternatively, $\mathbf{H}\delta\mathbf{x}$ in Eq. (1.2) can be replaced by the finite difference $H\mathbf{x} - H\mathbf{x}^b$, approximated at low resolution. The incremental formulation of 3D/4D-Var consists therefore of solving for $\delta\mathbf{x}$ the inverse problem defined by the (direct) observation operator $\mathbf{H}$, given the innovation vector $\mathbf{d}$ and the background constraint. The gradient of J is obtained by differentiating Eq. (1.2) with respect to $\delta\mathbf{x}$,

$$\nabla J = (\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})\delta\mathbf{x} - \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{d} \tag{1.5}$$

At the minimum, the gradient of the objective function vanishes, thus from Eq. (1.5) we obtain the classical result that minimizing the objective function defined by Eq. (1.2) is a way of computing the following equivalent matrix-vector products:

$$\delta\mathbf{x}^a = (\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{d} = \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\mathbf{d} \tag{1.6}$$

where $\mathbf{B}$ and $\mathbf{R}$ are positive definite, see e.g. *Lorenc* (1986) for this standard result. $\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$ may be interpreted as the square matrix of the covariances of background errors in observation space while $\mathbf{B}\mathbf{H}^{\mathrm{T}}$ is the rectangular matrix of the covariances between the background errors in model space and the background errors in observation space.

Most (if not all) implementations of OI rely on a statistical model for describing $\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$ and $\mathbf{B}\mathbf{H}^{\mathrm{T}}$ (*Hollingsworth and Lönnberg*, 1986; *Lönnberg and Hollingsworth*, 1986 and *Bartello and Mitchell*, 1992). 3D-Var uses the observation operator $\mathbf{H}$ explicitly and, as OI, if a statistical model is required it is only used for describing the statistics of the background errors in model space. Consequently, in 3D/4D-Var it turns out to be easier, from an algorithmic point of view, to make use of observations such as TOVS radiances, which have a quite complex dependence on the basic analysis variables.

1.3 PRACTICAL IMPLEMENTATION

As mentioned earlier in Section 1.2, the formulation used is incremental (*Courtier et al.* 1994). In the ECMWF implementaion two different resolutions are used—one for the comparison with observations, which is the same as the deterministic medium-range forecast model, and a lower resolution for the minimization. Several different job steps are performed:

- *(i)* Comparison of the observations with the background at high resolution to compute the innovation vectors Eq. (1.4). These are stored in the NCMIFC1-word of the ODB (the observation database) for later use in the minimization. This job step also performs **screening** (i.e. blacklisting, thinning and quality control against the background) of observations (see Chapter 10 'Observation screening'). The screening determines which observations will be passed for use in the main minimisation. Very large volumes of data are present during the screening run only, for the purpose of data monitoring,
- *(ii)* First minimization at low resolution to produce preliminary low-resolution analysis increments, using simplified physics
- *(iii)* Update of the high-resolution trajectory to take non-linear effects partly into account. Observed departures from this new atmospheric state are stored in the ODB and the analysis problem is re-linearized around the updated model state,
- *(iv)* Second main minimization at low resolution with tangent-linear physics,
- *(v)* Formation of the high-resolution analysis (described below) and a comparison of the analysis with all observations (also those not used by the analysis, for diagnostic purposes).

(vi) Computation of analysis and background errors, currently at T42L60, as described in Chapter 7 'Background, analysis and forecast errors'

Each of the job steps is carried out by a different configuration of IFS. They are commonly called:

(i) **The first trajectory run** (which includes screening and is sometimes called **the screening run**) – conf=2, LSCREEN=.T.
(ii) **The main minimization,** simplified physics, conf=131, LSPHLC=.T.,
(iii) **The trajectory update**, conf=1, LOBS=.T.,
(iv) **The main minimization with physics**, conf=131, LSPHLC=.F.,
(v) **The final trajectory runs**, conf=1, LOBS=.T., NUPTRA=NRESUPD, with verification screening,
(vi) **The background error minimization**, conf=131, LAVCGL=.T.

A truncation operator (the IFS full-pos post-processing package) allows one to go from high-resolution fields to low resolution, using appropriate grid-point interpolations. The steps *(iii)* and *(iv)* are referred to as the second iteration of the *outer loop*, and these can optionally be iterated further to incorporate additional nonlinear effects. The trajectory update is not normally done in 3D-Var. The *inner loop* takes place within the main minimization, job steps *(iii)* and *(v)*.

1.3.1 Data flow

All files containing model fields are coded in GRIB (the GRIB format is described in GRIB.ps). The high-resolution background $\mathbf{x}^{\mathrm{b}}_{\mathrm{HR}}$ (the input to the first trajectory run) is obtained from a standard MARS retrieve to files reftrajshml (model-level spectral fields), reftrajggml (model-level grid-point fields, i.e. q and clouds) and reftrajggsfc (surface grid-point fields).

The $\mathbf{x}^{\mathrm{b}}_{\mathrm{HR}}$ is truncated to the resolution of the minimization to form $\mathbf{x}^{\mathrm{b}}_{\mathrm{LR}}$ (the low-resolution background), which is the input to the main minimization. The low resolution file names are backgroundshml, backgroundggsfc and backgroundggml, for upper-air spectral data, surface grid-point data, and upper-air grid-point data (i.e. clouds), respectively. Specific humidity q and ozone are represented in spectral space as they are spectral variables in the variational analysis. The three files are linked to the names **ICMRFxxxx0000**, **ICMSHxxxxINIT** and **ICMGGxxxxIMIN**, **ICMGGxxxxINIT, ICMGGxxxxINIUA** (where **xxxx** is the 'expver' identifier of MARS) and read in by SUSPEC, SUGRIDF from SUECGES.

The main minimization job writes out the low-resolution background $\mathbf{x}^{\mathrm{b}}_{\mathrm{LR}}$ (the previous high-resolution trajectory in the second minimization) to files **MXVA00000+000hhmm** (where hhmm is the time of the field) and the low-resolution analysis $\mathbf{x}^{\mathrm{a}}_{\mathrm{LR}}$ to files **MXVA00999+00hhmm**. This is done in a call to STEPO near the end of SIM4D, Section 2.3, by the routine WRMLPP. Both these files are are saved under names starting respectively with **spfglr** and **spanlr** and later read by IFS in the trajectory runs (using SUINIF, called from RDFPINC) and transformed to the higher resolution by filling with zeroes (operator T^{-1}), and q is transformed to gridpoint space. The trajectory runs also read in $\mathbf{x}^{\mathrm{b}}_{\mathrm{HR}}$ (from the **reftraj** files) using SUINIF called from CSTA. The analysis increment is formed in RDFPIN:

$$\delta\mathbf{x}^{\mathrm{a}}_{\mathrm{HR}} \;=\; T^{-1}[\mathbf{x}^{\mathrm{a}}_{\mathrm{LR}}] - T^{-1}[\mathrm{NNMI}(\mathbf{x}^{\mathrm{b}}_{\mathrm{LR}})] \tag{1.7}$$

1.3.2 Formation of high-resolution analysis

The analysis field is the sum of the background and of the pseudo-inverse of the truncation operator applied to the low-resolution analysis and background. This pseudo-inverse comprises filling in the spectral wave-numbers greater than the minimization resolution by zeroes (T^{-1}, applied in Eq. (1.7)). At this stage temperature is converted to virtual temperature, as required by the model. (note that normal-mode initialization is no longer applied).

1.3.3 Humidity and ozone

The humidity control variable used in the minimization is specific humidity in spectral space (LSPQ, NAMDIM). There is no constraint forcing the minimization to produce positive and non supersaturated values for this quantity. However, before the computation of TOVS observation departures in the minimization stage, (low-resolution) gridpoint values are replaced by $f(q)$ where f is a differentiable function such that it results in positive humidity values (routine QNEGLIM, called from TOVCLR). Super–saturated humidity gridpoint values can optionally, under the switch LNEGHYP (=.false., namrinc), be modified to be below 1.2 (hard coded) in terms of relative humidity. This would be done in the routine QNEGHYP, called from SCAN2MDM.

The high resolution analysis of q in gridpoint space, is modified (in SUGPQLIMDM, called by RDFPINC) by resetting negative humidities to zero and supersaturated values to saturated values.

The ozone control variable used in the minimization is ozone in spectral space (LSPO3). The increment is converted to gridpoint space when computing the high resolution analysis. For the time being no special security is applied to the ozone increments.

1.4 PRECONDITIONING AND CONTROL VARIABLE

In practice, it is necessary to precondition the minimization problem in order to obtain a quick convergence. As the Hessian (the second derivative) of the objective function is not accessible, *Lorenc* (1988) suggested the use of the Hessian of the background term J_b. The Hessian of J_b is the matrix $\mathbf{B}$. Such a preconditioning may be implemented either by a change of metric (i.e. a modified innner product) in the space of the control variable, or by a change of control variable. As the minimization algorithms have generally to evaluate several inner products, it was found more efficient to implement a change of variable (under CHAVAR, CHAVARIN etc). Algebraically, this requires the introduction of a variable χ such that

$$J_b = \chi^T\chi \tag{1.8}$$

Comparing Eq. (1.2) and Eq. (1.8) shows that $\chi = \mathbf{B}^{-1/2}\delta\mathbf{x}$ satisfies the requirement. χ thus becomes the *control variable* of the preconditioned problem. This is indeed what has been implemented, as will be explained in Section 4.2. A single-observation analysis with such preconditioning converges in one iteration.

1.5 MINIMIZATION

The minimization problem involved in this 3D/4D-Var can be considered as large-scale, since the number of degrees of freedom in the control variable is of the order of 10^6. An efficient descent algorithm was provided by the Institut de Recherche en Informatique et Automatique (INRIA, France). It is a variable-storage quasi-Newton algorithm (M1QN3, auxlib) described in *Gilbert and Lemaréchal* (1989) and in an on-line postscript document. M1QN3 uses the available in-core memory to update an approximation of the Hessian of the cost function (the array ZVATRA, see CVA1). In practice, ten updates (NMUPD, namiomi) of this Hessian matrix are used. The approximation is modified during the minimization by deleting information from the oldest gradient and inserting information from the most recent one. Once per iteration M1QN3 calls the *simulator* SIM4D (Section 2.3). Sometimes extra simulations have to be performed in order to obtain a good step length for the descent. The number of iterations is limited to 70 (NITER, namvar) and the number of simulations to 80 (NSIMU, namvar). Normally, only very small adjustments of the analysis occur during the second half of the minimization. On the whole, the cost function is typically divided by a factor of two and the norm of the gradient by a factor of twenty (printed from

EVCOST and SIM4D, respectively).

The approximation of the Hessian computed during a 3D/4D-Var minimization (read in by SUHESS) is used as a first estimate for a subsequent analysis, if the switch LWARM=.true. (namiomi). LWARM is only used in the second minimization of 4D-Var (see Section 3.3).

Part II: DATA ASSIMILATION

CHAPTER 2 3D variational assimilation

Table of contents

2.1 Introduction

2.2 Top-level controls

2.2.1 Gradient test

2.2.2 Iterative solution

2.2.3 Last simulation

2.3 A simulation

2.3.1 Interface between control variable and model arrays

2.4 Interpolation to observation points

2.4.1 Method

2.4.2 Storage in GOM-arrays

2.5 Computation of the observation cost function

2.5.1 Organization in observation sets

2.5.2 Cost function

2.5.3 tables

2.5.4 Correlation of observation error

2.6 Variational quality control

2.6.1 Description of the method

2.6.2 Implementation

2.6.3 Correlated data

2.1 INTRODUCTION

This part of the documentation covers the top level controls of 3D-Var (CVA1) and gives a detailed description of a 3D-Var simulation (SIM4D) All of this chapter also applies to 4D-Var with some additions which will be detailed in Chapter 3 '4D variational assimilation' . The interpolation of model fields to observation points (OBSHOR) and the organization of the data in memory (yomsp, yommvo) are also descibed. We explain the structure of the computation of the observation cost function (FJO and FJOS in yomcosjo1) and its gradient, managed by the routines OBSV and TASKOB. The background term will be explained in Chapter 4 'Background term' .

2.2 TOP-LEVEL CONTROLS

The routine CVA1 controls the variational configuration of IFS—its flow diagram is shown in Fig. 2.1 . The first guess fields (FG) have been read in to the SP7-arrays (in YOMSP) by SUECGES, called from SUJBSTD within the J_b setup, see Subsection 4.3.3. The FG is optionally initialized to be consistent with the lower resolution orography. This is done in a call to CNMI from SUECGES, controlled by the switch LFGNMI (=false if L50 or L60, in namjg), see also Chapter 8 'Gravity-wave control' .

At the start of CVA1 additional setups for the variational configurations are done (SU1YOM). The SP3-arrays, i.e. the current model state, (in YOMSP) are filled by copying from SP7, using SP7TO3. A call to CNT2 computes $H\mathbf{x}^{\mathrm{b}}_{\mathrm{LR}}$, which is required for the finite-difference version of Eq. (1.4) of the incremental 3D-Var. This call to CNT2 is charaterized by LOBSREF=.true. (in YOMCT0). The result, stored in the NCMIFC2-word of the ODB, is the *low-resolution departure* from the FG, $\mathbf{y}^{\mathrm{o}}-H\mathbf{x}^{\mathrm{b}}_{\mathrm{LR}}$, and will be used in later iterations, Eq. (2.5). If, however, the tangent linear observation operators are used, Eq. (2.6), $\mathbf{y}^{\mathrm{o}}-H\mathbf{x}^{\mathrm{b}}_{\mathrm{LR}}$ is not needed. It can optionally be computed and stored, if LCALCFC2=.true. (in yomrinc).

2.2.1 Gradient test

If LTEST=.true. a gradient test will be performed both before and after minimization. This is done by the routine GRTEST. In the gradient test a test value t_1 is computed as the ratio between a perturbation of the co-t-function and its first order Taylor expansion:

$$t_1 = \lim_{\delta\chi \to 0} \frac{J(\chi + \delta\chi) - J(\chi)}{\langle \nabla J, \delta\chi \rangle} \tag{2.1}$$

with $\delta\chi = -\alpha\nabla J$. Repeatedly increasing α by one order of magnitude, printing t_1 at each step should show t_1 approaching one, by one order of magnitude at a time, provided $J(\chi)$ is approximately quadratic over the interval $[\chi,\chi + \delta\chi]$. The near linear increase in the number of 9's in the print of t_1 over a wide range of α (initially as well as after minimization) proves that the coded adjoint is the proper adjoint for the linearization around the given state χ .

The behaviour of the cost function in the vicinity of χ in the direction of the gradient ∇J is also diagnosed by several additional quantities for each α . The results are printed out on lines in the log-file starting with the string 'GRTEST:'. To test the continuity of J , for example, a test value t_0 is computed:

$$t_0 = \frac{J(\chi + \delta\chi)}{J(\chi)} - 1 \tag{2.2}$$

and printed. For explanation of other printed quantities see the routine GRTEST itself.

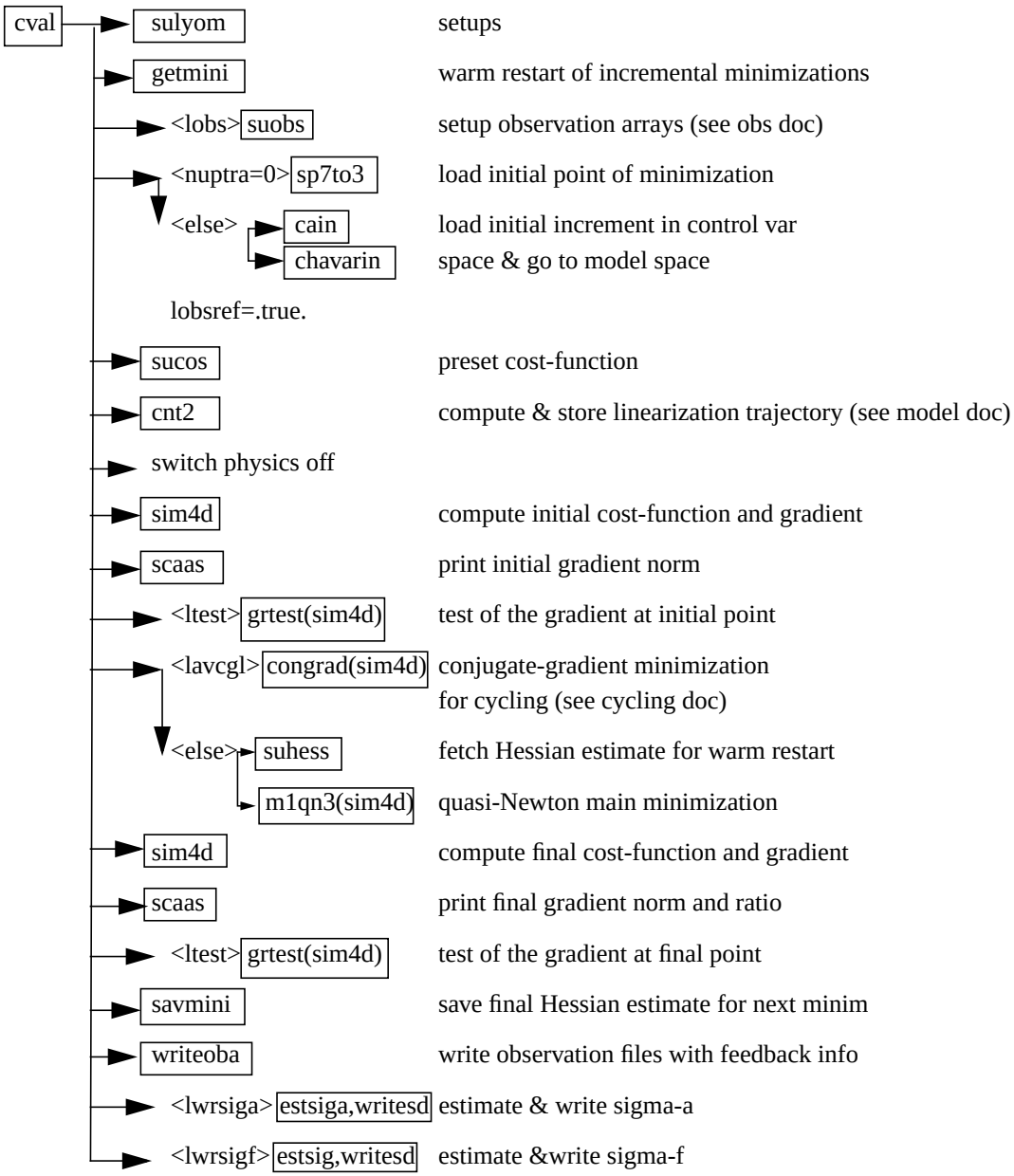

Figure 2.1 Flow diagram for subroutine cva1.

2.2.2 Iterative solution

When the cost function is exactly quadratic, as is the case in the background error estimation, conjugate gradient minimization (CONGRAD) can be used. This is controlled by the switch LAVCGL (namvar), and requires the use of tangent linear observation operators (LOBSTL=.true., namvrtl), tangent linear model (L131TL=.true., namvrtl) no VarQC (LVARQCG=.false., namjo) and de-aliased SCAT data (LQSCATT=.true., namjo).

In normal 3D/4D-Var, the cost function is allowed to be (weakly) nonlinear. The minimization algorithm used is M1QN3, see Section 1.5. The minimization software keeps calling the simulator (SIM4D) repeatedly until conver-

gence has been reached, or until the maximum number of iterations or simulations has been reached. The convergence criterion is given as a reduction in the norm of the gradient by a factor $10^{-\mathrm{NCVGE}}$, in namvar. The output mode of M1QN3 is printed in the log-file. The interpretation is:

1) Convergence reached, according to the above criterion.
2) M1QN3 called incorrectly.
3) Line search failed—step too big, > 1×10^{20}.
4) Maximum number of iterations (NITER) reached
5) Maximum number of simulations (NSIMU) reached
6) Line search failed—step too small, < RDX, (in namvar).
7) Impossible gradient value, 'descent' direction points uphill.

2.2.3 Last simulation

After M1QN3 has returned control to CVA1, one final simulation is performed. This simulation is diagnostic, and characterized by the simulation counter being set to 999, NSIM4D=NSIM4DL, yomvar. The observation departure from the low-resolution analysis, $\mathbf{y}^{\mathrm{o}}-H\mathbf{x}^{\mathrm{a}}_{\mathrm{LR}}$, is computed and stored in the NCMIOMN-word of the ODB. Finally at the end of CVA1, the updated ODB is written to disk, using the routine WRITEOBA.

2.3 A SIMULATION

A simulation consists of the computation of J and ∇J. This is the task of the routine SIM4D, see Fig. 2.2 for the flow diagram. The input is the latest value of the control variable χ in the array VAZX, computed by M1QN3, or CONGRAD. First J_{b} and its gradient are computed (see Sections 1.4 and 4.2):

$$\begin{aligned} J_{\mathrm{b}} &= \chi^{\mathrm{T}}\chi \\ \nabla_{\chi} J_{\mathrm{b}} &= 2\chi \end{aligned} \tag{2.3}$$

The gradient of J_{b} with respect to the control variable is stored in the array VAZG (YOMCVA).

- Copy χ from VAZX to SP3-arrays (YOMSP) using the routine YOMCAIN
- Compute $\mathbf{x}$, the physical model variables, using CHAVARIN:

$$\mathbf{x} = \delta\mathbf{x} + \mathbf{x}^{\mathrm{b}} = L\chi + \mathbf{x}^{\mathrm{b}}. \tag{2.4}$$

- Perform the direct integration of the model (if 4D-Var), using the routine CNT3, and compare with observations. See Section 2.5.
  Calculate J_{o} for which OBSV is the master routine.
- Perform the adjoint model integration (if 4D-Var) using CNT3AD, and observation operators' adjoint.
  Calculate $\nabla_{\mathbf{x}} J_{\mathrm{o}}$, and store it in SP3.
- J_{c} and its gradient are calculated in COSJC called from CNT3AD, if LJC is switched on (default) in namvar.
- Transform $\nabla_{\mathbf{x}} J_{\mathrm{o}} + \nabla_{\mathbf{x}} J_{\mathrm{c}}$ to control variable space by applying CHAVARINAD.
- Copy $\nabla_{\chi} J_{\mathrm{o}} + \nabla_{\chi} J_{\mathrm{c}}$ from SP3 and add to $\nabla_{\chi} J_{\mathrm{b}}$, already in the array VAZG, using YOMCAIN
- Add the various contributions to the cost function together, in EVCOST, and print to log file using prtjo.
- Increase the simulation counter NSIM4D by one.

The new J and $\nabla_\chi J$ are passed to the minimization algorithm to calculate the χ of the next iteration, and so on until convergence (or the maximum number of iterations) has been reached.

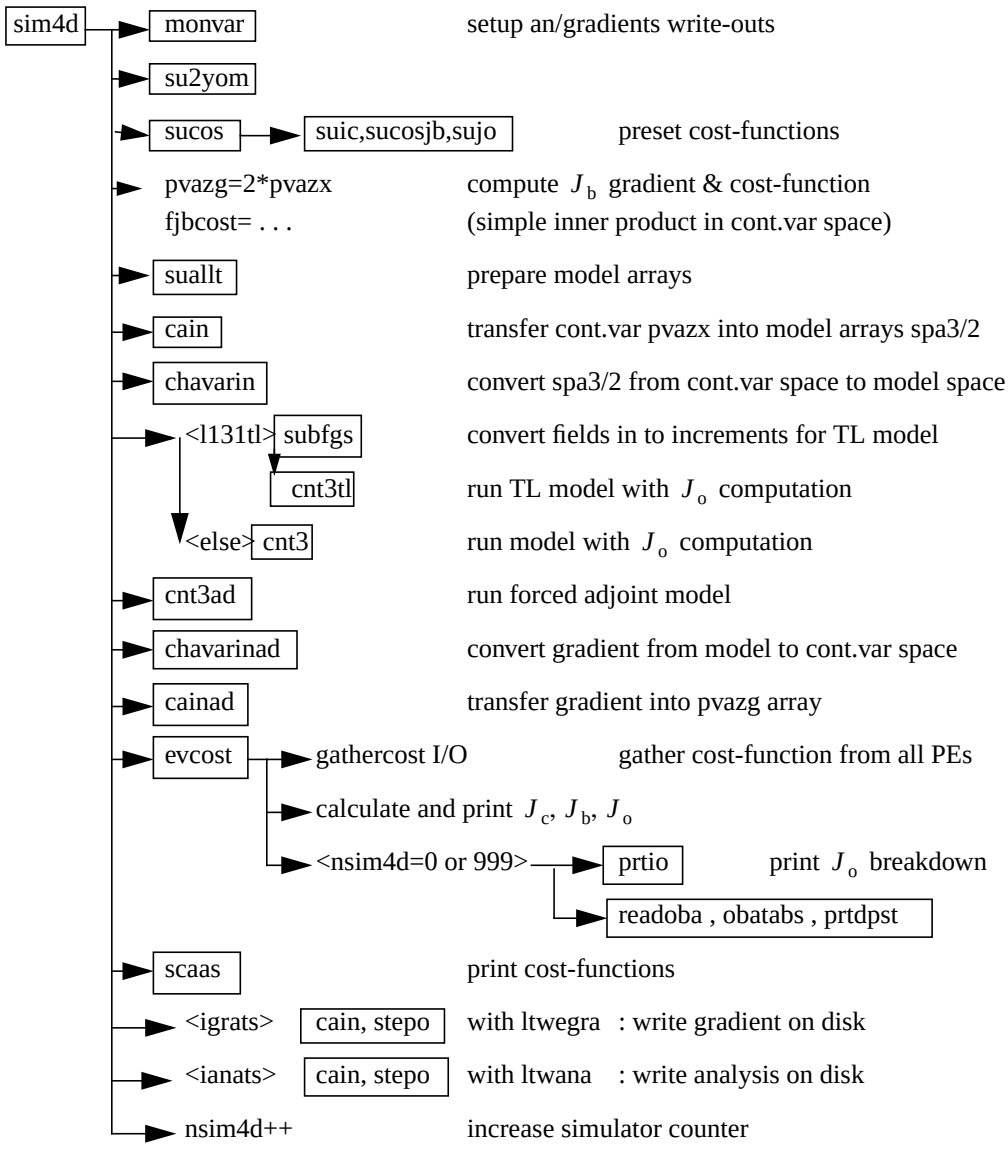

Figure 2.2 Flow diagram for the subroutine sim4d.

2.3.1 Interface between control variable and model arrays

The purpose of the routine CAIN (the canonical injection) is to identify those parts of the model state that should be included in the variational control variable. This is controlled by on/off switches such as NVA2D and NVA3D (yomcva) initialized in SUALCTV. The scalar product used is the one defined by the array SCALP (in yomcva, set up in the routine SCALJGS called from SUSCAL), which is 1 if $m = 0$, and 2 otherwise. This allows the compression of arrays of the type VAZX while using the L^2 norm on the sphere with real fields in spectral space.

CAIN is also the interface between the memory distributed spectral arrays and the non-distributed control variable.

The distributed spectral arrays SP3/2 are gathered with the routine GATHERSPA to form the control vector on each processor.

2.4 INTERPOLATION TO OBSERVATION POINTS

2.4.1 Method

COBSLAG is the master routine for the horizontal interpolation of model data to observation points. It is called after the inverse spectral transform in SCAN2MDM, and after the so-called *semi-Lagrangian buffers* have been prepared by COBS and SLCOMM, see the flow diagram in Fig. 2.3 . The interpolation code is shared with the semi-Lagrangian advection scheme of the dynamics. The buffers contain a *halo* of gridpoints big enough to enable interpolation to all observations within the grid-point domain belonging to the processor. COBSLAG calls OBSHOR which:

- Performs the interpolation, using SLINT
- Message-passes the result to the processors where the corresponding observations belong, using the routine MPOBSEQ.
- Copies the model data at observation points to the so-called GOM-arrays (yommvo, described below), in the routine INSOBSEQ.

There are three methods of horizontal interpolation:

1) LAIDDI: 12-point bi-cubic interpolation, used for all upper-air fields (if NOBSHOR=203) except clouds,

2) LAIDLI: Bi-linear interpolation, used for surface fields, and

3) LAIDLIC: Nearest gridpoint, used for cloud parameters.

The interpolation method for the upper-air fields can be switched to bi-linear by specifying NOBSHOR=201 in namobs. The default is NOBSHOR=203 (bi-cubic). Lists of interpolation points and weights are prepared by the routine LASCAW. In 4D-Var bi-cubic interpolation is used at high resolution (i.e. in the trajectory runs), and bi-linear is used at low resolution (i.e. in the main minimization). The interpolation is invoked once per 4D-Var time slot.

The adjoint (OBSHORAD) follows the same general pattern but gets further complicated by the fact that the gradient from several observations may contribute to the gradient at a given gridpoint. The summation of gradients is to done in the same order, irrespective of the number of processors, as reproducibility is desired.

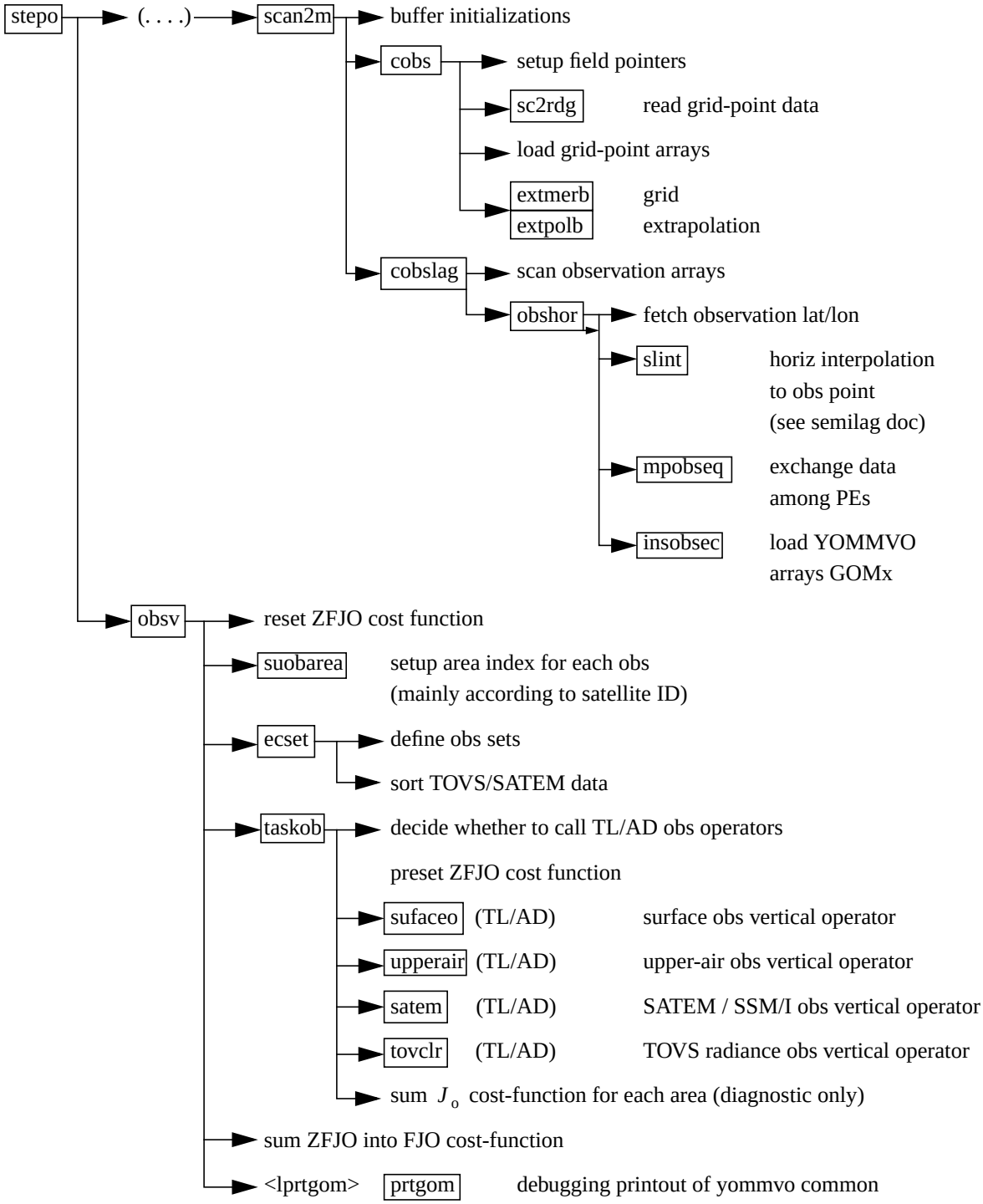

Figure 2.3 Flow diagram for subroutines scan2mdm and obsv.

2.4.2 Storage in GOM-arrays

The GOM arrays (YOMMVO) contain the model values at observation points. The list of upper-air model variables to appear in the GOM-arrays is under user control. There are five categories of GOM-arrays:

- GOMx for conventional data, containing full model profiles of optionally u, v, T, q, $O3$ (ozon), CLW (cloud liquid water), CLI (clod ice) and CC (cloud cover)
- GOSx for conventional data, containing surface data of p_{surf} (surface pressure), T_{skin} (skin temperature). w_{s} (soil water content), s_{n} (snow cover), z_0 (roughness length) and w_{l} (skin reservoir water content).
- GSMx for TOVS data, containing full model profiles similar to GOMx
- GSSx for TOVS data, containing surface data of p_s, T_s, w_s, s_n, z_0, w_l, u_l and v_l, where u_l and v_l are lowest model level wind components.
- GSCx for SCAT data, containing lowest model level data of u_l, v_l, T_l, and q_l, and surface data of p_s and T_s. z_0 (roughness length) is to be added shortly.

The reason for this split is purely to save space in memory. Model profiles of wind for example are not needed as inputs to the TOVS and SATEM operators, so those fields are not interpolated to the TOVS locations, and are not stored, unless requested. Upper-air profiles of model data at SCAT locations are also not computed. The selection of model variables to interpolate to GOMx and GSMX arrays, respectively, is flexible and is controlled through namelist switches LGOMx and LGSMx (in namdim). The default is that only LGOM-U/V/T/Q and LGSM-T/Q/O3 are ON, with the addition of LGSMCLW/CLI/CC in screening run to enable computation of cloudy radiances. The adressing of the GOM-arrays is done by referring to the MAPOMM (YOMOBA) and MABNOB (YOMOBSET) tables, e.g. ZPS(jobs) = GOSP(MAPOMM(iabnob)), where iabnob = MABNOB(jobs,kset) is an observation counter local to each processor.

The trajectory GOM5 arrays (identical to GOM) are allocated in the case that tangent linear observation operators are used. They are to hold the trajectory interpolated to the observation locations, and the GOM-arrays, in that case, hold the perturbations.

At the end of the adjoint observation operators the GOM-arrays are zeroed and overwritten by the gradient (in PREINTAD).

The r.m.s. of the GOM arrays is printed (by PRTGOM) if the switch LPRTGOM=.true., (in YOMOBS). The default is that the print is switched on. It can be located in the log file by searching for 'RMS OF GOM'. The printing is done from OBSV, 1) when the GOM arrays contain the background interpolated to the observation points, 2) when it contains ∇J_{o} of the first simulation, 3) when it contains first TL perturpations after the initial call to the minimizer and 4) when it contains ∇J_{o} at the final simulation.

2.5 COMPUTATION OF THE OBSERVATION COST FUNCTION

The cost function computation follows the same pattern for all observational data. This common structure is described in the following section. It is assumed that all observations are independent of each other, which means that the cost function contribution from each observation station can be computed independently of others. The specific observation operators for all data types and variables is detailed in Chapter 5 'Conventional observational constraints' and Chapter 6 'Satellite observational constraints' .

2.5.1 Organization in observation sets

The vertical observation operators are vectorized over NMXLEN (yomdimo) data. To achieve this the data first have to be sorted by type and subdivided into sets of lengths not exceeding that number. NMXLEN is currently set

to 511, in SUDIMO. The observation sets may span several 4D-Var time slots, as the input to the observation operators is the GOM-arrays which have been pre-prepared for all time slots during the tangent linear model integration. The organization of the sets is done in ECSET and SMTOV and the information about the sets is kept in yomobset. The only reason to have a separate routine for TOVS data (SMTOV) is that the TOVS sets must not contain data from more than one satellite. This is controlled by sorting according to the area-parameter, which for TOVS data is an indicator of satellite ID, prior to forming the sets. The area-parameter is determined in SUOBAREA, and is irrelevant for the observation processing for all data other than TOVS.

2.5.2 Cost function

The master routine controlling the calls to the individual observation operators is called HOP. This routine deals with all different types of observations.

The HOP/HOPTL/HOPAD routines are called from TASKOB/TASKOBTL/TASKOBAD (called from OBSV/OBSVTL/OBSVAD) in a loop over observation sets. The data type of each set is know from the information in tables such as MTYPOB(KSET) stored in yomobset.

The following describes HOP/HOPTL. The adjoint HOPAD follows the reverse order.

- First prepare for vertical interpolation using the routine PREINT. Data on model levels are extracted from the GOM-arrays (YOMMVO). Pressures of model levels are computed using GPPRE. Help arrays for the vertical interpolation are obtained (PPINIT) and $T*$ and T_0 are computed (CTSTAR). $T*$ and T_0 are later used for extrapolation of temperature below the model's orography, Subsection 5.3.2. The routine PREINTS deals with model surface fields needed for the near-surface observation operators and PREINTR deals with those fields that are specific to the radiance observation operators.
- The observation array is then searched to see what data is there. The 'body' of each observation report is scanned for data, and the vertical coordinate and the variable-number for each datum is retained in tables (ZVERTP and IVNMRQ). These tables will later constitute the 'request' for model equivalents to be computed by the various observation operators. Tables of pointers to data ('body' start addresses) and counters are stored (arrays IPOS and ICMBDY).
- Then the forward calculations are performed. There is an outer loop over all known 'variable numbers'. If there are any matching ocurrences of the loop-variable number with the content of IVNMRQ, then the relevant observation operator will be called. A variable-number and an observation operator are linked by a table set up in the routine HVNMTLT. The interface routines PPOBSA (upperair) and PPOBSAS (surface) are used, which in turn call PPFLEV and the individual operator routines. For radiance data the interface is RADTR which calls the radiative transfer code (Subsection 6.4.1).
- In HDEPART, calculate the departure $\mathbf{z}$ as

$$\mathbf{z} = \mathbf{y}^o - H\mathbf{x} + (\mathbf{y}^o - H\mathbf{x}^b_{\mathrm{HR}}) - (\mathbf{y}^o - H\mathbf{x}^b_{\mathrm{LR}}), \tag{2.5}$$

  where the two terms in brackets have been computed previously: the first one in the high resolution trajectory run (Section 1.3) and the second one in the LOBSREF call, described in Section 2.2.
  If LOBSTL then $\mathbf{z}$ is

$$\mathbf{z} = \mathbf{y}^o - \mathbf{H}\delta\mathbf{x} + (\mathbf{y}^o - H\mathbf{x}^b_{\mathrm{HR}}) - \mathbf{y}^o, \tag{2.6}$$

  which simplifies to what has been presented in Section 1.2.

The TOVS radiance bias correction is also carried out at this point by subtracting the bias estimate (kept in the NCMTORB-word of ODB) from the calculated departure.
Finally the departure is divided by the observation error σ_o (NCMFOE in ODB) to form the *normalized departure*.

- Departures of correlated data are multiplied by $\mathbf{R}^{-1}$, see 2.5.4. The division by σ_o has already taken place in HDEPART, so $\mathbf{R}$ at this point is in fact a correlation (not a covariance) matrix.
- The cost function is computed in HJO, as

$$J_o = \mathbf{z}^T\mathbf{z} \tag{2.7}$$

for all data except SCAT data. The SCAT cost function combines the two ambiguous winds (subscripts 1 and 2) in the following way (also in HJO)

$$J_{\mathrm{SCAT}} = \left[\frac{J_1^4 J_2^4}{J_1^4 + J_2^4}\right]^{1/4} \tag{2.8}$$

These expressions for the cost function are modified by variational quality control, see Section 2.6. The cost-function values are store in two tables, as detailed in 2.5.3.

- HJO also stores the resulting *effective departure* in the NCMIOM0-word of ODB, for reuse as the input to the adjoint. The effective departure is the normalized departure after the effects of observation error correlation and quality control have been taken into account, $\mathbf{z}_{\mathrm{eff}} = \mathbf{z}^T\mathbf{R}^{-1}[\mathbf{QC}_{\mathrm{weight}}]$, where the QC-weight will be defined below, Section 2.6.

2.5.2 (a) Adjoint. We have now reached the end of the forward operators. In the adjoint routine HOPAD some of the tasks listed above have to be repeated before the actual adjoint calculations can begin. The input to the adjoint (the effective departure) is read from the ODB. The expression for the gradient (with respect to the observed quantity) is then simply

$$\nabla_{\mathrm{obs}} J_o = -2\mathbf{z}_{\mathrm{eff}}/\sigma_o \tag{2.9}$$

which is calculated in HOPAD for all data. The gradient of J_{SCAT} is much more complicated and is calculated in a separate section of HOPAD. The adjoint code closely follows the structure of the direct code, with the adjoint operators applied in the reverse order.

2.5.3 J_o tables

There are two different tables for storing the J_o values. One is purely diagnostic (FJO, yomcosjo1), and is used for producing the printed J_o tables in the log-file (PRTJO called rom EVCOST). The other (FJOS) is the actual J_o-table. FJO is indexed by observation type, sub-obstype, variable and area. FJOS is indexed by the absolute observation number, iabnob=MABNOB(jobs,kset), so that the contributions from each individual observation can be summed up in a predetermined order (in EVCOST), to ensure reproducibility.

2.5.4 Correlation of observation error

The observation error is assumed uncorrelated (i.e. the matrix $\mathbf{R}$ is diagonal) for all data except time-sequences of SYNOP/DRIBU surface pressure and height data (used by default in 4D-Var, *Järvinen et al.* 1999). There is also code for vertical correlation of observation error for radiosonde geopotential data (not used by default) and SATEM thicknesses (not used by default). IN FACT, all vertcal correlations of observation error have been removed in 21r2,

but will be reintroduced again in a later cycle!

The serial correlation for SYNOP and DRIBU data is modelled by a continuous correlation function $a\,\mathrm{e}^{-b(t_1-t_2)^2}$ where a =RTCPART=0.3 and b =RTCEFT=6.0 hours, under the switch LTC (namjo). The remaining fraction $1-a$ of the error variance is assumed uncorrelated (see COMTC).

The radiosonde geopotential data are vertically correlated (under the switch LRSVCZ) using a continuous correlation function $a\,\mathrm{e}^{-b(x_1-x_2)^2}$ where a =RRSZPART=0.8, b a tuning constant close to 1 and x_1 and x_2 are transformation values, based on a sixth degree polynomial in $\ln p$ of the two pressures involved. The remaining fraction $1-a$ of the variance is assumed uncorrelated (see COMATP).

The vertical correlation of SATEM thickness data is as described in *Kelly and Pailleux* (1988) and is assigned in SURAD (and kept in YOMTVRAD). There is no horizontal correlation of SATEM and TOVS observation errors. The inter-channel correlation of radiance observation error is also assumed to be zero.

When $\mathbf{R}$ is non-diagonal, the 'effective departure' $\mathbf{z}_{\mathrm{eff}}$ is calculated by solving the linear system of equations $\mathbf{z}_{\mathrm{eff}}\mathbf{R} = \mathbf{z}$ for $\mathbf{z}_{\mathrm{eff}}$, using NAG routines F07FDF (Choleski decomposition) and F07FEF (backwards substitution), as is done in UPPERAIR, SATEM, COMTC and JOPDF. The NAG routines will shortly be replaced by the corresponding LAPACK routines SPOTRF and SPOTRS.

2.6 VARIATIONAL QUALITY CONTROL

The variational quality control, VarQC, has been described by *Andersson and Järvinen* (1999). It is a quality control mechanism which is incorporated within the variational analysis itself. A modification of the observation cost function to take into account the non-Gaussian nature of gross errors, has the effect of reducing the analysis weight given to data with large departures from the current iterand (or preliminary analysis). Data are not irrevocably rejected, but can regain influence on the analysis during later iterations if supported by surrounding data. VarQC is a type of buddy check, in that it rejects those data that have not been fitted by the preliminary analysis, often because it conflicts with surrounding data.

2.6.1 Description of the method

The method is based on Bayesian formalism. First, an *a priori* estimate of the probability of gross error $P(G)_{\mathrm{i}}$ is assigned to each datum, based on study of historical data. Then, at each iteration of the variational scheme, an *a posteriori* estimate of the probability of gross error $P(G)_{\mathrm{f}}$ is calculated (*Ingleby and Lorenc*, 1993), given the current value of the iterand (the preliminary analysis). VarQC modifies the gradient (of the observation cost function with respect to the observed quantity) by the factor $1-P(G)_{\mathrm{f}}$ (the QC-weight),which means that data which are almost certainly wrong ($P(G)_{\mathrm{f}} \approx 1$) are given near-zero weight in the analysis. Data with a $P(G)_{\mathrm{f}} > 0.75$ are considered 'rejected' and are flagged accordingly, for the purpose of diagnostics and feedback statistics, etc.

The normal definition of a cost function is

$$J_{\mathrm{o}} = -\ln p \tag{2.10}$$

where p is the probability density function. Instead of the normal assumption of Gaussian statistics, we assume that the error distribution can be modelled as a sum of two parts: one Gaussian, representing correct data and one flat distribution, representing data with gross errors. We write:

$$p_i = N_i[1-P(G_i)] + F_i P(G_i) \tag{2.11}$$

where subscript i refers to observation numer i. N and F are the Gaussian and the flat distributions, respectively:

$$N_i = \frac{1}{\sqrt{2\pi}\sigma_o} \exp\left[-\frac{1}{2}\left(\frac{y_i - Hx}{\sigma_o}\right)^2\right] \tag{2.12}$$

$$F_i = \frac{1}{L_i} = \frac{1}{2l_i\sigma_o} \tag{2.13}$$

The flat distribution is defined over an interval L_i which in Eq. (2.13) has been written as a multiple of the observation error standard deviation σ_o. Substituting Eqs. (2.11) to (2.13) into Eq. (2.10), we obtain after rearranging the terms, an expression for the QC-modified cost function J_o^{QC} and its gradient ∇J_o^{QC}, in terms of the normal cost function J_o^N

$$J_o^N = \frac{1}{2}\left(\frac{y_i - Hx}{\sigma_o}\right)^2 : \tag{2.14}$$

$$J_o^{QC} = -\ln\left(\frac{\gamma_i + \exp[-J_o^N]}{\gamma_i + 1}\right) \tag{2.15}$$

$$\nabla J_o^{QC} = \nabla J_o^N\left(1 - \frac{\gamma_i}{\gamma_i + \exp[-J_o^N]}\right) \tag{2.16}$$

where

$$\gamma_i = \frac{P(G_i)/(2l_i)}{[1 - P(G_i)]/\sqrt{2\pi}} \tag{2.17}$$

2.6.2 Implementation

The *a priori* information i.e. $P(G)_i$ and l_i are set during the screening, in the routine DEPART, and stored in the NCMFGC1 and NCMFGC2-words of the ODB. Default values are set in DEFRUN, and can be modified by the namelist namjo. VarQC can be switched on/off for each observation type and variable individually using LVARQC, or it can be switched off all together by setting the global switch LVARQCG=.false. Since an as good as possible 'preliminary analysis' is needed before VarQC starts, it is necessary to perform part of the minimization without VarQC, and then switch it on. This is controlled by NITERQC in yomcosjo, and is set to 40 by default. Printing of VarQC results is done by the routine PRTQC.

JOCOST computes J_o^{QC} according to Eq. (2.15) and the QC-weight—the factor within brackets in Eq. (2.16).

2.6.3 Correlated data

The quality control of radiosonde height data (if used) is more complex because of the correlation of observation error (see JOPDF). This is one of the reason why we changed to using temperature data instead, from cy18r6. VarQC for correlated data is no longer supported.

Part II: DATA ASSIMILATION

CHAPTER 3 4D variational assimilation

Table of contents

3.1. Introduction

3.2. Organization of data in time slots

3.2.1 Observation preprocessing.

3.2.2 Inside IFS.

3.2.3 Observation screening in 4D-Var

3.3. Inner and outer loops: practical implementation

3.4. Tangent linear physics

3.4.1 Set-up

3.4.2 Mixed-phase thermodynamics

3.4.3 Vertical diffusion

3.4.4 Sub-grid scale orographic effects

3.4.5 Large-scale precipitation

3.4.6 Long-wave radiation

3.4.7 Deep moist convection

3.4.8 Trajectory management

3.1 INTRODUCTION

4D-Var is a temporal extension of 3D-Var. Observations are organized in one-hour time-slots as described in Section 3.2. The cost-function now measures the distance between a model trajectory and the available information (background, observations) over an assimilation interval or window. For a 12-hour window (as currently used), it is either (03UTC–15UTC) or (15UTC–03UTC). Eq. (1.2) (see Chapter 1 'Incremental formulation of 3D/4D variational assimilation—an overview' is replaced by

$$J(\delta\mathbf{x}) = \frac{1}{2}\delta\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\delta\mathbf{x} + \frac{1}{2}\sum_{i=0}^{n}(\mathbf{H}_i\delta\mathbf{x}(t_i) - \mathbf{d}_i)^{\mathrm{T}}\mathbf{R}_i^{-1}(\mathbf{H}_i\delta\mathbf{x}(t_i) - \mathbf{d}_i) \tag{3.1}$$

with subscript i the time index. Each i corresponds to one-hour time slot. $\delta\mathbf{x}$ is as before the increment at low resolution at initial time, and $\delta\mathbf{x}(t_i)$ the increment evolved according to the tangent linear model from the initial time to time index i. $\mathbf{R}_i$ and $\mathbf{B}$ are the covariance matrices of observation errors at time index i and of background errors respectively. $\mathbf{H}_i$ is a suitable linear approximation at time index i of the observation operator H_i. The innovation vector is given at each time step by $\mathbf{d}_i = \mathbf{y}_i^{\mathrm{o}} - H_i\mathbf{x}^{\mathrm{b}}(t_i)$, where $\mathbf{x}^{\mathrm{b}}(t_i)$ is the background propagated in

time using the full nonlinear model and $\mathbf{y}_i^{\mathrm{o}}$ is the observation vector at time index *i*. As SYNOP and DRIBU time sequences of surface pressure and height data are now used **with** serial correlation of observation error, the observation costfunction computation for those data spans all time slots. Eq. (3.1) therefore needs generalising, as has been done in the paper by *Järvinen et al.* (1999).

The minimization is performed in the same way as in 3D-Var. However, it works fully in terms of increments, a configuration which is activated by the switches L131Tl and LOBSTL, and involves running the tangent-linear and adjoint models iteratively as explained in Section 2.3 of Chapter 2 '3D variational assimilation' , and using the tangent-linear observation operators.

A way to account in the final 4D-Var analysis for some non-linearities is to define a series of minimization problems

$$\begin{aligned} J(\delta\mathbf{x}^n) \;=\; & \frac{1}{2}(\delta\mathbf{x}^n + \mathbf{x}^{n-1} - \mathbf{x}^{\mathrm{b}})^{\mathrm{T}}\mathbf{B}^{-1}(\delta\mathbf{x}^n + \mathbf{x}^{n-1} - \mathbf{x}^{\mathrm{b}}) \\ & + \frac{1}{2}\sum_{i=0}^{n}(\mathbf{H}_i\delta\mathbf{x}^n(t_i) - \mathbf{d}_i^{n-1})^{\mathrm{T}}\mathbf{R}_i^{-1}(\mathbf{H}_i\delta\mathbf{x}^n(t_i) - \mathbf{d}_i^{n-1}) \end{aligned} \tag{3.2}$$

with superscript *n* the minimization index.

$\mathbf{x}^{n-1}$ is the current estimate of the atmospheric flow. It is equal to the background for the first minimization. $\mathbf{d}_i^{n-1} = \mathbf{y}_i^o - H_i\mathbf{x}^{n-1}(t_i)$ is the innovation vector, computed by integrating the model at high resolution from the current estimate. The way the increment is added to the current estimate is similar to that used in 3D-Var (see Chapter 1 'Incremental formulation of 3D/4D variational assimilation—an overview' .

$$\mathbf{x}_{\mathrm{HR}}^n \;=\; \mathbf{x}_{\mathrm{HR}}^{n-1} + \mathrm{NNMI}(\mathbf{x}_{\mathrm{HR}}^{n-1} + \delta\mathbf{x}_{\mathrm{HR}}^n) - \mathrm{NNMI}(\mathbf{x}_{\mathrm{HR}}^{n-1}) \tag{3.3}$$

The number of times the trajectory is updated, i.e. the number of outer-loops (which corresponds to the number of minimizations performed), is typically a number between one and four. In operational 4D-Var the number of outer loops is two.

This can be controlled in the prepIFS set-up, together with the number of inner-loops (iterations of m1qn3) within each minimization. One outer-loop corresponds to what is normally done in 3D-Var. The number of inner-loops should then be 70 as in 3D-Var. The most standard 4D-Var uses two outer-loops. The first minimization runs with the simplified physics on 50 inner-loops. The second minimization runs with the more complete linear physics on 25 inner-loops. Switches for the two sets of physics will be given in Section 3.4.

The variational quality-control (Chapter 2 '3D variational assimilation' Section 2.6) is switched on at the default iteration number (40) in the first minimization. It is activated from the first iteration in the subsequent minimizations.

The final 4D-Var trajectory is post-processed every 3 hours. Fields called 4v are created with initial date and time the start of the window (03UTC or 15UTC) and steps every 3 hours. The 4v field valid at 12UTC or 00UTC, is then renamed as the final analysis (type=an) for the atmospheric fields and the waves. The cycling from one cycle to the next is performed by taking these analysis fields, together with the surface fields updated by the SST, snow and soil moisture analyses as input to a 12-hour forecast which produces the background for the next cycle.

The analysis and forecast error calculations are performed as explained in Chapter 7 'Background, analysis and forecast errors' , with the inclusion of the time dimension in the minimization. The analysis error variances are available at the beginning of each window, and the forecast error variances at the end.

3.2 ORGANIZATION OF DATA IN TIME SLOTS

3.2.1 Observation preprocessing.

Observational input data (BUFR-format) is read in by means of 6-hour time-windows in OBSPROC preproc_mpp_makecma. Before input, each time-window has been organised into several BUFR-files based on major observation types. Input BUFR-files are labelled and split so that every processor can read one or more BUFR-files. The prefix of each file indicates the observation type. The suffix "<tw>.<proc>" defines which processor <proc> (here within a range [1..16]) will be responsible for inputting data for time-window <tw> (here within a range [01..02]). The number of files is not necessarily equal to the number of processors, it is really a matter of I/O-load balancing, and the end result is independent of the reading order.

In the case of 4D-Var there are NO6HTSL input time-windows. For 6 h, 12 h and 24 h 4D-Var analysis periods, the NO6HTSL will have values 1, 2 and 4, respectively. This can be set via the namelist namelist/namglp.h (see also yomglp).

Another affecting parameter (see discussion about reshuffle below) is NOSORTSL (set via OBSPROC namelist/nammkcma.h and declared in yommkcma). It defines how many time-slots will be used. The rule is to divide each input time-window into 1h out time-slots, but with half an hour lengths for start and end time-slots. Therefore, the value of NOSORTSL should be set to 7, 13 or 25 for 6 h, 12 h and 24 h 4D-Var analysis periods (i.e. one plus length of a 4D-Var period in hours), respectively.

Once all BUFR-data has been successfully read in, the unique sequence numbers for reports (before even having them around!) are generated in OBSORT-routine makeseqno_obsort called by preproc_mpp_makecma. These numbers are always independent of the number of processors in use. They form a basis for reproducibility of analysis results regardless of how many processors were used.

The sequence numbers are generated without honouring the input time-windows. Currently for CONV-data, sequence numbers start at offset 0, TOVS at offset 1,000,000, SCAT at 2,000,000 and SSMI 3,000,000. Thus, the increment is set to 1,000,000 meaning that we may not exceed more than one million reports per major observation type (CONV, TOVS, etc.) without making a small change into the local variable increment in routine preproc_mpp_makecma.

After the sequence-number generation, all BUFR data is read in and re-shuffled for better load balancing in report creation under the OBSPROC-routine MAKECMA makecma. Before that, the number of input time-windows NO6HTSL has already been reset to one in preproc_mpp_makecma, and all 4D-Var BUFR input data is regarded as a one 'supertime-window' for initial report creation. However, via NAMELIST parameters NANTIM, NANDAT, NTBMAR and NTFMAR defined in NAMGLP (namelist/namglp.h and yomglp), full control over valid 4D-Var analysis timerange is maintained. Therefore, observations not in this range will be discarded by the MAKECMA.

An essential step to organise observational data for 4D-Var purposes occurs in the OBSPROC routine postproc_mpp_makecma. The aim is to reshuffle and time-slot the initially created CMA files of which there are currently one CMA file per processor. The CMA data needs not only to be organised in time-slots, but they also need to obtain a better geographical distribution within a given time-slot to have a better load balancing in the subsequent IFS/Screening job.

Before the reshuffle of observations can take place, some crucial information about 4D-Var run characteristics needs to be passed on. Parameters NANTIM, NANDAT, NOSORTSL, NTBMAR and NTFMAR are transformed into the suitable constants for use by the OBSORT by use of SETPARAM_OBSORT in routine postproc_mpp_makecma. The following conversion takes place (OBSORT parameters in concern are declared in

yomstdin):

- NANTIM and NANDAT are used to calculate an absolute start time and date of an analysis period. The resulting OBSORT parameters are called TIME_INIT_YYYYMMDD and TIME_INIT_HHMMSS.
- NOSORTSL, NTBMAR and NTFMAR are used to get parameters NUM_TIME_SLOTS and TIME_DELTA_4DVAR in-line.
- OBSORT parameter vectors TIME_SLOT_YYYYMMDD and TIME_SLOT_HHMMSS to indicate start date and time of a particular time-slot will be implicitly generated upon start-up of the reshuffle in OBSORT gen_timeslot_data called by lib_obsort. This routine makes sure that the first and the last time-slot periods will have duration of half an hour (as discussed earlier).

The actual reshuffle is handled via OBSORT routine lib_obsort (in particular mapsort). The initial CMA-data is read back in and an internal global table (seen by every processor) is established. This table contains snapshot information about each CMA-report. There can be found things like which processor owns/should own the report before/after the re-shuffle, which 4D-Var time-slot observation belongs to, plus information to perform robustly the reshuffle itself.

The reshuffle of the CMA data is done per each time-slot. Currently all data is written into one CMA file per processor. Each time-slot is stacked after each other so that a particular time-slot could in principle be accessed by knowing its start address and data length. This offset information is available both in file obs_boxes (generated by OBSORT ifs_write) and CMA file's DDR (Data Description Records). The former one may become obsolete, so users should rely only to the information found in DDR number one (see also IFS yomcmddr, DDR#1 words 101–607).

When time-slot information has been once placed into the DDR#1, it will be propagated automatically into the subsequent CMA files (ECMA and CCMA) in a run cycle, and no regeneration is needed.

Finally, upon the CMA-data reshuffle also the BUFR data is re-shuffled to retain a one-to-one relationship with its sibling CMA reports. This is important, since the OBSPROC FEEDBACK (bufdback) relies on the order of these 'pseudo-original' BUFR-files to update observational data for archieving purposes. Actually, by aid of the OBSORT, we even manage to get this updated 'pseudo-original' BUFR-data back to its original input time-window frames (split by the major type CONV, TOVS, etc.), albeit that the original observation order cannot (and need not to) be preserved.

The CMA format is converted to an ODB database suitable for input to the IFS. This conversion is performed by utility ecma2odb. It will be converted back to CMA for BUFR feedback generation, but the ODB with feedback information is archived as such.

3.2.2 Inside IFS.

The timeslot information is read into IFS in RD_OBS_BOXES called from OBADAT. It is possible to run 3D-Var with an ODB prepared with timeslots, the timeslotting information is taken into account only if NSTOP > 1.The information that is extracted for each timeslot (only for your own processor) is,

- number of observations (NTSLTOB)
- length of observations (NTSLLEN)
- number of SCAT observations(NTSLSCA)
- number of TOVS observations(NTSLTOV)
- number of non-SCAT and non-TOVS observations(NTSLNTV)

The following global information regarding timeslots is extracted

- number of observations for each processor and time-slot (NTSLTOBP)

- global number of observations for each time-slot (NTSLTOBG)
- max (over processors) number of observations for each time-slot (NTSLTOBM)

The arrays to contain observation equivalents (the GOM-arrays) are allocated to be able to contain all time-slots. These arrays are then gradually filled during the forward integration. The reasons for allocating these arrays to contain all time-slots are:

1) that the trajectory is only run once

2) that they are used in screening. The tables needed to message pass the observation equivalents from the processor that 'owns' the part of the globe in grid-point space corresponding to the observation and the processor that 'owns' the observation is done in MKGLOBSTAB

3.2.3 Observation screening in 4D-Var

The trajectory integration can be performed in the observation screening mode. The part of the IFS code devoted to the observation screening is activated via namelist variable LSCREEN in NAMCT0 An array of good quality observations of desired variables is selected to be used in the minimization. Technically, the extended observation database (ECMA-ODB) becomes the compressed database (CCMA-ODB) which is the observational input for the minimization run. In 4D-Var, the observation screening can be applied either on an hourly or a 6-hourly basis. This selection is done via namelist variable LSCRE4D in NAMSCC. Hourly screening has been the default option since cy18r6.

At the end of the screening, the CCMA-ODBs are reshuffled for load-balancing in the subsequent minimizations, using MAPSORT.

Depending on whether the hourly or 6-hourly screening is applied, the division of observations into the sets and the appropriate pointers are updated accordingly (SCREEN, ECSET). The bulk of decisions (DECIS) is taken codewise just in the same way in both cases. In the hourly screening much more surface observations are retained for the assimilation. More details of the observation screening can be found in Chapter 10 'Observation screening' .

3.3 INNER AND OUTER LOOPS: PRACTICAL IMPLEMENTATION

Similarly to 3D-Var, job steps are carried out with different configurations of the IFS:

(i) **The first trajectory run** (which includes screening) – conf=2, LSCREEN=.T.

(ii) **The background error minimization**, conf=131, LAVCGL=.T.

(iii) **The main minimization,** conf=131

(iv) **The update of the trajectory** , conf=1, LOBS=.T.

Steps (*iii*) and (*iv*) are performed n times where n is the number of outer loops or, equivalently, of updates of the trajectory.

The first trajectory run (*i*), the background-error minimization (*ii*) and the first main minimization use the same input files as described for 3D-Var in Subsection 1.3.1 of Chapter 1 'Incremental formulation of 3D/4D variational assimilation—an overview' , the only difference being that the background field is a 3-hour forecast from the previous analysis at synoptic time, compared with a 6-hour forecast in 3D-Var.

The ouput of the minimization steps are the files **MXVAxx000+000000**, **MXVAxx999+000000** (as in 3D-Var), **trajxx+0000000** and **VATRH**. xx is an integer varying from 0 for the first minimization to (n-1) for the last minimization, where n is the number of updates of the trajectory. **VATRH** contains useful information for a warm restart of m1qn3 (including the diagonal of the Hessian). **trajxx+0000000** contains the control variable at the end of the minimization. The file **trajxx+0000000** is written out in SAVMINI called at the end of CVA1. This file will be

an input to the next minimization in addition to the background file used as in the first minimization. It is read in in GETMINI called from CVA1. The file **VATRH** is written out in SAVMIN, and read in SUHESS, both called by CVA1.

The input of the second trajectory is the same as in 3D-Var. The output is an analysis at the initial time of the trajectory (type = 4v, step = 0) written out on the FDB. It contains the current estimate of the flow at initial time. Another output are the updated observation files, as in 3D-Var. The 4v fields are used in the following trajectory, replacing the background in the input files **ICMSHxxxxINIT**, **ICMGGxxxxINIT** and **ICMGGxxxxINIUA** (where **xxxx** is the 'expver' identifier of MARS). Additional inputs are low resolution files (**MXVA**...) created during the previous minimization interpolated to high resolution as in 3D-Var. This data flow is represented in the diagram below.

In summary, the first two trajectories use the background as an input, and the following ones use the 4v fields created during the previous trajectory as reference files. All the trajectories except for the very first one add increments computed from the low-resolution files produced by the previous minimization, interpolated to high resolution. The first minimization uses only the background field, the following ones also use the control variable from the end of the previous minimization and some information for a warm restart of the minimization package.

The number of updates of the trajectory starting from 0 at the first minimization is carried inside the ODB files.

3.4 TANGENT LINEAR PHYSICS

The first minimization uses the simplified physics (vertical diffusion and surface drag) activated by the switches LSPHLC, LVDFDS, LSDRDS, LVDFLC, LSDRLC, LKEXP in namelist NAPHLC which is also activated for singular vector computations. A scientific description of the simplified physics is given in *Buizza* 1994) .

The following minimizations use a more complete linear physics activated by the switches LETRAJP, LEVDIF2, LEGWDG2, LECOND2, LERADI2, LERADS2, LECUMF2 in namelist NAMTRAJP, and described in this section. The description is focused on technical aspects, since scientific issues can be found elsewhere (*Mahfouf et al.*, 1997; *Rabier et al.* 1997; *Mahfouf* 1998).

3.4.1 Set-up

In order to activate the improved linear physics, the switch LSPHLC of the simplified linear physics inNAPHLC should be set to FALSE. In CVA1 when both logicals LSPHLC and LETRAJP are equal to TRUE, LSPHLC is reset to FALSE and a warning is written in the standard ouput (logical unit NULOUT).

The following switches must be set to TRUE : LEPHYS, LAGPHY (also necessary to activate the ECMWF nonlinear physics) and LETRAJP (to activate storage of the trajectory at $t - \Delta t$). The linear physics contains a set of five physical processes : vertical diffusion (LEVDIF2), sub-grid scale orographic effects (LEGWD2), large scale condensation (LECOND2), longwave radiation (LERADI2, LERADS2), and deep moist convection (LECUMF2).

Tunable parameters of the improved physics (which should not in principle be modified) are defined in SUPHLI. The logical LPHYLIN is used to activate the simplifications and/or modifications associated with the linear package in the non-linear physics. This variable is set to FALSE by default , but is forced to TRUE before calling the linear physics (CALLPARTL and CALLPARAD) in CPGLAGTL and CPGLAGAD whenever the logical LETRAJP is TRUE.

Diagram representing the input and output files during a standard 4D-Var analysis consisting of 3 trajectory steps and 2 minimisation steps.

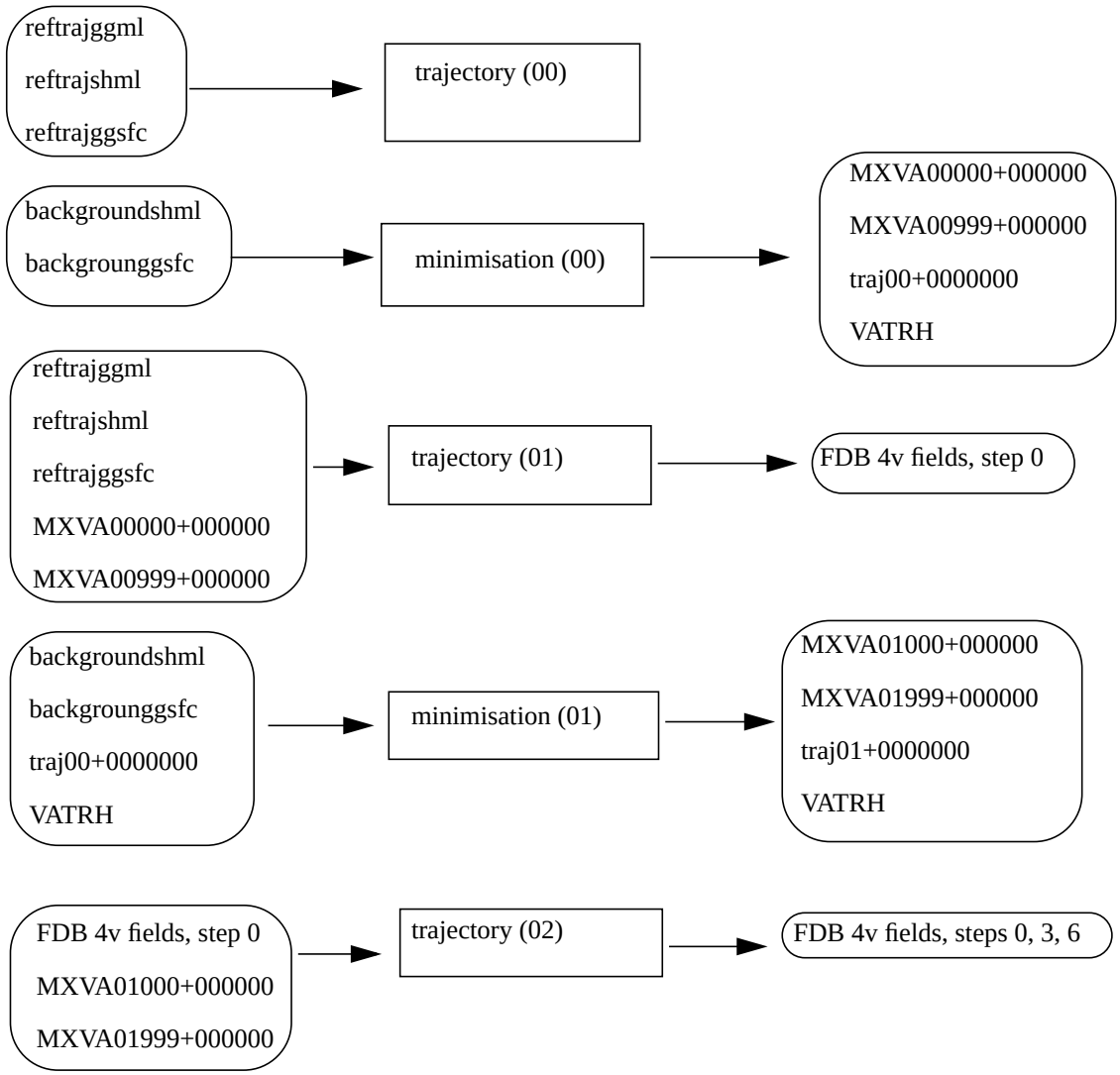

Figure 3.1 Diagram representing the input and output files during a standard 4D-Var analysis consisting of 3 trajectory steps and 2 minimisation steps.

3.4.2 Mixed-phase thermodynamics

The thermodynamical properties of the water mixed phase are represented by a differentiable weighting function between $T_0 = 0C$ and $T_{ice} = -23\mathrm{C}$:

$$\alpha(T) = \frac{1}{2}[1 + \tanh\{\mu(T - T_{crit})\}] \tag{3.4}$$

with $\mu = 0.15$ (RLPALP1) and $T_{crit} = T_{ice} + \frac{T_0 - T_{ice}}{\sqrt{2}}$ (RLPTRC).

The tuning parameter μ controls the intensity of the smoothing, and the temperature T_{crit} has been chosen to give $\alpha = 0.5$ for the same temperature as in the operational quadratic formulation (see function FCTTRE).

This weighting function is used by the large-scale condensation and moist-convection routines.

3.4.3 Vertical diffusion

The linear versions of the vertical diffusion scheme are called from the drivers VDFMAINTL and VDFMAINAD.

Vertical diffusion applies on wind components, dry static energy and specific humidity. The exchange coefficients in the planetary boundary layer and the drag coefficients in the surface layer are expressed as functions of the local Richardson number (*Louis et al.*, 1982). They differ from the operational formulation which uses the Monin–Obukhov length as a stability parameter in stable conditions and a K-profile approach for convective boundary layers (see the documentation of the ECMWF physics).

In stable conditions ($Ri > 0$), the drag coefficients are defined as :

$$C_{\mathrm{M}} = C_{\mathrm{MN}} \frac{1}{1 + \frac{2bRi}{\sqrt{1 + dRi}}} \tag{3.5}$$

and

$$C_{\mathrm{H}} = C_{\mathrm{HN}} \frac{1}{1 + 3bRi\sqrt{1 + dRi}} \tag{3.6}$$

with the following expressions for the neutral coefficients :

$$C_{\mathrm{MN}} = \frac{k^2}{\left[\log\left(\frac{z + z_{0\mathrm{M}}}{z_{0\mathrm{M}}}\right)\right]^2} \tag{3.7}$$

$$C_{\mathrm{HN}} = \frac{k^2}{\left[\log\left(\frac{z + z_{0\mathrm{M}}}{z_{0\mathrm{M}}}\right)\log\left(\frac{z + z_{0\mathrm{M}}}{z_{0\mathrm{H}}}\right)\right]} \tag{3.8}$$

In unstable conditions ($Ri < 0$), the drag coefficients are defined as:

$$C_{\mathrm{M}} = C_{\mathrm{MN}}\left(1 - \frac{2bRi}{1 + 3bcC_{\mathrm{MN}}\sqrt{\frac{z + z_{0\mathrm{M}}}{z_{0\mathrm{M}}}|Ri|}}\right) \tag{3.9}$$

$$C_{\mathrm{H}} = C_{\mathrm{HN}}\left(1 - \frac{3bRi}{1 + 3bcC_{\mathrm{HN}}\sqrt{\frac{z + z_{0\mathrm{M}}}{z_{0\mathrm{H}}}|Ri|}}\right) \tag{3.10}$$

The empirical coefficients b (RLPBB), c (RLPCC) and d (RLPDD) are set to 5 in SUPHLI.

In the planetary boundary layer, the exchange coefficients can formally be writen :

$$K = l^2 \left| \frac{\partial V}{\partial z} \right| f(Ri) \tag{3.11}$$

with the following mixing length vertical profile :

$$l = \frac{k(z+z_0)}{1+k\dfrac{(z+z_0)}{\lambda}} \left[\gamma + \frac{1-\gamma}{1+\dfrac{(z+z_0)^2}{L^2}} \right] \tag{3.12}$$

The asymptotic mixing lengh λ_M for momemtum is set to 150 m, whereas $\lambda_H = \lambda_M \sqrt{1.5d}$. The pseudo-depth of the boundary layer is defined by $L = 4$ km(RLPMIXL), and the reduction factor applied to the mixing length in the free atmosphere is $\gamma = 0.2$ (RLPBETA) [$l \to \gamma\lambda$ when $z \gg L$].

If this vertical-diffusion scheme is activated in the nonlinear model (LPHYLIN = .TRUE.), the post-processing of atmospheric parameters at observation level can be performed using the formulation of *Geleyn* (1988) in VDFPPCFLS (the tangent-linear and adjoint versions of VDFPPCFLS are not yet coded, but are already available elsewhere in the IFS for the observation operators).

This modified scheme make use of all the routines from the operational vertical diffusion, except VDFSFLX and VDFHGHT, however the exchange coefficients are computed in a different way in VDFEXCS and VDFEXCU. The linearization of the surface energy balance is also performed (VDFTSKTL, VDFTSKAD), but perturbations of the skin temperature are not evolved in time (section 4.4 in CPGLAGTL). This simplification should be relaxed when the skin temperature becomes part of the control variable.

The logical LEKPERT in NAMTRAJP controls the perturbations of the exchange and drag coefficients. It is set to FALSE by default, to prevent the growth of spurious instabilities in the tangent-linear model..

3.4.4 Sub-grid scale orographic effects

The subgrid-scale orographic scheme is a complete linearization of the operational ECMWF scheme described in *Lott and Miller* (1997). The linearized schemes are called from GWDRAGTL and GWDRAGAD. By setting the constant RLPDRAG to zero, the representation of wave breaking is not activated (The operational value GKDRAG is set to 0.3 in SUGWD and is used to compute the surface stress in the gravity-wave part of the scheme).

3.4.5 Large-scale precipitation

Linearized versions of large-scale condensation scheme are CONDTL and CONDAD. Local supersaturation is removed through a local moist-adjustment scheme (CUADJTQTL , CUADJTQAD). Supersaturation produces precipitation instantaneously (without a cloud stage). The effect of rainfall evaporation in sub-saturated layers is strongly reduced in the linearized versions of the scheme. The constant RLPEVAP is set to 0.05, instead of 0.95 in the nonlinear parametrization (it indicates that evaporation will take place as long as specific humidity in a given layer is below RLPEVAP times its saturation value).

3.4.6 Long-wave radiation

The linear long-wave radiation is based on a constant emissivity approach, where only perturbations on temperature are accounted for. Tendencies produced by the linearized long-wave radiation in RADHEATTL andRADHEATAD are damped above a pressure level p_c (RLPP00) set to 300 hPa in SUPHLI:

$$\frac{\partial T'}{\partial t} = \frac{1}{1+\left(\frac{p}{p_c}\right)^{10}} \frac{g}{c_p} \frac{\partial}{\partial p}\left(\frac{4F}{T}T'\right) \tag{3.13}$$

where the net flux arrays F (PEMTED5) computed from the full non-linear radiation scheme are stored as part of the trajectory during the non-linear integration (see description of the trajectory management).

3.4.7 Deep moist convection

The partial linearization of the ECMWF mass-flux scheme is performed, leading to the following tendencies for the perturbations of the prognostic variables (wind components, specific humidity, dry static energy):

$$\frac{\partial \psi'}{\partial t} = \frac{1}{\rho}\left[(M_{\text{up}} + M_{\text{down}})\frac{\partial \psi'}{\partial z}\right] \tag{3.14}$$

The mass-fluxes profiles associated with the updrafts and the downdrafts M_{up} and M_{down} are recomputed in the tangent-linear and adjoint integrations from the stored basic state. This partial linearization implies that all the non-linear routines for the convection scheme have their tangent-linear and adjoint counterparts (starting form the driving routines CUCALLNTL and CUCALLNAD). However, most of them are only used to recompute the trajectory. The only routines which contain linear statements are CUININTL (mean thermodynamical properties at half model levels), CUDUDVTL (tendencies for the wind components) and CUDTDQNTL (tendencies for dry static energy and specific humidity). Eq. (3.14) is solved numerically in the following form (see *Tiedtke*, 1989) :

$$\frac{\partial \psi'}{\partial t} = \frac{1}{\rho}\left[\frac{\partial (M_{\text{up}}\psi')}{\partial z} + \frac{\partial (M_{\text{down}}\psi')}{\partial z} + (D_{\text{up}} - E_{\text{up}} + D_{\text{down}} - E_{\text{down}})\psi'\right] \tag{3.15}$$

which requires extra local storage of the profiles of entrainment and detrainement rates E and D computed in CUASCN and

in CUDDRAFN (variables PDMFEN and PDMFDE). Eq. (3.15) is only applied when deep convection is diagnosed from the basic state.

3.4.8 Trajectory management

The ECMWF physics uses the tendencies from the dynamics, and variables at $t - \Delta t$ as input to compute the tendencies of a given process (represented by the operator P) for a prognostic variable ψ :

$$\frac{\psi^{n+1} - \psi_{\text{u}}{}^{n-1}}{2\Delta t} = \text{P}(\psi_{\text{u}}{}^{n-1}) \tag{3.16}$$

where the variable ψ_{u} has already been updated by the dynamics and by the previous physical processes (which are called in the following order: radiation; vertical diffusion; subgrid-scale orographic effects; moist convection; large-scale condensation).

Thus :

$$\psi_{\text{u}}^{n-1} = \psi^{n-1} + \left(\frac{\partial \psi}{\partial t}\right)_{\text{dyn}} + \left(\frac{\partial \psi}{\partial t}\right)_{\text{phys}} \tag{3.17}$$

In Eq. (3.16), if the operator P is nonlinear, its linearization around the basic state ψ_{u}^{n-1}, will require to store the model state at time step $n-1$ (trajectory at $t-\Delta t$) as well as the tendencies produced by the dynamics $(\partial\psi/\partial t)_{\mathrm{dyn}}$. The physical tendencies from the previous processes $(\partial\psi/\partial t)_{\mathrm{phys}}$, require an additional call to the nonlinear routines in the adjoint computations (CALLPARAD) and a local storage of the partial tendencies.

The storage of the trajectory at $t-\Delta t$ is performed in CPGLAG by the routine WRPHTRAJ called before the driver of the ECMWF physics CALLPAR. Fields are stored in grid-point space in an array TRAJPHYS allocated in SUSC2.

The following three-dimensional fields are stored :

- For the atmosphere: the prognostic variables (wind components, temperature, specific humidity) and their tendencies produced by adiabatic processes, the vertical velocity, the long-wave fluxes and the solar transmissivity (these two last fields allow the computation of the radiation tendencies from the trajectory in CALLPARTL and CALLPARAD)
- For the soil: the prognostic variables for temperature and moisture content (used to compute the surface fluxes from the trajectory in the linear vertical-diffusion scheme)

A number of two-dimensional fields used at time step $t-\Delta t$ need to be stored: surface pressure, surface fluxes, skin temperature, skin reservoir, snow reservoir, roughness lengths (mainly for the vertical diffusion).

The preliminary computations (pressure and geopotential at full and half model levels, astronomy parameters) are perfomed in CPGLAGTL and CPGLAGAD before calling the driver of the tangent-linear physics CALLPARTL or the driver of the adjoint physics CALLPARAD, and after reading the trajectory fields from RDPHTRAJ.

The number of fields to be stored is defined in SUTRAJP for 3-D atmospheric fields on full model levels (NG3D95), 3-D atmospheric fields on half model levels (NG3P95), 3-D soil fields (NG3S95), and surface fields (NG2D95).

The option to store the trajectory on disk (instead of in memory) also exists through the logical LIOTRPH defined in SUTRAJP but is not used anymore on FUJITSU VPP computers. Packing of the trajectory is also possible with the variable NPCKFT95 (set to 1 by default, which means no packing) and the packing parameter NEXPBT95, provided packing libraries are compiled with the IFS (routines EXPANDX1).

Part II: DATA ASSIMILATION

CHAPTER 4 Background term

Table of contents

4.1 Introduction

4.2 Description of the algorithm

4.3 Technical implementation

4.3.1 Input files

4.3.2 Namelist parameters of

4.3.3 IFS routines

4.3.4 Background error

4.1 INTRODUCTION

The background term described in *Courtier et al.* (1998) was in May 1997 replaced by a new formulation by *Bouttier et al.* (1997), available online as newjb.ps. The old code is still part of the IFS but will not be described in this documentation.

4.2 DESCRIPTION OF THE ALGORITHM

We use the following notation:

- $\delta\mathbf{x}$ is the low–resolution analysis increment, i.e. model field departures from the background,
- $\mathbf{B}$ is the assumed background error covariance matrix,
- ζ, η, (T, p_{surf}) and q are increments of vorticity, divergence, temperature and surface pressure, and specific humidity, respectively, on model levels.
- η_b and $(T, p_{\mathrm{surf}})_{\mathrm{bal}}$ are the *balanced* parts of the η and (T, p_{surf}) increments. The concept of balance will be defined below, and
- η_u and $(T, p_{\mathrm{surf}})_{\mathrm{unbal}}$ are the *unbalanced* parts of η and (T, p_{surf}), i.e. $\eta - \eta_{\mathrm{bal}}$ and $[(T, p_{\mathrm{surf}}) - (T, p_{\mathrm{surf}})_{\mathrm{bal}}]$, respectively.

The incremental variational analysis problem, Eq. (1.2) of Chapter 1 'Incremental formulation of 3D/4D variational assimilation—an overview' , is rewritten in the space defined by the change of variable $\delta\mathbf{x} = \mathbf{L}\chi$ (Section 1.4) where $\mathbf{L}$ satisfies $\mathbf{L}\mathbf{L}^{\mathrm{T}} = \mathbf{B}$ so that J_{b} takes the simple form of Eq. (1.8). In operational practice, the initial point of the minimization is the background, so that initially $\delta\mathbf{x} = \chi = 0$. The minimization can be carried out in the space of χ, where J_b is the euclidean inner product, Eq. (1.8). At the end of the minimization, the analysis increments are reconstructed in model space by $\delta\mathbf{x} = \mathbf{L}\chi$. In order to compare with observations $\mathbf{x}$ is reconstructed using Eq. (2.4), in each simulation. Thus the variational analysis can be done with L, the inverse change of variable from minimization space to model space (chavarin}, without ever using CHAVAR.

The background-error covariance matrix $\mathbf{B}$ is implied by the design of $\mathbf{L}$, which currently has the form

$$\mathbf{L} = \mathbf{K}\mathbf{B}_{\mathrm{u}}^{1/2} \tag{4.1}$$

where $\mathbf{K}$ is a balance operator going from the set of variables ζ, η_{u}, $(T, p_{\mathrm{surf}})_{\mathrm{u}}$ and q, to the model variables ζ, η, (T, p_{surf}) and q. The $\mathbf{B}_{\mathrm{u}}^{1/2}$ operator is the right–hand symmetric square root of the background-error covariances $\mathbf{B}_{\mathrm{u}}$ of ζ, η_{u}, $(T, p_{\mathrm{surf}})_{\mathrm{u}}$ and q, so that

$$\mathbf{B}_{\mathrm{u}} = (\mathbf{B}_{\mathrm{u}}^{1/2})^{\mathrm{T}}\mathbf{B}_{\mathrm{u}}^{1/2} \tag{4.2}$$

So far, the formulation is perfectly general. Now, we restrict $\mathbf{B}_{\mathrm{u}}$ to a simple form and choose a particular balance operator $\mathbf{K}$.

The covariance matrix $\mathbf{B}_{\mathrm{u}}$ is assumed to be block-diagonal, with no correlation between the parameters:

$$\mathbf{B}_{\mathrm{u}} = \begin{bmatrix} \mathbf{C}_{\zeta} & 0 & 0 & 0 \\ 0 & \mathbf{C}_{\eta_{\mathrm{u}}} & 0 & 0 \\ 0 & 0 & \mathbf{C}_{(T,\, p_{\mathrm{surf}})_{\mathrm{u}}} & 0 \\ 0 & 0 & 0 & \mathbf{C}_{q} \end{bmatrix} \tag{4.3}$$

It implies that the q analysis is independent from the other variables. However, assuming that the unbalanced variables are uncorrelated is not too restrictive because, as we shall see below, the design of the balance implies significant multivariate correlations between the meteorological variables.

Each autocovariance block in the above matrix is itself assumed to be block-diagonal in spectral space, with no correlation between different spectral coefficients, but a full vertical autocovariance matrix for each spectral coefficient. The vertical covariance matrices are assumed to depend only on the total wavenumber n. The resulting autocovariance model is homogeneous, isotropic and non-separable in grid-point space: the correlation structures do not depend on the geographical location, but they depend on the scale. The shape of the horizontal correlations is determined by the covariance spectra. The same representation was used in the previous J_{b} formulation (*Rabier and McNally* 1993, *Courtier et al.* 1998). The covariance coefficients are computed statistically using the NMC method (*Parrish and Derber* 1992, *Rabier et al.* 1998) on 24/48-hour forecast differences to estimate the total covariances for each total wavenumber n, and assuming an equipartition of errors between the $2n + 1$ associated spectral coefficients.

The balance relationship is arbitrarily restricted to the following form:

$$\begin{aligned} \eta_{\mathrm{b}} &= \mathbf{M}\zeta \\ (T, p_{\mathrm{surf}})_{\mathrm{b}} &= \mathbf{N}\zeta + \mathbf{P}\eta_{\mathrm{u}} \end{aligned} \tag{4.4}$$

So that the complete balance operator $\mathbf{K}$ is defined by:

$$\begin{aligned} \zeta &= \zeta \\ \eta &= \mathbf{M}\zeta + \eta_{\mathrm{u}} \\ (T, p_{\mathrm{surf}}) &= \mathbf{N}\zeta + \mathbf{P}\eta_{\mathrm{u}} + (T, p_{\mathrm{surf}})_{\mathrm{u}} \\ q &= q \end{aligned} \tag{4.5}$$

or equivalently, in matrix form:

$$\mathbf{K} = \begin{bmatrix} \mathbf{I} & 0 & 0 & 0 \\ \mathbf{M} & \mathbf{I} & 0 & 0 \\ \mathbf{N} & \mathbf{P} & \mathbf{I} & 0 \\ 0 & 0 & 0 & \mathbf{I} \end{bmatrix} \tag{4.6}$$

The matrix blocks $\mathbf{M}$, $\mathbf{N}$ and $\mathbf{P}$ are, in general, not invertible, but $\mathbf{K}$ is. As explained above, the inverse of $\mathbf{K}$ is not actually used in the variational analysis, because the initial point of the minimization is the background.

The matrix multiplication of $\mathbf{B}_u$ by $\mathbf{K}$ allows one to write explicitly the implied background error covariance matrix $\mathbf{B}$ in terms of the meteorological variables ζ, η, (T, p_{surf}) and q:

$$\mathbf{B} = \mathbf{K}\mathbf{B}_u\mathbf{K}^T = \begin{bmatrix} \mathbf{C}_\zeta & \mathbf{C}_\zeta\mathbf{M}^T & \mathbf{C}_\zeta\mathbf{N}^T & 0 \\ \mathbf{M}\mathbf{C}_\zeta & \mathbf{M}\mathbf{C}_\zeta\mathbf{M}^T + \mathbf{C}_{\eta_u} & \mathbf{M}\mathbf{C}_\zeta\mathbf{N}^T + \mathbf{C}_{\eta_u}\mathbf{P}^T & 0 \\ \mathbf{N}\mathbf{C}_\zeta & \mathbf{N}\mathbf{C}_\zeta\mathbf{M}^T + \mathbf{P}\mathbf{C}_{\eta_u} & \mathbf{N}\mathbf{C}_\zeta\mathbf{N}^T + \mathbf{P}\mathbf{C}_{\eta_u}\mathbf{P}^T + \mathbf{C}_{(T,\,p_{surf})_u} & 0 \\ 0 & 0 & 0 & \mathbf{C}_q \end{bmatrix} \tag{4.7}$$

The blocks implied by $\mathbf{C}_\zeta$ and its transforms by the balance operator blocks $\mathbf{M}$, $\mathbf{N}$ and $\mathbf{P}$ are the *balanced* parts of the covariances. For instance, the vorticity covariances $\mathbf{C}_\zeta$ and the unbalanced temperature covariances $\mathbf{C}_{(T,\,p_{surf})_u}$ are both homogeneous and isotropic, whereas the $\mathbf{N}\mathbf{C}_\zeta\mathbf{N}^T$ 'vorticity-balanced' (T, p_{surf}) matrix term depends on latitude—it is predominant in the extratropics, negligible near the equator. The $\mathbf{N}\mathbf{C}_\zeta$ term is responsible for the geostrophic mass/wind coupling.

The $\mathbf{M}$, $\mathbf{N}$ and $\mathbf{P}$ operators used to define the balance have a restricted algebraic structure. $\mathbf{M}$ and $\mathbf{N}$ are both the product of a so-called horizontal balance operator H by vertical balance operators M, N:

$$\begin{aligned} \mathbf{M} &= MH \\ \mathbf{N} &= NH \end{aligned} \tag{4.8}$$

The H operator is a block–diagonal matrix of identical horizontal operators transforming the spectral coefficients of vorticity, independently at each level, into an intermediate variable P_b which is a kind of linearized mass variable defined below. The horizontal operators in H have exactly the same algebraic structure as the standard analytical linear balance on the sphere, and this is where the latitudinal variations of the J_b structures come from: in spectral space,

$$P_b(n, m) = \beta_1(n, m)\varsigma(n, m+1) + \beta_2(n, m)\varsigma(n, m-1) \tag{4.9}$$

The M, N and $\mathbf{P}$ operators all have the same structure: block-diagonal, with one full vertical matrix per spectral component. The vertical matrices depend only on the total wavenumber n.

The actual calibration of the J_b operator requires the following 4 steps; each one uses a set of 24/48-hour-range forecast differences as surrogates to background error patterns in order to calculate the statistics:

1) H **operator**. The horizontal balance coefficients (β_1, β_2) of H are computed by a linear regression between the errors in vorticity and in linearized total mass P_{tot}, assuming the functional relationship defined by the above equation, and building P_{tot} from (T, p_{surf}) using the linearized hydrostatic relationship at level l,

$$P_{\mathrm{tot}}(l) = \sum_{i=L}^{l} R\,T_i \Delta \ln p_i + R\,T_{\mathrm{ref}} \ln p_{\mathrm{surf}} \tag{4.10}$$

which relies on the definition of the model vertical geometry and of reference values for (T, p_{surf}). We use (270 K, 800 hPa) currently. The sensitivity to the somewhat arbitrary choice of these parameters has been tested and it is negligible. Unlike in the previous J_{b} formulation, P_{b} is just an intermediate variable in the linear regression. Modifying the reference values, e.g. to (300 K, 1000 hPa), does change the scaling of H, but it is compensated by corresponding changes in the M and N operators, so that the effective covariances are virtually unchanged.

2) M **operator**. The vertical blocks $M(n)$ of this operator are computed for each wavenumber n by a linear regression between the spectral vertical profiles $[P_{\mathrm{b}}]_n^m$ and $[\eta]_n^m$, respectively, of balanced mass P_{b} (defined as H times the vorticity error patterns) and divergence. The relationship is assumed to be

$$[\eta]_n^m = M(n)[P_{\mathrm{b}}]_n^m \tag{4.11}$$

so that the statistical sampling is better for the small scales than for the large scales because there are $2n+1$ spectral profiles to be used per total wavenumber in each forecast error pattern. At least as many independent error patterns as number of model levels are needed in order to have a well-posed regression problem for the very large scales.

3) N **and P operators**. The vertical blocks are computed for each wavenumber exactly like M, except that now the linear regression goes from the vertical spectral profiles of $P_{\mathrm{b}} = H\zeta$ and $\eta_{\mathrm{u}} = \eta - M\zeta$ to the profiles of temperature concatenated with surface pressure:

$$[(T, p_{\mathrm{surf}})]_n^m = N_n[P_{\mathrm{b}}]_n^m + P_n[\eta_{\mathrm{u}}]_n^m \tag{4.12}$$

One notes that the N_n matrix is not square (the output is larger than the input because there is a kernel in the hydrostatic relationship) but the resulting (T, p_{surf}) covariances are still positive definite by construction thanks to the $C_{(T,\,p_{\mathrm{surf}})_{\mathrm{u}}}$ term in the expression of $\mathbf{B}$.

4) **Error covariances**. The vertical autocovariances of the ζ, η_{u}, $(T, p_{\mathrm{surf}})_{\mathrm{u}}$ and q, difference patterns are computed for each total wavenumber n. Again, since there are $2n+1$ wavenumbers for each n and each error pattern, at least as many linearly independent error patterns as model levels (plus one for p_{surf}) *must* be used in order to ensure that the autocovariances are positive definite at the very large scales. It is strongly advised to use several times more in order to reduce the sampling noise at large scales; this is important for the performance of the resulting assimilation/forecast system. In the May 1997 implementation of the 3D-Var system, about 180 forecast-difference patterns have been used for 31 levels.

In addition to these 4 steps, some minor preprocessing is performed on the covariances. The vertical correlations of humidity are set to zero above 100 hPa in order to avoid spurious stratospheric humidity increments because of the tropospheric observations. The ζ, η_{u} and $(T, p_{\mathrm{surf}})_{\mathrm{u}}$ vertical profiles of total variance are rescaled by an arbitrary factor of 0.9 in order to account for the mismatch between the amplitudes of the 24/48-hour-forecast differences and of the 6-hour forecast errors. In the future this factor will be recalculated more precisely using observation departures from the background in the assimilation, similarly to Hollingsworth and Lönnberg (1986). It may be different for 3D-Var than for 4D-Var. The variance spectra are slightly modified in order to ensure that the horizontal error correlations of ζ, η_{u} and $(T, p_{\mathrm{surf}})_{\mathrm{u}}$ are compactly supported (they are set to zero beyond

6000 km). This operation removes the residual sampling noise in the error covariances. No other processing is performed except for a spectral truncation if the analysis resolution is lower than the statistics resolution (currently T106). It would be easy to extrapolate the statistics to higher resolutions, but it would be very hazardous to alter the vertical geometry of the covariances and balance operators. Instead, it is recommended to run a set of forecasts using a model with the right vertical resolution, and recompute all the statistics from scratch.

4.3 TECHNICAL IMPLEMENTATION

The statistical calibration is done using dedicated scripts outside the IFS code. First, the 24/48-hour forecast error differences for a set of dates are constructed in terms of spectral ζ, η, $(T, p_{\mathrm{surf}})_{\mathrm{u}}$ and q. This involves running a set of MARS requests and building the required GRIB files. Then, the forecast-error differences are read and processed by a Fortran statistics program that finally writes two files in GSA format: one with the coefficients of the balance operator, one with the error covariances of ζ, η_{u}, $(T, p_{\mathrm{surf}})_{\mathrm{u}}$ and q. These files take up a couple of megabytes. They are computed for a given triangular truncation and number of levels (currently T106L31). In the covariance file there are 4 sets of vertical covariance matrices. The balance files contain one set of coefficients for the H operator, and three sets of vertical balance matrices for M, N and $\mathbf{P}$.

4.3.1 Input files

The IFS needs these two GSA files to use J_{b} in e.g. the incremental analysis jobs. The J_{b} configuration described here corresponds to namelist switch LSTABAL=.true. (NAMJG), and it is identified in the J_{b} code by the string CDJBTYPE='STABAL96'. LSTABAL=.false. would give the old J_{b} formulation. The input files must be named **stabal96.cv** and **stabal96.bal**. They are read in by sujbdat and sujbbal, respectively.

4.3.2 Namelist parameters of J_{b}

Some other important namelist options in NAMJG are LCFCE (to enforce uniform background errors), L3DBGERR (to have a 3D distribution vorticity background errors), and LCORCOSU (to enforce compactly supported horizontal correlations). The switch LGUESS in NAMCT0 can be used to switch J_{b} off altogether. The default is LGUESS=.true., i.e. J_{b} switched on.

(This is the J_b setup code tree in IFS cy16r3, option stabal96)
(namelist namjg has already been read into yomjg in routine sujb below su0yoma)

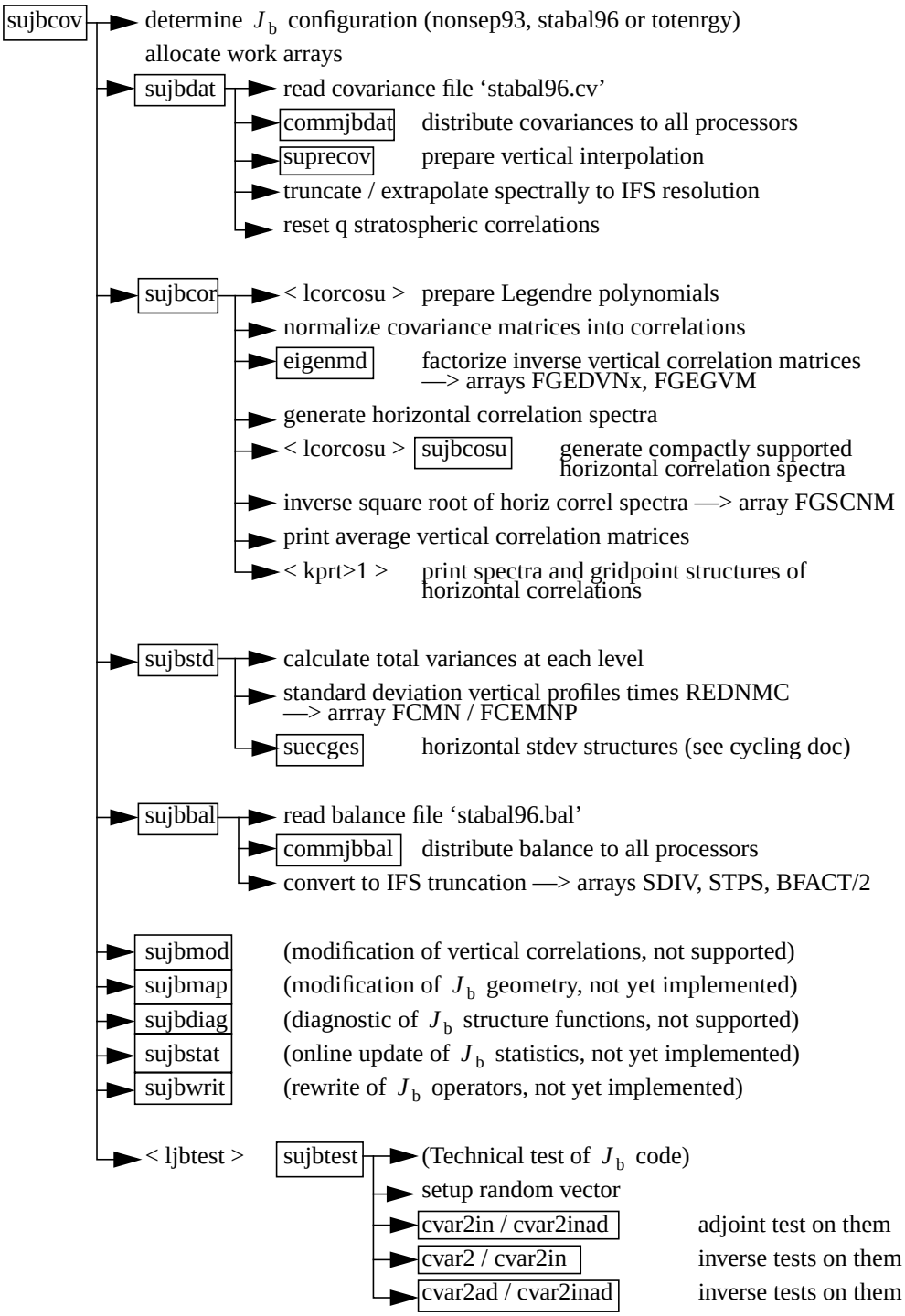

Figure 4.1 Calling tree for subroutine sujbcov.

4.3.3 IFS J_b routines

Inside the IFS code, J_b is localized in the setups below subroutine sujbcov and in the inverse change of variable cvar2in (and its adjoint and their inverses, CVAR2INAD, CVAR2 and CVAR2AD). Calling trees are shown in Fig. 4.1 and Fig. 4.2 . The computation of the cost function and its adjoint is done in sim4d(Section 2.3 in Chapter 2 '3D variational assimilation')---it is planned to move it to a dedicated subroutine. The sequence within the set-up routine is the following:

(i) SUJBDAT: Reads covariances from file **stabal96.cv,**
Interpolates in the vertical to the model's vertical levels (if necessary)
Sets humidity correlations to 0, for pressures less than 100 hPa.

(ii) SUJBCOR: Sets up spectral correlation operator
Covariance matrices (one per n) are converted to vertical correlation matrices and horizontal autocorrelation spectra. The eigenvectors and the eigenvalues of vertical correlations matrices are computed using EIGENMD and stored in FGEDVNS and FGEGVNS-arrays (yomjg), respectively, for later use in JGVCOR, JGVCORI, JGVCORAD and JGVCORIAD. The horizontal autocorrelation spectra are stored in the FGSCNM–array (yomjg), for later use in JGHCOR and JGHCORI.

(iii) SUJBSTD: Set up background error standard deviations, see Subsection 4.3.4.

(iv) SUJBBAL: Set up balance constraint. Read the file **stabal96.bal** and store in yomjg, for later use in balstat, balstatad, balvert, balvertad, balverti and balvertiad as part of the change of variable operator.

(v) SUJBTEST: Test of the adjoint of the change of variable, if LJBTEST=.true.

The distributed memory affects the setups below sujbdat and sujbbal when the data files are read in (by the master processor only). First, the resolution of the files is read, then the relevant arrays are allocated and the actual data is read, truncated if necessary, and broadcast. The code is designed to work at any resolution.

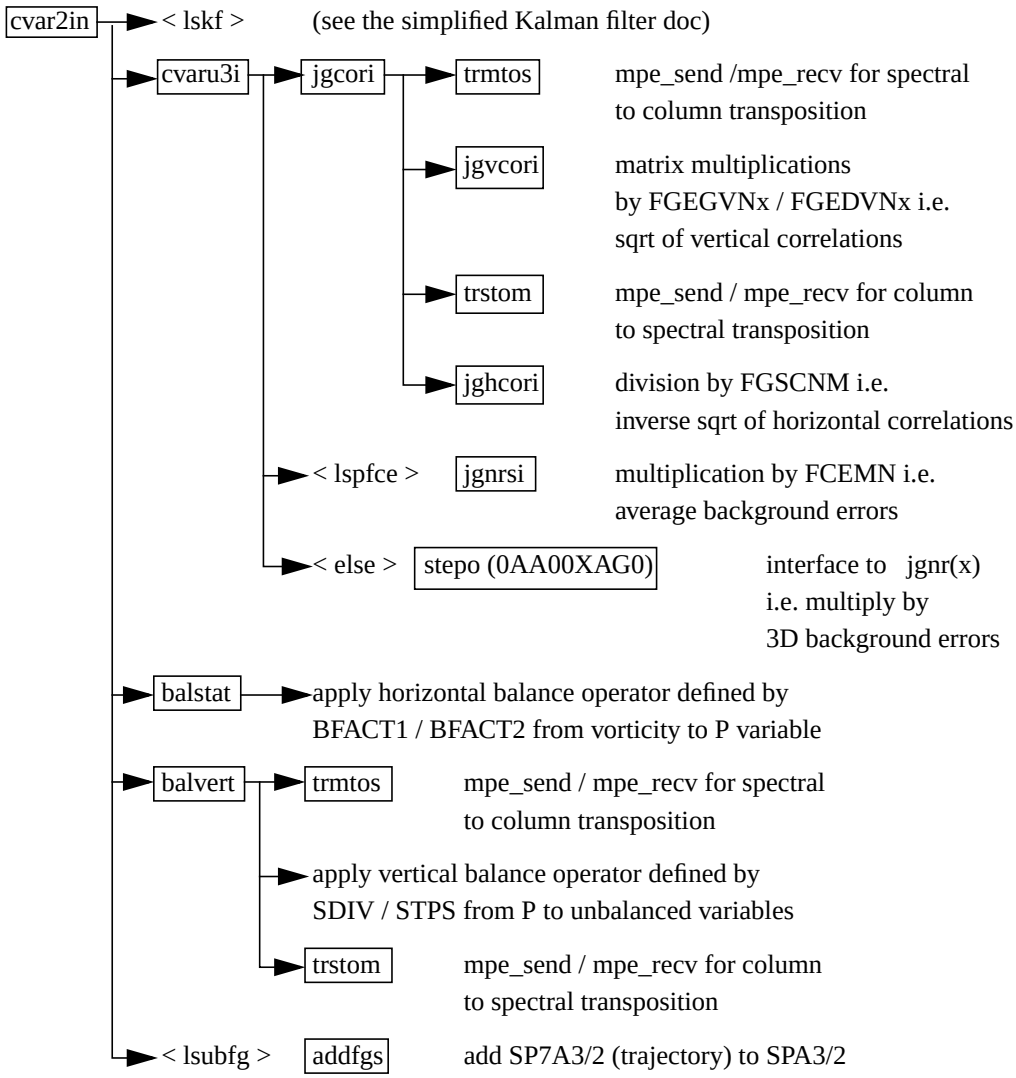

Figure 4.2 Calling tree for subroutine cvar2in.

In the change of variable, there is a transposition of the fields between the horizontal and vertical balance operators, balstat and balvert, repsectively. Note that the operator L is performed by calling cvar2in, so in IFS parlance L corresponds to the *inverse* change of variable.

4.3.4 Background error σ_b

The background standard errors are set up below sujbstd (in SUINFCE, called from SUECGES) and used in jgnr or jgnrs (and their adjoint and inverses, jgnrad and jgnrsi). In addition to the covariance files, they use a gridpoint GRIB file called **errgrib** in order to specify the three–dimensional error patterns. The data from the file is converted to the appropriate parameters and resolution if needed. The background error fields for some parameters (wind, height, temperature and surface pressure) are built for the screening job although they are not needed in the analysis itself. For more information, refer to the chapter on the cycling of background errors, Chapter 7 'Background, analysis and forecast errors' .

4.3.4 (a) Humidity. The humidity background errors are currently not cycled – they are computed (in SUSHFCE under JGNR) by a simple empirical formulation as a function of the temperature T^{b} and relative humidity U^{b} of the background:

$$\sigma_{\mathrm{b}} = -0.002\,T^{\mathrm{b}} - 0.0033|T^{\mathrm{b}} - 273| + 0.25\,U^{\mathrm{b}} - 0.35|U^{\mathrm{b}} - 0.4| + 0.70 \tag{4.13}$$

$$\sigma_{\mathrm{b}} = \min[0.18, \max(0.06, \sigma_{\mathrm{b}})] \tag{4.14}$$

The standard deviation in terms of relative humidity is then converted to specific humidity, taking the variation of q of the equation

$$q = \frac{U e_{\mathrm{sat}}}{\frac{1}{\varepsilon} p - U\left(\frac{1}{\varepsilon} - 1\right) e_{\mathrm{sat}}} \tag{4.15}$$

where U is the relative humidity, $\varepsilon = R_{\mathrm{dry}}/R_{\mathrm{vap}}$, e_{sat} is the saturation water-vapour pressure at the temperature in question (Tetens' formula, Eq. (5.11) in Chapter 5 'Conventional observational constraints') and p is pressure.

Humidity increments are forced to be negligibly small above the tropopause to avoid a systematic drift of stratospheric humidity over extended periods of data assimilation. This is achieved by setting a very low value of 10^{-8} for σ_{b} everywhere the pressure is lower than 70hPa, and at anyother point where the pressure is lower than 500hPa and the background temperature and pressure fields are such that the square of the buoyancy frequency exceeds $2 \times 10^{-4}\,\mathrm{s}^{-2}$ everywhere between that level and the 70hPa level.

More specifically, for each grid column σ_{b} is set to 10^{-8} for model levels k such that $k \le K$, where the level K is determined by requiring either that it is the highest level with $p^{\mathrm{b}}_K \ge 70$ hPa for which

$$\frac{p^{\mathrm{b}}_{K+1}(T^{\mathrm{b}}_{K+2} - T^{\mathrm{b}}_K)}{T^{\mathrm{b}}_{K+1}(p^{\mathrm{b}}_{K+2} - p^{\mathrm{b}}_K)} + \frac{R_{\mathrm{dry}} T^{\mathrm{b}}_{K+1}}{g^2}(2.5 \times 10^{-4}) > \kappa \tag{4.16}$$

or, if no such level can be found for p^{b}_K in the range from 500 to 70hPa, that it is the lowest level for which

$$p^{\mathrm{b}}_K \le 500 \text{ hPa}\,.$$

Here T^{b}_K and p^{b}_K are the background temperature and pressure at level K of the grid-column.

In addition, any values of σ_{b} lower than 10^{-8} are reset to 10^{-8}.

For pressures less than p_0 = 800 hPa, and over the sea, the model of background errors above is modified by

$$\sigma_{\mathrm{mod}} = \sigma_{\mathrm{b}}\left\{1 - a + a\exp\left[-\left(\frac{p^{\mathrm{b}} - p_0}{b}\right)^2\right]\right\} \tag{4.17}$$

where $\alpha = 0.5(1 - \mathrm{LSM})$ (where LSM = land–sea mask) and b =12500.

Part II: DATA ASSIMILATION

CHAPTER 5 Conventional observational constraints

Table of contents

5.1 Introduction

5.2 Data usage

5.2.1 Controls

5.2.2 Overview of observation operators

5.3 The observation operator for geopotential height

5.3.1 Quadratic vertical interpolation near the top of the model

5.3.2 Below the model's orography

5.4 The observation operator for wind

5.5 The observation operators for humidity

5.5.1 Saturation vapour pressure

5.5.2 Relative humidity

5.5.3 Precipitable water

5.5.4 Specific humidity

5.6 The observation operator for temperature

5.7 Surface observation operators

5.7.1 Mathematical formulation

5.7.2 Surface values of dry static energy

5.7.3 Transfer coefficients

5.7.4 Two-metre relative humidity

5.1 INTRODUCTION

The observation operators provide the link between the analysis variables and the observations (*Lorenc,* 1986, *Pailleux*, 1990). The operator H in Eq. (1.4) signifies the ensemble of operators transforming the control variable $\mathbf{x}$ into the equivalent of each observed quantity, $\mathbf{y}^o$, at observation locations. The 3D/4D–Var implementation allows H to be (weakly) non-linear, which is seen to be an advantage for the use of TOVS radiance data, for example. In this chapter we define the content of each of the observation operators and describe the observational data used in 3D/4D–Var. The use of satellite data is described in Chapter 6 'Satellite observational constraints'.

5.2 DATA USAGE

Observation operators for all observation types that were used by OI have also been implemented in 3D/4D–Var. In addition 3D/4D–Var uses TOVS cloud-cleared radiances, scatterometer ambiguous winds and SSMI total column water vapour. Table 10.6 lists the observing systems currently used by 4D–Var in ECMWF's operational data assimilation. The table also indicates important restrictions on data usage and thinning of data. TOVS, SCAT and SSMI data are further discussed in Chapter 6 'Satellite observational constraints' . Additional data types such as meteosat radiances and TOVS and ATOVS 1C-radiances are used experimentally. ATOVS 1C radiance usage is described in a separate chapter. 3D/4D–Var uses the data from a six-hour time window centred at the analysis time. In 3D-Var there is no interpolation in time of the background, which means that all data are used as if they were observed at the analysis time. In 4D-Var, on the other hand, the background trajectory is available on a one hourly interval (Chapter 3 '4D variational assimilation'). If there are multiple reports from the same fixed observing station within that time window (6 hours in 3D-Var, 1 hour in 4D-Var), the data nearest the analysis time are selected for use in the analysis. Some thinning is applied for the moving platforms reporting frequently. These tasks are performed in the screening configuration of IFS, see Chapter 10 'Observation screening' .

5.2.1 Controls

The **blacklist** mechanism is very flexible and allows the complete control of which data to use/not use in the variational assimilation. The blacklist is applied in the screening job, which removes the blacklisted data from the compressed CMA observation file. Data-selection rules should be coded in the blacklist files rather than in the IFS code itself.

Classes of data can also be switched on and off using the NOTVAR array in NAMJO, however it is preferable to use the blacklist mechanism for this purpose. The second dimension in thisarray is the observation type. The first dimension is variable number, area and subobs-type, respectively—see the documentation of obsproc for definitions. The elements of the NOTVAR array\ can take either of three values:

- 0, means that the data will be **used**,
- –1, means that the data will **not** be **used,** and
- –2, means that the data will be **passive,** i.e. departures will be calculated but there will be no contribution to J_o

5.2.2 Overview of observation operators

The operator H is subdivided into a sequence of operators, each one of which performs part of the transformation from control variable to observed quantity:

(i) The inverse change of variable (CHAVARIN) converts from control variables to model variables (see Section 2.3 in Chapter 2 '3D variational assimilation').

(ii) The inverse spectral transforms put the model variables on the model's reduced Gaussian grid (controlled by SCAN2MDM).

(iii) A 12-point bi-cubic or 4-point bi-linear horizontal interpolation gives vertical profiles of model variables at observation points (controlled by COBS, COBSLAG, Section 2.4). The surface fields are interpolated bi-linearly to avoid spurious maxima and minima. The three steps (*i*) to (*iii*) are common to all data types. Thereafter follows:

(iv) Vertical integration of, for example, the hydrostatic equation to form geopotential (Section 2.3), and of the radiative transfer equation to form radiances (if applicable, Subsection 6.4.1), and

(v) vertical interpolation to the level of the observations.

The vertical operations depend on the variable. The vertical interpolation is linear in pressure for temperature (PPT)

and specific humidity (PPQ), and it is linear in the logarithm of pressure for wind (PPUV). The vertical interpolation of geopotential (PPGEOP) is similar to wind (in order to preserve geostrophy) and is performed in terms of departures from the ICAO standard atmosphere for increased accuracy (*Simmons and Chen*, 1991, see Section 5.3 below). The current geopotential vertical interpolation together with the temperature vertical interpolation are not exactly consistent with hydrostatism. A new consistent and accurate vertical interpolation has been devised by Meteo-France, which may be important for intensive use of temperature information. The new routines have been tested by ECMWF and as the results were not unambiguously positive the new routines have not yet been adopted – and they are not described in this documentation. In the meantime, the old routines are still used (switch LOLDPP=.true. in namct0), under the names PPT_OLD, PPGEOP_OLD and PPUV_OLD, with tangent linear PPTTL_OLD, PPGEOPTL_OLD and PPUVTL_OLD and adjoint PPTAD_OLD, PPGEOPAD_OLD and PPUVAD_OLD.

The vertical interpolation operators for SYNOP 10 m wind (PPUV10M) and 2 m temperature (PPT2M) match an earlier version of the model's surface layer parametrisation. The vertical gradients of the model variables vary strongly in the lowest part of the boundary layer, where flow changes are induced on very short time and space scales, due to physical factors such as turbulence and terrain characteristics. The vertical interpolation operator for those data takes this into account following Monin–Obukhov similarity theory. Results using such operators, which follow *Geleyn* (1988) have been presented by *Cardinali et al.* (1994). It was found that 2-metre-temperature data could not be satisfactorily used in the absence of surface skin temperature as part of the control variable, as unrealistic analysis increments appeared in the near-surface temperature gradients. The Monin–Obukhov based observation operator for 10 m wind, on the other hand, is used for all 10 m winds (SYNOP, DRIBU, TEMP, PILOT and SCAT). SCAT 10 m winds may optionally (setting LSCASUR=.F. in namobs) be used through a simple logarithmic relationship between lowest model level wind (at approximately 32 m) and wind at 10 m (see PPOBSAS, Subsection 6.4.4).

Relative humidity is assumed constant in the lowest model layer to evaluate its 2 m value (PPRH2M), see Subsection 5.7.4. The model equivalent of SSMI total column water vapour data is obtained by vertical integration of q (in GPPWC and PPPWC). Observation operators also exist for SATEM precipitable water content (also PPPWC) and SATEM thicknesses (PPGEOP), but these data are currently not used operationally.

The variational analysis procedure requires the gradient of the objective function with respect to the control variable. This computation makes use of the adjoint of the individual tangent linear operators, applied in the reverse order. The details regarding observation operators for conventional data can be found in *Vasiljevic et al.* (1992), *Courtier et al.* (1998), and in the following sections.

5.3 THE OBSERVATION OPERATOR FOR GEOPOTENTIAL HEIGHT

The geopotential at a given pressure p is computed by integrating the hydrostatic equation analytically using the ICAO temperature profile and vertically interpolating $\Delta\phi$, the difference between the model level geopotential and the ICAO geopotential (*Simmons and Chen*, 1991). The ICAO temperature profile is defined as

$$T_{\mathrm{ICAO}} = T_0 - \frac{\Lambda}{g}\phi_{\mathrm{ICAO}} \tag{5.1}$$

where T_0 is 288 K, ϕ_{ICAO} is the geopotential above 1013.25 hPa and Λ is 0.0065 K m^{-1} in the ICAO troposphere and 0 in the ICAO stratosphere (the routine PPSTA). The ICAO tropopause is defined by the level where the ICAO temperature has reached 216.5 K (SUSTA). Using this temperature profile and integrating the hydrostatic equation provides T_{ICAO} and the geopotential ϕ_{ICAO} as a function of pressure (PPSTA). We may then evaluate the geo-

potential $\phi(p)$ at any pressure p following

$$\phi(p) - \phi_{\text{surf}} = \phi_{\text{ICAO}}(p) - \phi_{\text{ICAO}}(p_{\text{surf}}) + \Delta\phi \tag{5.2}$$

where p_{surf} is the model surface pressure and ϕ_{surf}, the model orography. $\Delta\phi$ is obtained by vertical interpolation from the full model level values $\Delta\phi_k$. The interpolation is linear in $\ln(p)$ up to the second model level (PPINTP) and quadratic in $\ln(p)$ for levels above it (PPITPQ, see below). The full model level values are obtained integrating the discretized hydrostatic equation using the routine GPGEO of the forecast model, following *Simmons and Burridge* (1981):

$$\Delta\phi_k = \sum_{j=L}^{k+1} R_{\text{dry}}(T_{\text{v}_j} - T_{\text{ICAO}_j}) \ln\left(\frac{p_{j+1/2}}{p_{j-1/2}}\right) + \alpha_k R_{\text{dry}}(T_{\text{v}_k} - T_{\text{ICAO}_k}) \tag{5.3}$$

with

$$\alpha_k = 1 - \frac{p_{k-1/2}}{p_{k+1/2} - p_{k-1/2}} \ln\left(\frac{p_{k+1/2}}{p_{k-1/2}}\right)$$

for $k > 1$ and $\alpha_1 = \ln(2)$.

5.3.1 Quadratic vertical interpolation near the top of the model

Above the second full level of the model, the linear interpolation (PPINTP)is replaced by a quadratic interpolation in $\ln p$, performed in the routine PPITPQ:

$$z(\ln p) = a + b(\ln p) + c(\ln p)^2 \tag{5.4}$$

where a, b and c are constants determined so that the above equation fits the heights at the top levels ($k = 1, 2$ and 3). The interpolation formula is:

$$\phi(\ln p) = z_2 + \frac{(z_2 - z_1)(\ln p - \ln p_2)(\ln p - \ln p_3)}{(\ln p_2 - \ln p_1)(\ln p_1 - \ln p_3)} - \frac{(z_2 - z_3)(\ln p - \ln p_1)(\ln p - \ln p_2)}{(\ln p_2 - \ln p_3)(\ln p_1 - \ln p_3)} \tag{5.5}$$

where 1,2 and 3 refer to levels $k = 1, 2$ and 3, respectively.

5.3.2 Below the model's orography

The extrapolation of the geopotential below the model's orography is carried out as follows: Find T^* (surface temperature) by assuming a constant lapse rate Λ, from the model level above the lowest model level (subscript $l-1$), see the routine CTSTAR,

$$T^* = T_{l-1} + \Lambda \frac{R_{\text{dry}}}{g} T_{l-1} \ln \frac{p_{\text{surf}}}{p_{l-1}} \tag{5.6}$$

$$T^* = \frac{\{T^* + \max[T_y, \min(T_x, T^*)]\}}{2} \tag{5.7}$$

Find the temperature at mean sea level, T_0 (also in CTSTAR)

$$T_0 = T^* + \Lambda\frac{\phi_{\text{surf}}}{g} \tag{5.8}$$

$$T_0 = \min[T_0, \max(T_x, T^*)] \tag{5.9}$$

where T_x is 290.5 K and T_y is 255 K. The geopotential under the model's orography is (in PPGEOP) calculated as:

$$\phi = \phi_{\text{surf}} - \frac{R_{\text{dry}}T^*}{\gamma}\left[\left(\frac{p}{p_{\text{surf}}}\right)^{\gamma} - 1\right] \tag{5.10}$$

where $\gamma = \frac{R_{\text{dry}}}{\phi_{\text{surf}}}(T_0 - T_{\text{surf}})$.

5.4 THE OBSERVATION OPERATOR FOR WIND

In PPUV a linear interpolation in $\ln p$ (PPINTP) is used to interpolate u and v to the observed pressure levels up to the second full model level, above which a quadratic interpolation is used (PPITPQ, see Subsection 5.3.1.1). Below the lowest model level wind components are assumed to be constant and equal to the values of the lowest model level.

5.5 THE OBSERVATION OPERATORS FOR HUMIDITY

Specific humidity q, relative humidity U and precipitable water content PWC are linearly interpolated in p, in PPQ, PPRH and PPPWC, respectively. Upper air relative humidity data are normally not used, but could be used, if required. The use of surface relative humidity data is described in Subsection 5.7.4.

5.5.1 Saturation vapour pressure

The saturation vapour pressure $e_{\text{sat}}(T)$ is calculated using Tetens's formula:

$$e_{\text{sat}}(T) = a_1 \exp^{a_3\left(\frac{T-T_3}{T-a_4}\right)} \tag{5.11}$$

using FOEEWM (mixed phases, water and ice) in the model and FOEEWMO (water only) for observations. The use of water-phase only is in accordance with the WMO rules for radiosonde and synop reporting practices. Note that these statement functions compute $(R_{\text{dry}}/R_{\text{vap}})e_{\text{sat}}(T)$, with the parameters set according to Buck (1981) and the AERKi formula of *Alduchov and Eskridge* (1996), i.e. a_1 = 611.21 hPa, a_3 = 17.502 and a_4 = 32.19 K over water, and for FOEEWM a_3 = 22.587 and a_4 = –0.7 K over ice, with T_3 = 273.16 K. Furthermore in FOEEWM the saturation value over water is taken for temperatures above 0°C and the value over ice is taken for temperatures below –23°C. For intermediate temperatures the saturation vapour pressure is computed as a combination of the values over water $e_{\text{sat(water)}}$ and ice $e_{\text{sat(ice)}}$ according to the formula

$$e_{\text{sat}}(T) = e_{\text{sat(ice)}}(T) + [e_{\text{sat(water)}}(T) - e_{\text{sat(ice)}}(T)]\left(\frac{T-T_{\text{i}}}{T_3-T_{\text{i}}}\right)^2 \tag{5.12}$$

with $T_3 - T_i = 23$ K.

5.5.2 Relative humidity

In GPRH relative humidity U is computed:

$$U = \frac{p\,q\,\frac{R_{vap}}{R_{dry}}}{\left[1 + \left(\frac{R_{vap}}{R_{dry}} - 1\right)q\right] e_{sat}(T)} \tag{5.13}$$

and then in PPRH interpolated to the required observed pressure levels (using PPINTP). Below the lowest model level and above the top of the model is U assumed to be constant. Saturation vapour pressure is calculated using FOEEWMO if GPRH has been called form the observation operator routines, and using FOEEWM if called from the model post processing.

5.5.3 Precipitable water

In GPPWC precipitable water is calculated as a vertical summation from the top of the model:

$$PWC_k = \frac{1}{g}\sum_{i=1}^{k} q_i (p_i - p_{i-1}) \tag{5.14}$$

and then in PPPWC interpolated to the required observed pressure levels (using PPINTP). PWC is assumed to be zero above the top of the model. Below the model's orography PWC is extrapolated assuming a constant $q = q_l$.

5.5.4 Specific humidity

Specific humidity q is in PPQ interpolated to the required observed pressure levels (using PPINTP). Below the lowest model level and above the top of the model is q assumed to be constant and equal to q_l and q_1, respectively.

5.6 THE OBSERVATION OPERATOR FOR TEMPERATURE

Temperature is interpolated linearly in pressure (PPINTP), in the routine PPT. Above the highest model level the temperature is kept constant and equal to the value of the highest model level. Between the lowest model level and the model's surface the temperature is interpolated linearly, using:

$$T = \frac{(p_{surf} - p)\,T_l + (p - p_l)\,T^*}{p_{surf} - p_l} \tag{5.15}$$

Below the lowest model level the temperature is extrapolated by

$$T = T^*\left[1 + \alpha \ln\frac{p}{p_{surf}} + \frac{1}{2}\left(\alpha \ln\frac{p}{p_{surf}}\right)^2 + \frac{1}{6}\left(\alpha \ln\frac{p}{p_{surf}}\right)^3\right] \tag{5.16}$$

with $\alpha = \Lambda R_{dry}/g$, for $\phi_{sat}/g < 2000$ m, but α is modified for high orography to $\alpha = R_{dry}(T_0' - T^*)/\phi_{surf}$,

where

$$T_0' = \min(T_0, 298) \tag{5.17}$$

for $\phi_{surf}/g > 2500$ m , and

$$T_0' = 0.002[(2500 - \phi_{surf}/g)T_0 + (\phi_{surf}/g - 2000)\min(T_0, 298)] \tag{5.18}$$

for $2000 < \phi_{surf}/g < 2500$ m . If $T_0' < T^*$ then α is reset to zero. The two temperatures T^* and T_0 are computed using Eqs. (5.6) to (5.9).

5.7 SURFACE OBSERVATION OPERATORS

All surface data are processed in the routine SURFACEO. Preparations for the vertical interpolation is done as for all other data in PREINT (see Subsection 5.3.2), and for surface data there are a few additional tasks which are performed in a separate routine, PREINTS. In PREINTS surface roughness over sea, dry static energy (SURBOUND), Richardson number, drag coefficients and stability functions (EXCHCO), are computed, as detailed in the following.

5.7.1 Mathematical formulation

An analytical technique (*Geleyn*, 1988) is used to interpolate values between the lowest model level and the surface. When Monin–Obukhov theory is applied:

$$\frac{\partial u}{\partial z} = \frac{u_*}{\kappa(z + z_0)}\phi_M\left(\frac{z + z_0}{L}\right) \tag{5.19}$$

$$\frac{\partial s}{\partial z} = \frac{s_*}{\kappa(z + z_0)}\phi_H\left(\frac{z + z_0}{L}\right) \tag{5.20}$$

$$L = \frac{c_p}{g}\frac{T}{\kappa}\frac{u_*^2}{s_*} \tag{5.21}$$

where u, s are wind and energy variables, u_*, s_* are friction values and $\kappa = 0.4$ is von Kármán's constant.

The temperature is linked to the dry static energy s by:

$$s = c_p T + \phi \tag{5.22}$$

$$c_p = c_{p_{dry}}\left[1 + \left(\frac{c_{p_{vap}}}{c_{p_{dry}}} - 1\right)q\right] \tag{5.23}$$

Defining the neutral surface exchange coefficient at the height z as:

$$C_N = \left[\frac{\kappa}{\ln\left(\frac{z+z_0}{z_0}\right)}\right]^2 \tag{5.24}$$

The drag and heat coefficients as:

$$C_M = \frac{u_*^2}{[u(z)]^2} \tag{5.25}$$

$$C_H = \frac{u_* s_*}{u(z)[s(z)-\tilde{s}]} \tag{5.26}$$

we can set the following quantities:

$$B_N = \frac{\kappa}{\sqrt{C_N}},\ B_M = \frac{\kappa}{\sqrt{C_M}},\ B_H = \frac{\kappa\sqrt{C_M}}{C_H} \tag{5.27}$$

and considering the stability function in stable conditions as:

$$\phi_{M/H} = 1 + \beta_{M/H}\frac{z}{L} \tag{5.28}$$

we obtain integrating Eqs. (5.19) and (5.20) from 0 to z_1 (the lowest model level):

$$u(z) = \frac{u(z_1)}{B_M}\left[\ln\left(1+\frac{z}{z_1}(e^{B_N}-1)\right) - \frac{z}{z_1}(B_N - B_M)\right] \tag{5.29}$$

$$s(z) = \tilde{s} + \frac{s(z_1)-\tilde{s}}{B_H}\left[\ln\left(1+\frac{z}{z_1}(e^{B_N}-1)\right) - \frac{z}{z_1}(B_N - B_H)\right] \tag{5.30}$$

In unstable conditions the stability function can be expressed as:

$$\phi_{M/H} = \left(1-\beta_{M/H}\frac{z}{L}\right)^{-1} \tag{5.31}$$

and the vertical profiles for wind and dry static energy are:

$$u(z) = \frac{u(z_1)}{B_M}\left[\ln\left(1+\frac{z}{z_1}(e^{B_N}-1)\right) - \ln\left(1+\frac{z}{z_1}(e^{B_N - B_M}-1)\right)\right] \tag{5.32}$$

$$s(z) = \tilde{s} + \frac{s(z_1)-\tilde{s}}{B_H}\left[\ln\left(1+\frac{z}{z_1}(e^{B_N}-1)\right) - \ln\left(1+\frac{z}{z_1}(e^{B_N - B_H}-1)\right)\right] \tag{5.33}$$

The temperature can then be obtained from s as:

$$T(z) = s(z) - \frac{zg}{c_p} \tag{5.34}$$

When z is set to the observation height, Eqs. (5.29) and (5.30) and Eqs. (5.32)–(5.34) give the postprocessed wind and temperature. To solve the problem, we have to compute the dry static energy at the surface $\tilde{s} = \tilde{s}(T_{\text{surf}}, q = 0)$ (Subsection 5.7.2), with B_{M}, B_{N} and B_{H} values depending on the drag and heat exchange coefficients Eq. as detailed in Subsection 5.7.3.

5.7.2 Surface values of dry static energy

To determine the dry static energy at the surface we use Eqs. (5.22) and (5.23) where the humidity at the surface is defined by:

$$\tilde{q} = q(z = 0) = h(C_{\text{snow}}, C_{\text{liq}}, C_{\text{veg}}) q_{\text{sat}}(T_{\text{surf}}, p_{\text{surf}}) \tag{5.35}$$

h is given by (*Blondin*, 1991):

$$h = C_{\text{snow}} + (1 - C_{\text{snow}})[C_{\text{liq}} + (1 - C_{\text{liq}})\bar{h}] \tag{5.36}$$

with

$$\bar{h} = \max\left\{0.5\left(1 - \cos\frac{\pi\vartheta_{\text{soil}}}{\vartheta_{\text{cap}}}\right), \min\left(1, \frac{q}{q_{\text{sat}}(T_{\text{surf}}, p_{\text{surf}})}\right)\right\} \tag{5.37}$$

where ϑ_{soil} is the soil moisture content and ϑ_{cap} is the soil moisture at field capacitiy (2/7 in volumetric units). Eq. (5.36) assigns a value of 1 to the surface relative humidity over the snow covered and wet fraction of the grid box. The snow-cover fraction C_{snow} depends on the snow amount W_{snow}:

$$C_{\text{snow}} = \min\left(1, \frac{W_{\text{snow}}}{W_{\text{snow}_{\text{cr}}}}\right)$$

where $W_{\text{snow}_{\text{cr}}} = 0.015$ m is a critical value. The wet skin fraction C_{liq} is derived from the skin-reservoir water content W_{liq}:

$$C_{\text{liq}} = \min\left(1, \frac{W_{\text{liq}}}{W_{\text{liq}_{\text{max}}}}\right),$$

where

$$W_{\text{liq}_{\text{max}}} = W_{\text{layer}_{\text{max}}}\{(1 - C_{\text{veg}}) + C_{\text{veg}} A_{\text{leaf}}\}$$

with $W_{\text{layer}_{\text{max}}} = 2 \times 10^{-4}$ m being the maximum amount of water that can be held on one layer of leaves, or as a film on bare soil, $A_{\text{leaf}} = 4$ is the leaf-area index, and C_{veg} is the vegetation fraction.

5.7.3 Transfer coefficients

Comparing the Eqs. (5.19) – (5.20) integrated from z_o to $z + z_0$ with Eqs. (5.24) to (5.26), C_{M} and C_{H} can be

analytically defined:

$$\frac{1}{C_{\mathrm{M}}} = \frac{1}{\kappa^2}\left[\int_{z_0}^{(z+z_0)} \frac{\phi_{\mathrm{M}}(z'/L)}{z'}\,\mathrm{d}z'\right]^2 \tag{5.38}$$

$$\frac{1}{C_{\mathrm{H}}} = \frac{1}{\kappa^2}\left[\int_{z_0}^{(z+z_0)} \frac{\phi_{\mathrm{M}}(z'/L)}{z'}\,\mathrm{d}z' \int_{z_0}^{(z+z_0)} \frac{\phi_{\mathrm{H}}(z'/L)}{z'}\,\mathrm{d}z'\right] \tag{5.39}$$

Because of the complicated form of the stability functions, the former integrals have been approximated by analytical expressions, formally given by:

$$\begin{aligned} C_{\mathrm{M}} &= C_{\mathrm{N}} f_{\mathrm{M}}\left(Ri, \frac{z}{z_0}\right) \\ C_{\mathrm{H}} &= C_{\mathrm{N}} f_{\mathrm{H}}\left(Ri, \frac{z}{z_0}\right) \end{aligned} \tag{5.40}$$

where C_{N} is given by Eq. (5.24). The bulk Richardson number Ri is defined as:

$$Ri = \frac{g\Delta z \Delta T_{\mathrm{v}}}{c_p T_{\mathrm{v}} |\Delta \underline{u}|^2} \tag{5.41}$$

where T_{v} is the virtual potential temperature. The functions f_{M} and f_{H} correspond to the model instability functions and have the correct behaviour near neutrality and in the cases of high stability (*Louis*, 1979; *Louis et al.* 1982)

(a) unstable case $Ri < 0$

$$f_{\mathrm{M}} = 1 - \frac{2b\ Ri}{1 + 3bCC_{\mathrm{N}}\sqrt{\left(1 + \frac{z}{z_0}\right)(-Ri)}} \tag{5.42}$$

$$f_{\mathrm{H}} = 1 - \frac{3b\ Ri}{1 + 3bCC_{\mathrm{N}}\sqrt{\left(1 + \frac{z}{z_0}\right)(-Ri)}} \tag{5.43}$$

C=5

(b) Stable case $Ri > 0$

$$f_{\mathrm{M}} = \frac{1}{1 + 2b\ Ri(1 + d\ Ri)^{1/2}} \tag{5.44}$$

$$f_{\mathrm{H}} = \frac{1}{1 + 3b\ Ri(1 + d\ Ri)^{1/2}} \tag{5.45}$$

d = 5

5.7.4 Two-metre relative humidity

In GPRH relative humidity is computed according to Eq. (5.13). The relative humidity depends on specific humidity, temperature and pressure (q, T and p , respectively) at the lowest model level. It is constant in the surface model layer, see PPRH2M.

Part II: DATA ASSIMILATION

CHAPTER 6 Satellite observational constraints

Table of contents

6.1 Introduction

6.2 Set-up of the radiative-transfer code

6.2.1 Satellite identifiers

6.2.3 Fixed pressure levels and RT validation bounds

6.2.4 Radiance observation errors, bias and emissivity

6.3 Set-up for geopotential thickness and PWC

6.3.1 Layers

6.3.2 Observation errors

6.4 Observation operators

6.4.1 Radiances

6.4.2 Thicknesses

6.4.3 Precipitable water from SATEM and SSM/I

6.4.4 Scatterometer winds

6.1 INTRODUCTION

The processing within 3D/4D–Var of satellite data follows the general layout presented in Sections 2.4 and 2.5 of Chapter 2 '3D variational assimilation' . The same vertical interpolation routines as in Chapter 5 'Conventional observational constraints' are used whenever possible. The main difference in the organization is that the radiative transfer code, RTTOV-5 (*Saunders and Matricardi* 1998), currently requires the model profiles to be interpolated to 43 fixed pressure levels from 1013.25 hPa to 0.1 hPa on which the radiative transfer coefficients are determined.

The current operational configuration uses TOVS radiances (*Andersson et al.* 1994), SCAT ambiguous surface winds (*Stoffelen and Anderson*, 1997), SSM/I total column water vapour and wind speed (Gerard and Saunders, 1999; *Phalippou*, 1996; *Phalippou and Gérard*, 1996) and SATOB cloud motion winds of various types (*Tomassini et al.* 1997). Operators also exist for SATEM thicknesses and PWC (*Kelly and Pailleux*, 1988; *Kelly et al.* 1991), and radiances from geostationary satellites. Cloud motion winds (SATOBs) are used just like any other upper level wind observations (Chapter 5 'Conventional observational constraints') and will not be discussed any further.

At the introduction of 21r1 (May 1999) we move from the use of RTOVS cloud-cleared radiances to 1C, or 'raw', radiances, which do not require the 1D-Var retrieval step.

6.2 SET-UP OF THE RADIATIVE-TRANSFER CODE

There are two set–up routines (GETSATID and RTSETUP) for the radiative transfer computations and both are

called from SURAD. The routine RTSETUP calls RTTVI (tovscode) which reads in the transmittance coefficients to memory for the satellites present. The file containing these coefficients is **rt_coef_ieee.dat** for TOVS and all satellites from VTPR through to NOAA-15 are supported.

6.2.1 Satellite identifiers

Satellite identifiers are dealt with in just one place in the IFS and that is in the routine GETSATID. The ODB contains the identifiers as given in the original BUFR messages. Lists of identifiers for which data exist in any given ODB are prepared in the routine MKGLOBSTAB. The routine GETSATID matches those BUFR satellite identifiers with the more traditional satellite numbers, used by the RT–code (e.g. 10 for NOAA–10 and 5 for METEOSAT). The id–conversion tables can be modified through a series of namelists: NAMTOVS; NAMDMSP; NAMMETEOSAT; NAMGOES; and NAMGMS. The satellites are furthermore associated with an 'area' number (between 1 and 20) to be used as an array index in the J_0 table (FJO, Subsection 2.5.3), and with a sequence number for addressing the transmittance coefficients within the RTTOVS code, for example. Note that this sequence number is universally determined across all processors, so if NOAA-12 (BUFR id=202) is satellite-1 on one processor then it will be satellite-1 on all other processors, too.

6.2.2 Satellite sensors

The various types of radiance data are classified by sensor. Each satellite sensor is assigned a number, in yomtvrad: currently HIRS=0, MSU=1, SSU=2, AMSUA=3, AMSUB=4, SSMI=6, VTPR1=7, VTPR2=8 and METEOSAT=20. The sensor number is used as index to various tables containing observation errors, BgQC thresholds, VarQC parameters, et cetera. See the routine DEFRUN.

6.2.3 Fixed pressure levels and RT validation bounds

The list of the 43 fixed pressure levels is passed from the RTTOV library (where they have been read from the transmittance coefficients file) to RTSETUP and SURAD and copied to yomtvrad. RTSETUP also similarly obtains (from RTTOV) lists of temperature, humidities and ozone limits indicating the vaild range of the RT transmittance regression. The RT code is not reliable in conditions beyond these limits. Checks are applied in RADTR.

6.2.4 Radiance observation errors, bias and emissivity

Observation errors and bias corrections for 1C radiances are written to the odb in a call to RAD1COBE (from HRETR). The bias correction is stored in the NCMTORB word and later applied at each iteration of 3D/4D–Var, in the routine HDEPART, Subsection 2.5.2 of Chapter 2 '3D variational assimilation' . Microwave (EMIS_MW) and infrared (EMIS_IR) emissivity are computed in RAD1CEMIS and stored in the odb for later use by RTTOV.

6.3 SET-UP FOR GEOPOTENTIAL THICKNESS AND PWC

NESDIS or 1D-Var thickness and/or PWC data are currently not used in operations (since August 1997, cy18r6), but may be switched on by blacklist changes. 1D–Var thicknesses were assimilated in 3D–Var north of $70°$ N in layers from 1000 hPa to 100 hPa and NESDIS thicknesses were assimilated above 100 hPa globally from January 1996 to August 1997. Lower resolution 500 km SATEM thicknesses are also available as a backup.

6.3.1 Layers

The extended odb (prior to screening) contains the reported layers of SATEM thickness and PWC. The specifica-

tion of the layers that will be used by 3D/4D-Var is given in SURAD. The reported layers may need splitting (THINUP) or summing up (SUMUP). This is done in the routine THICKPWC, called from SATEM, in the screening configuration of IFS. The compressed ODB after screening contains only those layers to be used by the analysis. The splitting and summing up of thickness layer is not desired unless the thickness data are actually going to be used actively. As the data are only used for monitoring, operationally, it is preferable to calculate departures for the reported layers. This is controlled by the switch LSUMUPTOVS=.F. (default since L50) in NAMSCC.

6.3.2 Observation errors

The observation errors are given in SURAD and assigned in THICKPWC. Observation errors are otherwise normally assigned in obsproc. Thicknesses and PWC are an exception because the layers to be used by the analysis are not known by the obsproc program. (TOVS radiances observation errors are another exception – they are assigned by 1D-Var, see Chapter 10 'Observation screening' for more details). The thickness errors (but not PWC) include a persistence error which is calculated using a routine of obsproc (PEREREV). There are no realistic thickness observation errors available for the 'un-summed up' reported layers (see LSUMUPTOVS=.F. above). For these data an ad-hoc thickness error corresponding to 1 K layer-mean temperature is inserted in ODB and used in the (diagnostic!) costfunction computation. Those data cannot be used actively unless realistic observation errors are set.

The PWC errors are given by

$$\sigma_{\mathrm{PWC}} = \sqrt{(\alpha_{\mathrm{rel}}\mathrm{PWC}_{\mathrm{sat}})^2 + \sigma_{\mathrm{trunc}}^2} \tag{6.1}$$

where $\mathrm{PWC}_{\mathrm{sat}}$ is the saturation PWC for the temperature profile of the background, σ_{trunc} is the truncation of the original BUFR report ($= 0.5\ \mathrm{kg\ m^{-2}}$), and α_{rel} is the relative accuracy of the PWC observation, set to 0.15 (kept in YOMTVRAD). $\mathrm{PWC}_{\mathrm{sat}}$ is calculated using Eq. (5.14) in a call to PPOBSA from SATEM with q replaced by the saturation specific humidity q_{sat} from FOQS:

$$q_{\mathrm{sat}} = \frac{\frac{R_{\mathrm{dry}}}{R_{\mathrm{vap}}}\frac{e_{\mathrm{sat}}(T)}{p}}{1-\left(\frac{R_{\mathrm{dry}}}{R_{\mathrm{vap}}}-1\right)\frac{R_{\mathrm{dry}}}{R_{\mathrm{vap}}}\frac{e_{\mathrm{sat}}(T)}{p}} \tag{6.2}$$

with e_{sat}, the saturation vapour pressure, computed by Eq. (5.11).

6.4 Observation operators

The computation of radiances is initiated and controlled by the HOP routine. Thicknesses, PWC and TPW are also computed in HOP and SCAT data too are processed in HOP. The general structure of HOP has been detailed in Subsection 2.5.2.

6.4.1 Radiances

The routine HOP interpolates the model profiles of temperature, humidity and ozone (T, q and oz) to the 43 RT levels ($\dot{T}$ and $\dot{q}$) and calls the interface RADTR to the RT code RTTOV. The standard routines PPT (Section 5.6 of Chapter 5 'Conventional observational constraints') and PPQ (Section 5.5) are used to carry out the vertical interpolation, and they are called through the PPOBSA interface, as usual. Various radiance preparations have been gathered in the new routine HRADP. In HRADP The model's pressure at the surface height of the observation lo-

cation (given in the report) is calculated, using PPPMER. For the purpose of radiance calculations $T_{2\mathrm{m}} = T_l$ and $q_{2\mathrm{m}} = \dot{q}_{40}$. These quantities represent a very shallow layer of air near the surface and contribute little to the calculated radiances—it was not considered necessary to use PPT2M and PPRH2M (Section 5.7) in this context. In order to make the radiance cost function continuous in p_{surf} it was necessary to ensure that $\dot{T}$ and $\dot{q}$ approach $T_{2\mathrm{m}}$ and $q_{2\mathrm{m}}$ as the pressure on any of the RT levels approches p_{surf}. This is done in a section of HRADP. More details on the radiative transfer code RTTOV can be found in *Eyre* (1991), updated by *Saunders* and Matricardi (1998), (available on-line ps–file).

Some of the radiance channels are highly sensitive to the surface skin temperature, which is also not part of the variational control variable when RTOVS data are used. It was found that the best results were obtained by replacing the model's T_{surf} with those retrieved by 1D-Var. The 1D-Var retrieval is carried out in a call to ADVAR from HRETR, called from TASKOB in the screening configuration only.

In the case of 1C, or 'raw' radiance data, as used since May 1999 (McNally *et al.* 1999) 1D-Var is no longer required. The radiance processing in HOP is similar for both 1C and RTOVS radiances, with the exception that surface skin temperature is retrieved by 4D-Var at each 1C-field of view, if the switch LTOVSCV is on (default is on).

In HOP the observation array is searched for radiance data. The compressed ODB (after screening) contains only those data to be used by the analysis. A list of existing channel numbers for each report is constructed. Model radiances for exactly those channels are then requested from the RT–code, via the interface RADTR. The routine RADTR checks that the input model profile is within the valid range of the transmittance regression. It packets the profiles into chunks of work of the appropriate maximum size for the RT–code (currently 65). The RT packet size has been communicated to IFS in the call to RTSETUP. The output is radiances for the channels requested.

The tangent linear HOPTL and the adjoint HOPAD follow the same pattern as HOP. In both the TL and the adjoint $\dot{T}$ and $\dot{q}$ have to be recomputed before the actual tangent linear and adjoint computations can start. The pointers to the radiance data in observation array are obtained just as it was done in the direct code. The input gradient to the adjoint is obtained as explained in Subsection 2.5.2.

6.4.2 Thicknesses

The pressures of layer bounds (top T, and bottom B) are found (in HOP) by scanning the observation array for thickness data. The geopotential for the top and the bottom of the layer are computed, using PPGEOP (Section 5.3), and the thickness is given by the difference $\phi_{\mathrm{T}} - \phi_{\mathrm{B}}$.

6.4.3 Precipitable water from SATEM and SSM/I

As for thicknesses, the pressures of layer bounds are found by scanning the observation array for TOVS PWC data. For SSMI TPW, the top pressure is set to the top of the model and the lower pressure bound is p_{s}. The PWC for the top and the bottom of the layer are computed, using PPPWC (Section 5.5), and the layer PWC is given by the difference $\mathrm{PWC}_{\mathrm{B}} - \mathrm{PWC}_{\mathrm{T}}$.

6.4.4 Scatterometer winds

In HOP, the observation array is scanned for SCAT data. Normally two ambiguous pairs of u –component and v -component observations are found at each SCAT location—with directions approximately 180 degrees apart. In 3D/4D-Var both winds are used and the ambiguity removal takes place implicitly through the special SCAT cost-function, Eq. (2.8), in HJO (*Stoffelen and Anderson*, 1997 ; *Gaffard et al.* 1997). If however LQSCATT=.true. (namjo), the normal quadratic J_{o} will b e used. In this case only the SCAT wind nearest the high resolution background will be used (which is determined in a section of HOP).

As PPUV10M (Section 5.7) is used also for SCAT data (since cy18r6), the observation operator is exactly the same as for SYNOP. SHIP and DRIBU winds. The z_0 (surface roughness) comes from the coupled wave model. The simpler logarithmic wind law can be used optionally under the switch LSCASUR=.F. in NAMOBS (true by default).

In the adjoint (SURFACAD) there is a separate section of HOP for the calculation of the $\nabla_{obs} J_{SCAT}$.

Part II: DATA ASSIMILATION

CHAPTER 7 Background, analysis and forecast errors

Table of contents

7.1 Nomenclature

7.2 Input and 'massaging' of background errors

7.3 Diagnosis of background error variances

7.4 Calculation of eigenvalues and eigenvectors of the Hessian

7.5 The Preconditioner

7.6 Calculation of analysis-error variances

7.7 Calculation of forecast error variances

7.1 NOMENCLATURE

The calculation of standard deviations of background errors is unfortunately an area where the use of inaccurate nomenclature is widespread. For example, standard deviations of background error are almost universally referred to as 'background errors'. Likewise, standard deviations of analysis and forecast error are referred to as 'analysis errors' and 'forecast errors'. Although inaccurate, this nomenclature has been adopted in the following for the sake of brevity.

A second source of confusion is that terms 'background error' and 'forecast error' are often used interchangeably. This confusion has even crept into the code, where the buffer which contains the standard deviations of background error is called FCEBUF. Such confusion is clearly unwise when discussing the calculation of forecast errors. The following sections will describe the processing of error variances during a single analysis cycle. The term 'background error' will refer exclusively to the standard deviations of background error which are used in the background cost function. The background errors are an input to the analysis. The term 'forecast error' will refer to an estimate of the standard deviation of error in a short-term forecast made from the current analysis. The forecast errors are calculated by inflating an estimate of the standard deviation of analysis error, and are an output from the analysis system.

7.2 INPUT AND 'MASSAGING' OF BACKGROUND ERRORS

Background errors for use in J_b are initialised by a call to SUINFCE. This is part of the J_b set-up descibed in Subsection 4.3.3. First, a call to INQGRIB is made. This returns a description of the data in the background error file (filename **errgrib**). COMMFCE1 communicates the description of the data to other processors. After checking some parameters and allocating arrays to hold the background errors, a call to READGRIB reads the errors into a local array. The errors are communicated to the other processors by a call to COMMFCE2. Optionally (under the control of LFACHR) the errors may be increased in the tropics at this stage. (This is not done by default, and is not recommended.) The background errors may be on a regular latitude–longitude, or reduced Gaussian grid. They are interpolated bilinearly in latitude and longitude onto the reduced Gaussian analysis grid by a call to SUHIFCE.

At this stage, all processors have complete fields of background error. Each processor now allocates a buffer (confusingly called FCEBUF) in yomfceb to hold background errors for those gridpoints which are local to the processor.

A large loop over variables follows. For each variable, the GRIB parameter code is examined. Depending on the setting of LSTABAL, LRDQERR, and on the presence or absence of vorticity errors in the background error file, the variable may be ignored (by cycling VARIABLE_LOOP) or an offset, IOFF, into the background error buffer is calculated.

The background errors are interpolated onto the model levels by a call to SUVIFCE. A number of variable-dependent things now happen. First, geopotential height errors are converted to geopotential by multiplying by g . Second, wind component errors are converted to vorticity errors by an *ad hoc* scaling. (Note that if vorticity errors are available in the file, then these will be used by preference. Wind component errors will be ignored.) Finally, if errors for the unbalanced components of temperature, divergence, ozone or surface pressure are not present in the file, the corresponding elements of the background error buffer are initialized to sensible values.

Background errors for specific humidity are read from the background-error file if the namelist variable LRDQERR is set. Currently, it is usual to calculate specific humidity errors as a function of background humidity and various other parameters. This is done by a call to STEPO('0IB00Z000'), which in turn calls SUSHFCE. The calculation of background errors for specific humidity is described in Subsection 4.3.4.

Next, one of two routines is called. SUMDFCE calculates a vertically average 'pattern' of background error. This is required if the background errors are to be represented as a product of a vertical profile of global mean error and a horizontal pattern, and was the default with the 'old' J_b . The pattern is stored in FGMWNE. Note in particular that SUMDFCE is called if horizontally-constant background errors are requested by setting LCFCE. In this case, all elements of FGMWNE are set to one.

Alternatively, SUPRFFCE is called to calculate global mean profiles of the input background errors. This is the default. The profiles are stored in FCEIMN.

The final step in processing the background errors is to call STEPO('00000Y000'). This, in turn, calls SUSEPFCE to modify the background errors. The modification takes one of two forms. If separable background errors have been requested, the contents of the background error buffer are replaced by the product of the vertical profile stored in FCEMN and the horizontal pattern stored in FGMWNE. Otherwise, the background errors for each variable at each level are multiplied by the ratio of the corresponding elements of FCEMN and FCEIMN. The result of this operation is to adjust the global mean profiles of background error to match those stored in FCEMN.

7.3 DIAGNOSIS OF BACKGROUND ERROR VARIANCES

The analysis errors are calculated by subtracting a correction from the variances of background error. The first stage in the calculation is therefore to determine the background error variances. This is done by subroutine BGVECS, which is called from CVA1. One of two methods may be employed, depending on whether NBGVECS is equal to, or greater than, zero. In either case, the estimated variances of background error are stored in the analysis error buffer, ANEBUF (in yomaneb).

If NBGVECS is zero, as it is by default, then background errors for variables which are explicitly present in the background error buffer, FCEBUF, are copied into ANEBUF and squared. Errors for those variables whose background errors are defined implicitly through the change of variable are estimated using simple scaling of appropriate explicit errors. This scaling is performed by a call to ESTSIGA.

If NBGVECS is non-zero, then the variances of background error are estimated using randomization. This method

assumes that the change of variable transforms the background error covariance matrix into the identity matrix. A sample of NBGVECS vectors drawn from a multi-dimensional Gaussian distribution with zero mean and identity covariance matrix is generated by calls to the Gaussian random number generator GASDEV. These vectors are transformed to the space of physical variables by CHAVARIN. The transformed variables form a sample drawn from the distribution of background errors. A call to STEPO('0AA00A000') transforms each vector to gridpoint space and accumulates the sums of squares in ANEBUF. Finally, the sums of squares are divided by the number of vectors by a call to SCALEAE to provide a noisy, but unbiassed estimate of the variances of background error actually used in the analysis. The noise may be filtered by a cal to FLTBGERR, which transforms the variances to spectral coefficients, multiplies each coefficient by $\cos^2(\min((n/\mathrm{NBGTRUNC}),1)\pi/2)$, and then transforms to grid space. The default is to filter with a very large value of NBGTRUNC. Effectively, the background errors are simply spectrally truncated. It is highly recommended that the filtering is performed, since it prevents a grid-scale numerical instability which occurs when the error growth model introduces spatial features which cannot be resolved by the spectral control variable.

The code allows two further configurations of the background error estimation. Neither is operational at present. The two configurations are controlled by switches LBGOBS and LBGM (namvar), respectively. If LBGOBS=.T. then the full set of tangent-linear observation operators will be applied to the NBGVECS random vectors, in model grid point space. This is done in the routine BGOBS called from VEC2GP, under SCAN2MTL. The TL routines are required as the observation operators have been linearized around the background state. The result is **background errors in observation space**. They are stored and accumulated in ANEBUF and written out as grib-fields, for geopotential, temperature, wind, humidity, total ozone, total column water vapour, TOVS and ATOVS radiance channels, 10 metre wind and 2 metre temperature. If in addition LBGM=.T. then the randomized estimate of background error will be propagated in time, using the adiabatic tangent linear model, i.e. a call to CNT3TL from BGVECS. The eigenvectors of the analysis Hessian (next section) are also propagated similarly in time, by a call to CNT3TL from XFORMEV, to obtain **flow dependent background errors**. The number of model integrations required by LBGM is NBGVECS+invtot, which is typically 50+100=150. If LBGM=.T. then the simplified error growth model (Section 7.7) is not used. In that case the routine ESTSIG is used only to limit the error growth produced by the model to within 10 and 90 % of the climate variance for vorticity.

The background errors diagnosed by BGVECS may be written out for diagnostic purposes by setting LWRISIGB. The errors are written by a call to WRITESD (called from CVA1)..

7.4 Calculation of eigenvalues and eigenvectors of the Hessian

The second stage in the calculation of analysis errors is to determine eigenvalues and eigenvectors of the Hessian of the cost function. This is done using a combined Lanczos and conjugate-gradient algorithm, CONGRAD, called from CVA1 under the control of LAVCGL. Note that CONGRAD requires that the cost function is strictly quadratic. The tangent linear model and observation operators must be invoked by setting L131TL and LOBSTL. (L131TL should be set even in 3D-Var.) Variational quality control must be disabled by unsetting LVARQCG and LQSCATT must be set to request a quadratic cost function for scatterometer observations.

CONGRAD starts by transforming the initial control variable and gradient to a space with euclidian inner product. Typically, this transformation is simply a multiplication by SCALPSQRT, but may also involve preconditioning via calls to PRECOND. The transformed initial gradient is normalized to give the first Lanczos vector. Depending on the setting of LIOWKCGL, the Lanczos vectors are stored either on the MIO file associated with unit NWKCGL, or in the allocated array VCGLWK.

Each iteration of the conjugate-gradient/Lanczos algorithm starts by calculating the product of the Hessian and the latest search direction. This is calculated as $J''\mathrm{d} = \|\mathrm{d}\|(\nabla J(\mathrm{x}_0 + \mathrm{d}/\|\mathrm{d}\|) - \nabla J(\mathrm{x}_0))$. This finite difference formu-

la is exact, since the cost function is quadratic.

The optimal step is calculated as the point at which the gradient is orthogonal to the search direction. The control variable and gradient at the optimal point are also calculated. Once the gradient at the optimal point is known, it is orthogonalized with respect to previous gradients, and the search direction and gradient for the next iteration are calculated. The tridiagonal matrix of the Lanczos algorithm is initialized and its eigenvalues and eigenvectors are determined by a call to the NAG routine F08JEF.

The leading eigenvalue of the tridiagonal system is compared against the leading converged eigenvalue of the Hessian matrix. This provides a sensitive test that the algorithm is behaving correctly. Any increase in the leading eigenvalue provides an early indication of failure (for example, due to a bad gradient) and the algorithm is immediately terminated. The calculation is not aborted, since the test detects the failure of the algorithm before the converged eigenvalues and eigenvectors become corrupted.

The new Lanczos vector is calculated by normalizing the gradient and the subroutine loops back to perform the next iteration.

After the last iteration, the converged eigenvectors of the Hessian are calculated by calling WREVECS. Note that the criterion used to decide which eigenvalues have converged is relaxed at this stage to $\|J''\mathbf{v} - \lambda\mathbf{v}\| < \varepsilon\|\mathbf{v}\|$, where ε is given by EVBCGL. The default vaule for EVBCGL is 0.1 .The eigenvectors are passed to XFORMEV, which calculates the analysis errors. This part of the calculation is described in Section 7.6.

Finally, CONGRAD transforms the control vector and gradient from the euclidian space used internally to the usual space of the control variable.

7.5 THE PRECONDITIONER

CONGRAD allows the use of a preconditioner. The preconditioner is a matrix which approximates the Hessian matrix of the cost function. The preconditioner used in CONGRAD is a matrix of the form

$$\mathbf{I} + \sum_{i-1}^{L} (\mu_i - 1)\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} \tag{7.1}$$

where the vectors $\mathbf{w}_i$ are orthogonal. The pairs $\{\mu_i,\mathbf{w}_i\}$ are calculated in PREPPCM, and are intended to approximate some of the eigenpairs (i.e. eigenvalues and associated eigenvectors) of the Hessian matrix of the cost function. They are calculated as follows.

A set of L vectors, $\mathbf{u}_i$, is read in using READVEC. These vectors are assumed to satisfy

$$\mathbf{B} - \sum_{i=1}^{L} \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} \approx \mathbf{P}_{\mathrm{a}} \tag{7.2}$$

where $\mathbf{B}$ is the background-error covariance matrix, and $\mathbf{P}_{\mathrm{a}}$ is the analysis-error covariance matrix. Vectors which meet this criterion can be written out from an earlier forecast error calculation by setting LWRIEVEC. The vectors are transformed to the space of the control vector by calls to CHAVAR to give an approximation to the inverse of the Hessian matrix

$$\mathbf{I} - \sum_{i=1}^{L} (\mathbf{L}\mathbf{u}_i)(\mathbf{L}\mathbf{u}_i)^{\mathrm{T}} \approx (J'')^{-1} \tag{7.3}$$

(Here, $\mathbf{L}$ denotes the change-of-variable operator implemented by CHAVAR.)

Let us denote by $\mathbf{U}$ the matrix whose columns are the vectors $\mathbf{u}_i$. A sequence of Householder transformations is now performed to transform $\mathbf{LU}$ to upper triangular. Let us represent this sequence of Householder transformations by the matrix $\mathbf{Q}$. Then $\mathbf{QLU}$ is upper triangular, which means that $(\mathbf{QLU})(\mathbf{QLU})^{\mathrm{T}}$ is zero except for an $L \times L$ block in the top left hand corner.

It is clear that $(\mathbf{QLU})(\mathbf{QLU})^{\mathrm{T}}$ has only L non-zero eigenvalues. Moreover, the non-zero eigenvalues are the eigenvalues of the $L \times L$ block matrix, and the eigenvectors of $(\mathbf{QLU})(\mathbf{QLU})^{\mathrm{T}}$ are the eigenvectors of the block matrix, appended by zeroes. These eigenvalues and eigenvectors are calculated by a call to the NAG routine F02FAF.

Now, since $\mathbf{Q}$ is an orthogonal matrix, we have $\mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathrm{I}$. So, we may write Eq. (7.3) as

$$\mathbf{I} - \mathbf{Q}^{\mathrm{T}}(\mathbf{QLU})(\mathbf{QLU})^{\mathrm{T}}\mathbf{Q} \approx (J'')^{-1} \tag{7.4}$$

Let us denote the eigenpairs of $(\mathbf{QLU})(\mathbf{QLU})^{\mathrm{T}}$ by $\{\rho_i, \mathbf{v}_i\}$. Then we may write Eq. (7.4) as

$$\mathbf{I} - \sum_{i=1}^{L} \rho_i (\mathbf{Q}^{\mathrm{T}}\mathbf{v}_i)(\mathbf{Q}^{\mathrm{T}}\mathbf{v}_i)^{\mathrm{T}} \approx (J'')^{-1} \tag{7.5}$$

The orthogonality of $\mathbf{Q}$ and the orthonormality of the eigenvectors $\mathbf{v}_i$, means that the vectors $\mathbf{Q}^{\mathrm{T}}\mathbf{v}_i$ are orthonormal. They are, in fact, the required vectors, $\mathbf{w}_i$ of the preconditioner matrix.

Inverting Eq. (7.5) gives

$$\mathbf{I} - \sum_{i=1}^{L} \frac{1}{\rho_i} \mathbf{w}_i \mathbf{w}_i^{\mathrm{T}} \approx J'' \tag{7.6}$$

Defining $\mu_i = 1 - 1/\rho_i$ gives the required approximation to the Hessian matrix.

The preconditioner vectors are sorted in decreasing order of μ_i, and all vectors for which $\mu_i < 1$ are rejected. These vectors cannot be good approximations to eigenvectors of the Hessian matrix, since the eigenvalues of the Hessian matrix are all greater than or equal to one. A final refinement to the calculation is to reduce large values of μ_i to a maximum of 10. This was found to be necessary in practice to avoid ill-conditioning the minimization.

The numbers μ_i are stored in RCGLPC. The vectors, $\mathbf{w}_i$ are stored in VCGLPC.

Application of the preconditioner is straightforward, and is performed by subroutine PRECOND. This routine can also apply the inverse, the symmetric square root, or the inverse of the symmetric square root of the preconditioner matrix. Application of the latter matrices relies on the observation that if

$$\mathbf{M} = \mathbf{I} + \sum_{i-1}^{L} (\mu_i - 1)\mathbf{w}_i \mathrm{w}_i^{\mathrm{T}} \tag{7.7}$$

with orthonormal $\mathbf{w}_i$, then the expressions for $\mathbf{M}^{-1}$, $\mathbf{M}^{1/2}$ and $\mathbf{M}^{-1/2}$ result from replacing μ_i in Eq. (7.7) by $1/\mu_i$, $\sqrt{\mu_i}$ and $1/(\sqrt{\mu_i})$ respectively.

7.6 CALCULATION OF ANALYSIS-ERROR VARIANCES

The eigenvectors and eigenvalues of the Hessian matrix calculated by CONGRAD are passed to XFORMEV, which uses them to estimate the analysis error variances. If preconditioning has been employed, then the eigenvectors and eigenvalues provide an approximation to the preconditioned Hessian, $\mathbf{M}^{-1/2} J'' \mathrm{M}^{-1/2}$, of the form

$$\mathbf{M}^{-1/2} J'' \mathbf{M}^{-1/2} \approx \mathbf{I} + \sum_{i=1}^{K} (\lambda_i - 1)\mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} \tag{7.8}$$

The approximation is equivalent to setting to one all but the leading K eigenvalues of the preconditioned Hessian.

The first step is to undo the preconditioning. Multiplying to the left and right by $\mathbf{M}^{1/2}$, gives

$$J'' \approx \mathbf{M} + \sum_{i=1}^{K} (\lambda_i - 1)(\mathbf{M}^{1/2}\mathbf{v}_i)(\mathbf{M}^{1/2}\mathbf{v}_i)^{\mathrm{T}} \tag{7.9}$$

Substituting for the preconditioner matrix from Eq. (7.7), gives the following

$$J'' \approx \mathbf{I} + \sum_{i=1}^{L+K} \mathbf{s}_i \mathbf{s}_i^{\mathrm{T}} \tag{7.10}$$

where

$$\mathbf{s}_i = \begin{cases} (\mu_i - 1)^{1/2}\mathbf{w}_i & \text{for } i = 1 \ldots L \\ (\lambda_{i-L} - 1)^{1/2}\mathbf{M}^{1/2}\mathbf{v}_{i-L} & \text{for } i = L+1 \ldots L+K \end{cases} \tag{7.11}$$

Operationally, preconditioning is not used. However XFORMEV makes no particular use of this fact. It simply sets L to zero in Eqs. (7.10) and (7.11).

The first step in XFORMEV is to calculate the vectors $\mathbf{s}_i$. They are stored in VCGLWK.

The next step is to invert the approximate Hessian matrix defined by Eq. (7.10). Let $\mathbf{S}$ be the matrix whose columns are the vectors $\mathbf{s}_i$. Then, applying the Shermann–Morrison–Woodbury formula, the inverse of the approximate Hessian matrix is

$$(J'')^{-1} \approx \mathbf{I} - \mathbf{S}(\mathbf{I} + \mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}} \tag{7.12}$$

The matrix $(\mathbf{I} + \mathbf{S}^{\mathrm{T}}\mathbf{S})$ is formed and its Cholesky decomposition is calculated using the NAG routine F07FDF. This gives a lower triangular matrix $\mathbf{C}$ such that

$$(J'')^{-1} \approx \mathbf{I} - (\mathbf{S}\mathbf{C}^{-1})(\mathbf{S}\mathbf{C}^{-1})^{\mathrm{T}} \tag{7.13}$$

The matrix $(\mathbf{S}\mathbf{C}^{-1})$ is calculated by back-substitution.

The final stage in the calculation of the analysis errors is to transform the columns of the matrix $(\mathbf{S}\mathbf{C}^{-1})$ to the space of model variables by appplying the inverse change of variable, CHAVARIN. This gives the required approximation to the analysis error covariance matrix

$$\mathbf{P}_{\mathrm{a}} \approx \mathbf{B} - \mathbf{V}\mathbf{V}^{T} \tag{7.14}$$

where $\mathbf{V} = \mathbf{L}^{-1}\mathbf{S}\mathbf{C}^{-1}$, and where $\mathbf{L}^{-1}$ represents the inverse of the change of variable. The columns of $\mathbf{V}$ may be written out (e.g. for diagnostic purposes, or to form the preconditioner for a subsequent minimization) by setting LWRIEVEC. The columns of $\mathbf{V}$ are then transformed to gridpoint space, and their sums of squares (i.e. the diagonal elements of $\mathbf{V}\mathbf{V}^{\mathrm{T}}$ in gridpoint space) are subtracted from the variances of background error which were stored in ANEBUF before the minimization by BGVECS.

The analysis errors are calculated as the difference between the background errors and a correction derived from the eigenvectors of the Hessian. If the background errors are underestimated, there is a danger that the correction will be larger than the background error, giving negative variances of analysis error. This is unlikely to happen if the background errors are estimated using randomization, or for variables whose background errors are explicitly specified in the background cost function, but is possible for variables such as temperature whose background errors are not explicitly specified. To guard against this eventuality, if NBGVECS is zero, then the variances of analysis error for variables whose background errors are not explicit are estimated by applying a scaling to the explicit variables by a call to ESTSIGA from CVA1. The variances are then converted to standard deviations and written out by a call to WRITESD.

7.7 CALCULATION OF FORECAST ERROR VARIANCES

The analysis errors are inflated according to the error growth model of *Savijärvi* (1995) to provide estimates of short term forecast error. This is done by a call to ESTSIG. There is also an option to advect the background errors for vorticity as if they were a passive tracer. The advection is performed by ADVSIGA.

The error growth model is

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t} = (a + b\sigma)\left(1 - \frac{\sigma}{\sigma_{\infty}}\right) \tag{7.15}$$

Here, a represents growth due to model errors, b represents the exponential growth rate of small errors, and σ_{∞} represents the standard deviation of saturated forecast errors.

The saturation standard deviations are calculated as $\sqrt{2}$ times the standard deviation of each field. The standard deviations have been calculated for each month from the re-analysis dataset. ESTSIG reads these climatological error fields from file 'stdev_of_climate' by calling READGRIB, and interpolates them in the horizontal and vertical

using SUHIFCE and SUVIFCE. The climatological errors may also be artificially increased in the tropics under the control of LFACHRO. This is the default, and is recommended in preference to using LFACHR, since it means that the forecast errors that are archived, and are used to screen observations, are closer to those used to formulate the background cost function. If climate standard deviations are not available for any field, they are estimated as 10 times the global mean background error for the field.

The growth due to model error is set to 0.1 times the global mean background error per day. The exponential growth rate, b , is set to 0.4 per day.

The error growth model is integrated for a period of NFGFCLEN hours. The integration is done analytically using the expression given by *Savijärvi* (1995). Two precautions are taken in integrating the error growth model. First, negative analysis error variances are set to zero. Second, the growth rate due to model error is limited to a sensible value with respect to the saturation errors. This was found to be necessary to prevent numerical problems when calculating specific humidity errors for the upper levels of the model.

ESTSIG overwrites the contents of ANEBUF with the estimated variances of forecast error. The variances are converted to standard deviations and written out by WRITESD.

Part II: DATA ASSIMILATION

CHAPTER 8 Gravity-wave control

Table of contents

8.1 Introduction

8.2 Normal-mode initialization

8.3 Computation of normal modes

8.3.1 Vertical modes

8.3.2 Horizontal modes and help arrays

8.4 Implementation of NMI

8.5 Computation of

8.6 Digital filter initialization

8.7 Implementation of DFI as a weak constraint in 4D-Var

8.1 INTRODUCTION

In 3D-Var, gravity-wave control is achieved via the techniques of normal-mode initialization (NMI), in 4D-Var a weak constraint digital filter is used. The construction of a high-resolution analysis by combining a high-resolution background with increments defined at lower resolution makes direct use of an incremental form of nonlinear NMI, as in Chapter 1 'Incremental formulation of 3D/4D variational assimilation—an overview' for 3D-Var and Chapter 3 '4D variational assimilation' Eq. (3.3) for 4D-Var. There is an initialization step in creating the low-resolution background field, see Section 2.2. Computation of the penalty term J_c (see Eq. (1.1)) is based on NMI methods for 3D-Var, on DFI for 4D-Var.

Section 8.2 provides a brief overview of NMI techniques, together with references to scientific papers in which further details can be found. Section 8.3 describes the computation of normal modes and related arrays. Section 8.4 documents the implementation of nonlinear NMI in 3D- and 4D-Var, while Section 8.5 describes the computation of J_c. Section 8.6 gives an overview of digital filter initialization techniques while Section 8.7 describes its implementation.

8.2 NORMAL-MODE INITIALIZATION

If the model equations are linearized about a state of rest, the solutions can (with a certain amount of arbitrariness) be classified into 'slow' (Rossby) and 'fast' (gravity) modes. This classification defines two mutually orthogonal subspaces of the finite-dimensional vector space containing the model state $\mathbf{x}$. Thus, the model state can be written as

$$\mathbf{x} = \mathbf{x}_R + \mathbf{x}_G \tag{8.1}$$

where $\mathbf{x}_R$ is the 'slow' component and $\mathbf{x}_G$ the 'fast' component. *Linear* NMI consists of removing the fast com-

ponent altogether ($\mathbf{x}_G = 0$). Since the model is nonlinear, a much better balance is obtained by setting the *tendency* of the fast component to zero ($\dot{\mathbf{x}}_G = 0$); it is this balance condition which *nonlinear* NMI seeks to impose.

Nonlinear NMI was first demonstrated by *Machenhauer* (1977), in the context of a spectral shallow-water model. For a multi-level model, the first stage in the modal decomposition is a vertical transform; each vertical mode then has its own set of horizontal slow and fast modes (for the shallower vertical modes, all the corresponding horizontal modes can be considered as 'slow'). In the case of a multi-level spectral model using the ECMWF hybrid vertical coordinate the details may be found in the report by *Wergen* (1987), which also describes techniques for taking into account forcing by physical (non-adiabatic) processes and the diurnal and semi-diurnal tidal signals. Although these options are still coded in the IFS, they are no longer used operationally at ECMWF and will not be described in this documentation.

Implicit normal mode initialization (*Temperton* 1988) is based on the observation that, except at the largest horizontal scales, the results of NMI can be reproduced almost exactly without computing the horizontal normal modes at all. The calculation reduces to solving sets of elliptic equations. In the case of a spectral model (*Temperton* 1989), these sets of equations are tridiagonal in spectral space. The IFS code includes the option of 'partially implicit NMI', in which the initialization increments are computed using the full 'explicit' NMI procedure for large horizontal scales while the remaining increments at smaller horizontal scales are computed using the simpler implicit procedure.

8.3 COMPUTATION OF NORMAL MODES

8.3.1 Vertical modes

The vertical normal modes depend on the number of levels in the model and on their vertical distribution. They also depend on the choice of reference temperature SITR (assumed isothermal) and reference surface pressure (SIPR). The vertical modes used by the initialization routines are also used in the semi-implicit scheme for the forward integration of the model. The computation of J_b and J_c also uses the vertical normal modes, but for these purposes different values of SITR and SIPR may be selected. Thus the vertical modes are computed both in SUDYN and SUSINMI, the latter being used especially in 4D-Var where it is necessary to alternate between applications using different choices of SITR and SIPR. The vertical modes are computed by first calling SUBMAT to set up a vertical structure matrix and then calling an eigenvalue/eigenvector routine EIGSOL (at the end of SUDYN, it calls routine RG in the auxiliary library). After reordering and normalization, the eigenvectors (vertical modes) are stored in the matrix SIMO, while the corresponding eigenvalues (equivalent depths) are stored in the array SIVP. The inverse of SIMO is computed and stored in SIMI.

8.3.2 Horizontal modes and help arrays

The horizontal normal modes depend on the equivalent depths (see above) and the chosen spectral truncation NXMAX. For 'explicit' NMI, NXMAX is equal to the model's spectral truncation NSMAX. Normally, 'partially implicit NMI' is chosen by setting the switch LRPIMP to .TRUE. In this case the explicit NMI increments are used only up to spectral truncation NLEX (21 by default) but in order to blend the explicit and implicit increments smoothly, explicit increments are computed up to a slightly higher resolution. By default, $\mathrm{NXMAX} = \mathrm{NLEX} + 5$.

For most applications of the NMI procedure in the operational suite, it is considered that the larger horizontal scales are best left uninitialized (they include, for example, atmospheric tidal signals and large-scale tropical circulations driven by diabatic processes). To cater for this option there is another logical switch, LASSI ('adiabatic small-scale

initialization'), which sets to zero all the initialization increments for total wavenumbers up to NFILTM (= 19 by default). Since only the small-scale increments are used, the NMI can be completely implicit: NLEX is set to 0 and there is no need to calculate the 'explicit' horizontal normal modes.

All the horizontal-normal-mode computations are carried out only for the first NVMOD vertical modes. By default, NVMOD = 5.

The horizontal modes are computed by calling SUMODE3. In turn, SUMODE3E computes the explicit modes and their frequencies while SUMODE3I computes the 'help' arrays required to invert the tridiagonal systems encountered in implicit NMI.

8.4 IMPLEMENTATION OF NMI

Nonlinear NMI is invoked by calling NNMI3. Model tendencies are computed by calling STEPO to perform one (forward) timestep. The tendencies are then supplied to MO3DPRJ which computes the required increments, using the 'explicit' (Machenhauer) or the 'implicit' scheme (or both, after which the results are merged). The increments are added to the original spectral fields and the process is iterated NITNMI (by default 2) times.

8.5 COMPUTATION OF J_c

In the notation of Eq. (8.1), the penalty term J_c is defined by

$$J_c = \varepsilon \left\| (\dot{\mathbf{x}} - \dot{\mathbf{x}}_b)_G \right\|^2 \tag{8.2}$$

where ε is an empirically chosen weighting factor, $\mathbf{x}$ is the current state of the control variable and $\mathbf{x}_b$ is the background. The norm $\| \ \|^2$ is based on a weighted sum of squares of spectral coefficients. Only the first NVMOD vertical modes are included in the evaluation of (8.2).

J_c is computed by calling the routine COSJC. Control passes through JCCOMP to NMIJCTL, where J_c is evaluated by calling STEPO twice, then projecting the differences in the tendencies on to the gravity modes via MO3DPRJ, and finally computing J_c in NMICOST.

8.6 DIGITAL FILTER INITIALIZATION

Digital filter initialization consists in removing high frequency oscillations from the temporal signal represented by the meteorological fields. A general description of digital filter initialization can be found in Lynch (1993). It can be implemented as a strong constraint by filtering the model fields at the beginning of each forecast or as a weak constraint as described in Gauthier and Thepaut (2000).

Time oscillations exceeding a cut-off frequency $\omega_c = (2\pi)/T_c$ can be filtered by applying a digital filter to a time series $f_k = f(t_k)$ for $t_k = k\Delta t$, Δt being the timestep. This proceeds by doing a convolution of $f(t)$ with a step function $h(t)$ so that

$$f \bullet h(t_N) = \sum_{k=-\infty}^{\infty} h_k f_{N-k}$$

The step function h_k is found to be

$$h_k = \frac{\sin(\omega_c k\Delta t)}{k\pi}$$

In practice, the convolution is restricted to a finite time interval of time span T_s. We can write $T_s = 2M\Delta t$ and

$$f \bullet h(t_0) = \sum_{k=-M}^{M} \alpha_k f_k$$

with $\alpha_k = -h_{-k}$. This truncation introduces Gibbs oscillations which can be attenuated by introducing a Lanczos window which implies that the weights α_k are defined as $\alpha_k = -h_{-k}W_k$ with

$$W_k = \frac{\sin((k\pi)/(M+1))}{(k\pi)/(M+1)}$$

An alternative which is used at ECMWF has been proposed by Lynch to use a Dolph-Chebyshev window in which case

$$W_k = \frac{1}{2M+1}\left[1 + 2r\sum_{m=0}^{M} T_{2M}(x_0\cos\theta_m/2)\cos m\theta_k\right]$$

where $1/x_0 = \cos(\pi\Delta t)/\tau_s$, $1/r = \cosh(2M\,\mathrm{acosh}\,x_0)$, $\theta_k = (k2\pi)/M$ and T_{2M} is the Chebyshev polynomial of degree $2M$. The time span of the window is chosen so that $\tau_s = M\Delta t$.

8.7 IMPLEMENTATION OF DFI AS A WEAK CONSTRAINT IN 4D-VAR

In the context of variational data assimilation, the digital filter is used as a weak constraint. A penalty term is added to the cost function and replaces the NMI based penalty term.

During each integration of the tangent linear model in the inner loop of the 4D-Var, the digital filter is applied to the increments. This gives a filtered increment valid at the mid-point of the assimilation window (array RACCSPA). The value of the non-filtered increment valid at the same time is also stored (array RSTOSPA).

The weak constraint term which is added to the cost function is the moist energy norm of the departure between those two states times a weight factor. All these computations are conducted in spectral space and applied to the spectral fields. The norm of the departure is computed in two steps. In EVJCDFI, the difference between RACCSPA and RSTOSPA is computed and summed in array RSUMJCDFI for each wavenumber. Then, in EVCOST, the contributions from each wavenumbers and variables are added to obtain the final value of the penalty term.

Part II: DATA ASSIMILATION

CHAPTER 9 Data partitioning (OBSORT)

Table of contents

9.1 Introduction

9.2 Data flow with the analysis components

9.3 Observational data partitioning

9.4 Data partitioning scheme

9.5 The parallel data flow of the OBSORT

9.6 OBSORT calling tree

9.1 INTRODUCTION

The observational data partitioning scheme has been encapsulated into a separate module called OBSORT. The program OBSORT redistributes the observational data across the available processors. Data supported must be either in the CMA and/or BUFR formats. As a result the subsequent steps in the general analysis data flow will be well load-balanced with respect to observation handling and the total elapsed time for the analysis is reduced. The facilities offered by the OBSORT can be used as a stand-alone executable, or used via calling the LIB_OBSORT subroutine. The latter form is normally used, and is found in the OBSPROC (MAKECMA and FEEDBACK) and IFS/Screening. We have five different modes of the OBSORT:

- *Mode 0, submodule BUFRsort.* Partitioning and splitting of the BUFR data among the available processors (no geographical order, though).
- *Mode 1, submodule CMA+BUFRsort.* Geographical re-ordering of the CMA data in conjunction with the counterpart BUFR data; it is consequently called the CMA-data-driven BUFR sort.
- *Mode 2, submodule CMAsort.* Geographical re-ordering of the CMA data.; it also copes with the virtual-processor case where fewer processors than are required by the main analysis can produce more CMA files than the actual number of processors used by the OBSORT.
- *Mode 3, submodule MATCHUP*. Matching up and updating the ECMA data present in one geographical distribution with the CCMA data present in another distribution.
- *Mode 4, submodule VMATCHUP*. The same as MATCHUP, but for virtual processors. More than `NPROC` ECMA files are brought back to the `NPROC` files in the same order as the BUFR counterparts were left after the MAKECMA.

9.2 DATA FLOW WITH THE ANALYSIS COMPONENTS

This section describes OBSORT as a part of the analysis pre- and postprocessing (OBSPROC) and the main analysis; it illustrates why we need the OBSORT. The following discussion applies to the *single-processor* implementation only. In a later section, where the module OBSORT is introduced, the parallel aspects are covered in more detail.

Data-assimilation cycle (see Fig. 9.1) starts by retrieval of BUFR data. Currently there are four different BUFR

files involved: conventional observations, or GTS, and TOVS, SCAT and SSMI satellite observations. Files are prepared for a 6-hour data assimilation period both in the 3D- and 4D-VAR contexts. In the near future BUFR files for 12- and even 24-hour periods will be prepared for the 4D-VAR purposes. Storagewise BUFR files occupy from a few megabytes to tens of megabytes, depending on a chosen BUFR compression scheme.

In the first step MAKECMA picks up prepared BUFR files and decodes them. All, except formally erroneous observational data, are transformed to the CMA format and written to a so-called Extended CMA-file (ECMA). Furthermore, the unpacked BUFR data are also written into another BUFR-format-conforming file, where the high BUFR compression rate is relaxed. The output BUFR file maintains the same order of the observations as its counterpart CMA file, which in turn contains only about 80% of the information that is present in the BUFR file. The reduction in information content is the main reason for creating a new BUFR file. However it should be emphasized that the BUFR data, as such, are not used in the main analysis.

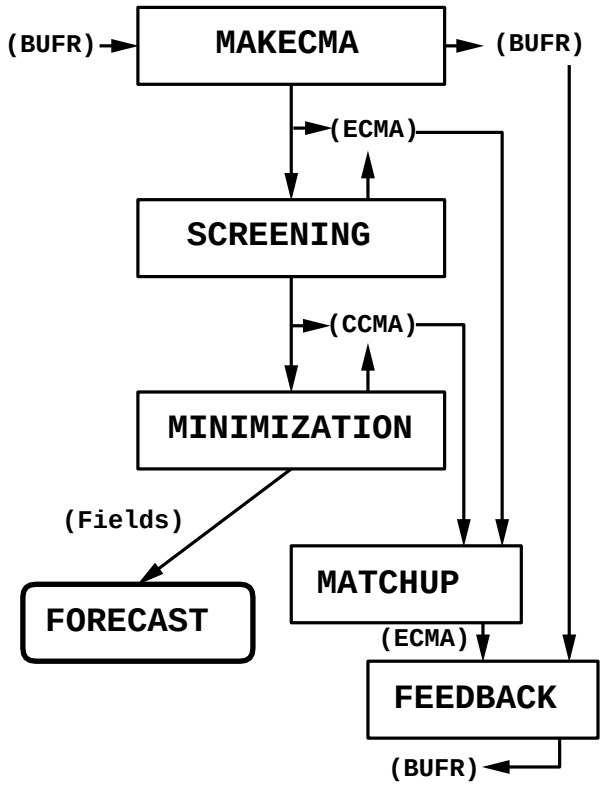

Figure 9.1 Dataflow in observation processing at the ECMWF on a single processor implementation.

In the next step, the ECMA file is passed to the process for screening the observations; this rejects undesirable observations, using information from the meteorological fields within the IFS. At the start-up time the current state of the atmosphere is known in terms of the observations, and in terms of the best *a priori* estimate of the atmosphere, together with the estimate of its uncertainty (or error). The quality of the observations is checked against a 6-hour forecast valid at the analysis time. The screening scheme picks up the best set of observations by rejecting erroneous, duplicated, blacklisted etc. observations, so that they are omitted from further processing.

All the observations passing this test are written to a so-called Compressed CMA-file (CCMA); this conforms to the CMA format, but occupies much less space—about one tenth of the original ECMA. The ECMA file is also updated, since there is a need for data monitoring, even of the rejected observations.

The observations for the actual data assimilation (the minimization) are passed through a CCMA file. The file is read into the memory in one large chunk to make subsequent operations on the data as fast as possible. Interpolation and extrapolation from the forecast-model grid-point space to the observation space can take place in a flexible manner, without the need to carry out complicated data conversions that were a feature of the older systems.

The minimization provides the initial-condition fields for the subsequent global forecast. It also updates the CCMA file with the observation departures from the analysis fields. At this point it is possible to start feeding information back to the archive, but in practice this step is postponed until the time-critical forecast-model run has been finished.

The subsequent observation-processing step to the minimization is called MATCHUP; this is part of the OBSORT library, that is described in Section 9.3. The purpose of MATCHUP is to read the latest CCMA and ECMA data and return the extra information found in the CCMA file back to the ECMA file. This enables the data, that existed only as a reduced set after the minimization, to be included within an ECMA set.

When the ECMA file is up to date, the last part of the process starts with FEEDBACK, the purpose of which is to encode the ECMA file back into a highly compressed BUFR format. It also separates data back into the groups of GTS, TOVS, SCAT and SSMI and retains the original input time periods (typically 6 hours). For this purpose the FEEDBACK needs not only the updated ECMA but also the (semi-)original BUFR file from the MAKECMA output. This is necessary because this BUFR file contains some information not present in the ECMA. After FEEDBACK has run the resulting BUFR files are ready for archiving.

9.3 OBSERVATIONAL DATA PARTITIONING

Due to the high data volume and the time-critical scheduling in operations, it is necessary to parallelize all observation-processing components. Most of the modules described in the section about the data flow rely on availability of large memory, since the CMA data are brought into core for efficient data-management reasons. This eliminates the need for slow random-access I/O operations when the contents of a particular observation message are needed.

The fact that there is always a limit for the maximum available memory per processor, forces us to look at parallel processing almost immediately. At the time of writing, a typical set of the CMA data for a 6-hour data assimilation period consumes a half gigabyte of disk space (or 100–200 MB if packed CMA format is used; see below). When brought into the memory the total CMA array size is often doubled due to the data the shuffling algorithm (see the section on the reshuffling), for which both incoming and outgoing CMA data are kept in memory. Also, some non-neglible in-core space is needed to hold certain global table information.

Disk-space consumption by the CMA files is greatly reduced by the introduction of various packed CMA formats. The implementation of the packing is such that when CMA data are written out they are packed as a part of the I/O process using novel vectorizable packing algorithms. The reverse naturally holds when reading the data. As a result, file-size reductions, from 50 to 85%, are not uncommon with neglible cost in the 'on-the-fly' packing/unpacking.

Although the screening module effectively reduces the size of the CMA data by an order of magnitude, it is that very module which requires maximum amount of in-core memory as well as disk-space time during the course of a data-assimilation cycle. Fortunately the screening is a part of the IFS and, thus, parallelization in grid-point space has been present there from day one. Some additions apart from the OBSORT are needed, though, to accommodate full handling of observational data by the IFS.

9.4 DATA PARTITIONING SCHEME

In the parallel implementation it might, at first sight, look feasible to split up observational data set evenly (and geographically randomly) among the available processors. However, there are considerable differencies in the distribution of observations over the globe, and thus certain regions might be more 'expensive to process' than others, because of the variable amount of floating-point operations per observation type when calculating contributions to the cost function in the minimization.

Furthermore, it would be convenient to have approximately the same geographical distribution of observations as in the grid-point space in order to reduce communication between processors during the interpolation and extrapolation phases. This has led us to assign an observation-type-dependent weight for each CMA report, and to have the same processing grid (of size NPROCA times NPROCB) as the main analysis scheme.

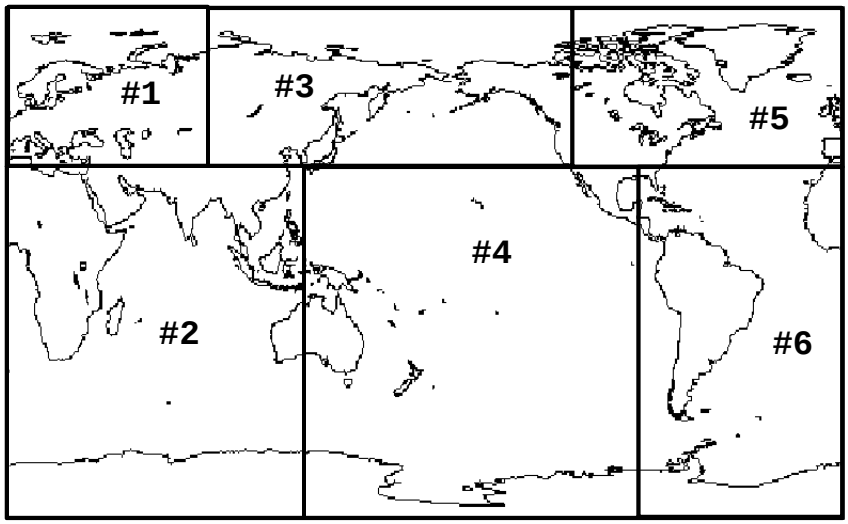

Figure 9.2 A typical split of the globe containing observational data, produced by the OBSORT for 2x3 processor grid. Each box contains approximately the same amount of observations (in a weighted sense). A number denotes the processor identifier to which observations in a particular box belong. Partition boxes are different from those used in the grid-point calculations.

We have found the following geographical partitioning of observational data works well, and is practical for observation handling (see Fig. 9.2):

1) Set the origin for observational-data space on the Greenwich meridian, since this obeys the convention chosen in the GRIB-field origin definition.

2) Choose NPROCA to be the number of latitudinal bands from north to south, and NPROCB to be the number of longitudinal boxes within each latitude band.

3) Set the number of processors to NPROC = NPROCA $\times$ NPROCB.

4) Read the local CMA files and build up a table of the observational data that contains the geographical and processor location, the unique sequence number, the time stamp, the 4D-VAR time slot, the relative weight, the pointers to local CMA data etc., for each observation

5) Communicate the local table, and sort locally, the resulting global table with respect to time slot (4D-VAR) and time stamp, and unique sequence number.

6) For each time slot, sort the locally available global-information table with respect to latitude (from north to south). Subdivide the table into the suitable parts, so that every latitude band contains about the same amount of observations in a weighted sense.

7) Continue in similar fashion for each latitude band to resolve the final longitudinal boxes.

8) Assign one box for each processor and update the destination-processor information (the processor where each particular observation ends up) into the global table.

9) Shuffle the actual CMA data, based on the information in the global table. This step involves essentially all-to-all communication, where every processor (very likely) sends a few CMA reports to every other processor (including itself, but via local copies rather than by message passing). An efficient way to communicate data is to use parallel data-transfer channels with a tournament table-like approach, where each processor communicates with each other processor in turn forming a parallel pattern of communication.

Because of inadequate load balancing in the main analysis process, the requirement for strict geographical partitioning has recently been relaxed. In the enhanced scheme, the already resolved partitions are broken up by reassigning the location of the observations found in a processor box to a new processor. This remapping is done in round-robin fashion where, for example, the observations 1, 2, 3, 4 in the box#1 are destined to the boxes 1, 2, 3, 4, respectively.

9.5 THE PARALLEL DATA FLOW OF THE OBSORT

In the parallel implementation (Fig. 9.3) we have to revise the dataflow diagram described in a previous section, which was meant only for the single-threaded execution. Program modules MAKECMA, SCREENING and FEEDBACK had to be tied to OBSORT in order to run them in parallel with observations. It turns out, that a call to the OBSORT library entry point is the only notable change in introducing parallelism in the observation processing process.

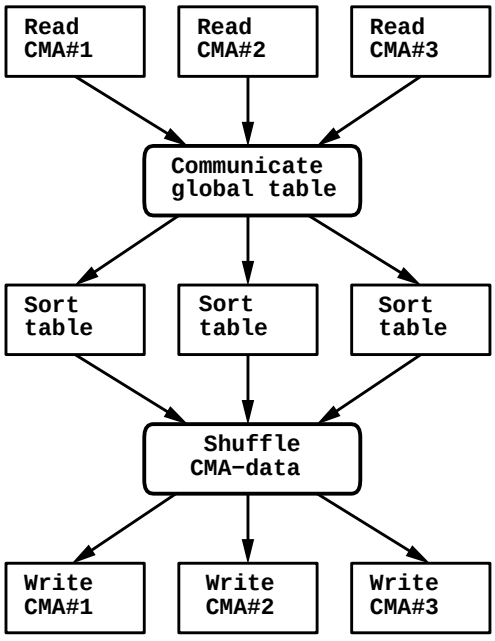

Figure 9.3 Dataflow in observation processing in the parallel scheme.

Before going on, it is advisable to explain some internal details in re-ordering (or shuffling) of the observational data. We concentrate on CMAsort, where no BUFR data are present. The extra cost of having BUFR data in CMA+BUFRsort, would not alter the generality expressed here, though.

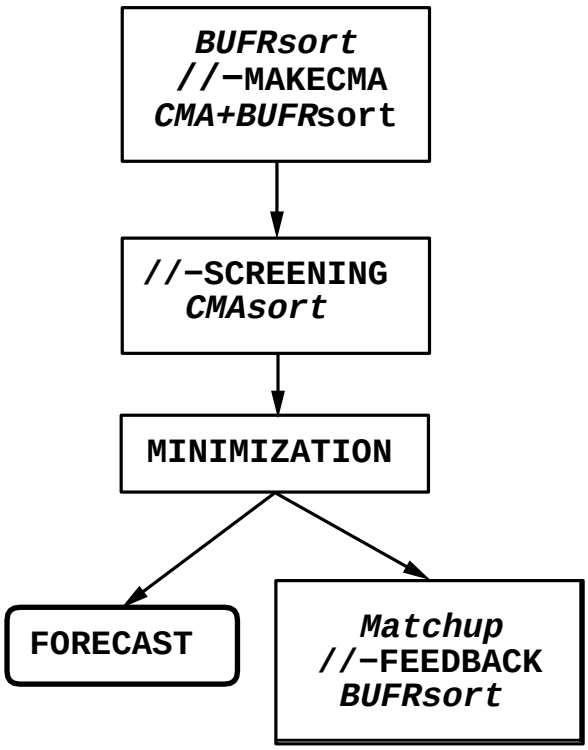

Figure 9.4 The five main stages when redistributing CMA data among processors.

Fig. 9.4 shows the five main stages in the geographical data re-ordering. Firstly, every processor reads the local CMA-files independently of each other, and establishes local tables that the contain the necessary information for the subsequent data shuffle. The second stage makes these tables global through defining the all-to-all communication steps; thus every processor obtains a local copy of a potentially large table. As a result, every processor knows all about the initial distribution of observations among the other processors. The third phase re-orders the global tables (now physically local) independently on each processor. In the fourth step the actual shuffle of CMA data takes place. This is also an all-to-all operation. In the worst case scenario, every CMA report would end up in a new owner processor. Finally, the new distribution is written out, again one CMA file per processor indepedently of the others.

A recent change to OBSORT has been to enable more CMA files to be written than processors were available to OBSORT. This so-called virtual-processor approach allows us to release more resources from the computer system to the main analysis running, possibly with considerably larger number of processors than OBSORT.

The I/O, in OBSORT, to the CMA file is done in two chunks. Firstly for the DDR sections, and secondly for the large bulk-data part. All the I/O uses pre-allocation and buffered I/O in addition to a possible 'on-the-fly' CMA packing. Wherever possible the memory-resident file system is used extensively. Furthermore, a so-called concatenated CMA file has been introduced that contains *all* (packed) CMA files together in one file. This option has reduced the I/O contention.

Prior to the actual MAKECMA, we wanted to add functionality that redistributes the few input BUFR files among

the available processors. Despite an extra step, it boosts overall execution performance of the parallel MAKECMA, since all processors—not only the first few—will enter into the BUFR data decoding phase. Currently each processor tries to read in one or more BUFR files in a cyclic fashion. If there were fewer files than processors, then only the few first processors would get a single file; others would remain idle for a while. For each BUFR message, an artificial latitude and longitude is assigned. Also, the basic contents of a BUFR message is checked to get a proper weight assigned. After this, the BUFR data can be partitioned in a manner of the 'geographical' re-ordering explained earlier. Despite the tricks involved, this method gives good results and offers nearly maximal re-use of the existing source code in OBSORT.

After receiving re-distributed BUFR data, MAKECMA continues to perform its decoding functions, but now in a fully parallel mode. Every MAKECMA process(or) unpacks and writes both a new CMA report and the counterpart BUFR message to a temporary file, one by one. However, the geographical ordering is not optimal for the screening purposes. Therefore, the OBSORT needs to be consulted and a CMA-data shuffle followed by a CMA-driven BUFR sort (i.e. CMA+BUFRsort) will be performed as a final substep in MAKECMA.

The screening proceeds in parallel mode by reading local ECMA files, one per processor, and performing its data validation functions. Once the resulting CCMA data set is ready, it may occur that in some of the processors CCMA data has almost vanished due to different kinds of rejections. Therefore, it is necessary to create new geographical distribution of observational data, now applied to the CCMA data. The new reordering is done as a final step in the screening.

There is a special function present in the OBSORT for post-adjustment of clustered data. It was found that with a different number of processors more data got rejected in one case than in another one. Certain observation types, like AIREPs, tended to cluster around partitions' boundaries. In the first case, a processor owned all the corresponding observations, since there were no partition boundaries in the neighbourhood; in the second case an AIREP cluster got effectively split among two adjacent processors. For reproducibility reasons, it was necessary to check the existence of such clusters and to move the clustered observations completely into the lowest numbered processor around the processors' boundary

For a better parallel performance in the minimization it is crucial to obtain a good load balance. Therefore, the OBSORT's unique feature that assigns weights for each observation report plays an important role. There are a few built-in weight functions in OBSORT. Depending on the contents of the observation, a different weight gets assigned. The weight functions are controlled by an external input file, where separate weight coefficients can be assigned for each observation type.

Finally, it was soon clear, that MATCHUP could be integrated inthe FEEDBACK. As a result MATCHUP is performed quickly as the first parallel task in FEEDBACK. The updated local ECMA data is then passed back in a form of CMA array to the actual FEEDBACK which encodes information back to a highly compressed BUFR format. Furthermore, the FEEDBACK benefits also from OBSORT's built-in feature to collect BUFR (or CMA) data into a single processor. This way we were able to get from the final, but still distributed, BUFR data back to the non-distributed and ready-to-archive BUFR data in a rather elegant manner. And the fact that the original BUFR-message order was not necessary (or even possible ?) to be preserved, has simplified the programming.

To accomodate the virtual-processor approach, a recent change to MATCHUP has been introduced to perform so-called virtual MATCHUP (or VMATCHUP) prior to the genuine MATCHUP. This option enables us to bring the full set of ECMA files used in the main analysis back to fewer ECMA files, and to preserve the same CMA report order as in the BUFR message counterparts after MAKECMA.

9.6 OBSORT CALLING TREE

- LIB_OBSORT
  - SWAP_FWD
  - INIT_COMMON
    - SUNUMC
    - SETCOMBU
    - SUBUOCTP
    - SETBUFR
    - BUPRQ
    - SETCOMCM
    - SUCMOCTP
    - SUCMA
  - CHECK_NAMELIST
  - BUFRSORT
    - EXPAND_STRING
    - DUMP_NAMELIST
    - PRECHECK
      - PRECHECK_CMA_ARRAY
        - SYNC_TIMESLOT_DATA
          - GLOBAL
          - INIT_COMMON
          - GEN_TIMESLOT_DATA
            - UPDCAL2
          - REF_TIME
      - PRECHECK_CMA
        - CMA_ATTACH
        - CMA_INFO
        - CMA_GET_DDRS
        - SYNC_TIMESLOT_DATA
        - CMA_DETACH
      - PRECHECK_BUFR_ARRAY
        - UTIL_NUMPROD_ARRAY
      - PRECHECK_BUFR
        - OLDBUFR_OPEN
        - UTIL_NUMPROD
        - OLDBUFR_CLOSE
    - GLOBAL
    - READ_BUFR
      - UTIL_PRODLENGTH
      - OLDBUFR_OPEN
      - OLDBUFR_READ
      - OLDBUFR_CLOSE
      - GLOBAL
    - MAKESEQNO_OBSORT
      - PRECHECK
      - FILL_SEQNOLIST
        - READ_BUFR

- CRACK_BUFR_HDR
  - BUS012
- BUFRSORT_PREPARE
  - CRACK_BUFR_HDR
- OBS_SORT
  - COLLECT
  - KEYSORT
  - REF_TIME
  - CHECK_DUPLICATES
  - GLOBE_SPLIT
    - KEYSORT
    - MERGE_CLUSTERS
- BUFR_TRAFFIC_INFO
- IFS_WRITE
  - GLOBAL
- .BUFR_SHUFFLE
  - SETUP_COMM_DATA
  - BUFR_COPY_BUFFER
- WRITE_BUFR
  - OLDBUFR_OPEN
  - UTIL_ALLOC_IOBUF
  - OLDBUFR_WRITE
  - OLDBUFR_CLOSE
- BUFR_FORM_EXPLIST

- GEN_TIMESLOT_DATA
- MAPSORT
  - EXPAND_STRING
  - DUMP_NAMELIST
  - PRECHECK
  - READ_CMA
    - CMA_ATTACH
    - CMA_GET_DDRS
    - . CMA_BIN_INFO
    - CMA_READ
    - CMA_CLOSE
    - GLOBAL
    - REF_TIME
  - SORT_PREPARE
    - REF_TIME
    - GLOBAL
  - OBS_SORT
  - VMAPSORT
    - EXPAND_STRING
    - . VCMA_TRAFFIC_INFO
      - GLOBAL
    - IFS_WRITE
    - OPTIMAL_MSGLEN

- UTIL_IGETENV
- GLOBAL
- VCMA_SHUFFLE
  - SETUP_COMM_DATA
  - DUMP_INFO
  - CMA_COPY_BUFFER
- GLOBAL
- WRITE_OBSMAP
  - CMA_STAT
  - CMA_OPEN
  - DATE_AND_TIME
  - CMA_WRITEI
  - CMA_CLOSE
- UPDATE_DDR
  - GLOBAL
  - DATE_AND_TIME
  - ADD_SATELLITE
  - . REF_TIME
- WRITE_CMA
  - CMA_GET_CONCAT_BYNAME
  - CMA_OPEN
  - CMA_BIN_INFO
  - CMA_IS_PACKED
  - GLOBAL
  - CMA2PCMA
  - CMA_INFO
  - OPTIMAL_MSGLEN
  - COMM_WRITE
  - CMA_WRITE
  - CMA_CLOSE

- CMA_TRAFFIC_INFO
  - GLOBAL
- IFS_WRITE
- OPTIMAL_MSGLEN
- CMA_SHUFFLE
  - SETUP_COMM_DATA
  - CMA_COPY_BUFFER
- UPDATE_DDR
- WRITE_CMA
- READ_BUFR
- BUFR_PTRLEN_UPDATE
  - COLLECT
  - KEYSORT
- BUFR_TRAFFIC_INFO
- BUFR_SHUFFLE
- WRITE_BUFR
- CMA_GET_CONCAT_BYNAME

- EXPAND_STRING
- MATCHUP
  - EXPAND_STRING
  - DUMP_NAMELIST
  - PRECHECK
  - READ_CMA
  - MATCHUP_PREPARE
  - OBS_MATCHUP
    - COLLECT
    - . KEYSORT
    - CHECK_DUPLICATES
    - . KEYSEARCH
  - CMA_TRAFFIC_INFO
  - OPTIMAL_MSGLEN
  - CMA_SHUFFLE
  - MATCHUP_PERFORM
    - KEYSEARCH
    - MATCHUP_REPORT
  - UPDATE_DDR
  - WRITE_CMA
- VMATCHUP
  - EXPAND_STRING
  - DUMP_NAMELIST
  - READ_OBSMAP
    - CMA_STAT
    - CMA_OPEN
    - . CMA_READI
    - DISTR
    - CMA_DETACH
    - CMA_ATTACH
    - CMA_CLOSE
    - DISTR_OBS
    - GLOBAL
    - COLLECT
  - PRECHECK
  - READ_CMA
  - VMATCHUP_PREPARE
    - COLLECT
  - KEYSORT
  - VMATCHUP_TRAFFIC_INFO
    - GLOBAL
  - OPTIMAL_MSGLEN
  - XCMA_SHUFFLE
    - SETUP_COMM_DATA
  - UPDATE_DDR
  - WRITE_CMA
- SWAP_BWD

Part II: DATA ASSIMILATION

CHAPTER 10 Observation screening

Table of contents

10.1 Introduction

10.2 The structure of the observation screening

10.2.1 The incoming observations

10.2.2 The screening run

10.2.3 General rationale of the observation screening

10.2.4 3D- versus 4D-Var screening

10.3 The independent observation screening decisions

10.3.1 Preliminary check of observations

10.3.2 Blacklisting

10.3.3 Background quality control

10.4 Screening of satellite radiances

10.4.1 General

10.4.2 Input

10.4.3 Bias correction

10.4.4 Quality control

10.4.5 Retrieval

10.4.6 SSM/I radiances

10.5 Scatterometer processing

10.5.1 Introduction

10.5.2 Background

10.5.3 ERS Wind scatterometer processing

10.5.4 NASA scatterometer (NSCAT) processing

10.5.5 NASA ``QUIK'' scatterometer (QuikSCAT) processing

10.6 The dependent observation screening decisions

10.6.1 Update of the observations

10.6.2 Global time–location arrays

10.6.3 Vertical consistency of multilevel reports

10.6.4 Removal of duplicated reports

10.6.5 Redundancy check

10.6.6 Thinning

10.6.7 A summary of the current use of observations

10.6.8 Compression of the ODB

10.7 A massively-parallel computing environment

10.1 INTRODUCTION

This chapter describes the observation screening in the ECMWF 3D/4D-Var data assimilation. A more general description can be found in *Järvinen and Undén* (1997). The purpose of the observation screening is to select a clean array of observations to be used in the data assimilation. This selection involves quality checks, removal of duplicated observations, thinning of their resolution etc.. The new selection algorithm was implemented operationally in September 1996 and was designed to reproduce (to a large extent) the selection of observations that the old screening of the ECMWF OI analysis code used to make (*Lönnberg and Shaw*, 1985 and 1987; *Lönnberg*, 1989).

This chapter was prepared in September 1997 by Heikki Järvinen, Roger Saunders and Didier Lemeur, and updated in February 1999 by Roger Saunders for TOVS processing, by Elias Holm and Francois Bouttier for the remainder.

10.2 THE STRUCTURE OF THE OBSERVATION SCREENING

10.2.1 The incoming observations

Before the first trajectory integration, the observations are extracted from a data base of observations converted from the BUFR archive to a set of CMA files (by programs **obsproc** and **obsort** described in the 'observation' part of the documentation). These data have already undergone some rudimentary quality control, e.g. a check for the observation format and position, for the climatological and hydrostatic limits, as well as for the internal and temporal consistency (*Norris*, 1990). The so-called RDB (Report Data Base) flag is assigned at this stage. Then a set of observation files suitable for assimilation is created in a separate observation preprocessing module. This entails format conversions, changes of some observed variables (like calculation of relative humidity from dry and wet bulb temperatures), as well as assignment of observation error statistics. The resulting 'extended' CMA file set (ecma.**#** ; # = processor id) contains all the observational information from the six-hour data window available at the cut-off time, and is an input for the IFS. The observation screening then selects the best quality and unique observations, preferably close to the middle of the data window because the background is not interpolated to the exact time of the observation. Unlike the OI, the 3D- and 4D-Var data assimilation is global and, therefore, no separate data selection for analysis boxes is needed.

10.2.2 The screening run

The ECMWF 3D/4D-Var data assimilation system makes use of an incremental minimization scheme (*Courtier et al*. 1994 and 1997) to reduce the computational cost. The variational data assimilation starts with the first (high resolution) trajectory run (CONF = 2, LSCREEN = .TRUE.). During this run the model counterparts for all the observations are calculated through the nonlinear observation operators (controlled by taskob). As soon as these background departures are available for observations, the screening can be performed. For the observation screening, the background errors (errgrib - file) are interpolated to the observation locations for the observed variables (sufger).

Technically, the final result of the observation screening is a pair of observation arrays. The original 'extended' observation array now contains observations complemented by the background departures, together with quality control information for most of the observations. This array is stored for later feedback (ecma.# - set of files). The compressed observation array is a subset of the original array of observations, and is passed for the subsequent minimization job. The compressed array (ccma.# - set of files) contains only the observations to be used in the variational assimilation. Memory wise, the first trajectory run is a demanding one as all the observations are kept in memory. Prior to the screening the model fields are deallocated (dealmod) as most of the information necessary in the screening is stored in the observation data base (ODB).

10.2.3 General rationale of the observation screening

The general logic in the 3D/4D-Var observation screening algorithm is to make the *independent* decisions first, i.e. the ones that do not depend on any other observations or decisions (decis). One example is the background quality control for one observed variable. These can be carried out in any order without affecting the result of any other independent decision. The rest of the decisions are considered as mutually *dependent* on other observations or decisions, and they are taken next, following a certain logical order. For instance, the horizontal thinning of TOVS reports is only performed for the subset of reports that passed the background quality control. Finally, the observation array is compressed (compres) for the minimization in such a way that it only contains the data that will be used.

10.2.4 3D- versus 4D-Var screening

In the 3D-Var assimilation system, the observations processed have been gathered over a 6-hour long time window (from 3 hours before to 3 hours after the nominal analysis time) ; all the screening is performed as if they have actually been performed simultaneously at the central analysis time. In the early implementation of the 4D-Var assimilation system, the same processing called 3D-screening was applied over the 6-hour long 4D-Var time window, which resulted in essentially the same screening decisions as in 3D-Var, except that the model comparison with the observation was performed at almost the appropriate time with no more than a 30-minute approximation.

In summer 1997, a new screening procedure called 4D-screening was implemented that took into account the time dimension of the observations. The time window was divided in timeslots of typically 1-hour length (30mn for the first and the last time slot). The 3D-screening algorithm was then applied separately to observations in each timeslot. This allowed more data to be used by 4D-Var, for instance, all messages from an hourly reporting station can now be used, whereas only one (closest to central time) would have been allowed by the redundancy check in the 3D-screening. The 4D-screening behaviour is activated by switch LSCRE4D ; it is meant to be used in conjunction with time correlation of observation errors where appropriate, as explained in Järvinen *et al* (1999) and in the chapter on conventional obs error constraints.

10.3 THE INDEPENDENT OBSERVATION SCREENING DECISIONS

10.3.1 Preliminary check of observations

The observation screening begins with a preliminary check of the completeness of the reports (prech). None of the following values should be missing from a report: observed value, background departure, observation error and vertical coordinate of observation. Also a check for a missing station altitude is performed for synop, temp and pilot reports. The reporting practice for synop and temp mass observations (surface pressure and geopotential height) is checked (repra), as explained in the appendix. At this stage also, the observation error for synop geopotential observations is inflated if the reported level is far from the true station level (addoer). The inflation is defined as a

proportion of the difference between the reported level and the true station altitude by adding 2% of the height difference to the observation error.

10.3.2 Blacklisting

Next, the observations are scanned through for blacklisting. At the set-up stage the blacklist interface is initialized (blinit) to the external blacklist library (libbl95.a). The interface between the IFS and the black-list is in the IFS routine BLACK. The blacklist itself consists formally of two parts. Firstly, the selection of variables for assimilation is done using the data selection part of the blacklist file. This controls which observation types, variables, vertical ranges etc. will be selected for the assimilation. Some more complicated decisions are also performed through the data selection file; for instance, an orographic rejection limit is applied in the case of the observation being too deep inside the model orography. This part of the blacklist also provides a handy tool for experimentation with the observing system, as well as with the assimilation system itself. Secondly, a normal monthly monitoring blacklist is applied for discarding the stations that have recently been reporting in an excessively noisy or biased manner compared with the ECMWF background field. A full documentation of the new blacklisting mechanism is found in *Järvinen et al.* (1996).

10.3.3 Background quality control

The background quality control (FIRST) is performed for all the variables that are intended to be used in the assimilation. The procedure is as follows. The variance of the background departure $y - H(x_b)$ can be estimated as a sum of observation and background error variances $\sigma_o^2 + \sigma_b^2$, assuming that the observation and the background errors are uncorrelated. After normalizing with σ_b, the estimate of variance for the normalized departure is given by $1 + \sigma_o^2/\sigma_b^2$. In the background quality control, the square of the normalized background departure is considered as suspect when it exceeds its expected variance more than by a predefined multiple (FGCHK, SUFGLIM). For the wind observations, the background quality control is performed simultaneously for both wind components (FGWND). In practice, there is an associated background quality-control flag with four possible values, namely 0 for a correct, 1 for a probably correct, 2 for a probably incorrect and 3 for an incorrect observation, respectively (SUSCRE0). Table 10.1 gives the predefined limits for the background quality control in terms of multiples of the expected variance of the normalized background departure. These values can be changed in namelist NAMJO. For satob winds the background error limits are modified as explained in Appendix A.

TABLE 10.1 THE PREDEFINED LIMITS FOR THE BACKGROUND QUALITY CONTROL, GIVEN IN TERMS OF MULTIPLES OF THE EXPECTED VARIANCE OF THE NORMALIZED BACKGROUND DEPARTURE.

Variable	Flag 1	Flag 2	Flag 3
u, v	9.00	16.00	25.00
z, ps	12.25	25.00	36.00
dz	x	x	x
T	9.00	16.00	25.00
rh, q	9.00	16.00	25.00

Flag values are denoted by 1 for a probably correct, 2 for a probably incorrect and 3 for an incorrect observation. The variables are denoted by u and v for wind components, z for geopotential height, ps for surface pressure, dz for thickness, T for temperature, rh for relative humidity and q for specific humidity, respectively.

There is also a background quality control for the observed wind direction (FGWIND). The predefined error limits

of 60^o, 90^o and 120^o apply for flag values 1, 2 and 3, respectively. The background quality control for the wind direction is applied only above 700 hPa for upper-air observations for wind speeds larger than 15 m s^{-1}. If the wind-direction background quality-control flag has been set to a value that is greater than or equal to 2, the background quality-control flag for the wind observations is increased by 1. For scatterometer winds, a test for high wind speeds and cold SST is applied in the IFS routine FGWIND.

10.4 SCREENING OF SATELLITE RADIANCES

This section describes the use of RTOVS, valid in April 1999. At the time of writing it was planned to switch to using 1-C radiances for which the processing is rather different; it is described in a separate chapter of the documentation.

10.4.1 General

The radiances from the RTOVS 120 km BUFR data received from NESDIS are preprocessed in a dedicated module which performs several functions to allow the assimilation of TOVS radiances in 4D-Var (the NESDIS retrievals are not used in 4D-Var, but are only monitored with the background profiles). This module is called ADVAR and is part of the TOVSCODE library (libtovscode.a). ADVAR is called for each TOVS observation with the model background temperature, specific-humidity and ozone profiles, and surface parameters interpolated to the location of the observations. For each analysis cycle there are typically 22,000 TOVS observations in total, for a dual polar orbiter system. ADVAR performs the following functions described below, dependent on the setting of a switch IS which determines the mode of operation of ADVAR

. In the screening pass ADVAR is called twice by TOVCLR, once with IS set to 1 (when all the operations described below are performed) and once with IS set to –1 (when only the high resolution radiance departures for 4D-Var are computed). When IS is set to –1 the profile extrapolation and 1D-Var retrieval is not performed. Finally if IS is set to 0 then the background radiances are computed from the profiles, but the 1D-Var retrieval is not performed, an option only used for offline tests. Several input files are required for ADVAR which are listed in Table 0.2. A set up routine for ADVAR (SUADVAR) is called within the IFS by SURAD to open the necessary files and fill common arrays for ADVAR and the fast radiative transfer model RTTOV-5. The various operations performed by ADVAR are described below. The full scientific description of ADVAR is described in the paper by *Eyre et. al.* (1993).

10.4.2 Input

The fast radiative-transfer model RTTOV-5 for TOVS radiances requires an input profile of 40 levels from 1013.25 to 0.1 hPa. RTTOV-5 has been described in detail by *Saunders et al.* (1999). The original forecast model temperature, specific-humidity and ozone profiles are interpolated onto the fixed pressure levels required by RTTOV-5 before they are input to ADVAR. For the 31 level model, the background profiles are only available up to 10 hPa, and so an extrapolation has to be performed up to 0.1 hPa for temperature using the NESDIS retrievals to 1 hPa, and a simple extrapolation based on model atmospheres above this level. Climatological mean profiles are assumed for water vapour and ozone. For the 50- and 60-level versions of the model with levels in the stratosphere this extrapolation is not necessary. Once the full profile is defined and checked (see below) RTTOV-5 is called to compute the background radiances from the background profiles. Background radiances are computed for all the TOVS channels listed in Table 10.2, but only a subset of the channels are subsequently used in the 1D-Var retrieval.

TABLE 10.2 TOVS CHANNEL USAGE AND (**O+F**) ERRORS ASSUMED IN 1D-VAR (THE ERRORS USED IN 4D-VAR ARE INFLATED BY 50%)

Channel Number	1/4D-Var usage	Clear (K)	Cloudy (K)	Land/mixed (K)	Sea-ice (K)
1	All/global	1.40	1.40	1.40	1.40
2	All/global	0.35	0.35	0.35	0.35
3	All/global	0.30	0.30	0.30	0.30
4	Clear/sea/global	0.20		0.20	0.20
5	Clear/sea/global	0.30		0.30	0.30
6	Clear/sea/global	0.40		0.80	0.80
7	Clear/sea	0.60			1.20
8	Clear/sea	1.00			2.00
9	FG only				
10	Clear/sea	0.80			1.60
11	Clear/global	1.10		1.10	1.10
12	Clear/global	1.50		1.50	1.50
13	Clear/sea	0.50			1.00
14	Clear/sea	0.35			0.70
15	Clear/sea	0.30			0.60
16	FG only				
17	FG only				
18	FG only				
19	FG only				
21	QC check				
22	All/sea*	0.30	0.30		1.00
23	All/global*	0.22	0.22	0.22	0.22
24	All/global	0.25	0.25	0.25	0.25
25	All/global	0.60	0.60	0.60	0.60
26	All/global	1.00	1.00	1.00	1.00
27	All/global	1.80	1.80	1.80	1.80

*Cloudy data not used in tropics

10.4.3 Bias correction

The next step is to apply the bias correction to the NESDIS radiances. The details of the bias correction for TOVS radiances is given in *Eyre* (1992) and *Harris* (1997). An update to the bias correction coefficients is performed once a month on the past two to four weeks of radiance-departure statistics, the exact period depending on how rapidly the biases have changed during the period. The bias-correction coefficients are stored in a file for all of the satellites, and this is used in ADVAR. The bias correction code is in the BIASCOR subdirectory of TOVSCODE.

10.4.4 Quality control

Several quality checks are then applied to the measured and background radiances, and ADVAR returns a flag (IFAIL); this is zero if all the checks are passed, but is set to a specific value if a problem is detected. The values for IFAIL and their meaning are given in Table 10.3. The radiances or retrievals are only used in 4D-Var if IFAIL is zero. The gross checks applied are:

(i) Check that the background profile vector is within realistic limits (e.g. temperature is within the range 150–350 K, specific humidity is positive and not supersaturated, ozone is within climatological extremes). ADVAR terminates with a severe error flag if this test fails.

(ii) The measured and background brightness temperatures are present for all required channels and are within the range 150–350 K.

TABLE 10.3 DEFINITION OF 1D-VAR FAILURE FLAGS AND TYPICAL RATES IN THE IFS.

IFAIL	Typical %	Comment
0	80%	Retrieval OK
nn	1.0%	Measurement cost too high for channel *nn*
55	17%	At edge of scan, otherwise OK
66	0%	Failed stability check (not applied)
99	0.5%	Minimization failed to converge
100	1.0%	Failed window channel cloud test
5*nn*	0.1%	Channel *nn* failed fine background check
6*nn*	0.3%	Channel *nn* failed gross background check
7nn	0%	Bad background radiance for channel nn
887	0.3%	background profile outside RTTOV limits
888	0%	background profile corrupt
9nn	0%	Radiances for channel nn corrupt
999	0%	No valid scan or valid satid or bias coeffs

A series of more critical tests are then applied where ADVAR continues even if the test fails but returns a non-zero IFAIL value.

(i) Gross background check (i.e. the measured radiance departures from the background are less than 20 K).

(ii) The background temperature, specific humidity and ozone profiles are checked to make sure they are close to, or within, the range encompassed by the diverse 32 (or 35 for ozone) profile data set for which the RTTOV is valid.

(iii) A fine background check where the square of the radiance departures are flagged if they are greater than $16 \times [\mathbf{KBK}^{\mathrm{T}} + \mathbf{O} + \mathbf{F}]$ (see below for definitions).

(iv) A check for cloud contamination for the HIRS channels is included by checking that the radiance departure for HIRS channel 8 is inside the range –4 to +8 K over the sea and south of 20°N. Over land the thresholds are brighter north of 20°N. IFAIL is set to 100 if outside this range.

(v) Radiances at the two extreme edge positions of the swath are flagged at present and not used in 4D-Var.

(vi) Checks are also made that the bias-correction coefficients, satellite id, and scan position are all valid before proceeding.

10.4.5 Retrieval

TABLE 10.4 FILES REQUIRED BY ADVAR

Filename	Contents
chanspec.dat	Specifies channel usage
rmtberr.dat	Specifies radiance observation errors (O+F)
fcbkerr.dat	Specifies 1D-Var background error covariances (B)
bcor.dat	Bias correction coefficients
rt_coef_ieee.dat or rt_coef_fmt.dat	RTTOV coefficients in binary or ascii format.

The main task for ADVAR is to perform a 1D-Var retrieval of temperature, water vapour and ozone profiles. Details on the theory of 1D-Var retrieval are described by *Eyre et al.* (1993), and so only the technical details described here are concerned with the implementation at ECMWF. Each radiance profile is assigned to be clear, partly cloudy or cloudy by NESDIS, and different TOVS channels and observation errors are used for each type as listed in Table 10.2. The files defining the channel selection and observation plus forward model error **(O+F)** covariances are given in Table 10.4 and no interchannel correlations are assumed (i.e. a diagonal matrix). The **(O+F)** errors specified here are subsequently used in 4D-Var, but are inflated by 50%. The background-error covariances **B** for all 43 levels are also specified in a file, and for temperature are close to the global-mean background errors assumed in 4D-Var. For specific humidity the background errors assumed in 1D-Var follow the same formulation as in 4D-Var (*Rabier et al.* 1997) and the correlations are the same as in 4D-Var.

The minimization of the cost function is performed using the method of Newtonian iteration, and up to 5 iterations are allowed before the minimization fails. Convergence is obtained when the profile departures are less than 0.4 times σ_b at every level. If the cost function of the observed radiance in any of the channels exceeds a predefined threshold then a flag is set indicating an inconsistent set of radiances. The output of 1D-Var includes background and retrieved temperature, water-vapour and ozone profiles, together with several retrieved surface parameters also included in the 1D-Var control vector. The retrieved profiles are output both on the 43 levels and as virtual layer-mean temperatures on 15 levels and layer-mean column water vapour on 3 levels to match the NESDIS retrievals.

A final check on the stability of the retrieved profile is provided in the code, but is not implemented as the profiles are not used in 4D-Var.

10.4.6 SSM/I radiances

SSM/I radiances are also screened in a similar module DVSSMI, which performs a similar set of functions to ADVAR B by retrieving the total column water vapour, surface wind speed and cloud liquid-water path. The total cloud water vapour retrievals have been activated operationally in Spring 1998, with an horizontal thinning to 250km.

A specialized library, **ssmicode** is used for the retrievals. Some documentation can be found in Gerard and Saunders (1999). At the time of writing it is envisaged to start using surface wind speed retrievals over sea in summer 1999.

10.5 SCATTEROMETER PROCESSING

10.5.1 Introduction

This section describes the flow of ERS, NSCAT, and QuikSCAT scatterometer data through the assimilation system. Some tasks like thinning of ERS data and retrieval of 50 km QuikSCAT winds are performed in modules before the screening but it is most natural to describe the whole processing step here. This section provides a working knowledge of the software, and guidance on possible modifications and updates. It is not intended to explain the scientific background of microwave remote sensing or scatterometry and assumes some knowledge of these topics (see Stoffelen (1999), Freilich and Dunbar (1999)).

This section is broken into five subsections. The first is the introduction, which you are reading now. The second is background information about scatterometer processing at ECMWF. The background includes a brief history of the software, lists persons who contributed the changes and outlines the structure and function of the whole library. Subsections 3-5 describe processing for scatterometers used or currently in use at ECMWF, i.e. ERS-1 and ERS-2, NSCAT and QuikSCAT.

10.5.2 Background

ESA's ERS-1 scatterometer was launched in July 1991 and stopped operating in June 2000. The successor ERS-2 was launched in 1995 and is still functioning well. Data from ERS-2 were introduced into operations at ECMWF in January 1996. Scatterometer data from ERS-2 have been used in operations since that time. Ad Stoffelen, David Anderson and Ross Hoffman were the first to work on the problem at ECMWF. Stoffelen and Anderson worked on QC and wind retrieval issues in the OI system of the day. Hoffman looked at assimilation of sigma0's directly in 3D-Var. Once in operations, several others (Herve Roquet, Catherine Gaffard, Didier LeMeur and Lars Isaksen) took turns monitoring and improving the use of the data. Lately Mark Leidner worked on the use of data from NASA scatterometers (NSCAT and QuikSCAT).

Source code for scatterometer processing resides in ClearCase under the project name scat. The library contains the following directories:

etimesort/ source code for pre-processing ERS data
module/ shared modules
qbukey/ source code for adding RDB info to QuikSCAT 50km BUFR
qfilter/ source code for pre-processing QuikSCAT 25km BUFR
qretrieve/ source code for SeaWinds wind retrieval
test/ empty directory for future test code

e* and q* directories contain processing software specific to ERS and QuikSCAT, respectively. NSCAT-specific codes have not been put in ClearCase, because the satellite stopped operating in June 1997, i.e. its data will never be used in operations.

10.5.3 ERS Wind scatterometer processing

Fig. 10.1 shows a simple flow chart for ERS processing at ECMWF. Below the processing chain in described in general and the functionality of each executable in the scat library in particular.

The MARS archive definitions for the different wind scatterometer observations are:

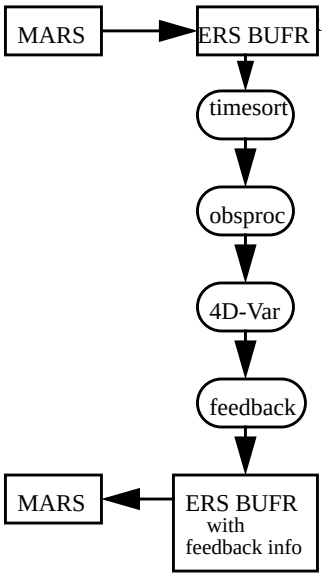

Figure 10.1 ERS processing

TABLE 10.5

	BUFR/MARS obstype	CMA code type	Satellite ID
ERS-1	122	122	1
ERS-2	122	122	2
NSCAT	136	210	280
QuikSCAT	137*	300	281
QuikSCAT	138	301	281

* For QuikSCAT the BUFR format changed January 2000 which is reflected in the change of MARS obstype from 137 to 138 and change in BUFR sequence descriptor.

Data for a given time window are retrieved from MARS. These data are then input to program timesort, which sorts the observations by time and removes duplicate records. Duplicate records occur in the input data because more than one ground station may receive the same ERS data. Duplicated data are almost identical and their time stamps will differ by a small margin (< 4 seconds). Duplicates are rejected on this basis.

ERS winds are retrieved as part of the IFS observation pre-processor, OBSPROC (IFS Documentation Part I: Observation Processing). Within OBSPROC, the three sigma0's are decoded from BUFR, and used to retrieve winds for 50 km diameter foot prints. The ERS winds are available at 25 km resolution, i.e. over sampled. It is not possible to determine a unique wind vector solution, at least two ambiguous solutions will be found. Only the most likely wind and the first one in the opposite direction are kept and written to the observation data base (ODB) file . ERS winds are retrieved instead of using the unique wind distributed by ESA to be able to supply two winds to the variational data assimilation system at ECMWF, and to be able to apply bias corrections to sigma0 before wind re-

trieval, and finally to be able to apply quality controls to the retrieved winds via the retrieval residual.

A horizontal thinning is performed on the 19x19 data layout of the ERS scatterometer reports. In that case the process is defined with respect to the particular measurement geometry of the instrument. Indeed the backscatter data are acquired within individual cells related to a 450 km wide grid with a mesh of 25 km in the across and along track directions. 19 measurement nodes are thus defined across the scatterometer´s swath, numbered from 1 to 19 as the incidence angle increases, while 19 rows are also considered in the along track direction to gather the data in squares of 19 by 19 points. The thinning factor is controlled by LSCATTHI and NTHINSCA in namelist NAMMKCMA. The inner two nodes are skipped (smallest incidence angles) because the scatterometer operates best at larger incidence angles. Then, every NTHINSCA node is used (default NTHINSCA = 4 results in use of node 3, 7, 11, 15, and 19) in every NTHINSCA row. The data are thus by default used at a resolution of 100 km instead of the original 25 km sampling distance. This simple way of thinning is preferable because ERS wind measurements typically always are of the same high quality as they are not affected by rain or clouds. It should be noted that the thinning process is actually set up in the observation pre-processing (OBSPROC), but that only a flag is assigned at that stage, which is then applied in the screening. In that way all the data are completely pre-processed before assimilation, and the subsequent information kept in the feedback files allows to perform their monitoring at the full resolution.

In the IFS, the two retrieved winds are used in an obs cost function with 2 minima (see pp_obs/hjo.F90). The function is not quadratic, but depends on the 4th power and root of the product of the departures of u and v (Stoffelen and Anderson (1997)). It is very similar to the sum of two quadratic cost functions.

Quality control decisions made by the IFS screening run are:

High wind speed check: Data rejected if observed or first guess wind speeds are above 25 m/s (RSCAWLI). Performed by **obs_preproc**/fgwnd.F90

Sea ice check: Data rejected if sea ice fraction is greater than 0.1 (RSCATLI). Performed by **obs_preproc**/fgwnd.F90

Global Quality Control: If the average distance-to-the-cone residual for the backscatter measurements during a (1 hour) time slot for any of the active nodes is above the QC threshold all ERS data for that time slot is blacklisted. This is done by the routine **obs_preproc**/scaqc.F90.

There is no back ground wind check performed on scatterometer data, but data may be de-weighted or effectively removed from the analysis during the minimisation in 4D-Var by variational quality control (Andersson and Järvinen (1999)).

Quality control decisions and departures from background and analyses are appended to each subset in the feedback BUFR message.

ERS feedback messages have a PRESCAT section sandwiched between the original ERS and the feedback data. The PRESCAT section contains outputs about the quality of the winds from the retrieval.

Here are some of the key words and bits to examine in the ERS feedback message (these are in the order in which they are encountered in the processing):

Winds retrieved at ESA: BUFR descriptor 11012 for speed and 11011 for direction winds available in observation part of BUFR file.

Winds retrieved at ECMWF: BUFR descriptor 11192 for u and 11193 for v winds retrieved in program OBSPROC.

Report rejected by thinning if BUFR descriptor 33229 (Report Event Word 2) = 1. QC decision made by program OBSPROC in subroutine scatsin.F90.

Background departures x 2 ambiguities: BUFR descriptor 224255 for u ('U - COMPONENT AT 10 M') and for v ('V - COMPONENT AT 10 M'). BUFR descriptor 8024 = 33, BUFR descriptor 33210 = 1, BUFR descriptor 33211 = 1001.

Report rejected by high wind speed check if BUFR descriptor 33233 (Report Status Word 1) = 16. QC decision made by program IFS in subroutine **obs_preproc**/fgwnd.F90.

Report rejected if Sea Ice faction > 0.1: BUFR descriptor 33220 (Report Event Word 1) = 12. QC decision made by program IFS in subroutine **obs_preproc**/fgwnd.F90.

Report rejected if global QC fails: BUFR descriptor 33220 (Report Event Word 1) = 16. QC decision made by program IFS in subroutine **obs_preproc**/ersqc.F90.

Datum 4D-Var quality control status: BUFR descriptor 33233 (Report Status Word 1) = 1/2/4/8 1 - active, 2 - passive, 4 - rejected, 8 - blacklisted Datum. 4D-Var variational quality control rejection BUFR descriptor 33236 (Datum event Word 1) bit 27 = 1

Analysis departures x 2 ambiguities: BUFR descriptor 224255 for u ('U - COMPONENT AT 10 M') and for v ('V - COMPONENT AT 10 M'). BUFR descriptor 8024 = 33, BUFR descriptor 33210 = 9, BUFR descriptor 33211 = 999.

10.5.4 NASA scatterometer (NSCAT) processing

NSCAT data has been used experimentally for impact experiments in 4D-Var as well as a surrogate for QuikSCAT data (another Ku-band scatterometer). The processing is not automatic in IFS , as is the case for ERS and QuikSCAT. The NSCAT data quality is more consistent compared to ERS and QuikSCAT, because the archived NSCAT data are a re-processed science product, not an ``as-is'' real-time product.

Data for the whole 9-month mission are stored on ecfs in HDF format archived in ecfs:/oparch/nscat/50km/L17 - Level 1.7 files (sigma0 data) and ecfs:/oparch/nscat/50km/L20 - Level 2.0 files (wind data).

The format and content of HDF NSCAT files are thoroughly documented in QuikSCAT Science Data Product User's Manual (avaialble from ECMWF or JPL). Level 1.7 and 2.0 files are present for each orbit in the mission. Each sigma0 file has a corresponding wind file.

Assimilation experiments with NSCAT data are only possible after offline processing of the data. Please contact the research department for further information.

10.5.5 NASA ``QUIK'' scatterometer (QuikSCAT) processing

The implementation of QuikSCAT data processing borrowed many lessons from the use of NSCAT data. QuikSCAT, however, was implemented to be used operationally, so the process is more streamlined (see Fig. 10.2).

The processing of QuikSCAT data will now be described.

Data for a given time window are retrieved from MARS. These data are then fed to program **qfilter**/qscat_filter, which sorts the observations by time, and removes duplicate/incomplete records

Duplicate and incomplete records are part of the QuikSCAT real-time data stream because of Seawinds' conically-scanning geometry. See Leidner et al. (1999) for a discussion of duplicate and incomplete records introduced by the scanning geometry.

QuikSCAT winds are retrieved with program **qretrieve**/qscat25to50km. The input is 25-km QuikSCAT BUFR messages. These are decoded, consecutive rows are paired together, sigma0's are grouped into 50-km boxes, and

winds are retrieved at this resolution. The output is 50-km BUFR messages, including sigma0's and winds. The 50 km resolution is more representative of the scales resolved by the increments in 4D-Var.

The winds are used just as in NSCAT. The winds are re-ordered (most likely first and its 180-degree opposite is next), and only the first two are used in 4D-Var.

Here are some of the key words and bits to examine in the QuikSCAT feedback message (these are in the order in which they are encountered in the processing):

Background departures x 2 ambiguities: Like for ERS described above.

Report rejected if sea ice fraction is > 0.1: Like for ERS described above.

Report rejected if data not in the sweet spots: when BUFR descriptor 33229 (Report Event Word 2) = 3. QC decision made by program IFS in subroutine **obs_preproc**/qscatqc.F90.

Report rejected if number of winds is < 2: when BUFR descriptor 33220 (Report Event Word 1) = 3. QC decision made by program IFS in subroutine **obs_preproc**/qscatqc.F90.

Report rejected if wind directions are too close: when BUFR descriptor 33229 (Report Event Word 2) = 2. QC decision made by program IFS in subroutine **obs_preproc**/qscatqc.F90.

Datum rejected if number of ambiguities > 2: when BUFR descriptor 33236 (Datum Event Word 1) = 19. QC decision made by program IFS in subroutine **obs_preproc**/qscatqc.F90.

Report rejected if global QC fails: BUFR descriptor 33220 (Report Event Word 1) = 16. QC decision made by program IFS in subroutine obs_preproc/qscatqc.F90.

Datum 4D-Var quality control: Like for ERS described above.

Analysis departures x 2 ambiguities: Like for ERS described above.

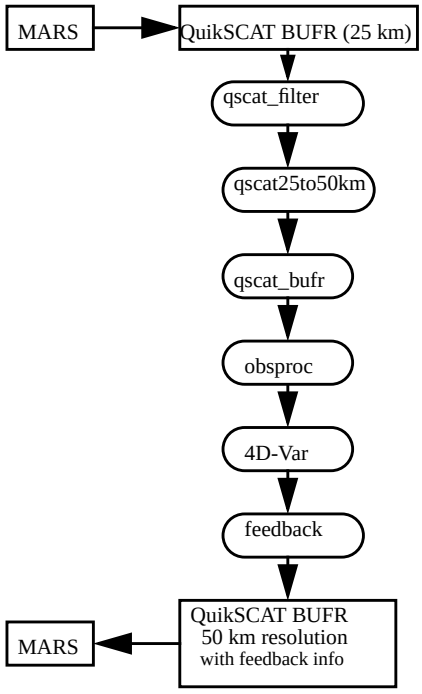

Figure 10.2 QuickSCAT processing

10.6 THE DEPENDENT OBSERVATION SCREENING DECISIONS

10.6.1 Update of the observations

Just before performing the dependent screening decisions, the flag information gathered so far is converted into a status of the reports, namely active, passive, rejected or blacklisted, and also into a status of the data in the reports (FLGTST). The reports with a RDB report flag value 2 (probably incorrect) or higher for latitude, longitude, date and time are rejected. For the observed data there are RDB datum flags for the variable and for the pressure, i.e. the pressure level of the observation. The rejection limits for these are as follows: all data are rejected for the maximum RDB datum flag value 3 (incorrect), non-standard-level data are rejected for the maximum RDB datum flag value 2, and for the pressure RDB datum flag the rejection limit is 1 (probably correct). The background quality control rejection limits are flag value 3 for all the data, and flag value 2 for the non-standard-level data.

10.6.2 Global time–location arrays

Some of the dependent decisions require a global view to the data which is not available as the memory is distributed. Therefore *ad hoc* global time–location arrays are formed and broadcast in order to provide this view (GLOBA, DISTR).

10.6.3 Vertical consistency of multilevel reports

The first dependent decisions are the vertical-consistency check of multilevel reports (VERCO), and the removal of duplicated levels from the reports. The vertical-consistency check of multilevel reports is applied in such a way that if four consecutive layers are found to be of suspicious quality, even having a flag value one, then these layers are rejected, and also all the layers above these four are rejected in the case of geopotential observations. These decisions clearly require the quality-control information, and they are therefore 'dependent' on the preceding decisions.

10.6.4 Removal of duplicated reports

The duplicated reports will be removed next. That is performed (MISCE, DUPLI, REDSL) by searching pairs of collocated reports of the same observation types, and then checking the content of these reports. It may, for instance, happen that an airep report is formally duplicated by having a slightly different station identifier but with the observed variables inside these reports being exactly the same, or partially duplicated. The pair-wise checking of duplicates results in a rejection of some or all of the content of one of the reports.

10.6.5 Redundancy check

The redundancy check of the reports, together with the level selection of multilevel reports, is performed next for the active reports that are collocated and that originate from the same station (REDUN). In 3D-screening, this check applies to the whole observation time window. In 4D-screening (LSCRE4D=.true.), this check applies separately in each timeslot.

For **land synop and paob** reports, the report closest to the analysis time with most active data is retained, whereas the other reports from that station are considered as redundant and are therefore rejected from the assimilation (REDRP, REDMO). For **ship synop and dribu** observations the redundancy check is done in a slightly modified fashion (REDGL). These observations are considered as potentially redundant if the moving platforms are within a circle with a radius of 1° latitude. Also in this case only the report closest to the analysis time with most active data is retained. All the data from the multilevel **temp and pilot** reports from same station are considered at the same

time in the redundancy check (REDOR, SELEC). The principle is to retain the best quality data in the vicinity of standard levels and closest to the analysis time. One such datum will, however, only be retained in one of the reports. A wind observation, for instance, from a sounding station may therefore be retained either in a temp or in a pilot report, depending on which one happens to be of a better quality. A **synop mass** observation, if made at the same time and at the same station as the temp report, is redundant if there are any temp geopotential height observations that are no more than 50 hPa above the synop mass observation (REDSM).

10.6.6 Thinning

Finally, a horizontal thinning is performed for the airep, TOVS, SSM/I and SATOB reports. The horizontal thinning of reports means that a predefined minimum horizontal distance between the nearby reports from the same platform is enforced. For airep reports the free distance between reports is currently enforced to about 125 km. The thinning of the airep data is performed with respect to one aircraft at a time (MOVPL, THIAIR). Reports from different aircraft may however be very close to each other. In this removal of redundant reports the best quality data is retained as the preceding quality control is taken into account. In vertical, the thinning is performed for layers around standard pressure levels, thus allowing more reports for ascending and descending flight paths.

Thinning of TOVS, SSM/I and SATOB reports are each done in two stages controlled by THINN. For TOVS (THINNER), a minimum distance of about 70 km is enforced and, thereafter, a repeated scan is performed to achieve the final separation of roughly 250 km between reports from one platform. The thinning algorithm is the same as used for aireps except that for TOVS a different preference order is applied: a sea sounding is preferred over a land one, a clear sounding is preferred over a cloudy one and, finally, the closest observation time to the analysis time is preferred. A similar thinning technique is applied to SSM/I data and SATOB high-density data (THINNER).

The screening of SATOB data has been extended for atmospheric motion wind observations, including individual quality estimate. The quality information from the quality control performed by the producer at extraction time is appended to each wind observation. This Quality Indicator (QI) is introduced as an additional criterion in the thinning step; priority is given to the observation with the highest QI value.

TABLE 10.6 A SUMMARY OF THE CURRENT USE OF OBSERVATIONS IN THE 3D/4D-VAR DATA ASSIMILATION AT THE ECMWF.

Observation type	Variables used	Remarks
synop	u, v, ps (or z), rh	u and v used only over sea, in the tropics also over low terrain (< 150 m). Orographic rejection limit 6 hPa for rh, 100 hPa for z and 800 m for ps
airep	u, v, T	Not used in full resolution. Used only below 50 hPa
satob	u, v	Selected areas and levels. thinning of high-density winds.
dribu	u, v, ps	Orographic rejection limit 800m for ps
temp	u, v, T, q	Used on all reported levels. q only below 300 hPa. 10 m u and v used over land only in tropics over low terrain (< 150 m). Orographic rejection limit 10 hPa for u, v and T, and -4 hPa for q

TABLE 10.6 A SUMMARY OF THE CURRENT USE OF OBSERVATIONS IN THE 3D/4D-VAR DATA ASSIMILATION AT THE ECMWF.

Observation type	Variables used	Remarks
pilot	u, v	Used on or closest to standard pressure levels. 10 m u and v used over land only in tropics over low terrain (< 150 m). Orographic rejection limit 10hPa for u and v
tovs	Tb	For TOVS radiance usage see Table 0.3 and the chapter on 1C radiance processing.
paob	ps	Used south of 19°S. Orographic rejection limit 800 m for ps
scatt	u, v	Not used in full resolution. Used if SST is warmer than 273 K or if both observed and background wind less than 25 m s^{-1}
ssm/i	$tcwv$	Thinned, used over sea.

The variables are as in Table 10.1, with the addition that *Tb* stands for brightness temperature and *tcwv* stands for total cloud water vapour. The observation types are shortened by *synop* for synoptic surface observations, *airep* for aircraft reports, *satob* for satellite cloud track winds, *dribu* for drifting buoy reports, *temp* for radiosonde soundings, *pilot* for wind soundings, *tovs* for satellite temperature soundings, *paob* for pseudo observations of surface pressure and *scatt* for scatterometer reports

Apart from this thinning, the other observation dependent decisions involved by the screening of the scatterometer data come essentially from the application of a sea-ice contamination test from the model sea-surface-temperature analysis, using a minimum threshold of 273 K, and a high-wind rejection test with an upper-wind speed limit set to 25 m s^{-1} for the higher of the scatterometer and background winds (FGWIND).

In addition, the quality flag set in OBSPROC is also applied, and an extra quality control is done on the wind retrieval residual, or the so-called 'normalized distance to the cone'. After being implicitly checked for each report through the OBSPROC flag (different from 0 if its value is larger than 3), this quantity is tested in global average over the 6 hours of the analysis cycle for each of the 19 measurement nodes across the swath. All the data are then rejected in bulk if an excessive value is found for any node (more than 1.3 times the expected average) where the number of data taken into account is judged to be significant (i.e. more than 500). While the first check, performed locally, aims at avoiding geophysical effects not explained by the transfer function CMOD4, such as rain or sea-state effects in the vicinity of deep lows, this global quality control on the distance to the cone allows the detection of technical anomalies not reported in real time by ESA that are likely to affect the measurements in a correlated way and at larger scales. Such anomalies, which occur typically in the case of orbital manoeuvres, are missed by the preliminary test of the instrumental quality flag in OBSPROC.

10.6.7 A summary of the current use of observations

A summary of the current status of use of observations in the 3D-Var data assimilation is given in Table 10.6.

10.6.8 Compression of the ODB

After the observation screening roughly a fraction of 1/10 of all the observed data are active and so the compressed observation array for the minimization run only contains those data (COMPRES). The large compression rate is mainly driven by the number of TOVS data, since after the screening there are only 10–20% of the TOVS reports

left, whereas for the conventional observations the figure is around 40%. As a part of the compression, the observations are re-sorted amongst the processors for the minimization job in order to achieve a more optimal load balancing of the parallel computer.

10.7 A MASSIVELY-PARALLEL COMPUTING ENVIRONMENT

The migration of operational codes at the ECMWF to support a massively-parallel computing environment has set a requirement for reproducibility. The observation screening needs to result in exactly the same selection of observations when different numbers of processors are used for the computations. As mentioned earlier, in the observation screening there are the two basic types of decision to be made. Independent decisions, on one hand, are those where no information concerning any other observation or decision is needed. In a parallel-computing environment these decisions can be happily made by different processors fully in parallel. For dependent decisions, on the other hand, a global view of the observations is needed which implies that some communication between the processors is required. The observation array is, however, far too large to be copied for each individual processor. Therefore, the implementation of observation screening at the ECMWF is such that only the minimum necessary information concerning the reports is communicated globally.

The global view of the observations is provided in the form of a global 'time–location' array for selected observation types. That array contains compact information concerning the reports that are still active at this stage. For instance, the observation time, location and station identifier as well as the owner processor of that report are included. The time–location array is composed at each processor locally and then collected for merging and redistribution to each processor. After the redistribution, the array is sorted locally within the processors according to the unique sequence number. Thus, every processor has exactly the same information to start with, and the dependent decisions can be performed in a reproducible manner independently of the computer configuration.

The time–location array is just large enough for all the dependent decisions, except for the redundancy checking of the multilevel temp and pilot reports. This is a special case, in the sense that the information concerning each and every observed variable from each level is needed. Hence, the whole multilevel report has to be communicated. The alternative to this would be to force the observation clusters of the multilevel reports always into one processor without splitting them. In that case the codes responsible for the creation of the observation arrays for assimilation would need to ensure the geographical integrity of the observation arrays distributed amongst the processors. This is, however, not possible in all the cases, and the observation screening has to be able to cope with this. Currently, it is coded in such a way that only a limited number of multilevel temp and pilot reports, based on the time–location array, are communicated between the appropriate processors as copies of these common stations.

APPENDIX A

A.1 BAD REPORTING PRACTICE OF SYNOP AND TEMP REPORTS

The way the synoptic surface stations report mass observations (pressure or geopotential height) is considered as bad if the

- station altitude is above 800 m and station reports mean sea level pressure
- station altitude is above 800 m and station reports 1000 hpa level
- station altitude is above 1700 m and station reports 900 hpa level
- station altitude is below 300 m and station reports 900 hpa level
- station altitude is above 2300 m and station reports 850 hpa level

- station altitude is below 800 m and station reports 850 hpa level
- station altitude is above 3700 m and station reports 700 hpa level
- station altitude is below 2300 m and station reports 700 hpa level
- station altitude is below 3700 m and station reports 500 hpa level

The reporting practice is also considered as bad if the station reports 500 gpm, 1000 gpm, 2000 gpm, 3000 gpm or 4000 gpm level pressure, respectively, and station altitude is more than 800 m different from the reported level.

For temp geopotentials the reporting practice is considered as bad if the

- station altitude is above 800 m and station reports 1000 hpa level
- station altitude is above 2300 m and station reports 850 hpa level
- station altitude is above 3700 m and station reports 700 hpa level

A.2 REVISED BACKGROUND QUALITY CONTROL FOR SELECTED OBSERVATIONS

The background quality-control rejection limits are applied more strictly for some observation types than stated in Table 10.1. The special cases are the following ones

- airep wind observations with zero wind speed are rejected if the background wind exceeds 5 m/s
- for airep and dribu wind observations the rejection limit is multiplied by 0.5, and for pilot wind by 0.8
- for satob wind observations the rejection limit is multiplied by 0.1, except below 700 hPa level where it is multiplied by 0.2
- no background quality control is applied for scatt winds
- for dribu surface pressure observations the rejection limit is multiplied by 0.9, and for paob surface pressure by 0.7
- for airep temperature observations the rejection limit is multiplied by 1.6

A.3 USE OF ATMOSPHERIC MOTION WINDS

This appendix describes those parts of the ECMWF assimilation system which involves some special code for the AMW case, i.e. the data selection and the FG quality check. It refers to the operational status as from December 1996. A thinning procedure was introduced for high-density winds in Spring 1998.

A.3.1 Data selection

There are several model independent checks which AMW data have to pass in order to be considered for the assimilation process:

Check on longitude/latitude

- AMW must be within a circle of 55˚ from the sub-satellite point

Check on levels depending on the computational method

- WW CMW and WVMW must be above 400 hPa
- VIS CMW must be below 700 hPa
- IR CMW can be used at all levels.

Check on land/sea

- All AMW over sea are used
- AMW over land is not used north of 20ºN. .

- For Meteosat (0º mission) instead of 20ºN this threshold is 35ºN to allow usage of AMW over north Africa.
- For Meteost (63º mission) the use of AMW has been extended over Asia if above 500 hPa. This is restriced for longitudes east of 30ºE.
- AMW are blacklisted over the Himalayas as a precautionary measure.
- AMW over land south of 20ºN (35ºN for Meteosat) is used if above 500 hPa.

Check on satellite (35ºN for Meteosat) is used if above 500 hPa.

This is a temporary selection on certain channels or satellites. At present channels and satellite used are:

- METEOSAT cloud tracked winds with 90 min temporal sampling
- METEOSAT IR (not at medium level), VIS, WV
- METEOSAT HVIS, also at asynoptic times, only if $QI_2 \equiv 0$ (Automatic Quality Control ≡ PASSED
- GOES IR & WV (NOT at asynoptic times)
- GMS IR & VIS

A.3.2 Background quality check

The background quality check is based on a comparison of the AMW deviation from the background. Observed wind components are checked together. The AMW is flagged with j = 1 or 2 or 3 if this deviation squared is greater than a predetermined multiple ERRLIM * ZREJMOD of its estimated variance, as given by the following expression:

if [D2 > (sfg 2 + sobs 2) * ERRLIMj * ZREJMOD] then flag= j where D 2 = 1/2 (Du2 +D v2) with Du, Dv wind component deviations from background; sfg std of the background wind component error (mean for u and v); sobs std of the observation wind component error, 2 m s^{-1} for levels below 700 hPa included, 3.5 m s^{-1} at 500 hPa, 4.3 m s^{-1} at 400 hPa and 5 m s^{-1} for all levels above; ERRLIMj is 8 for j=1, 18 for j=2 and 20 for j=3. The value of ZREJMOD depends on the level of AMW and normally its value is:

- ZREJMOD = 0.2 for low level
- ZREJMOD = 0.1 for all others levels

A special check or asymmetric check is applied when the observed speed is more than 4 m s^{-1} slower than the background speed SPDfg. This check has a more restrictive rejection limit:

- ZREJMOD = 0.15 at low level
- ZREJMOD = 0.07 in the tropics
- ZREJMOD = 0.075 – 0.00125 * SPDfg all others
- ZREJMOD = 0.0 if SPDfg > 60 m s^{-1} (observation gets always flag j = 3)

When the data is passed to the following variational quality control its probability of being used depend on the

flag j. With flag j = 1 the data will be assimilated, with flag j =2 it will be given an intermediate probability and might be used or not and finally the analysis will reject all data with j = 3

Part II: DATA ASSIMILATION

CHAPTER 11 Analysis of snow

Snow depth is a model prognostic variable that needs to be analysed. Its analysis is performed in a module that is currently separated from the analysis of the atmosphere and of the soil wetness. This module includes also the sea-surface temperature, sea-ice fraction and screen-level temperature and relative humidity.

Table of contents

11.1 Organization

11.2 Snow-depth analysis

11.3 Technical aspects

11.1 ORGANIZATION

The snow analysis is a 3-D sequential analysis performed every 6 hours using a successive correction method. The **snow-depth background** S^b (units: m) is estimated from the short-range forecast of snow water equivalent W_s^b (units: m of water equivalent) and snow density ρ_s^b (units : kg m^{-3}):

$$S^b = \frac{1000 \times W_s^b}{\rho_s^b}$$

The **snow-depth analysis** S^a is performed using snow-depth observations and the snow-depth background field. If snow-depth observations are not available, the snow accumulation/melting is simulated from the model 6-hour forecast and a weak relaxation towards climatology is added. The **snow climate** is used to ensure the stability of the scheme and to give a seasonal snow trend in areas without any snow observations.

11.2 SNOW-DEPTH ANALYSIS

The observations S^o are snow depths from SYNOP reports. The background is S^b defined above. The analysis is done using a Cressman spatial interpolation:

$$S^a = S^b + \frac{\sum_{n=1}^{N} w_n S_n^o}{\sum_{n=1}^{N} w_n}$$

The weight function w_n is the product of functions of the horizontal distance r and vertical displacement h (model minus obs height) between the observation and analysis points:

$$w = H(r)\nu(h),$$

where

$$H(r) = max\left(\frac{r_{max}^2 - r^2}{r_{max}^2 + r^2}, 0\right)$$

and

$$\begin{aligned} \nu(h) &= 1 && \text{if } 0 < h \\ \nu(h) &= \frac{h_{max}^2 - h^2}{h_{max}^2 + h^2} && \text{if } -h_{max} < h < 0 \\ \nu(h) &= 0 && \text{if } h < -h_{max} \end{aligned}$$

The snow depth is preserved when the model height is above the observing station, but it is severely reduced below. The influence distances are set to $r_{max} = 250$ km and $h_{max} = 300$ m.

In addition to the preliminary quality control in the observation data base, the following checks are applied for each grid point :

- if $T_{2m}^b < 8°C$ only snow depth observations below 140 cm are accepted.
- this limit is reduced to 70 cm if $T_{2m}^b > 8°C$.
- snow-depth observations are rejected if they differ by more than 50 cm from the background.
- when only one snow-depth observation is available within the influence radius r_{max}, the snow depth increments are set to zero.
- snow-depth analysis is limited to 140 cm.
- snow-depth increments are set to zero when larger than $(160 - 16\,T_{2m}^b)$ mm (where T_{2m}^b is expressed in Celsius)
- snow-depth analysis is set to zero if below 0.04 cm
- if there is no snow in the background and in more than half of the observations within a circle of radius r_{max}, the snow-depth increment is kept to zero.

The analysis of snow depth is finally weighted with climatological values S^{clim} to provide the final analysis:

$$S^a = (1 - a)S^a + a\,S^{clim}$$

The relaxation coefficient a is set to 0.02 corresponding to a time scale of 12.5 days. The global snow depth climatology is taken from Foster and Davy (1988). Finally the snow density from the background is used to archive the analysis in terms of snow water equivalent :

$$W_s^a = \frac{\rho_s^b \times S^a}{1000}$$

The snow density is unchanged in the analysis process : $\rho_s^a = \rho_s^b$

Areas with permanent snow and ice (defined using the Global Land Cover Characterization product) are set to an arbitrary high value at each analysis cycle ($W_s^a = 10\,m$).

11.3 TECHNICAL ASPECTS

The snow analysis software is implemented as a branch of the more comprehensive surface and screen-level analysis (SSA) package. The other branches currently include two-metre temperature and relative humidity analysis, and also sea surface temperature and sea-ice fraction analyses. The program organization when performing snow analysis is roughly as follows:

- SSA
  - CONTROL_SSA
    - INISNW
      - SCAN_DDR
      - COORDINATES
      - GETFIELDS
      - SCAN_CMA
        - SCAN_OBS
          - LAND_OBS
          - INITIAL_REJECTION
          - REDUNDANT_OBS
    - SNOW_ANALYSIS
      - SUCSNW
      - SCAN_OBS
        - FG2OBS
      - SUCSNW
      - SNOW_FG
    - FDB_OUTPUT
    - PRINT_SUMMARY
    - PLOTDATA
    - FEEDBACK

The main program SSA calls CONTROL_SSA where most of the setup and namelist handling are done. Routine INISNW performs initialization of the actual snow analysis by sensing the size of the observation array file (CMA-file) in SCAN_DDR and generating latitudinal coordinates that stem from the model resolution in concern and zeros of the Bessel function.

After this, all input fields are read into memory in GETFIELDS. They consist of the snow water equivalent and snow density from the first-guess (6-hour forecast), 2 m temperature first guess, snow-depth climate (varies monthly with a linear temporal interpolation), land/sea mask and finally the orography in a form of the geopotential.

In SCAN_CMA observations are read into memory and a quick validity check of the non-applicable observations for this analysis is performed. Furthermore, the land/sea mask is calculated in LAND_OBS for the retained snow depth observation points.

Additional screening is done in INITIAL_REJECTION and in REDUNDANT_OBS. The former one sets up an internal table where all the observations which survived from the quick screening are placed with a minimum context information. This routine rejects some of the observations entered into the table due to inconsistencies.

The routine REDUNDANT_OBS removes time duplicates and retains the observations of the station in concern with the closest (and the most recent) to the analysis time. Since only synoptic observations are considered, slowly moving platform handling present in the REDUNDANT_OBS is not applicable to the snow analysis.

The actual snow analysis is performed under SNOW_ANALYSIS. The analysis technique is Cressman's successive correction method (routine SUCSNW). The structure functions are set to be separable in horizontal and verti-

cal directions. A special mountain region handling is performed, depending whether the datum or grid point is in the valley or at high altitudes, as explained before.

The snow-depth background (i.e. first guess) field is constructed from the model first-guess snow water equivalent and snow density. Once the snow-depth first guess field is present, it is used to calculate the first guess departure at snow-depth observation points. This increment is finally added to the snow depth fields at grid points producing the final snow depth output field, which is output in routine FDB_OUTPUT.

The accuracy of the analysis is estimated in PRINT_SUMMARY where some important statistics are summarized. The internal observation table can be printed if requested from PLOTDATA and an updated observation file for feedback purposes can be created in routine FEEDBACK.

The main logicals of the namelist NAMSSA are :

- L_SNOW_ANALYSIS : When set to TRUE, the snow analysis is performed.
- L_SNOW_DEPTH_ANA : When set to TRUE, the snow analysis is performed in snow depth (in opposition to snow water equivalent assuming a constant value of 250 kg m^{-2} for observed snow density).
- L_USE_SNOW_CLIMATE : When set to TRUE, a relaxation of the snow analysis towards a monthly climatology is performed with a time scale of 12.5 days (this constant is hard coded in SNOW_FG).
- L_USE_FG_FIELD : When set to TRUE the snow analysis is set to the first-guess value (no use of observations) and there is no relaxation to climatology.

CHAPTER 12 Land surface analysis

12.1 INTRODUCTION

Soil temperature and soil water content are prognostic variables of the forecasting system and, as a consequence, they need to be initialised at each analysis cycle. Currently the land surface analysis is performed every 6 hours and is decoupled from the atmospheric analysis. The absence of routine observations on soil moisture and soil temperature requires to use proxy data. The ECMWF soil analysis relies on SYNOP temperature and relative humidity at screen-level (2 m) available on the GTS (around 12000 reports over the globe are provided every 6 hours). Firstly, a screen-level analysis is performed for temperature and humidity. Secondly, the screen-level analysis increments are used as inputs to perform the analysis in the soil.

12.2 SCREEN-LEVEL ANALYSIS

12.2.1 Methodology

Two independent analyses are performed for 2 m temperature and 2 m relative humidity. The method used is a two-dimensional univariate statistical interpolation. In a first step, the background field (6 h or 12 h forecast) is interpolated horizontally to the observation locations using a bilinear interpolation scheme and background increments ΔX_i are estimated at each observation location i.

The analysis increments ΔX_j^a at each model grid-point j are then expressed as a linear combination of the first-guess increments (up to N values) :

$$\Delta X_j^{\mathrm{a}} = \sum_{i=1}^{N} W_i \times \Delta X_i \tag{12.1}$$

where W_i are optimum weights given (in matrix form) by :

$$(\mathbf{B} + \mathbf{O})\mathbf{W} = \mathbf{b} \tag{12.2}$$

The column vector $\mathbf{b}$ (dimension N) represents the background error covariance between the observation i and the model grid-point j. The $N \times N$ matrix $\mathbf{B}$ describes the error covariances of background fields between pairs of observations. The horizontal correlation coefficients (structure functions) of $\mathbf{b}$ and $\mathbf{B}$ are assumed to have the following form:

$$\mu(i, j) = \exp\left(-\frac{1}{2}\left[\frac{r_{ij}}{d}\right]^2\right) \tag{12.3}$$

where r_{ij} is the horizontal separation between points i and j and d the e-folding distance taken to 300 km (hard coded in subroutine OIINC).

Therefore :

$$B(i, j) = \sigma_{\mathrm{b}}^2 \times \mu(i, j) \tag{12.4}$$

with σ_b the standard deviation of background errors.

The covariance matrix of observation errors $\mathbf{O}$ is set to $\sigma_o^2 \times \mathbf{I}$ where σ_o is the standard deviation of observation errors and $\mathbf{I}$ the identity matrix.

The standard deviations of background and observation errors are set respectively to 1.5 K and 2 K for temperature and 5% and 10% for relative humidity. The number of observations closest to a given grid point that are considered to solve (12.1) is $N = 50$ (scanned within a radius of 1000 km). The analysis is performed over land and ocean but only land (ocean) observations are used for model land (ocean) grid points.

12.2.2 Quality controls

Gross quality checks are first applied to the observations such as $RH \in [2, 100]$ and $T > T^d$ where T^d is the dewpoint temperature. Redundant observations are also removed by keeping only the closest (and more recent) to the analysis time.

Observation points that differ by more than 300 m from the model orography are rejected.

For each datum a check is applied based on statistical interpolation methodology. An observation is rejected if it satisfies :

$$|\Delta X_i| > \gamma \sqrt{\sigma_o^2 + \sigma_b^2} \tag{12.5}$$

where γ has been set to 3, both for temperature and humidity analyses.

The number of used observations every 6 hours varies between 4000 and 6000 corresponding to around 40% of the available observations.

The final relative humidity analysis is bounded between 2% and 100%. The final MARS archived product is dew-point temperature that uses the 2 m temperature analysis T_a to perform the conversion :

$$T^d = \frac{17.502 \times 273.16 - 32.19 \times \Psi}{17.05 - \Psi} \tag{12.6}$$

with

$$\Psi = \log(RH_a) + 17.502 \times \frac{T_a - 273.16}{T_a - 32.19} \tag{12.7}$$

12.2.3 Technical aspects

The technical aspects are similar to the snow analysis (see Chapter 11) expect for the computation of the analysis increments obtained from the subroutine OIUPD instead of SUCSNW (Cressman interpolation).

Subroutine OISET selects the N closest observations from a given grid-point.

Subroutine OIINC provides the analysis increments from Equations (12.1) and (12.2), by first computing $\mathbf{q} = (\mathbf{B} + \mathbf{O})^{-1}\Delta\mathbf{X}$ (in subroutine EQUSOLVE - inversion of a linear system) which does not depend upon the position of the analysis gridpoint and then estimating $\mathbf{b}^T\mathbf{q}$ (in subroutine DOT_PRODUCT).

Most of the control parameters of the screen-level analysis are defined in the namelist NAMSSA:

1) C_SSA_TYPE : ‘t2m’ for temperature analysis and ‘rh2m’ for relative humidity analysis

2) L_OI : ‘ true’ for statistical interpolation and ‘false’ for Cressman interpolation

3) N_OISET : number of observations (parameter N)
4) SIGMAB : standard deviation of background error (parameter σ_b)
5) SIGMAO : standard deviation of observation error (parameter σ_o)
6) TOL_RH : Tolerance criteria for RH observations (parameter γ in Equation (12.5))
7) TOL_T : Tolerance criteria for T observations (parameter γ in Equation (12.5))
8) SCAN_RAD_2M(1) : Scanning radius for available observations (set to 1000 km)

12.3 SOIL ANALYSIS

The soil analysis scheme is based on an "local" optimum interpolation technique as described in *Mahfouf* (1991) and *Douville et al.* (2001). The analysis increments from the screen-level analysis are used to produce increments for the water content in the first three soil layers (correponding to the root zone) :

$$\Delta\theta = a \times (T_a - T_b) + b \times \left[100 \frac{e_s(T_a^d) - e_s(T_b^d)}{e_s(T_a)} \right] \tag{12.8}$$

and for the first soil temperature layer :

$$\Delta T = c \times (T_a - T_b) \tag{12.9}$$

The coefficients a and b are defined as the product of optimum coefficients α and β minimising the variance of analysis error and of empirical functions F_1 , F_2 and F_3 reducing the size of the optimum coefficients when the coupling between the soil and the lower boundary layer is weak.

$$\alpha = \frac{\sigma_\theta}{\phi \sigma_b^T} \left\{ \left[1 + \left(\frac{\sigma_a^{RH}}{\sigma_b^{RH}} \right)^2 \right] \rho_{T\theta} - \rho_{RHT} \rho_{RH\theta} \right\} \tag{12.10}$$

and

$$\beta = \frac{\sigma_\theta}{\phi \sigma_b^{RH}} \left\{ \left[1 + \left(\frac{\sigma_a^{RH}}{\sigma_b^{RH}} \right)^2 \right] \rho_{RH\theta} - \rho_{RHT} \rho_{T\theta} \right\} \tag{12.11}$$

with

$$\phi = \left[1 + \left(\frac{\sigma_a^T}{\sigma_b^T} \right)^2 \right] \left[1 + \left(\frac{\sigma_a^{RH}}{\sigma_b^{RH}} \right)^2 \right] - \rho_{RHT}^2 \tag{12.12}$$

where ρ_{xy} represents the correlation of background errors between parameters x and y .

The statistics of background errors have been obtained from a series of Monte-Carlo experiments with a single-column version of the atmospheric model where initial conditions for soil moisture have been perturbed randomly. They were obtained for a clear-sky situation with strong solar insolation. Empirical functions are aimed to reduce soil increments when atmospheric forecast errors contain less information about soil moisture. To obtain negligible

soil-moisture corrections during the night and in winter, F_1 is a function of the cosine of the mean solar zenith angle μ_M, averaged over the 6 h prior to the analysis time :

$$F_1 = \frac{1}{2}\{1 + \tanh[\lambda(\mu_M - 0.5)]\} \qquad \lambda = 7 \tag{12.13}$$

The optimum coefficients are also reduced when the radiative forcing at the surface is weak (cloudy or rainy situtations). For this purpose, the atmospheric transmittance τ_r is computed from the mean downward surface solar radiation forecasted during the previous 6 hours $\langle R_g \rangle$ as :

$$\tau_r = \left(\frac{\langle R_g \rangle}{S_o \mu_M}\right)^{\mu_M} \tag{12.14}$$

where S_o is the solar constant.

The empirical function F_2 is expressed as :

$$F_2 = \begin{cases} 0 & \tau_r < \tau_{rmin} \\ \dfrac{\tau_r - \tau_{rmin}}{\tau_{rmax} - \tau_{rmin}} & \tau_{rmin} < \tau_r < \tau_{rmax} \\ 1 & \tau_r > \tau_{rmax} \end{cases} \tag{12.15}$$

with $\tau_{rmin} = 0.2$ and $\tau_{rmax} = 0.9$.

The empirical function F_3 reduces soil moisture increments over mountainous areas :

$$F_3 = \begin{cases} 0 & Z > Z_{max} \\ \left(\dfrac{Z - Z_{max}}{Z_{min} - Z_{max}}\right)^2 & Z_{min} < Z < Z_{max} \\ 1 & Z < Z_{min} \end{cases} \tag{12.16}$$

where Z is the model orography, Z_{min} =500 m and Z_{max} =3000 m.

Furthermore, soil moisture increments are set to zero if one of the following conditions is fufilled:

1) The last 6 h precipitation exceeds 0.6 mm
2) The instantaneous wind speed exceeds 10 m s^{-1}
3) The air temperature is below freezing
4) There is snow on the ground

To reduce soil moisture increments over bare soil surfaces, the standard deviations and the correlations coefficients are also weighted by the vegetation fraction $C_v = c_L + c_H$, where low and high vegetation cover are defined in Chapter 7 of the Physics Documentation.

The statistics of forecast errors necessary to compute the optimum coefficients are given in Table 12.1.

The correlations have been produced from the Monte-Carlo experiments. The standard deviation of background error for soil moisture σ_θ is set to 0.01 m^3m^{-3} on the basis of ECMWF forecasts differences between day 1 and day 2 of the net surface water budget (precipitation minus evaporation minus runoff).

The standard deviation of analysis error σ_a is given by the screen-level analysis from :

$$\frac{1}{\sigma_a^2} = \frac{1}{\sigma_b^2} + \frac{1}{\sigma_o^2} \tag{12.17}$$

From the values chosen for the screen-level analyis $\sigma_a^T = 1.2\text{K}$ and $\sigma_a^{RH} = 4.47$ %.

Soil moisture increments $\Delta\theta$ are such that they keep soil moisture within the wilting point θ_{pwp} and the field capacity θ_{cap} values, i.e. :

- if $\theta_b < \theta_{cap}$ then $\theta_a = \min(\theta_{cap}, \theta_b + \Delta\theta)$
- if $\theta_b > \theta_{pwp}$ then $\theta_a = \max(\theta_{pwp}, \theta_b + \Delta\theta)$

Finally the coefficients providing the analysis increments are :

$$\begin{aligned} a &= C_v \times \alpha \times F_1 F_2 F_3 \\ b &= C_v \times \beta \times F_1 F_2 F_3 \end{aligned} \tag{12.18}$$

and

$$c = (1 - F_2) F_3 \tag{12.19}$$

The coefficient c is such that soil temperature is more effective during night and in winter, when the temperature errors are less likely to be related to soil moisture. This way, 2 m temperature errors are not used to correct soil moisture and soil temperature at the same time.

TABLE 12.1 STATISTICS OF BACKGROUND ERRORS FOR SOIL MOISTURE DERIVED FROM MONTE-CARLO EXPERIMENTS

Coefficient	Value
$\rho_{T\theta 1}$	–0.82
$\rho_{T\theta 2}$	–0.92
$\rho_{T\theta 3}$	–0.90
$\rho_{RH\theta 1}$	0.83
$\rho_{RH\theta 2}$	0.93
$\rho_{RH\theta 3}$	0.91
σ_b^T	1.25 K
σ_b^{RH}	9.5 %
ρ_{RHT}	–0.99

In the 12 h 4D-Var configuration, the soil analysis is performed twice during the assimilation window and the sum of the increments is added to the background values at analysis time.

REFERENCES

Douville, H., Viterbo, P., Mahfouf, J.-F. and Beljaars, A. C. M. (2001): "Evaluation of the optimum interpolation and nudging techniques for soil moisture analysis using FIFE data" *Mon. Wea. Rev.*, **128**, 1733-1756

Mahfouf, J.-F. (1991): "Analysis of soil moisture from near surface parameters : a feasibility study", *J. Appl. Meteor.*, **30**, 1534-1547

Part II: DATA ASSIMILATION

CHAPTER 13 Sea surface temperature and sea-ice analysis

THIS CHAPTER IS NOT YET AVAILABLE

Part II: DATA ASSIMILATION

CHAPTER 14 Reduced-rank Kalman filter

Table of contents

14.1 The modified change-of-variable

14.2 The Hessian singular vector calculation

14.1 THE MODIFIED CHANGE-OF-VARIABLE

From the point of view of the analysis, the reduced-rank Kalman filter (sometimes known as the "simplified" Kalman filter) consists of a modification to the change-of-variable. The control variable for the analysis is defined as

$$\chi = \mathbf{X}^{\mathrm{T}} \begin{bmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{F} & \mathbf{I} \end{bmatrix} \mathbf{X}\mathbf{L}\delta\mathbf{x} \tag{14.1}$$

where $\mathbf{L}$ is the static change of variable used in 3D- and 4DVar; $\mathbf{U}$ is a small, square, upper-triangular matrix; and $\mathbf{X}$ is an orthogonal matrix which rotates the control variable so that the leading few elements correspond to a subspace of interest, such as the space spanned by a set of singular vectors.

The background cost function corresponding to the change of variable defined by Eq. (14.1) is

$$J_{\mathrm{b}} = \delta\mathbf{x}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} \begin{bmatrix} \mathbf{E} & \mathbf{F}^{\mathrm{T}} \\ \mathbf{F} & \mathbf{I} \end{bmatrix} \mathbf{X}\mathbf{L}\delta\mathbf{x} \tag{14.2}$$

where $\mathbf{E} = \mathbf{U}^{\mathrm{T}}\mathbf{U} + \mathbf{F}^{\mathrm{T}}\mathbf{F}$.

The aim of the reduced-rank Kalman filter is to choose the matrices $\mathbf{E}$ and $\mathbf{F}$ to be good approximations to the corresponding sub-matrices from the true covariance matrix of background error (in the space defined by the matrices $\mathbf{L}$ and $\mathbf{X}$). The algorithm is described in detail by Fisher (1998) [ECMWF rd Tech Memo 260]. The modified change-of-variable is completely specified by two sets of vectors $\mathbf{s}_k$ and $\mathbf{z}_k$ for $k = 1 \ldots \mathrm{NSKFVECS}$. The vectors $\mathbf{s}_k$ define the subspace of interest, while for each k, the vector $\mathbf{z}_k$ defines the action of the inverse of the true covariance matrix of background error on $\mathbf{s}_k$.

The main namelist or the reduced rank Kalman filter is NAMSKF. This contains LSKF, the global switch for the modified change of variable; NSKFVECS, the number of pairs of vectors $\mathbf{s}_k$ and $\mathbf{z}_k$; and CINSKFY and CINSKFZ, which are the names of the files containing the vectors $\mathbf{s}_k$ and $\mathbf{z}_k$ respectively. The remaining variables in the namelist are used in the Hessian simplified vector calculation, and are described later.

The setup routine for the reduced-rank Kalman filter is SUSKF. After reading the namelist NAMSKF to find how many vectors to read, and from which files, SUSKF reads the vectors $\mathbf{s}_k$ into SKFROT (in yomskf) and $\mathbf{z}_k$ into ZZVECS. Both sets of vectors are read using the routine READVEC, which expects spectral fields of vorticity, divergence, specific humidity, temperature and LNSP. The fields must be on model levels, but may be of a different spectral truncation to that of the analysis increments, in which case they are truncated or padded with zeroes as appropriate. The $\mathbf{z}_k$ vectors are usually produced by the Hessian singular vector calculation described below, and

must be scaled by the reciprocal of the eigenvalue which is stored in the GRIB header for each field, and is returned in the optional argument of READVEC. The vectors are transformed to control vector space and then to a space with a Euclidean inner product by calls to CHAVAR, CHAVARINAD and LCZTOIFS.

Next, the orthogonal transformation represented by the matrix $\mathbf{X}$ in Eq. (14.1) is constructed. This transformation consists of a sequence of Householder matrices (i.e. matrices of the form $(\mathrm{I} - 2\mathrm{uu}^T)$, where $\mathbf{u}$ is a normalized vector). The transformation is constructed so that it sets to zero all but the first NSKFVECS elements of each of the vectors in SKFROT. The vectors which define the transformation are stored in SKFROT (in yomskf), overwtriting the previous content. The non-zero elements of the transformed vectors $\mathbf{X}\mathbf{s}_k$ are retained in the array ZU. The transpose of the orthogonal transform is applied to the vectors in ZZVEC. Since the Householder matrices are symmetric, the transpose of $\mathbf{X}$ is equivalent to applying the sequence of Householder matrices in reverse order.

At this stage, SKFROT contains the matrix $\mathbf{X}$ and ZU contains the matrix $\mathbf{S}$ in the following equation (equation 11 of Fisher, 1998)

$$\mathbf{Z} = \begin{bmatrix} \mathbf{E} & \mathbf{F}^T \\ \mathbf{F} & \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{S} \\ \mathbf{0} \end{bmatrix} \tag{14.3}$$

The matrix $\mathbf{S}$ is upper-triangular, so the elements of $\mathbf{E}$ and $\mathbf{F}$ may be determined by back-substitution. Following this calculation, $\mathbf{E}$ may not be exactly symmetric, due to rounding errors and the fact that the change-of-variable is not exactly invertible. It is explicitly symmetrized by replacing each element $\mathbf{E}_{ij}$ by $(\mathbf{E}_{ij} + \mathbf{E}_{ji})/2$.

The matrix $\mathbf{U}$ in Eq. (14.1) is the Cholesky square root of $(\mathbf{E} - \mathbf{F}^{\mathrm{T}}\mathbf{F})$. The decomposition requires that the latter matrix is positive definite. This is also the condition for positive definiteness of the background error covariance matrix implied by the change of variable. The Cholesky decomposition of the matrix $(\mathbf{E} - \mathbf{F}^{\mathrm{T}}\mathbf{F})$ is performed using the NAG routine F07FDF. If the decomposition fails due to an indefinite matrix, then the elements of the matrix $\mathbf{F}$ are reduced by a factor of 2 and the Cholesky decomposition is attempted again. A maximum of 4 attempts are made.

The elements of the matrix $\mathbf{U}$ are stored in the leading NSKFVECS elements of each vector of SKFMAT. The remaining elements contain the matrix $\mathbf{F}$.

The modified change-of-variable is applied in CVAR2, CVAR2IN, CVAR2AD, and CVAR2INAD. In the case of CVAR2, the code corresponds exactly to the change of variable defined in Eq. (14.1). The inverse, adjoint and inverse-adjoint of the change of CVAR2 are similar. The inverse makes use of the following equation, and uses back-substitution to apply the matrix $\mathbf{U}^{-1}$.

$$\begin{bmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{F} & \mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{U}^{-1} & \mathbf{0} \\ \mathbf{F}\mathbf{U}^{-1} & \mathbf{I} \end{bmatrix}. \tag{14.4}$$

14.2 THE HESSIAN SINGULAR VECTOR CALCULATION

The reduced-rank Kalman filter requires as input pairs of vectors which satisfy $\mathbf{z}_k = (\mathbf{P}^f)^{-1}\mathbf{s}_k$, where $\mathbf{P}^f$ is a flow-dependent approximation to the true covariance matrix of background error. Fisher (1998) [ECMWF rd tech memo 260] describes how pairs of vectors satisfying this requirement may be calculated during the course of a "Hessian singular vector" calculation. That is, a singular vector calculation in which the inner product at initial time is defined by the Hessian matrix of an analysis cost function. The vectors $\mathbf{s}_k$ are partially-evolved singular vectors. The vectors $\mathbf{z}_k$ are produced during the adjoint model integration.

The Hessian singular vector calculation is controlled using the namelist NAMLCZ. The global switch for the calculation is LJACDAV. Initial and final time inner products are selected by NLANTYPE. For NLANTYPE=6 or 8, the initial time inner product is defined by the analysis Hessian. Otherwise, the spectral inner product is used. For NLANTYPE=8 or 9, the final time inner product is defined by the background error covariance matrix. Otherwise, the energy inner product is used. The final time inner product may be restricted to a given geographic area using the variables ALAT1, ALON1, ALAT and ALON3. The optimization time is specified in units of timesteps by NJDSTOP. The maximum number of iterations to be performed by JACDAV (see below) is specifed by NITERL. The calculation will stop when this number of iterations has been performed, or when NEIGEVO singular vectors have been calculated.

The control-level routine is CUN3, which is called directly from CNT0. Much of the first part of CUN3 is concerned with initialization of observations, etc. which are needed by the analysis Hessian calculation. This part of the code is essentially the same as the corresponding part of CVA1, and will not be described here.

The Hessian singular vector calculation is unusual in that it explicitly changes the values of NCONF and NSTOP during the calculation. During the parts of the calculation which resemble an analysis, NCONF is set to 131 and NSTOP is set to zero. When the calculation resembles the ordinary singular vector calculation, NCONF is set to 601 and NSTOP is set to the optimization time for the singular vector calculation defined by NJDSTOP. The scalar product is also recalculated at various times in the code. In general, however, a spectral inner product is used for most of the calculation. When other inner products are required, they are calculated using an explicit weight matrix rather than by resetting the SCALP array.

After the initializations for the Hessian calculation, the trajectory for the singular vector calculation is created. A starting vector for the singular vector calculation is initialized, and the gradient for zero control variable is calculated and saved in VAZG (in yomcva).

The singular vectors are calculated by a call to NALAN2, which writes them to the file svifs. CUN3 reads the vectors and calls CNT3TL to give the singular vectors at final time. These are written to the file svevo. The vectors $\mathbf{s}_k$ required for the reduced rank Kalman filter are written during the tangent linear integration at a step specified by NWRISKF (in NAMSKF). The vectors $\mathbf{z}_k$ required by the reduced rank Kalman filter are produced by a call to CNT3AD. The switch LWRISKF (also in NAMSKF) determines whether the vectors $\mathbf{s}_k$ and $\mathbf{z}_k$ are written.

NALAN2 provides an interface to the main generalized eigenvector solver, JACDAV. The main task ofNALAN2 is to write the singular vectors to the file svifs, and to perform some diagnostics. JACDAV calculates the Hessian singular vectors as the solutions to the following generalized eigenvector equation

$$\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{M}\mathbf{x} = \lambda \mathbf{J}''\mathbf{x} \tag{14.5}$$

where $\mathbf{M}$ denotes the tangent linear model, $\mathbf{W}$ defines the inner product at optimization time, and $\mathbf{J}''$ is the Hessian of an analysis cost function. The algorithm requires operators which apply $\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{M}$ and $\mathbf{J}''$ to arbitrary vectors. These operations are represented in the code by the subroutines OPK and OPM respectively. Subroutine OPM calculates a Hessian-vector product as a finite difference between the gradient for the input vector, and the gradient for zero control vector which is in VAZG. The gradient for the input vector is calculated by a call to SIM4D.

JACDAV starts with an initial matrix $\mathbf{V}$ of KSTART vectors. The columns of $\mathbf{V}$ are orthonormalized with respect to the initial time inner product by a call to MORTHODM. That is, they are made to satisfy $\mathbf{V}^{\mathrm{T}}\mathbf{J}''\mathbf{V} = \mathbf{I}$. MORTHODM also applies OPK and OPM to the vectors.

Next, the following small ordinary eigenvalue problem is solved

$$\mathbf{V}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{WMVy} = \theta\mathbf{y} \tag{14.6}$$

The eigenvalues of this problem are the Ritz values (i.e. approximations to the eigenvalues) of Eq. (14.5). The residual, $\mathbf{r} = \mathbf{M}^{\mathrm{T}}\mathbf{WMVy} - \theta\mathbf{J''Vy}$, for the leading unconverged Ritz value is selected. The residual is orthogonal to the columns of $\mathbf{V}$ in the Euclidean sense. A vector which is orthogonal with respect to the Hessian is produced by first calculating an approximate solution to the linear equation $\mathbf{J''v} = \mathbf{r}$, and then explicitly orthonormalizing $\mathbf{v}$ by a call to MORTHODM. The linear equation is solved by a call to PCGBFGS, which implements a preconditioned conjugate gradient algorithm. The accuracy of the solution is determined by GREDBFGS (the required reduction in the norm of the error), and NINNER (the maximum number of iterations to be performed). A limited memory BFGS preconditioner is used, which is applied by the routine BFGS. The memory size (in pairs of vectors) is given by MEMBFGS.

Once the vector $\mathbf{v}$ has been determined, it is included as a new column of $\mathbf{V}$, and the process is repeated. It can be shown that if the linear equation $\mathbf{J''v} = \mathbf{r}$ is solved exactly, then the algorithm is equivalent to a Lanczos algorithm. If it is solved approximately, the algorithm resembles the Jacobi-Davidson method.

Part II: DATA ASSIMILATION

REFERENCES

Alduchov, O. A. and Eskridge, R.E., 1996: Improved Magnus form approximation of saturation vapor pressure. *J. Appl. Met.*, **35**, 601–609.

Andersson, E., Pailleux, J., Thépaut, J.–N., Eyre, J. R., McNally, A. P., Kelly, G. A. and Courtier, P., 1994: Use of cloud–cleared radiances in three/four–dimensional variational data assimilation. Q. *J. R. Meteorol. Soc.*, **120**, 627–653

Andersson, E., Haseler, J., Undén, P., Courtier, P., Kelly, G., Vasiljevic, D., Brankovic, C., Cardinali, C., Gaffard, C., Hollingsworth, A., Jakob, C., Janssen, P., Klinker, E., Lanzinger, A., Miller, M., Rabier, F., Simmons, A., Strauss, B., Thépaut, J.–N. and Viterbo, P., 1998: The ECMWF implementation of three dimensional variational assimilation (3D–Var). Part III: Experimental results. To appear in *Q. J. R. Meteorol. Soc.*

Andersson, E., 1997: Implementation of variational quality control. Proc. ECMWF workshop on "Non–linear aspects of data assimilation", Reading, 9–11 September 1996.

Andersson, E. and Järvinen, H., 1999: Variational quality control. *Q. J. R. Meteorol. Soc.*, 125, 697–722

Bartello, P. and Mitchell, H. L., 1992: A continuous three–dimensional model of short–range forecast error covariance. *Tellus*, **44A**, 217–235.

Blondin, C., 1991: 'Parametrization of land surface processes in numerical weather prediction'. Pp. 31–54 in *Land surface evaporation: measurement and parametrization*, T. J. Schmugge and J.–C. André, Eds., Springer–Verlag

Bouttier, F., Derber, J. and Fisher, M., 1997: The 1997 revision of the Jb term in 3D/4D–Var. ECMWF Tech. Memo. 238.

Buck, A.L., 1981: New equations for computing vapor pressure and enhancement factor. *J. Appl. Met.*, **20**, 1527–1532.

Buizza, R., 1994: Sensitivity of optimal unstable structures. *Q. J. R. Meteorol. Soc.*, **120**, 429–451

Cardinali, C., Andersson, E., Viterbo, P., Thépaut, J.–N. and Vasiljevic, D., 1994: Use of conventional surface observations in three–dimensional variational data assimilation. ECMWF Tech. Memo. 205.

Courtier, P., Thépaut, J.–N. and Hollingsworth, A., 1994: A strategy for operational implementation of 4D–Var, using an incremental approach. *Q. J. R. 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). I: Formulation. *Q. J. R. Meteorol. Soc.*, **124,** 1783-1807.

Eyre, J. R., 1989: Inversion of cloudy satellite sounding radiances by nonlinear optimal estimation. *Q. J. R. Meteorol. Soc.*, **115**, 1001–1037.

Eyre, J. R., 1991: A fast radiative transfer model for satellite sounding systems. ECMWF Tech. Memo. 176.

Eyre, J.R. 1992: A bias correction scheme for simulated TOVS brightness temperatures. ECMWF Research Dept. Tech. Memo. 186.

Eyre, J. R., Kelly, G. A., McNally, A. P., Andersson, E. and Persson, A.,1993: Assimilation of TOVS radiance information through one–dimensional variational analysis. *Q. J. R. Meteorol. Soc.*, **119**, 1427–1463.

Fisher, M. and Courtier, P., 1995: Estimating the covariance matrices of analysis and forecast error in variational data assimilation, ECMWF Tech. Memo. 220.

Freilich, M. S. and Anderson, D., 1997: Ambiguity removal and assimilation of scatterometer data. *Q. J. R. Meteorol. Soc.*, 123, 491–518

Foster , D.J. and R.D. Davy, 1988 : Global snow depth climatology. U.S. Air Force Environmental Tech. Applications Center/TN-88/006, 48 pp. [Available from National Climate Data Center, 151 Patton Avenue, Asheville, NC 28801]

Gaffard, C., Roquet, H., Hansen, B., Andersson, E. and Anderson, D., 1997: Impact of the ERS–1 scatterometer wind data on the ECMWF 3D–Var assimilation system. To appear in *Q. J. R. Meteorol. Soc.*

Gauthier, P. and Thepaut, J.-N., 2000: Impact of the digital filter as a weak constraint in the pre-operational 4D-Var assimilation system of Meteo-France. Submitted to Mon. Weather Rev.

Geleyn, J.–F., 1988: Interpolation of wind, temperature and humidity values from the model levels to the height of meaurement. *Tellus*, **40**, 347–351

Gerard, E, and R. Saunders, 1999: 4D-Var assimilation of SSM/I total column water vapour in the ECMWF model. ECMWF RD Tech Memo no.270.

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

Harris, B. 1997: A revised bias correction scheme for TOVS radiances. ECMWF Technical Memorandum No. ?.

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

Ingleby, N. B. and Lorenc, A. C., 1993: Bayesian quality control using multivariate normal distributions. *Q. J. R. Meteorol. Soc.*,**119**, 1195–1225.

Järvinen, H. and Undén, P., 1997: Observation screening and first guess quality control in the ECMWF 3D-Var data assimilation system. ECMWF Tech. Memo. 236.

Järvinen, H, S Saarinen and P Undén, 1996: User's guide for blacklisting. 51pp. Available on request from ECMWF, Shinfield Park, RG2 9AX, Reading, Berkshire, England.

Järvinen, H., E. Andersson and F. Bouttier, 1999: Variational assimilation of time sequences of surface observations with serially correlated errors. submitted to Tellus, 28pp. also, RD Tech Memo no.266.

Kelly, G. and Pailleux, J., 1988: Use of satellite vertical sounder data in the ECMWF analysis system. ECMWF Tech. Memo. 143.

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

Leidner, S. M., Hofman, R. N. and Augenbaum, J., 1999: SeaWinds Satterometer Real-Time BUFR Geophysical Data Product User's Guide Version 1.0, available from ECMWF and AER

Lönnberg, P, 1989: Developments in the ECMWF analysis system. ECMWF Seminar on Data assimilation and the use of satellite data. 5-9 September 1988, 75-119.

Lönnberg, P. and Hollingsworth, A., 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part II: The covariance of height and wind errors. *Tellus*, **38A**, 137–161.

Lönnberg, P and D Shaw, 1985: Data selection and quality control in the ECMWF analysis system. ECMWF

Workshop on The Use And Quality Control of Meteorological Observations, 6-9 November 1984, 225-254.

Lönnberg, P and D Shaw (Eds.), 1987: ECMWF Data Assimilation Scientific Documentation. Research Manual 1.

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

Lorenc, A. C., 1988: Optimal nonlinear objective analysis. *Q. J. R. Meteorol. Soc*., **114**, 205–240.

Lott ,F. and Miller M.J., 1997 : A new subgrid-scale orographic drag parametrization : its formulation and testing. *Q.J.R. Meteorol. Soc.*, **123**, 101-127.

Louis, J.–F., 1979: A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteorol*., **17**. 187–202.

Louis, J.–F., Tiedtke, M. and Geleyn, J.–F., 1982: ‘A short history of the PBL parametrization at ECMWF’. Pp. 59–80 in Proc. ECMWF Workshop on Planetary boundary layer parameterization, Reading, 25–27 November, 1981

Lynch, P., 1993: Digital Filters for Numerical Weather Prediction. HIRLAM Technical Report No 10.

Lynch, P., 1996: The Dolph-Chebyshev Window: A Simple Optimal Filter. Mon. Weather Rev., 125, 655–660

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

McNally, A. P. and Vesperini, M., 1996: Variational analysis of humidity information from TOVS radiances. Q. *J. R. Meteorol. Soc.*, **122**, 1521–1544.

Mahfouf, J.-F., Buizza, R., and Errico, R. M., 1997: Strategy for including physical processes in the ECMWF data assimilation system. In Proceedings of the ECMWF Workshop on non-linear aspects of data assimilation, Shinfield Park, Reading, RG2 9AX, 9–11 September 1996

Mahfouf, J.-F., 1998: Influence of physical processes on the tangent-linear approximation

Pailleux, J.,1990: ‘A global variational assimilation scheme and its application for using TOVS radiances’. Pp. 325–328 in Proc. WMO International Symposium on Assimilation of observations in meteorology and oceanography”, Clermont–Ferrand, France

Parrish, D. F. and Derber, J. C., 1992: The National Meteorological Center’s spectral statistical interpolation analysis system. *Mon. Weather Rev.*, **120**, 1747–1763.

Phalippou, L., 1996: Variational retrieval of humidity profile, wind speed and cloud liquid–water path with the SSM/I: Potential for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **122**, 327–355.

Phalippou, L. and Gérard, É., 1996: ‘Use of precise microwave imagery in numerical weather forecasting’. Study report to the European Space Agency. Available from ECMWF.

Rabier, F. and McNally, A., 1993: Evaluation of forecast error covariance matrix. ECMWF Tech. Memo. 195.

Rabier, F., McNally, A., Andersson, E., Courtier, P., Undén, P., Eyre, J., Hollingsworth, A. and Bouttier, F., 1998: The ECMWF implementation of three dimensional variational assimilation (3D–Var). II: Structure funtions. To appear in *Q. J. R. Meteorol. Soc*.

Rabier, F., Mahfouf, J.-F., Fisher, M., Järvinen, H. , Simmons. A., Andersson, E., Bouttier, F., Courtier, P., Hamrud, M., Haseler, J., Hollingsworth, A., Isaksen, L., Klinker, E., Saarinen, S., Temperton, C., Thépaut, J.–N. , Undén, P., and Vasiljevic, D., 1997a: Recent experimentation on 4D-Var and first results from a simplified Kalman filter. ECMWF Tech. Memo. 240.

Rabier, F., Thépaut, J.–N. and Courtier, P., 1997b: Four-dimensional variational assimilation at ECMWF. In Proceedings of the ECMWF Seminar on data assimilation, Shinfield Park, Reading, RG2 9AX, September 1996.

Saunders, R. W. and Matricardi, M., 1998: 'A fast forward model for ATOVS (RTATOV)'. Tech. Proc. 9th International TOVS Study Conf., Igls, Austria, 20–26 February, 1997. 11 pp.

Savijärvi, H., 1995: Error growth in a large numerical forecast system. *Mon. Wea. Rev.*, **123**, 212–221.

Simmons, A. J. and Burridge, D., 1981: An energy and angular momentum conserving vertical finite difference scheme and hybrid coordinate. *Mon. Weather Rev.*, **109**, 758–766.

Simmons, A. J. and Chen, J., 1991: The calculation of geopotential and the pressure gradient in the ECMWF atmospheric model: Influence on the simulation of the polar atmosphere and on temperature analyses. *Q. J. R. Meteorol. Soc.*, **117**, 29–58.

Stoffelen, A. 1999: Scatterometry. PhD Thesis, available from KNMI

Stoffelen, A. and Anderson, D., 1997: Ambiguity removal and assimilation of scatterometer data. *Q. J. R. Meteorol. Soc.*, **123**, 491–518.

Temperton, C., 1988: Implicit normal mode initialization. *Mon. Weather Rev.*, **116**, 1013-1031.

Temperton, C., 1989: Implicit normal mode initialization for spectral models. *Mon. Weather Rev.*, **117**, 436-451.

Thépaut, J.–N., Courtier, P., Belaud, G., and Lemaitre, G., 1996: Dynamical structure functions in a four-dimensional variational assimilation: a case study. Q. *J. R. Meteorol. Soc.*, **122**, 535–561.

Tiedtke, M., 1989 : A comprehensive massflux scheme for cumulus parametrization in large-scale models. *Mon. Weather Rev.*, **117**, 1779–1800.

Tomassini, M., LeMeur, D. and Saunders, R., 1997: Satellite wind observations of hurricanes and their impact on NWP model analyses and forecasts. To appear in *Mon. Weather Rev.*

Vasiljevic, D., Cardinali, C. and Undén, P., 1992: 'ECMWF 3D–Variational assimilation of conventional observations'. In Proc. ECMWF workshop on Variational assimilation with emphasis on three–dimensional aspects. Reading, 9–12 November 1992.

Wergen, W., 1987: Diabatic nonlinear normal mode initialisation for a spectral model with a hybrid vertical coordinate. ECMWF Tech. Report 59.