Geosci. Model Dev. Discuss., 7, 4291–4352, 2014
www.geosci-model-dev-discuss.net/7/4291/2014/
doi:10.5194/gmdd-7-4291-2014

Geoscientific Model Development Discussions
Open Access

Generalized Background Error covariance matrix model (GEN_BE v2.0)

G. Descombes[1], **T. Auligné**[1], **F. Vandenberghe**[2], **and D. M. Barker**[3]

[1]National Center for Atmospheric Research/MMM, Boulder, Colorado, USA
[2]National Center for Atmospheric Research/RAL, Boulder, Colorado, USA
[3]Met Office, Exeter, UK

Received: 13 May 2014 – Accepted: 12 June 2014 – Published: 10 July 2014

Correspondence to: G. Descombes (gael@ucar.edu)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

The specification of state background error statistics is a key component of data assimilation since it affects the impact observations will have on the analysis. In the variational data assimilation approach, applied in geophysical sciences, the dimensions of the background error covariance matrix (**B**) are usually too large to be explicitly determined and **B** needs to be modeled. Recent efforts to include new variables in the analysis such as cloud parameters and chemical species have required the development of the code to GENerate the Background Errors (GEN_BE) version 2.0 for the Weather Research and Forecasting (WRF) community model to allow for a simpler, flexible, robust, and community-oriented framework that gathers methods used by meteorological operational centers and researchers.

We present the advantages of this new design for the data assimilation community by performing benchmarks and showing some of the new features on data assimilation test cases. As data assimilation for clouds remains a challenge, we present a multivariate approach that includes hydrometeors in the control variables and new correlated errors. In addition, the GEN_BE v2.0 code is employed to diagnose error parameter statistics for chemical species, which shows that it is a tool flexible enough to involve new control variables. While the generation of the background errors statistics code has been first developed for atmospheric research, the new version (GEN_BE v2.0) can be easily extended to other domains of science and be chosen as a testbed for diagnostic and new modeling of **B**. Initially developed for variational data assimilation, the model of the **B** matrix may be useful for variational ensemble hybrid methods as well.

1 Introduction

Since improvements in data assimilation cannot be done without the best estimate of background error covariances (**B**), various meteorological operational centers such as the European Centre for Medium-Range Weather Forecast (ECMWF), the National Centers for Environmental Prediction (NCEP), or the UK Met office, continue to develop new algorithms, techniques and tools (Bannister, 2008a, b) to model **B** within a variational framework assuming that the underlaying probability errors are normally distributed. Statistics of the background error covariance matrix **B** are usually determined for a limited set of variables, called control variables that minimize the error covariance between variables. Then, several parameters need to be diagnosed to drive the series of operators that model **B**.

However, as more and more dataset observations coming from satellites, airplanes and ground stations become available in real time, there is a tendency to generalize data assimilation to a large set of sensors that involves more variables, which are present in geophysical numerical models. Necessities for cloud data assimilation (Auligné et al., 2011) have required to redesign the GEN_BE code by extending its capabilities to investigate and to estimate new error covariances. Originally, the GEN_BE code was developed by Barker et al. (2004, 2012) as a component of a three-dimentional variational data assimilation (3DVAR) method to estimate the background error of MM5 for a limited-area system. Since this initial version, various branches of code have been developed at NCAR and at the UK Met Office to address specific needs using different models such as (WRF) and the Unified Model (UM) on different data assimilation platforms such as the Weather Research Forecast Data Assimilation system (WRFDA) and the Gridpoint Statistical Interpolation system (GSI, Kleist et al., 2009). The framework of the GEN_BE code version 2.0 has been designed to merge this different efforts, to read input from different models and to provide output for different data assimilation platforms. The possibility to define the set of control variables and their covariance errors as an input should reduce considerably future developments of the code and unite them.

This document is organized as follows: the two first sections present the role of the background error covariance and how a series of different operators can model **B**. The third section describes the general structure of the code and gives key information to model **B** for a specific application. The estimation of the different parameters and their role in the data assimilation processes are discussed. It contains technical information that explains how to modify and extend the control variables and their error covariances. Section 4 presents results of a benchmark performed on two different systems of data assimilation (WRFDA and GSI) using different transforms involving the same set of five control variables (CV5) as defined for real time at NCEP on the rapid refresh domain. Finally, Sect. 5 presents an expansion of the control variable set on a test case that includes cloud hydrometeors in a multivariate approach (CV9). All the results presented in the differents sections were obtained from a numerical experiment with the WRF model involving an ensemble of 50 members over the CONUS domain at 15 km resolution. Figure 1 shows the extension of the WRF computational domain. Each member is a six hour forecast valid at 12:00z on 3 June 2012. The community system Data Assimilation Research Test (DART) was used to generate the ensemble (Romine et al., 2012). Appendix A contains results coming from a different dataset and it illustrates an application to chemical species.

2 Role of the background error covariance matrix in the variational data assimilation method

2.1 The variational method

The solution of three-dimensional variational data assimilation (3DVAR) is sought as the minimum of the following cost function (Courtier et al., 1994):

$$J(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}_\mathrm{b} - \boldsymbol{x})^\mathrm{T}\mathbf{B}(\boldsymbol{x}_\mathrm{b} - \boldsymbol{x}) + \frac{1}{2}[\boldsymbol{y}_\mathrm{o} - H(\boldsymbol{x})]^\mathrm{T}\mathbf{R}^{-1}[\boldsymbol{y}_\mathrm{o} - H(\boldsymbol{x})] \quad (1)$$

Where $\boldsymbol{x}$ is the state vector composed of the model variables to analyse, at every grid point of the 3-dimensional (3-D) model computational grid. $\boldsymbol{x}_\mathrm{b}$ is the background state vector, and usually provided by a previous forecast. $\boldsymbol{y}_\mathrm{o}$ is the vector of observations

GMDD
7, 4291–4352, 2014

and H called the observation operator, is a mapper from the gridded model variables to the irregularly distributed observation locations. **R** is the observational error covariance matrix. **B** is the background error covariance matrix. Note that exact knowledge of **R** and **B** would theoretically require the knowledge of the true state of the atmosphere at all times and everywhere on the model computational grid. This is not possible, and both matrices have to be estimated in practice. In general, the **R** matrix is assumed to be diagonal, i.e. uncorrelated observations, with empirically prescribed variances. Notice also that the dimension of the **B** matrix is the square of the 3-D model grid multiplied by the number of analyzed variables. For typical geophysical applications as in meteorology, the size of the **B** matrix, comprised of nearly $10^8 \times 10^8 = 10^{16}$ entries, is too large to be calculate explicitly nor be stored in present computer memories. As a result, the **B** matrix needs to be parameterized.

2.2 Modelling of the background error covariance matrix

2.2.1 Control variable transform

The cost function as defined in Eq. (1) is usually minimized after applying the change of a variable:

$$\delta \boldsymbol{x} = \mathbf{B}^{1/2}\boldsymbol{u}, \tag{2}$$

as it improves the conditioning and therefore accelerates the convergence. $\mathbf{B}^{1/2}$ is the square root of the background error covariance matrix. The variable $\boldsymbol{u}$ is called the control variable and the cost function becomes:

$$J(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} + \frac{1}{2}(\boldsymbol{d} - \mathrm{H}\mathbf{B}^{1/2}\boldsymbol{u})^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{d} - \mathrm{H}\mathbf{B}^{1/2}\boldsymbol{u}) \tag{3}$$

Where $\boldsymbol{d}$ is the innovation vector defined as $\boldsymbol{d} = (\boldsymbol{y}_{\mathrm{o}} - H(\boldsymbol{x}_{\mathrm{b}}))$ and it represents the difference between observations and their modeled values using a non-linear observation operator. H is the tangent linear operator.

GMDD
7, 4291–4352, 2014

2.2.2 Background errors covariance matrix modelled by a succession of operators

The square root of the **B** matrix as defined in Eq. (2) is decomposed to a series of sub-matrices, each corresponding to an elemental transform that can be individually modeled:

$$\delta \boldsymbol{x} = \mathbf{S}\mathbf{U}_{\mathrm{p}}\mathbf{U}_{\mathrm{v}}\mathbf{U}_{\mathrm{h}}\boldsymbol{u} \tag{4}$$

where:

- the matrix **S** is diagonal and composed of the standard deviations of the background errors,
- the $\mathbf{U}_{\mathrm{p}}$ matrix defines the cross-correlations between different control variables. In practice the unbalanced variables are obtained by removing the balance parts through linear regressions. The idea is that those new variables are less correlated with each other and so the corresponding cross-correlations entries in the matrix vanish,
- the matrix $\mathbf{U}_{\mathrm{h}}$ defines the horizontal auto-correlations for the $\boldsymbol{u}$ control variables. It is modeled through successive applications of recursive filters (Purser et al., 2003a and 2003b), which are affordable approximations of horizontal diffusion,
- the matrix $\mathbf{U}_{\mathrm{v}}$ defines the vertical auto-correlations for each of the $\boldsymbol{u}$ control variables. It is modeled by either homogeneous Empirical Orthogonal Functions (EOFs) or the applications of recursive filter.

Wu et al. (2002), Barker et al. (2004), and Michel and Aulign√© (2010) explain in more detail the methods used to construct these operators.

2.2.3 Motivation for an updated background error covariance model

The addition of new control variables in data assimilation systems requires the estimation of the error variances for each field, the calculation of the regression coefficients to derive uncorrelated (and unbalanced) control variables, and the estimation of the parameters to model vertical and horizontal correlated errors.

The structure of the GEN_BE code version 2.0 has been designed to perform those operations efficiently, to gather different methods to model **B** and make to additional developments easier. The new version of the code allows modeling a real time configuration of **B** like NCEP does using five control variables (CV5, e.g. Sect. 4), as well as, diagnosing and implementing a new model of **B**. The set of control variables has been expanded to include hydrometeors (CV9, e.g. Sect. 5) in order to assimilate cloudy radiances. An analysis increment for cloud hydrometeors cannot exist without any control variables representing them. The multivariate approach is used to balance them along other variables. Finally, statitics of chemisty species to model **B** have been evaluated in Appendix A. These different experiments show the possibility to use the GEN_BE code as a diagnostic tool, or to implement new modelling of background errors.

3 GEN_BE code version 2.0

The general structure of the code has been designed to split the input, output, and algorithms in independent stages. In the version 2.0 of the code, the five steps, from stage 0 to 4, that lead to modeling of the error covariance matrix become independent of the choice of control variables and model input (Fig. 2). The namelist input file, defined by the user, drives these different stages to determine the parameters of the physicals transforms $\mathbf{U}_\mathrm{p}$, $\mathbf{U}_\mathrm{v}$ and $\mathbf{U}_\mathrm{h}$ as shown in Sect. 2.2.2. The version 2.0 of the code includes more physics options and flexibility has been added making all the algorithms in the different stages independent of the choice of control variables and model input.

The new features of the code help to experiment and implement new modeling of **B** on different data assimilation systems.

3.1 Five stages to generate the background error covariance statistics

Stages 0 and 1 compute the raw model perturbations of the analysis variables that are used as a proxi of modeled background error. Stage 2 calculates the covariance between the control variables by estimating their regression coefficients. Stages 3 and 4 estimate the necessary parameters to spread out the information in data assimilation processes using Empirical Orthogonal Functions (EOFs) and recursive filters.

3.1.1 Sampling and binning (stage 0 and stage 1)

Since the background error covariance matrix is a statistical entity, samples of model forecasts are required to estimate the associated variances and correlations. Traditionally, two distinct techniques are used and available in stage 0 to compute the perturbations:

- Differences between two forecasts valid at the same time but initiated at different dates (time lagged forecast, e.g. 24-minus 12 h forecasts), can be used to represent a sample of model background errors. This is an ad hoc technique, called the NMC (named for the National Meteorological Center) method (Parish and Derber, 1992), which has been widely used in operational centers where large databases of historical forecasts are available.

- Background error statistics can be evaluated from an ensemble of previous forecasts valid at the same time. This method tends to be more accurate because it better represents the background error of the day, rather than a climatological error, as with the NMC method. However, more computational resources are required to run an ensemble simulation.

Pereira and Berre (2006) highlight the consequences of the evaluation of perturbations using the NMC method vs. an ensemble approach (called ensemble of the day, D-ensemble). The authors point out that the NMC method tends to underestimate the background errors in data-sparse areas (when the forecast comes from cycling analysis). They show that correlation length scales as described by Daley (1991) are smaller in D-ensemble methods compared to NMC. Table 1 summarizes the different options to compute these raw perturbations.

Some kind of spatial averaging needs to be performed to increase the number of samples as the number of perturbations available are limited and to reduce the dimensional of statistical output parameter. The different options available for this technique, referred as binning, are described in the Table 2 and can be setup in the namelist input file (Table 3). Their goals are to add heterogenity and anisotropy in the application of the operators $\mathbf{U}_\mathrm{v}$, $\mathbf{U}_\mathrm{h}$ and $\mathbf{S}$ to specify natural phenomena more accourately. Options *bin_type* = 2, 3, 4 compute statitics across the zonally averaged ensemble perturbations, to create a latitude-dependent correlation function usually used for large and global domains where latitude flow dependency occurs. Morever, binning can become tricky when it is applied to represent meteorological events at small scales. For example, the statistics of hydrometeors, as cloud liquid water, which are characterized by a high spatial and temporal variability can be skewed if, at a given grid point, only few members of the D-ensemble indicate the presence of clouds. For that reason, it may be preferable to use a cloud mask in the hydrometeor cloud calculations, which is referred as “geographical binning”. Montmerle and Berre (2010) and Michel et al. (2011) show improvements using rain mask (option 7) with the vorticity and divergence control variables to characterize convection events.

For this reason, the GEN_BE code has been modified to facilitate the introduction of new binning options for specific applications. All the algorithms of the different stages from 1 to 4 do not make any specific assumption on the binning option used. Stage 1 creates the NetCDF file bin.nc that contains all the information to define the binning option. The *bin_type* variable encapsulates this information and makes it available for

GMDD
7, 4291–4352, 2014

all the different programs. This modification simplifies the introduction of a new binning option, as it needs to be defined just once in the da_create_bins FORTRAN routine. In case of a new dynamical or geographical mask, the developer has to introduce the method to update the mask of binning in the routine update_dynamical_mask of the module io_input.f90. The binning definition is an important component in the model of **B** as it is applied in the following stages, especially in stage 2 for the balance operator.

3.1.2 Balance through linear regressions (stage 2)

The estimation error for one analysis variable may affect the value of another if they are correlated. The simplest way to model them is to use linear regression. Firstly, the regression coefficients between variables can be calculated in stage 2 using two differents methods. The original GEN_BE code first inverts the variance matrix and then directly calculates the regression coefficient as a product. The NCEP method, that uses a Choleski decomposition for the GSI system, has been merged into version 2.0. A broad set of options for binning, described in Table 2, can be applied to specify and differentiate statistical covariance errors.

Secondly, linear regressions are performed to derive uncorrelated control variables and then remove the balanced part for each other variable as shown in Sect. 3.2.2. This part achieves the $\mathbf{U}_\mathrm{p}$ transform: it models correlations between variables and allows to transform the matrix as a diagonal bloc in the control (uncorrelated) space. The structure of the code, and specially stage 2, have changed significantly. One of the goals is to have a code flexible enough to diagnose model background errors for a large set of different control variables, shown in Table 4, that models specific covariance errors by using a namelist input file. Adding new control variables or defining new error covariances is straight forward, as all the algorithms do not depend of the control variables.

Stage 2, which removes the linear cross-covariances between control variables (the balanced part), is the preliminary step before estimating the vertical and horizontal auto-correlation parameters for each control variable.

GMDD
7, 4291–4352, 2014

3.1.3 Estimation of the vertical correlation and the variance (stage3)

After calculating the vertical auto-covariance matrix (VACM), two techniques are currently available in stage 3 to compute the parameters useful to model the mean vertical auto-correlation transform ($\mathbf{U}_\text{v}$). The first method diagonalizes the VACM, computing eigenvectors (aka. EOFs) and eigenvalues. The variable is re-written in this new base for each EOF. Stage 4 will later evaluate a length scale for each EOF mode. The vertical transform occurs with the change of base EOF-physical space and the variances are represented by the eigenvalues. The second method estimates, a vertical length scale from the vertical auto-correlation matrix directly in the physical space, to propagate the increment via recursive filters. The diagnostic of the vertical length scale (L_v) comes from the Daley's formula (1991, p. 110) for a one dimension homogeneous and isotropic case:

$$L_\text{v} = \sqrt{\frac{1}{\nabla^2 \rho(0)}} \tag{5}$$

with $\rho(0)$ the correlation taken at the origin.

Doing a Taylor development of $\rho(\delta \boldsymbol{x})$ at the second order and subsisting $\rho(0)$ in Eq. (5), it results in:

$$L_\text{vp} = \frac{\delta \boldsymbol{x}}{\sqrt{2[1 - \rho(\delta \boldsymbol{x})]}} \tag{6}$$

We named L_vp the vertical length scale with the parabolic approximation. If the correlation can be approximated at the origin by a Gaussian function as following,

$$\rho(\boldsymbol{x}) = \exp\left(-\frac{\delta \boldsymbol{x}^2}{2L_\text{vg}}\right)$$

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |
GMDD
7, 4291–4352, 2014
Generalized Background Error covariance matrix model (GEN_BE v2.0)
G. Descombes et al.

the length scale expression for the Gaussian approximation can be deduced from Eq. (6):

$$L_{\mathrm{vg}} = \frac{\delta \boldsymbol{x}}{\sqrt{-2 \ln \rho(\delta \boldsymbol{x})}} \tag{7}$$

Pannekoucke et al. (2008) studied the sensitivity of sampling errors of these formulae and shows that the Gaussian and the parabolic approximation give close results. Furthermore, the vertical length scale can be binned and be handled by inhomogeneous recursive filters, which is not the case with the vertical transform defined by the EOF decomposition. The local eigenvectors and eigenvalues computed by bins are mostly useful as a diagnostic, but not for data assimilation. Finally, the user can decide, through the namelist described in Table 5, to utilize the same binning as defined by *bin_type* for the regression coefficient if the flag *global_bin* is set to false. In the reverse case, the global binning is applied by vertical level (equivalent to *bin_type* = 5). Stage 3 can be applied independently and simultaneously to each variable as well as stage 4 to determine the parameters of the $\mathbf{U}_\mathrm{h}$ transform.

3.1.4 Estimation of the horizontal correlation (stage 4)

Horizontal auto-correlations can be computed for each control variable at each grib point. Figure 3 shows a diagnostic of correlation for a few selected points of the WRF computational domain at level 5 (~ 500 m above the ground). The stream function (Fig. 3a) and potential velocity control variables have larger and more isotropic spatial correlations while the temperature (Fig. 3b) and the humidity (Fig. 3c) control variables show smaller and anisotropic correlations at different locations. Hydrometeors mixing ratio show even more local structures due to their sparse repartition on the horizontal and the vertical (Fig. 3d).

In stage 4, we estimate length scales averaged by vertical level or EOF mode for a field analysis in a 2-D plan. It represents the radius of influence, calculated in grid point, around the position of an observation and is an input parameter for recursive filters to spread out horizontally the increment ($\mathbf{U}_\mathrm{h}$). The different options available, as described below, are also contained in Table 5.

The first method (*ls_method* = 1) employs a distribution function to fit the correlation for a 2-D field by EOF mode or by vertical level as explained in Sect. 3.1.2. If the parameter *horizfunc* = *"gauss"* is selected, the length scale L is determined by solving the Eq. (8):

$$\rho(r) = \rho(0) \cdot \exp\left(-\frac{r^2}{8 \cdot L}\right) \tag{8}$$

where ρ is the correlation.

If the parameter *horizfunc* = *"SOAR"* is selected, the length scale is modeled by fitting the pseudo correlation of a field through a soar function. The length scale L is determined by solving the Eq. (9):

$$\rho(r) = \rho(0) \cdot \left(1 + \frac{r}{L}\right) \cdot \exp\left(-\frac{r^2}{L}\right) \tag{9}$$

However, as this procedure proved both computationally expensive and prone to sampling errors, a second option (*ls_method* = 2) based on the ratio of the variance of a field (φ) and the variance of its laplacian, has been added:

$$L = \frac{8 \cdot \mathrm{Variance}(\phi)}{\mathrm{Variance}(\nabla^2 \phi)} \tag{10}$$

The formula Eq. (10) was used by Wu et al. (2002) and is similar to the diagnostic of Peireira and Berre (2006), which was analyzed in Panckoucke et al. (2008).

Usually, the horizontal length scale is uniform over a vertical model level; at best it can be statistically binned. Inhomogeneous recursive filters, as implemented in the GSI, are able to handle the binned length scale generated with the flag *global_bin* set to false. In this case, the increment is spreaded out with a length scale according to the bin class of each grid point. Morever, inhomogeneous recursive filters could produce poor results if the binned length scale is not smooth enough. This last criterion is hard to define as it depends directly on the length scale values themselves over the different bins (Pannekoucke et al., 2008; Michel and Auligné, 2010). Otherwise, if the flag *global_bin* is set to true (equivalent to *bin_type* = 5), homogeneous recursive filters are able to handle a unique length scale defined by model vertical level, or EOF mode.

3.2 Framework of GEN_BE code version 2.0

The framework of the GEN_BE version 2.0 code generalizes the use of the five stages to a larger set of potential control variables. Input and output format have been chosen to allow flexibility for input coming from different numerical models and for output that is potentially useful for different data assimilation systems.

3.2.1 FORTRAN code and input/output

New FORTRAN modules have been developed to generalize the calculation of the error covariance matrix from different input models and for new controls variables. Table B1 contains a complete list of these modules and their contents. All the algorithms from stage 1 to stage 4 are now independent of the choice of control variables and driven by a unique namelist file, called namelist.input, and read by the FORTRAN module configure.f90. Flexibility has been added for future experiments. Only few modifications are needed in stage 0 to add new control variables. The FORTRAN module io_input_models.f90 converts the standard variables from a given model to the analysis variables. The interface is already made with the WRF model. Only the FORTRAN module io_input_model.f90 needs to be updated to implement new model input and to run the different stages.

GMDD
7, 4291–4352, 2014

The NetCDF format has been chosen to improve robustness and flexibility in the input and output of the different stages as shown in Table B2. The final NetCDF output file be.nc contains all the information needed for variational a data assimilation system, as shown in Table B3. Several converters from NetCDF format to binary have been developed to ensure backward compatibility to another data assimilation system. A binary file be.dat can be generated for the WRFDA application using the program gen_be_diags.f90 and a binary file be_gsi.dat can be created for GSI using the converter gen_be_nc2gsi.f90.

3.2.2 Generalized control variables and error covariances

The framework of the GEN_BE code version 2.0 allows to use a broad set of control variables, as shown in Table 4, and the $\mathbf{U}_\mathrm{p}$ transform is fully driven by the namelist file. For example, NCEP operates for its real time operations a physical transformation along with pressure, temperature, water vapor mixing ratio and wind components are converted into stream function, unbalanced velocity potential, unbalanced temperature, unbalanced surface pressure and relative humidity (univariate). The part of the namelist file presented in Table 6 summarizes the content of this transform precising through the parameters *covar* that the unbalanced part of the velocity potential, the temperature, and the surface pressure are calculated removing their balanced part with the stream function. This model of $\mathbf{U}_\mathrm{p}$ is referenced as CV5 in this document and benchmark results are shown Sect. 4, as the same $\mathbf{U}_\mathrm{p}$ transform can be combined with different other series of operators to model **B**.

Recent studies that propose a new definition for a multivariate humidity balance approach show potential improvements for cloud data assimilation. Diagnostics such as vertical cross-covariance or vertical cross-correlation can be done by using stage 2. Figure 4 displays the correlated errors between temperature with specific humidity (Fig. 4a) and relative humidity (Fig. 4b). While the errors between temperature and specific humidity are highly correlated close to saturation, they become anti-correlated for a drier and mixed atmosphere (Ménétrier and Montmerle, 2011). At saturation, these statistics probably rely on processes of condensation or precipitation when the released latent flux warms the atmosphere. For a winter test-case where stratiform-type precipitation is predominant, Carron and Fillon (2010) use psi regression in addition of mass field temperature (t) and ps to characterize balance operator of specific humidity in the precipitating area. They explain that imbalance in precipitating areas comes from condensation processes and probably from dynamical processes that deviate from geostrophic balance. The balance applied for specific humidity is represented by Eq. (11):

$$\begin{aligned} \mathrm{qs_u}(i,j,k) = \mathrm{qs}(i,j,k) &- \sum_{l=1}^{N_k} \alpha_{\mathrm{qs,\,psi}}(b,k,l)\mathrm{psi}(i,j,l) \\ &- \sum_{l=1}^{N_k} \alpha_{\mathrm{qs},t_\mathrm{u}}(b,k,l)t_\mathrm{u}(i,j,l) - \alpha_{\mathrm{qs,\,ps_u}}(b,k)\mathrm{ps_u}(i,j) \end{aligned} \tag{11}$$

where the α regression coefficients are estimated to minimize the correlations with other variables, the triplet (i,j,k) are the indexes of the grid point, the index b defines the binning class for this position and N_k is the total number of vertical levels. The variable $\mathrm{qs_u}$ is called the unbalanced part of qs.

Ménétrier and Montmerle (2011) show the benefit of balancing the specific humidity control variable only with the mass fields of temperature and surface pressure for fog data assimilation purposes. As the dynamic control variable of vorticity and divergence do not explain statically the presence of fog, they are not involved in the balance humidity operator. The geographical mask used is based on the diagnositc of nebulosity. While the balance of the specific humidity and temperature highly depend on the relative humidity rate, the correlation of relative humidity itself and temperature remains mainly negative for a dry and humid atmosphere. A univariate version of the relative humidity balance would not be able to handle increments of cloud hydrometeors from data assimilation of satellite cloudy radiances. Thus, the NCEP $\mathbf{U}_\mathrm{p}$ transform used in real time, is modified. In a first approach, relative humidity is balanced with the mass

GMDD
7, 4291–4352, 2014

fields and does not include dynamic variables such as the stream function and potential velocity as in the following Eq. (12a):

$$\mathrm{rh_u}(i,j,k) = \mathrm{rh}(i,j,k) - \sum_{l=1}^{N_k} \alpha_{\mathrm{rh},t_\mathrm{u}}(b,k,l)t_\mathrm{u}(i,j,l) - \alpha_{\mathrm{rh,ps_u}}(b,k)\mathrm{ps_u}(i,j) \tag{12a}$$

This multivariate $\mathbf{U}_\mathrm{p}$ transform can be defined using the same code without any new development. In this case, the line describing covariances with the humidity becomes: *covar5* = 0, 0, 1, 1, 0, 0, 0, 0, 0, 0. In the meantime, the control variables are expanded to include the mixing ratios of cloud water condensate (q_cloud), rain (q_rain), ice (q_ice) and snow (q_snow). The hydrometeors q_cloud and q_ice are balanced with respect to relative humidity as their presence or absence is directly related to the humidity rate. The regression coefficients can be computed directly, without any assumptions (Fig. 5a and b), or filtered to take into account only the perturbations that represent the transition of a non-cloudy to a cloudy area (Fig. 5c and d). This choice is made to intensify the statistical relationship of the statistical balance to be able to remove misplaced clouds, or to create clouds. However, such filter may overestimate the vertical correlation around a given vertical model level. For this reason, the line *covar6* = 0, 0, 0, 0, 1, 0, 0, 0, 0, 0 represented by Eq. (12b) can be replaced by the line *covar6* = 0, 0, 0, 0, 2, 0, 0, 0, 0, 0 represented by the Eq. (12c). In this case, only the diagonal terms of the regression coefficient are calculated and the increment is spreaded out by the recursive filters.

$$q_{\mathrm{cloud_u}}(i,j,k) = q_\mathrm{cloud}(i,j,k) - \sum_{l=1}^{N_k} \alpha_{q_{\mathrm{cloud,rh_u}}}(b,k,l)\mathrm{rh_u}(i,j,l) \tag{12b}$$

$$q_{\mathrm{cloud_u}}(i,j,k) = q_\mathrm{cloud}(i,j,k) - \alpha_{q_{\mathrm{cloud,rh_u}}}(b,k)\mathrm{rh_u}(i,j,k) \tag{12c}$$

Similar balance is applied to q_ice. Table 7 summarizes the definition of this balance operator called CV9 and Sect. 5 contains the result of experiments.

3.2.3 Installation, compilation, set up and vizualization

The GEN_BE code version 2.0 is a standalone package that can be installed on different UNIX/LINUX systems. It has been tested with Intel FORTRAN compiler, Portland Group FORTRAN compiler, and GNU FORTRAN compiler. It requires compilation of NetCDF libraries. First, a configuration file needs to be created using the command *configure* in the main directory of the code. Then, the compilation, is launched by the command *compile gen_be*. Once successfuly completed, the executables are created in the src directory.

Korn-shell scripts available in the scripts directory allow to setup the experiment. The wrapper script, named gen_be_wrapper.ksh, sets up some global variables and launches the main script gen_be.ksh. The user needs to setup most of the other options that determine the way to model the **B** matrix in the namelist.template file. The gen_be.ksh script fills out the initial date and the final dates, the frequency of date available (interval) coming from the global variables setup in the wrapper script and in the gen_be_set_defaults.ksh script, and generates a namelist.input file in the working directory during the first stage. The namelist.input file contains four main parts presented Tables 1, 3, 5, and 6. Each stage can then be run successively by setting the environmental variable RUN_GEN_BE_STAGE [0, 1, 2, 3, 4] to true in the gen_be_set_defaults.ksh script. The output of the stages 0, 1, 2, 3 and the be.nc file can be easily visualized with existing tools (Ncview, NCL, Python, MatLab).

4 Comparison of different modelling of B for two data assimilation systems

The real time configuration of **B** used at NCEP, that includes the set of five control variables (CV5) and their covariance errors (Table 6), can be used in both GSI and WRFDA data assimilation systems. In the following, we present first the different parameters that define the vertical transform $\mathbf{U}_v$ by using EOF decomposition for WRFDA ($\mathbf{B}_{eof}$) and by using recursive filter for GSI ($\mathbf{B}_{rcf}$). Finally, the results of data assimilation obtained with

$\mathbf{B}_{eof}$ and $\mathbf{B}_{rcf}$, determined for the CONUS domain at 15 km of resolution, are compared with the background error $\mathbf{B}_{nam}$ that operates in real time on the rapid refresh domain using GSI. $\mathbf{B}_{nam}$ statistics are based on NAM forecast at 0.1° of resolution and using the NMC method.

4.1 Statistics of the background error covariance matrix for different transforms

4.1.1 Decomposition by EOF and length scale

If the EOF decomposition is used, the eigenvectors model the vertical transform ($\mathbf{U}_v$) and the associated eigenvalues represent the variance. The length scale is estimated in the EOF space and represents the horizontal transform ($\mathbf{U}_h$). In the data assimilation process, the eigenvalues weight the analysis increment and the recursive filter first spreads out the information in the EOF space according to length scale value. Then, the transformation from EOF mode to physical space spreads out the information vertically. The first five eigenvectors are shown Fig. 6 for the control variables (CV5) and Fig. 7 shows the associated eigenvalues. 99 % of the variance of the stream function and the potential velocity are represented by the first ten and twenty modes respectively, while more than 30 modes are useful for temperature and relative humidity.

The horizontal length scales, estimated by Eq. (10), are presented in Fig. 8. The stream function and the potential velocity have the largest length scale value reaching 39 grid points for the first EOF mode, i.e. close to 600 km. While, the unbalanced temperature length scale has a strong variation for the three first EOF passing approximately from 9 to 2 grid points and from there, slightly decreases from 30 km to reach 15 km for the last EOF mode. Relative humidity is decreasing more monotonically from approximately 30 km to 15 km as a function of the EOF mode. The unbalanced temperature and the relative humidity have a small length scale, which means that they have more local features represented by a small radius of influence. Thus, the analysis increment from these variables will remain closer to the observation.

GMDD
7, 4291–4352, 2014

4.1.2 Horizontal and vertical length scales defined in physical space

The horizontal correlation is modeled by the application of recursive filters based on the estimation of the horizontal length scale solving Eq. (10), applied by vertical model level for each variable, as shown in Fig. 9. The diagnostic of horizontal length scale shows similar characteristics to the one presented on the precedent paragraph performed by EOF modes. The length scales of the stream function and the potential velocity control variables have the largest values above 150 km for all the vertical model levels, while the length scales of temperature and relative humidity remain in a range of one to two grid points under 200 hPa, (i.e. 15 km and 30 km). Temperature and humidity, which have more local structures, are modeled with smaller length scales. Globally, the horizontal length scales of different variables increase from the bottom to the top of the model as they represent more synoptic events at high altitude.

The vertical correlation is modeled by the application of recursive filters based on the estimation of the vertical length scale coming from the formula of Daley (1991, p. 110) and using the parabolic approximation Eq. (6). The stream function and the velocity potential in Fig. 10 that represent large scale horizontal flow have a bigger vertical length scale than those of temperature and humidity. The vertical gradients of temperature and humidity can vary strongly locally, lowering down the vertical correlation.

4.2 Pseudo single observation test on WRFDA and GSI data assimilation systems

The single pseudo-observation is a powerful way to make a benchmark as it allows visualizing the increment of an isolated observation and its impact on other variables. Thus, the following are a series of plots for a pseudo observation test of temperature with an innovation and the observation error of 1 K. The position of the pseudo-observation is arbitrarily taken at the center of the domain at 500 hPa.

As expected, the horizontal slice done at the 500 hPa for the temperature shows an isotropic response to the pseudo observation of 1 K. The maximum of intensity simulated with this pseudo observation depends on the variance value coming from the **B** matrix. When the operator ($\mathbf{U}_v$) employed the EOF decomposition, the J_b term of the cost function is computed by the variance that comes from the eigenvalues of $\mathbf{B}_{eof}$ (Fig. 11). While the operator ($\mathbf{U}_v$) is modeled by the estimation of a length scale and the recursive filters applied on the vertical in $\mathbf{B}_{rcf}$, the analysis is weighted by the variance computed directly on the model's mesh grid binned by vertical level as in this case the domain is a limited area (Fig. 12). For data assimilation of global models, the variance is classically binned by vertical level and latitude band, which is the case for the Bnam matrix coming from NAM (Fig. 13). Figures 11–13 show close results regarding the intensity on the horizontal slice and the differences are mainly due to the length scale value. The temperature innovation is spreaded out by the recursive filter over a larger area in Fig. 11 because the length scale computed in the EOF mode is larger.

On the vertical slice XZ, the temperature innovation has a larger impact on the vertical using $\mathbf{B}_{eof}$ and $\mathbf{B}_{rcf}$ than $\mathbf{B}_{nam}$. These differences come from the dataset used to model these **B** matrices. By construction, the statistics from $\mathbf{B}_{nam}$ are more climatological as they are averaged over time and they are interpolated on the mesh grid domain of our test case during the data assimilation process. In addition, they are computed from the background of another model. The statistics coming from $\mathbf{B}_{eof}$ and $\mathbf{B}_{rcf}$ represent the statistics coming from an ensemble of the day directly related to the meteorological events and using the same model. This kind of background error statistics has potentially more skills to estimate correlated errors.

The horizontal cross-section (XY) plotted for U and V showed dipole lobes, which can be explained by the geostrophic balance adjustement that the covariances statistics reproduce. The XZ plan follows the isocontour of $0\,\mathrm{m\,s^{-1}}$ for U while more complex structures can be observed on the slices for V.

Finally, the temperature pseudo observation test performed using the $\mathbf{B}_{rcf}$ and $\mathbf{B}_{nam}$ under the GSI system, and $\mathbf{B}_{eof}$ under WRFDA system show comparable results and differences that can be explained. The data assimilation of real observations performed

GMDD
7, 4291–4352, 2014

on this domain using $\mathbf{B}_{eof}$ and $\mathbf{B}_{rcf}$ may provide better results than the one using $\mathbf{B}_{nam}$ as the statistics of the background error have been specifically determined.

5 Generation of a multivariate background error covariance for hydrometeors

Modifications code have been done to in WRFDA to add a multivariate balance operator for the hydrometeor variables: cloud liquid water mixing ratio (q_{cloud}), rain mixing ratio (q_{rain}), ice mixing ratio (q_{ice}), snow mixing ratio (q_{snow}), so that the WRFDA minimization is now performed over nine 3-D fields instead of the five previously included. The main scientific issue in this task is to define a proper **B** matrix and particularly, the cross-correlation terms that will ensure that the analysis is multivariate (Table 7), i.e. the observed and unobserved model fields are modified simultaneously and consistently during the analysis. The question of the estimation of the forecast error covariance matrix is the focus of this section. Figure 14 provides the conversion from vertical model level to pressure level.

5.1 Statistics of the background error covariance matrix for hydrometeors

The vertical and horizontal transforms retained are the recursive filters making the analysis of the length scale parameter easier. The four main hydrometeors have been added in this study, as they could be useful for data assimilation in remote sensing such as satellite cloudy radiances.

The horizontal length scale values of the different hydrometeors shown in Fig. 15a do not overpass two grid points, i.e. 30 km, which is smaller than that of other control variables. Significant values of length scale, that overpass one grid point (15 km), are related to the presence of hydrometeors: it occurs under 150 hPa for q_{ice} and q_{snow} and under 400 hPa for q_{cloud} and q_{ice}. The maximum of q_{cloud} length scale, located around 950 hPa, can be associated to the presence of low maritime clouds above the Pacific ocean remarkable by the high standard deviation in Fig. 18a and b. In the lower levels of the model, the length scale of q_{ice} vanishes as expected.

The vertical correlation maxima of the precipitating hydrometeors are higher compared to that of cloud water, or cloud ice hydrometeors as they can drop freely through multiple levels (Fig. 16a). The vertical length scale of q_{rain} increases regularly from around 500 hPa (level 18) until reaching a maximum at the ground. As the length scale increases fast after 800 hPa, where the highest density of the lower levels occurs, an arbitrary cut-off equal to one third of the total vertical grid point value is applied in order to avoid spreading out increment information outside the area of potential presence of rain with the recursive filter. The length scale of q_{snow} has two local maxima. The first one happens where the precipitating hydrometeors have the highest density at around 400 hPa. A steep increase occurs from 950 hPa until reaching the highest value close to the ground. The high rate of presence of snow mixing ratio equal to zero at these low levels tends to artificially enforce vertical correlation as well.

5.2 Example of a pseudo single observation of cloud mixing ratio in a multivariate approach

To verify that our analysis is multivariate, we conducted a series of tests in which pseudo observations of hydrometeors were assimilated into WRFDA and the corresponding analysis increment was plotted. Figure 17 shows the analysis response for the q_{cloud} and q_{vapor} model variables when three simulated observations of cloud liquid water are assimilated. One obervation is taken over the Pacific ocean, a second one over Texas and the last one in Canada.

The intensity of the increment can be weighted by the 1-D variance or by the 3-D variance (S operator) coming from the ensemble. The 1-D variance, displayed in Fig. 18a, gives a general information by vertical level and binning type without any assumption of horizontal location. It is most of the time used when the perturbations come from the NMC method or when the variance is not diagnosed for the analysis time. In our test case, the increment is modulated by the 3-D variance computed from a 6 h ensemble forecast with 50 members. The cloudy area coming from the background of the different members is represented by a high value of variance in Fig. 18b while low variance takes place in the dry area. The increment is most likely important where the variability of cloud presence exists, as over the Pacific Ocean (Fig. 17). A minimum value would probably need to be set to retain the possibility of increments in the dry area.

The covariance between the mixing ratio of cloud water condensate and relative humidity, described in Sect. 3.2.2, can reinforce the possibility to add clouds in the dry area or to remove clouds in the cloudy background area. The univariate version of the balance for hydrometeors is beneficial at the analysis time as it allows including increments of hydrometeors directly at the analysis time. The multivariate balance is present to help to propagate the q_{cloud} increment in the forecast by balancing it with a q_{vapor} increment. The increments of temperature, due to the multivariate balance between humidity and temperature, are not significant.

The determination of the balance of humidity and hydrometeors is a difficult task as it involves the microphysical processes of meteorological NWP models and different local phenomena. The use of local covariances coming from D-ensemble may help to balance those high sensible variables. Furthermore, operational centers, such as Météo-France and the Met Office, already use ensemble forecasts at high resolution to more accurately characterize specific meteorological events, such as precipitation and convection. Nowadays, their ensemble size remains small (often less then 10 members) because the cost in CPU time is still elevated. Studies have been dedicated to evaluate the sampling errors in the ensemble method and in the parameters, such as correlation length scales, that usually model the background errors. Methods that combine general statistics of the background errors and local balance are found to perform better when the ensemble size is small (Hamil and Snyder, 2000). Figures 15a, b and 16a, b, that display horizontal and vertical length scales parameters respectively, for the hydrometeors in regards of the number of members, show stable results.

GMDD
7, 4291–4352, 2014

6 Conclusions

While variational methods have been successfully used in operational centers for a long time, the estimation of background errors needs to be continuously improved to assimilate new variables and to provide more accurate statistics. The GEN_BE v2.0 code has been developed to investigate and model univariate or multivariate covariance errors from control variables defined by a user as an input. It gathers some methods and options that can be easily applied to different model inputs and used on different data assimilation platforms by extending its former capabilities. The flexibility of the framework of the GEN_BE V2.0 code should help the diagnostics of correlated errors and the implementation of new background error modeling.

This document describes the different stages and transforms that lead to the modeling of the background error covariance matrix **B** by performing benchmark tests and showing examples that use these new functionalities. First, the GEN_BE v2.0 code has been validated through single observation tests on two different platforms using the EOF decomposition (WRFDA) and the recursive filters (GSI) to model the vertical transform. The benchmark test shows similar results with comprehensive differences for the set of five control variables used at NCEP for real time purposes. Second, the GEN_BE v2.0 code has been used to experiment with a multivariate approach that includes new control variables for cloud data assimilation. The precedent set of control variables used by NCEP has been expanded from five to nine to include a multivariate approach for humidity and mixing ratio for hydrometeors. As clouds have an intermittent presence, the 3-D variance coming from an ensemble of the day gives a spatial envelope useful to weight the analysis relatively to the observation and the background confidence. The next step is to test cloudy radiance data assimilation using a new definition of the **B** matrix that includes hydrometeors as control variables. Finally, statistics of background are estimated for chemical species (shown in Appendix A) such as carbon monoxide (CO), nitrogen oxides (NO_x) and ozone (O_3) even if data assimilation of chemical species and aerosols remains difficult due to strong non-linearities.

GMDD
7, 4291–4352, 2014

The trend is to model more complex background error expanding the control variables and correlated errors by using techniques for more heterogeneity and anisotrpy. The geographical binning and the 3-D variance available in the GEN_BE v2.0 code can be utilized with new data assimilation algorithms. Hybrid data assimilation that combines variational and ensemble methods may be helpful especially to add some flow dependence in the estimation of the background error and to reduce the ensemble size due to CPU time constraints (Hamil and Snyder, 2000). In addition, the GEN_BE code can be a tool to diagnose inhomogeneous 3-D localization parameters in the ensemble methods. The GEN_BE v2.0 code has been tested in atmospheric science but the flexibility of the code may be useful in other geophysical applications.

Appendix A: Background error for chemical species

This text is just included to show the applicability of the GEN_BE v2.0 code as a diagnostic tool for other topics than meteorlogy. However further testing needs to be performed to demonstrate real benefits. In recent decades, a large amount of studies that investigate chemical data assimilation have been conducted. Some of the first studies on stratospheric and tropospheric chemistry data assimilation were performed roughly two decades ago (e.g. Austin, 1992; Fisher and Lary, 1995; Elbern et al., 1997). During the last two decades, efforts have been made in order to improve atmospheric chemical modeling and data assimilation scheme performances. Recently Barré et al. (2013) assimilated stratospheric and tropospheric ozone observations simultaneously in a fully resolved chemical scheme. Whereas Massart et al. (2012) pointed out the importance of using ensemble estimated background error covariance in chemical data assimilation. Due to the increasing complexity and accuracy of new chemical model schemes and data assimilation systems, it appears necessary to take a realistic background error characterization into account.

Benedetti and Fisher (2007) defined the background errors for some aerosols species, which include sea salt, desert dust and continental particulate using aerosol

GMDD
7, 4291–4352, 2014

the optical depth. Statistics were analyzed in detail to ensure that the **B** matrix reproduced relevant correlation structures during the data assimilation process. Since data assimilation of chemical species is more recent than for meteorology, the GEN_BE code version 2.0 may be useful to test new definitions of background error covariance matrices and to allow the user to utylize it on different platforms. Dust, sea salt, particulate matter (PM), and several other chemical species have been already included as new possible control variables in the GEN_BE code. Results for some of them: CO (carbon monoxide), NO_x (nitrogen oxides) and O_3 (ozone) are shown next.

The statistics are estimated using 20 members over the CONUS domain. Each member comes from a 12 h forecast of WRF-CHEM (WRF model coupled with Chemistry) at 36 km of horizontal resolution and 33 vertical levels. The boundary conditions coming from MOZART (Model for OZone And Related chemical Tracers) and the emissions factors coming from MEGAN (Model of Emissions of Gases and Aerosols from Nature) are perturbed. Most of the ozone variability takes place in the middle atmosphere (stratosphere) on the ozone layer around 100 hPa where the NO_x concentration fluctuates as well, due to photochemistry (Fig. A1a–c). The NO_x emitted from the ground are reactive species with a short lifetime. The ozone, as a secondary product and less reactive than NO_x, has a larger length scale for all the pressure levels of the model above 950 hPa (Fig. A2). CO, which is a precursor for tropospheric ozone and aerosols, has a high variability in the boundary layer due to the combined effect of emissions and transport. Another variability maximum is observed at the top of the troposphere in Fig. A1(d) where strong gradients of CO exist at the tropopause levels. The largest vertical length scales of these species are diagnosed close to the surface where they are emitted or secondarily produced for ozone (Fig. A3). Then, they sharply decrease between 1000 hPa and 850 hPa because of the mixing that occurs inside the boundary layer: their correlation with the lower levels decreases. The recursive filters associated with these vertical length-scale values would better predict vertical diffusion close to the ground. Above 850 hPa, which is around the top of the boundary layer, the evolution of the vertical length scale decreases slowly from approximately two to one grid point. Vertical diffusion will be less significant for these levels.

The diagnostics of simple statistics of the background for chemical species are straight forward with the GEN_BE code version 2.0. Moreover, data assimilation of chemistry components remains a challenge because of the uncertainties of various parameters that predict chemical processes as emission factors, deposition velocity and (photochemical) reaction constant. For these reasons, the analysis may fit the observation not for the good reason if the data assimilation does not involve the origin of the mismatch. In addition, chemical processes can be highly non-linear and may be hard to model. Hybrid and ensemble methods may help to diagnose complex covariance structures in future work.

Appendix B: Code information

The Tables B1–B3 of this Appendix describe the FORTRAN sources, input and output of the GEN_BE v2.0 code.

The Supplement related to this article is available online at doi:10.5194/gmdd-7-4291-2014-supplement.

Acknowledgements. Funding for this work was provided by the US Air Force Weather Agency. The authors benefited from numerous discussions with Yann Michel. Glen Romine and Jerome Barré are thanked for providing the ensemble over the conus domain. Syed Rizvi is thanked for discussions concerning the previous version of the code.

References

Aulighé, T., Lorenc, A., Michel, Y., Montmerle, T., Jones, A., Hu, M., and Dudhia, J.: Toward a new cloud analysis and prediction system, B. Am. Meteorol. Soc., 92, 207–210, doi:10.1175/2010BAMS2978.1, 2011.

Austin, J.: Toward the 4-dimensional assimilation of stratospheric chemical-constituents, J. Geophys. Res., 97, 2569–2588, 1992.

Bannister, R. N.: A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characterisitics and measurements of forecast error covariances, Q. J. Roy. Meteor. Soc., 134, 1951–1970, doi:10.1002/qj.339, 2008a.

Bannister, R. N.: A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error statistics, Q. J. Roy. Meteor. Soc., 134, 1971–1996, doi:10.1002/qj.340, 2008b.

Barker, D. M., Huang, W., Guo, Y. R., and Xiao, Q. N.: A three-dimensional (3DVAR) data assimilation system for use with MM5: implementation and initial results, Mon. Weather Rev., 132, 897–914, 2004.

Barker, D. M., Huang, X. Y., Liu, Z., Auligné, T., Zhang, X., Rugg, S., Ajjaji, R., Bourgeois, A., Bray, J., Chen, Y., Demirtas, M., Guo, Y. R., Henderson, T., Huang, W. H., Lin, C., Michalakes, J., Rizvi, S., and Zhang, X.: The Weather Research and Forecasting Model's Community Variational/Ensemble Data Assimilation System: WRFDA, B. Am. Meteorol. Soc., 93, 831–843, doi:10.1175/BAMS-D-11-00167.1, 2012.

Benedetti, A. and Fisher, M.: Background error statistics for aerosols, Q. J. Roy. Meteor. Soc., 133, 391–405, doi:10.1002/qj.37, 2007.

Caron, J. F. and Fillion, L.: An examination of background error correlations between mass and rotational wind over precipitation regions, Mon. Weather Rev., 138, 563–578, doi:10.1175/2009MWR2998.1, 2010.

Courtier, P., Thépaut, J. N., and Hollingsworth, A.: A strategy for operational implementation of 4D-Var, using an incremental approach, Q. J. Roy. Meteor. Soc., 120, 1367–1387, 1994.

Daley, R.: Atmospheric Data Analysis, Cambridge Univeristy Press, 1991.

Elbern, H., Schimdt, H., and Elbel, A.: Variational data assimilation for tropospheric chemistry modeling, J. Geophys Res. Rev., 102, 15967–15985, 1997.

Fisher, M. and Lary, D. J.: Lagrangian four-dimensional variational data assimilation of chemical species, Q. J. Roy. Meteor. Soc., 121, 1681–1704, 1995.

Hamill, T. M. and Snyder, C.: A hybrid ensemble Kalman filter–3D variational analysis scheme, Mon. Weather Rev., 128, 2905–2919, doi:10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2, 2000.

Klesit, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Wu, W. S., and Lord, S.: Introduction of the GSI into the NCEP Global Data Assimilation System, Mon. Weather Rev., 24, 1691–1705, doi:10.1175/2009WAF2222201.1, 2009.

Massart, S., Piacentini, A., and Pannekoucke, O.: Importance of using ensemble estimated background error covariances for the quality of atmospheric ozone analyses, Q. J. Roy. Meteor. Soc., 138, 889–905, doi:10.1002/qj.971, 2012.

Ménétrier, B. and Montmerle, T.: Heterogeneous background-error covariances for the analysis and forecast of fog events, Q. J. Roy. Meteor. Soc., 137, 2004–2013, doi:10.1002/qj.802, 2011.

Michel, Y. and Auligné T.: Inhomogeneous background error modeling over Antarctica, Mon. Weather Rev., 138, 2229–2252, doi:10.1175/2009MWR3139.1, 2010.

Michel, Y., Auligné T., and Montmerle, T.: Heterogeneous convective-scale Background Error Covariances with the inclusion of hydrometeor variables, Mon. Weather Rev., 139, 2994–3015, doi:10.1175/2011MWR3632.1, 2011.

Montmerle, T. and Berre, L.: Diagnosis and formulation of heterogeneous background error covariances at mesoscale, Q. J. Roy. Meteor. Soc., 136, 1408–1420, doi:10.1002/qj.655, 2010.

Pannekoucke, O., Berre, L., and Desroziers, G.: Background-error correlation length-scale estimates and their sampling statistics, Q. J. Roy. Meteor. Soc., 134, 497–508, doi:10.1002/qj.212, 2008.

Parrish, D. F. and Derber, J. C.: The National Meteorological Center's spectral statistical-interpolation analysis system, Mon. Weather Rev., 120, 1747–1763, 1992.

Pereira, M. B. and Berre, L.: The use of an ensemble approach to study the background error covariances in a global NWP model, Mon. Weather Rev., 134, 2466–2489, doi:10.1175/MWR3189.1, 2006.

Purser, R. J., Wu, W. S., Parrish, D. F., and Roberts, N. M.: Numerical aspects of the application of recursive filters to variational statistical analysis, Part I: Spatially homogeneous and isotropic Gaussian covariances, Mon. Weather Rev., 131, 1524–1535, doi:10.1175/1520-0493(2003)131<1524:NAOTAO>2.0.CO;2, 2003a.

GMDD

7, 4291–4352, 2014

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀I | ▶I
◀ | ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

Purser, R. J., Wu, W. S., Parrish, D. F., and Roberts, N. M.: Numerical aspects of the application of recursive filters to variational statistical analysis, Part II: Spatially inhomogeneous and anisotropic general covariances, Mon. Weather Rev., 131, 1536–1548, doi:10.1175//2543.1, 2003b.

Romine, G., Weisman, M., Manning, K., Wang, W., Schwartz, C., Anderson, J., and Snyder, C.: The Use of WRF-DART Analyses for 3 km Explicit Convective Forecasts in Support of the 2012 DC3 Field Program, 13th WRF Users Workshop, Boulder, CO, 26–29 June 2012, WS2012/5.2, 2012.

Wu, W. S., Purser, R. J., and Parrish, D. F.: Three-dimensional variational analysis with spatially inhomogeneous covariances, Mon. Weather Rev., 130, 2905–2916, doi:10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2, 2002.

GMDD

7, 4291–4352, 2014

Table 1. General information defining the experiment in the namelist input file (&gen_be_info part).

&gen_be_info	Namelist options	Description
model	"WRF"	Set up the acronym for the model input. Allow GEN_BE to read different input model in the stage0.
application	"WRFDA"	"WRFDA" and "GSI" interface have been developed and tested.
be_method	"ENS" or "NMC"	Compute perturbations from an ensemble or from different time lagged forecast.
ne	Number of members	If NMC method ne = 1.
cut	0, 0, 0, 0, 0, 0	Allow to subset an area of a domain, defined in grid points. imin, imax, jmin, jmax, kmin, kmax.
use_mean_ens	"false"	If be_method = "ENS" is selected, the perturbation can be calculated from the mean of all the members or from 2 different members.
start_date	"_START_DATE_"	Initial date, format ccyymmddhh.
end_date	"_END_DATE_"	Final date, format ccyymmddhh.
interval	"hh"	Frequency of the historical date available, defined in hour (useful for the NMC method only).

GMDD
7, 4291–4352, 2014

Table 2. Description of the binning options.

Bin_type	Description
0	Binning by grid point.
1	Binning by vertical level along the x direction point of the model.
2	Binning by vertical heights and by latitude num_bins_lat. The parameters binwidth_lat and binwidth_hgt defined the width that splits the bins.
3	Binning by vertical level model and latitude dependent. The parameters lat_min, lat_max are computed from the model input data and the parameter binwidth_lat is defined in the namelist.input file.
4	Binning by model vertical level and along the y direction.
5	Binning on vertical level model including all the horizontal point.
6	Average over all points.
7	Binning rain/no-rain by vertical levels and based on thresholds in the model background (Michel et al., 2011).

GMDD
7, 4291–4352, 2014

Generalized Background Error covariance matrix model (GEN_BE v2.0)

G. Descombes et al.

Table 3. Parameters defining the binning options of the namelist input file (&gen_be_bin part).

&gen_be_bin	Namelist options	Description
bin_type	1–8	Bin type option
lat_min, lat_max		Minimum and Maximum of latitude defined in degree. Used if bin_type = 2
binwidth_lat	5.0	Width of the bins defines by latitude in degree Used if bin_type = 2, 3, 4
hgt_min	1000.0	Used if bin_type = 2 (height, meter)
binwidth_hgt	2000.0	Width of bins defines by height in meter Used if bin_type = 2 (meter)

GMDD
7, 4291–4352, 2014

Table 4. Description of the control variables available for the meteorology.

Nomenclature of the control variables	Description
psi	Stream function (ψ)
chi	Velocity potential (χ)
vor	Vorticity
div	Divergence
u	Horizontal wind component in x direction
v	Horizontal wind component in the y direction
t	Temperature
ps	Surface pressure
rh	Relative humidity
qs	Specific humidity
q_{cloud}	Cloud mixing ratio
q_{rain}	Rain mixing ratio
q_{ice}	Ice mixing ratio
q_{snow}	Snow mixing ratio
sst	Sea Surface Temperature

GMDD
7, 4291–4352, 2014

Table 5. Description of the options available in the namelist input file (&gen_be_lenscale part) to diagnose length scale parameter.

&gen_be_lenscale	Namelist options	Description
data_on_levels	"true"	The statistics can be computed by vertical model level (GSI) or by EOF mode (WRFDA)
vert_ls_method	1, 2	Estimate the vertical length scale (stage 3) Option 1: parabolic approximation formula Option 2: gaussian approximation formula
ls_method	1, 2	Estimate length scale (stage 3) See the paragraph for more details
use_med_ls	"true"	Estimate the length using the median value or not.
stride	1	Subset of point to speed up the stage 4
n_smth_ls	2	Number of point to smooth the length scale
use_global_bin	"true"	The statistics can be binned or not in the stages 3 and 4. Only inhomogeneous recursive filters can handle binned length scale.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

GMDD

7, 4291–4352, 2014

Generalized Background Error covariance matrix model (GEN_BE v2.0)

G. Descombes et al.

Table 6. Information related to the control variables and their covariance errors in the namelist input file (&gen_be_cv part, example CV5).

&gen_be_cv	Namelist options	Description
nb_cv	5,	Number of control variables
cv_list	'psi', 'chi','t', 'ps', 'rh'	Variables used for the analysis
fft_method	1,2	Conversion of u and v to psi and chi 1 = Cosine, 2 = Sine transform
covar1	0, 0, 0, 0, 0, 0, 0, 0, 0, 0	First variable do not have covariance
covar2	1, 0, 0, 0, 0, 0, 0, 0, 0, 0	Covariance of variable 1 (psi) and variable 2 (chi)
covar3	1, 0, 0, 0, 0, 0, 0, 0, 0, 0	Covariance of variable 1 (psi) with variable 3 (t)
covar4	1, 0, 0, 0, 0, 0, 0, 0, 0, 0	Covariance of variable 1 (psi) with variable 3 (ps)
covar5	0, 0, 0, 0, 0, 0, 0, 0, 0, 0	Relative humidity univariate
covar6	0, 0, 0, 0, 0, 0, 0, 0, 0, 0	Other possible variable
use_chol_reg	.false.	by default, compute the regression coefficient as a ratio of covariance by variance. If true, use a choleski decomposition (specific to GSI).

GMDD
7, 4291–4352, 2014

Table 7. Information related to the control variables and their covariance errors in the namelist input file (&gen_be_cv part, example CV9, definition of multivariate humidity and hydrometeors error covariance matrix).

&gen_be_lenscale	Namelist Options
nb_cv	9
cv_list	'psi', 'chi', 't', 'ps', 'rh', 'q_{cloud}', 'q_{ice}', 'q_{rain}', 'q_{snow}'
covar1	0, 0, 0, 0, 0, 0, 0, 0, 0, 0
covar2	1, 0, 0, 0, 0, 0, 0, 0, 0, 0
covar3	1, 0, 0, 0, 0, 0, 0, 0, 0, 0
covar4	1, 0, 0, 0, 0, 0, 0, 0, 0, 0
covar5	0, 0, 1, 1, 0, 0, 0, 0, 0, 0
covar6	0, 0, 0, 0, 1, 0, 0, 0, 0, 0
Covar7	0, 0, 0, 0, 1, 0, 0, 0, 0, 0
Covar8	0, 0, 0, 0, 0, 0, 0, 0, 0, 0
Covar9	0, 0, 0, 0, 0, 0, 0, 0, 0, 0

GMDD
7, 4291–4352, 2014

Table B1. FORTRAN code description of the GEN_BE v2.0 framework.

FORTRAN modules	Comments
variables_types.f90	It defines, declares and allocates new types as state_type, mesh_type, bin_type, state_matrix. Some basics operations as addition substraction, calculation of variance, covariance are available.
configure.f90	It reads the namelist.input file and initialize the variables
io_input_models.f90	It reads input standard variables from a model define by the user and convert them into control variables. If the user needs to introduce new input model, only this module needs to be updated to read and transform the data.
io_input.f90	It reads NetCDF input data and initialize new types
io_output.f90	It writes NetCDF output format for all new types
io_output_applications.f90	It writes output for different application needs

GMDD
7, 4291–4352, 2014

Table B2. Input and output of the different components of the GEN_BE v2.0 code.

Programs	Input	output	comments
gen_be_stage0.F	Various models (ex: WRF)	pert.ccyymmddhh	It contains the perturbations for all the control variables defined in the namelist
		mesh_grid.nc All_mesh.grid.nc	It contains all the static data as latitude array, longitude array, map factors
		mask.ccyymmddhh	This file exist only with the option dynamical_mask which is activated with bin_type=7 or bin_type=8
		standard_variable.txt control_variable.txt	It contains the list of the control variables in ASCII format.
gen_be_stage1.F	pert.ccyymmddhh	var.ccyymmddhh	The input file is splitted per variables
		bins.nc	All the information related to the binning options are included in this file.
gen_be_stage2.F	var.ccyymmddhh	gen_be_stage2_regcoeff.nc	All the regression coefficients are included in this file
		var(_u) ccyymmddhh	If a linear regression is applied to the current variable to remove its balanced part, an unbalanced output variable is written under this nomenclature
gen_be_stage3.F	var(_u).ccyymmddhh	gen_be_stage3_vert_lenscale.var(_u).nc	It contains the vertical length scale parameter for the full or unbalanced part of the variable
		gen_be_stage3_varce.var(_u).nc	Variance 3 dimensions by grid point
		gen_be_stage3_vert_varce(_u).nc	Binned vertical variance.
		var(_u).ccyymmddhh.ennn.kkk	Intermediate binary files splitted by vertical level.
gen_be_stage4.F	var(_u).ccyymmddhh.ennn.kkk	sl_print.blll.qcloud	Intermediate ASCII file format that contain the horizontal lenscale.
gen_be_diags.F	Results of the precedents stages from 2 to 4	be.nc	Final netcdf file that contains all the information to model **B**.
gen_be_nc2gsi.F	be.nc	be_gsi_little_endian.gcv be_gsi_big_endian.gcv	Binary format directly readble by GSI.

GMDD
7, 4291–4352, 2014
Generalized Background Error covariance matrix model (GEN_BE v2.0)
G. Descombes et al.

Table B3. Content of the final output file be.nc (NetCDF format) of the GEN_BE v2.0 code.

Name of the field	Description
Fields defined by control variable name (e.g. cv1)	
Lenscale_cv1	Horizontal lenscale in EOFs space or physical space
vert_lenscale_cv1	Vertical lenscale available only if the flag data_on_levels is true and the control variable number 1 is 3-D.
vert_variance_cv1	Vertical variance of the control variable number 1 per bin
eigen_value_cv1	Eigen value of the control variable number 1 only available if the flag data_on_levels is false
eigen_vector_cv1	Eigen vector of the control variable number 1 only available if the flag data_on_levels is false
varce_cv1	Variance 3-D
Regression coefficients	
list_regcoeff	Complete list of the regression coefficients used in the balance constraint.
regcoeff_cv1_cv2	Example of regression coefficient between the control variable 1 and 2. It can be 1-D, 2-D or 3-D
vert_autocov_cv1	Vertical autocovariance of the control variable number 1
Binning parameters	
bin_type	Bin_type option selected
bin2d	Binning field 2-D array
bins	Binning field 3-D array

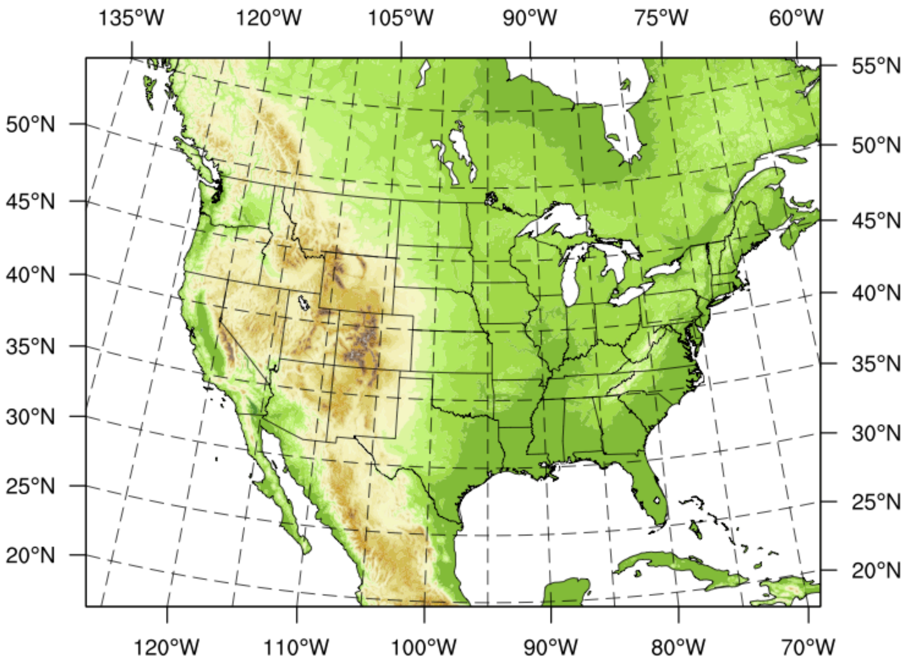

Figure 1. WRF domain over the conus area at the resolution of 15 km. Based on this configuration, the 50 members coming from a 6 h forecast (DART, experiment DC3) are used to generate background error stsatistics.

GMDD
7, 4291–4352, 2014

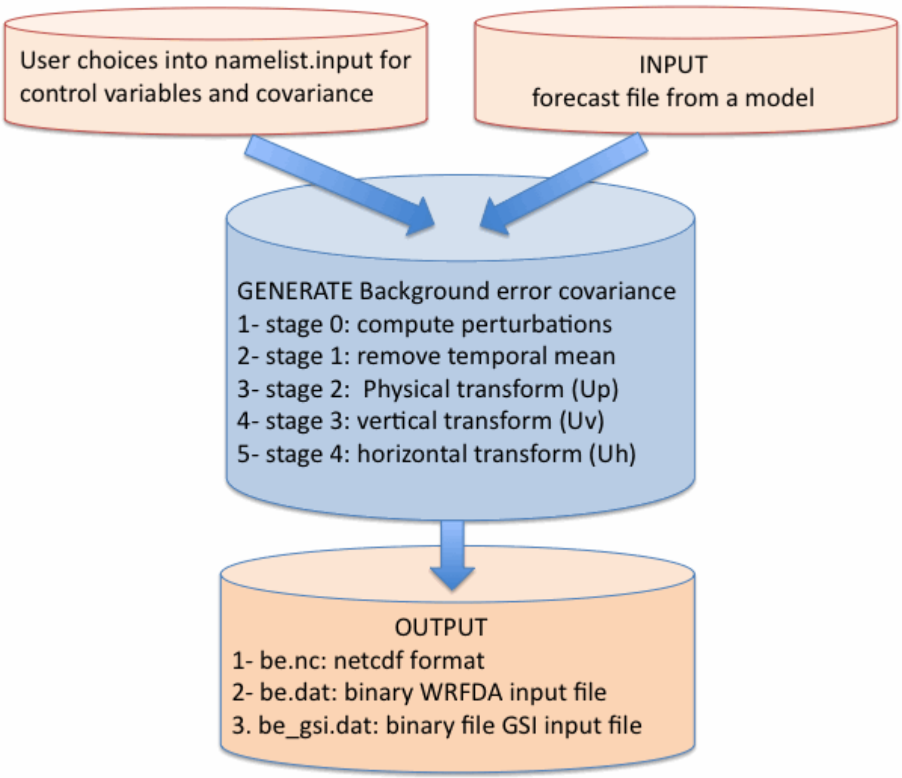

Figure 2. General structure of the code to generate a background error covariance matrix. The input and output are represented by the orange boxes and the five main stages that lead to model **B** is in blue.

GMDD
7, 4291–4352, 2014

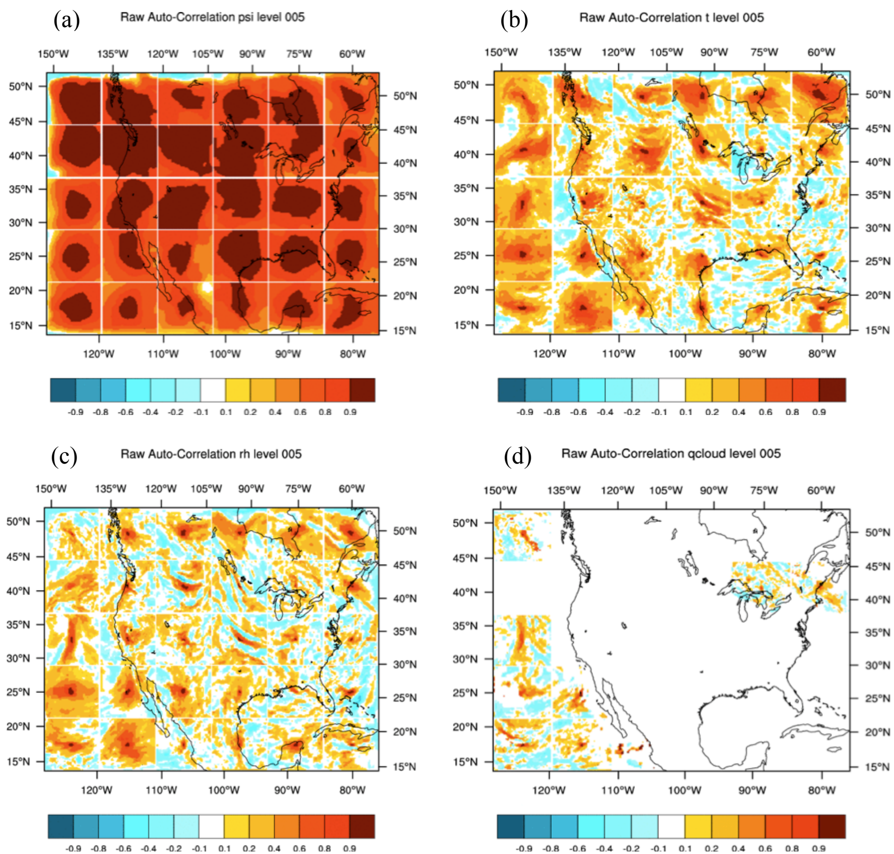

Figure 3. Horizontal autocorrelation performed at the center of each square grid over vertical model level 5, around 950 hPa, for the control variables **(a)** psi, **(b)** t, **(c)** rh, and **(d)** q_{cloud}. Larger correlations are observed for psi compared to t and rh. q_{cloud} has the smallest correlation sparecly distrubuted.

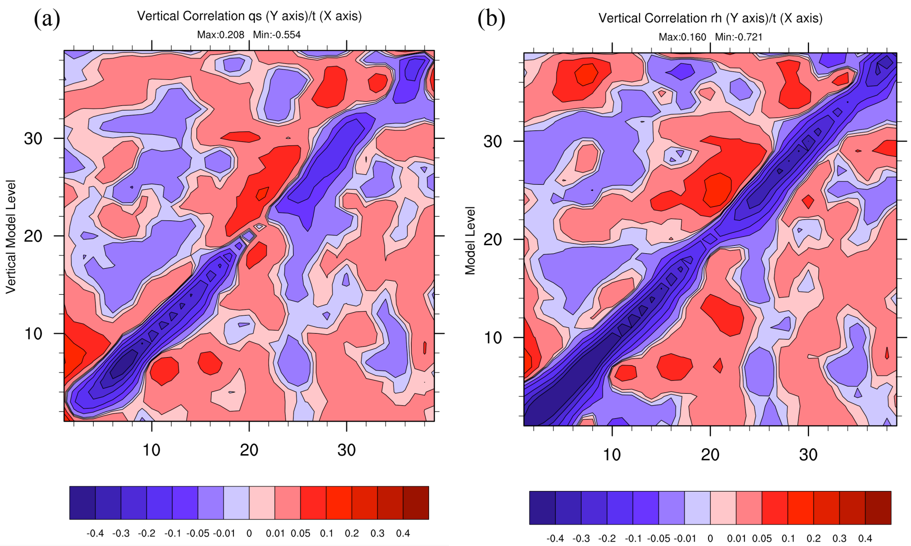

Figure 4. (a) Vertical cross-correlation between temperature and specific humidity, **(b)** vertical cross-correlation between temperature and relative humidity.

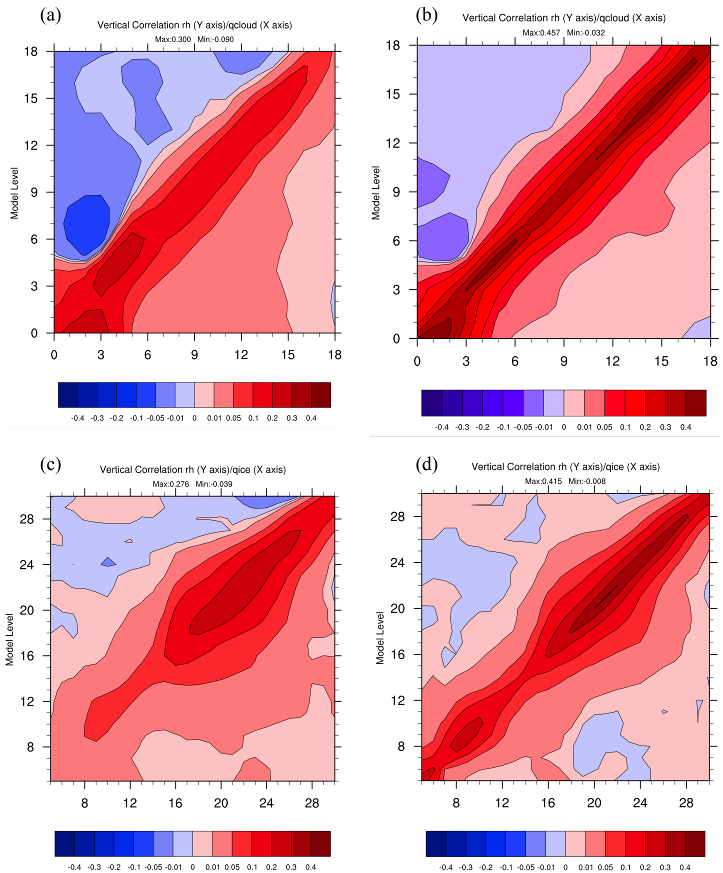

Figure 5. (a) Raw vertical cross-correlations between q_{cloud} and rh, **(b)** filtered vertical cross-correlations between q_{cloud} and rh, **(c)** raw vertical cross-correlations between q_{ice} and rh, **(d)** filtered vertical cross-correlations between q_{ice} and rh. Taking into account only the perturbations coming from the transition of a cloudy to a non-cloudy area intensify the vertical correlation.

GMDD
7, 4291–4352, 2014

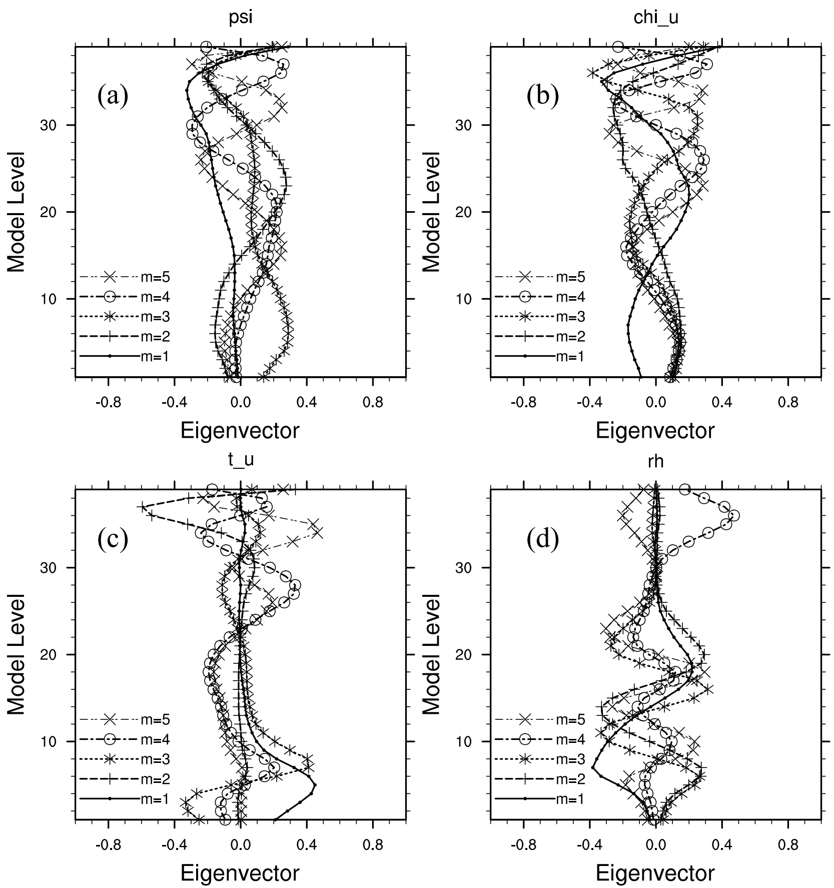

Figure 6. Representation of the first five eigenvectors resulting of the EOF decomposition of the vertical autocovariance matrix, eigenvectors of **(a)** psi, **(b)** $\mathrm{chi_u}$, **(c)** t_u, and **(d)** rh. The eigenvectors are parameters that define the vertical transform ($\mathbf{U}_v$).

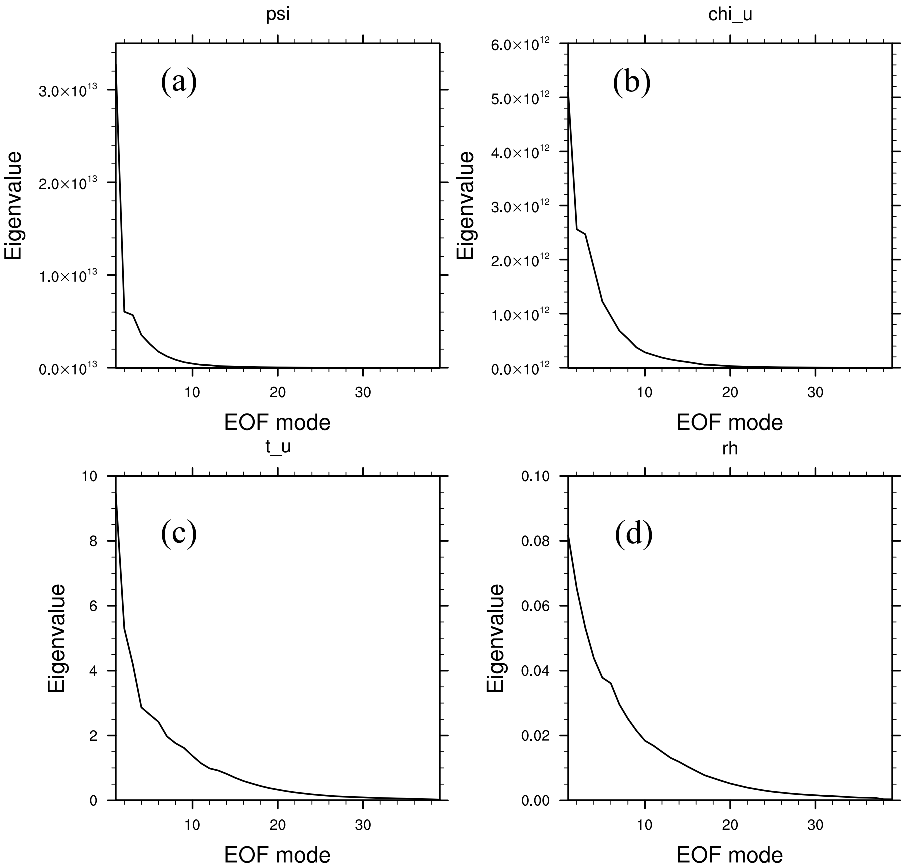

Figure 7. Eigenvalues computed by EOF mode for **(a)** psi, **(b)** chi_u, **(c)** t_u and **(d)** rh. They represent the variance of the former variables.

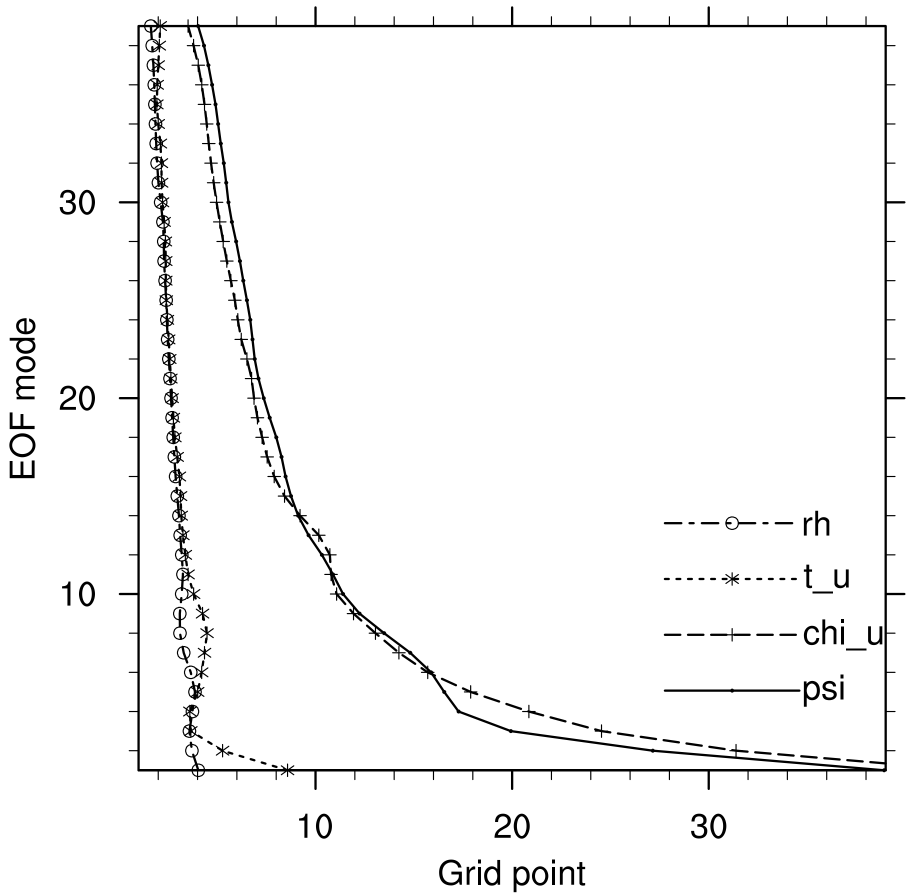

Figure 8. Length scales defined in grid point through EOF mode for CV5. The analysis control variables representating the dynamical variables, psi and $\mathrm{chi_u}$, have larger length scales than t_u, and rh.

GMDD
7, 4291–4352, 2014

Generalized Background Error covariance matrix model (GEN_BE v2.0)

G. Descombes et al.

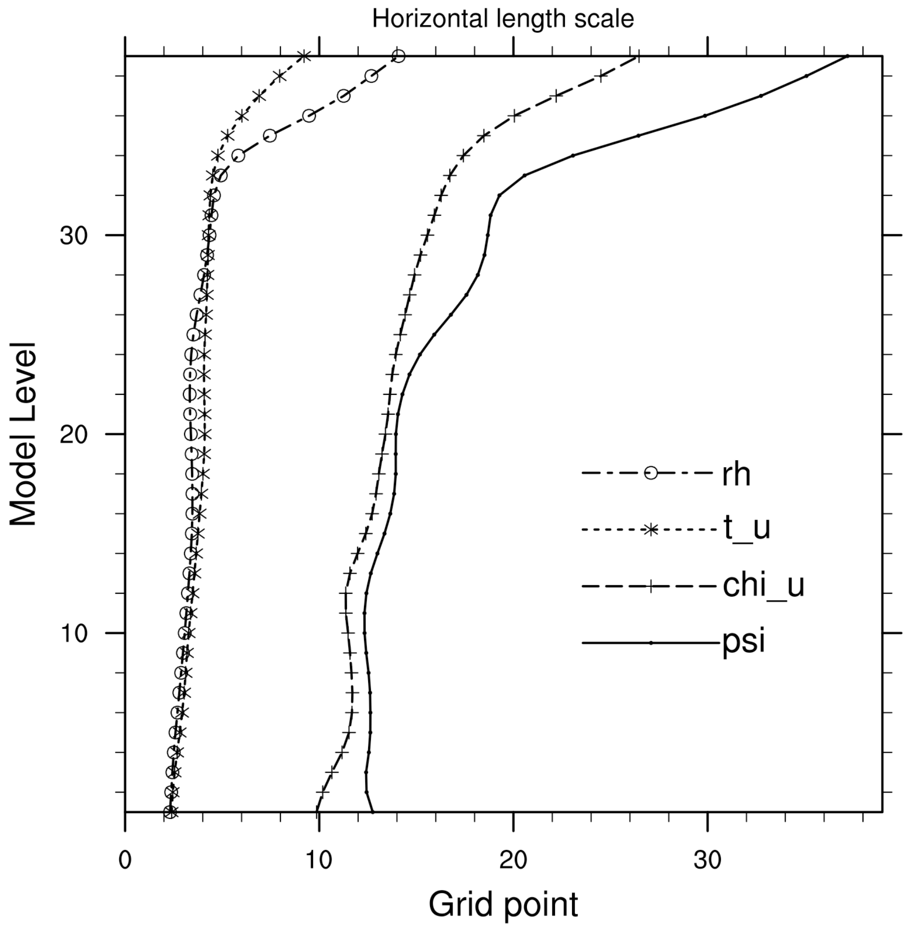

Figure 9. Horizontal length scales for CV5. t_u and rh, which have more local structures, are modeled by smaller length scales.

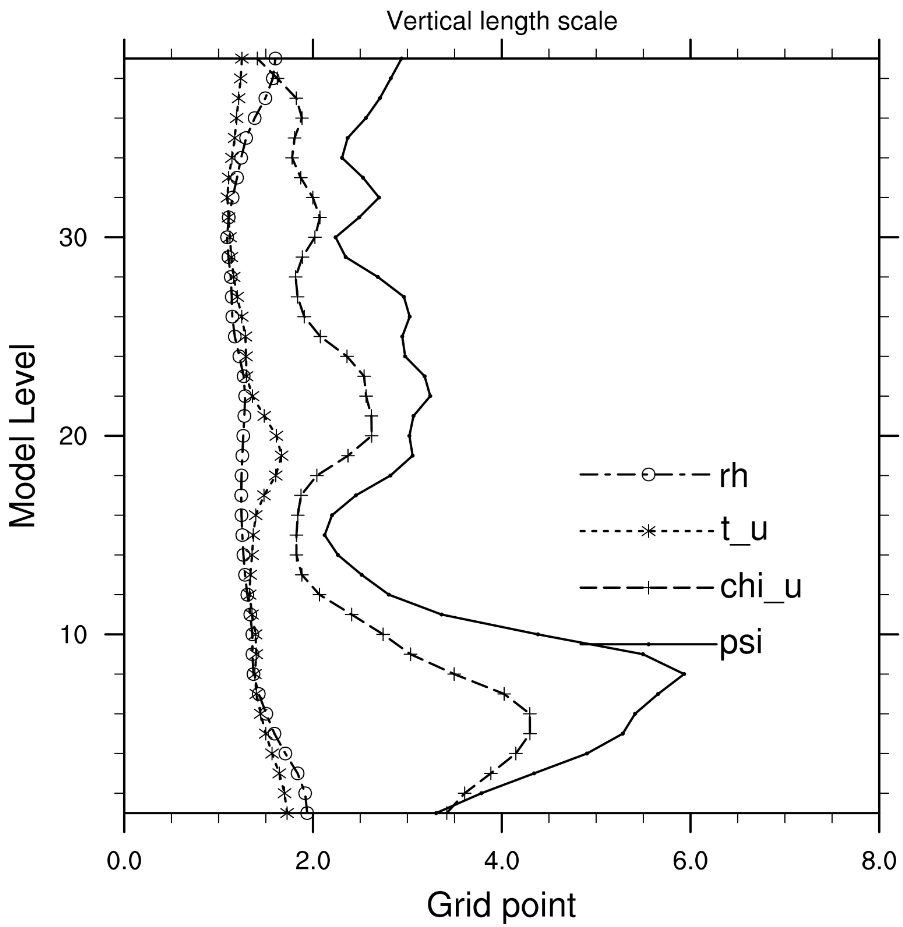

Figure 10. Vertical length scale for CV5.

GMDD
7, 4291–4352, 2014

Generalized Background Error covariance matrix model (GEN_BE v2.0)

G. Descombes et al.

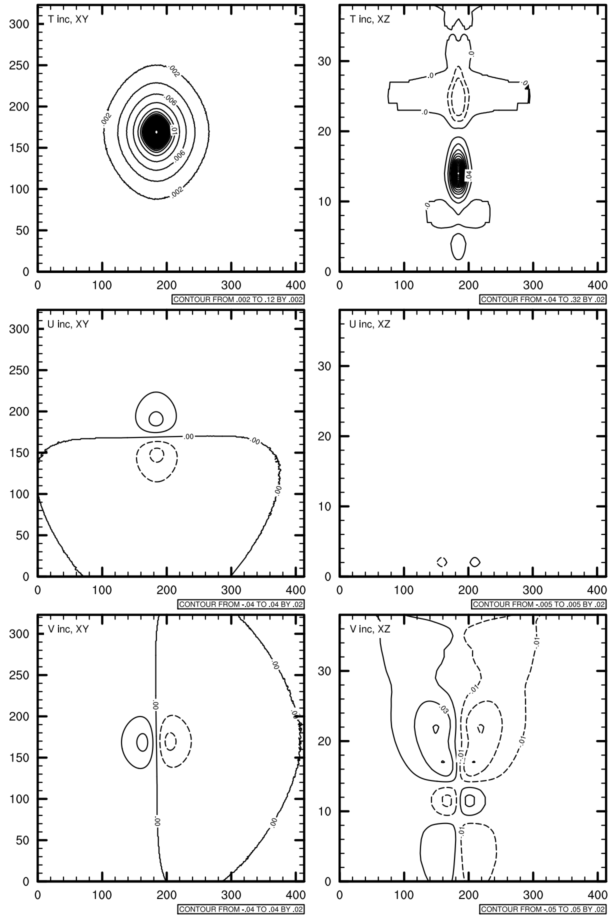

Figure 11. Pseudo observation test of temperature (+1 K) under the WRFDA application and using a background error covariance matrix diagnostic from GEN_BE v2.0 with EOF decomposition ($\mathbf{U}_\mathrm{v}$).

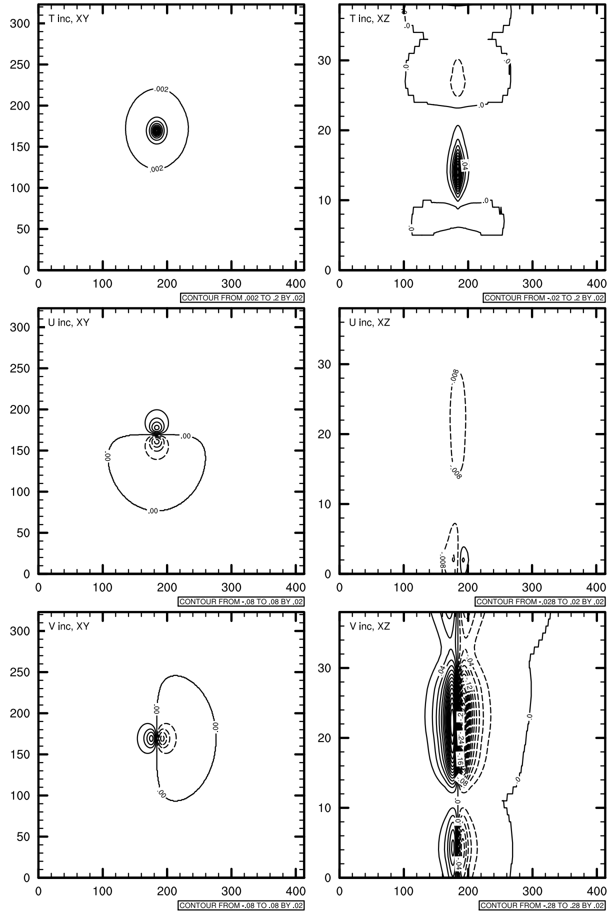

Figure 12. Pseudo observation test of temperature (+1 K) using recursive filters ($\mathbf{U}_v$) under the GSI application and a background error covariance matrix diagnostic from GEN_BE v2.0 ($\mathbf{B}_{rcf}$).

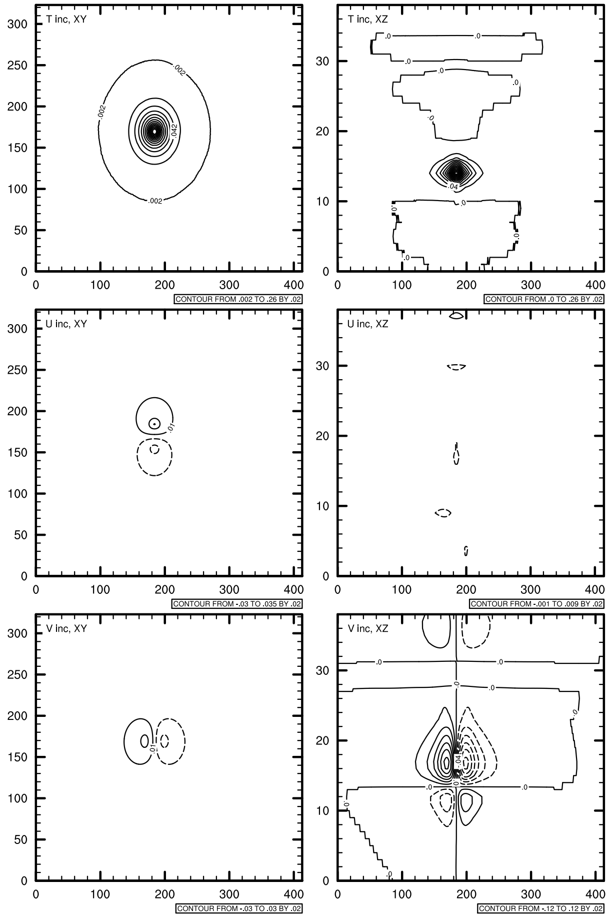

Figure 13. Pseudo observation of temperature (+1 K) using recursive filters under the GSI application and the NAM background error covariance ($\mathbf{B}_{\mathrm{nam}}$).

GMDD
7, 4291–4352, 2014

Generalized Background Error covariance matrix model (GEN_BE v2.0)

G. Descombes et al.

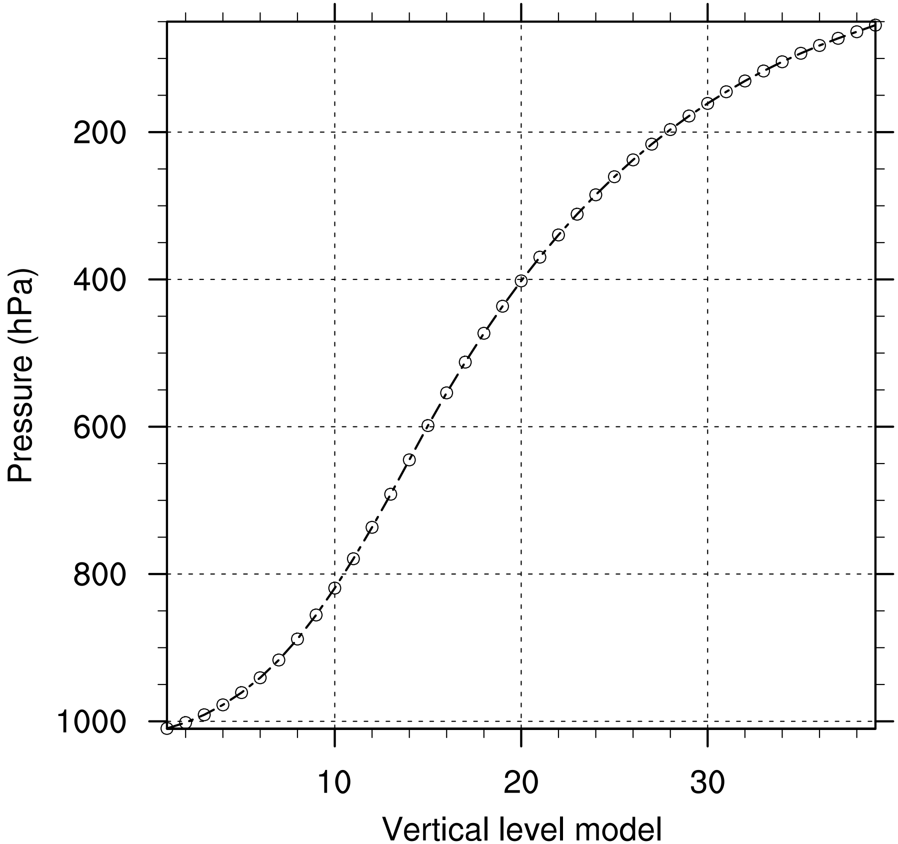

Figure 14. Plot of pressure (hPa) against vertical model levels.

GMDD
7, 4291–4352, 2014

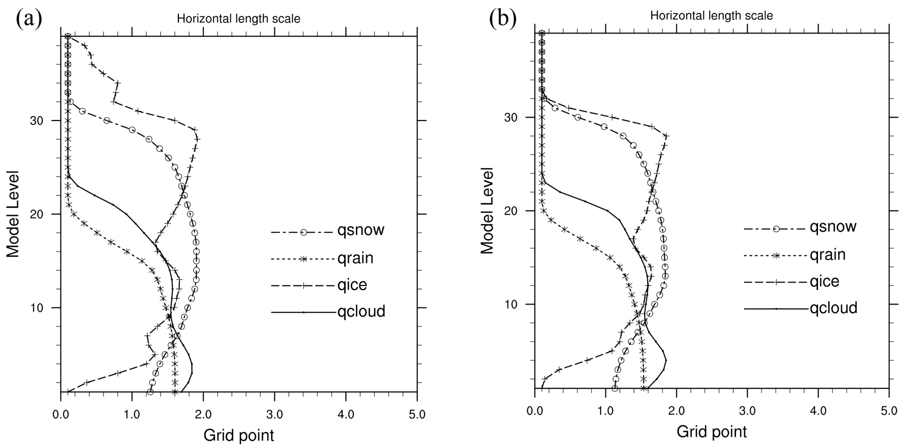

Figure 15. Horizontal length scale for the hydrometeors using **(a)** 50 members and **(b)** using 5 members show similar behavior.

GMDD
7, 4291–4352, 2014

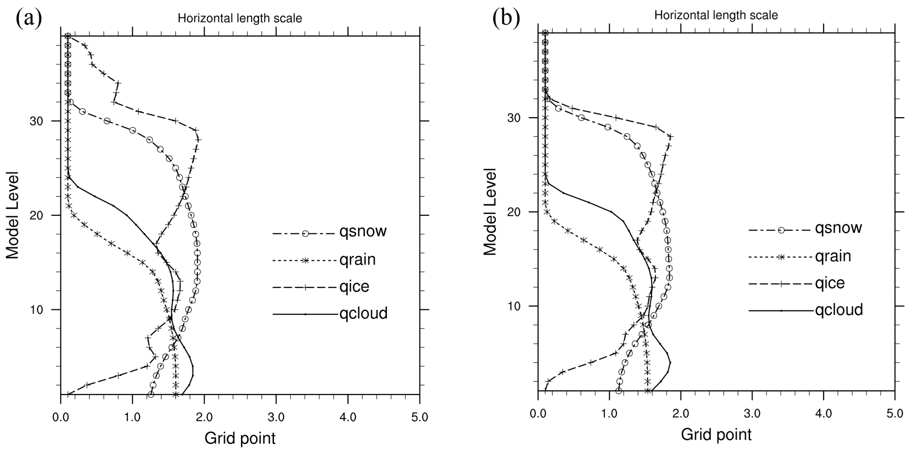

Figure 16. Vertical length scale for the hydrometeors using **(a)** 50 members and **(b)** using 5 members **(b)** show a similar evolution along the model levels.

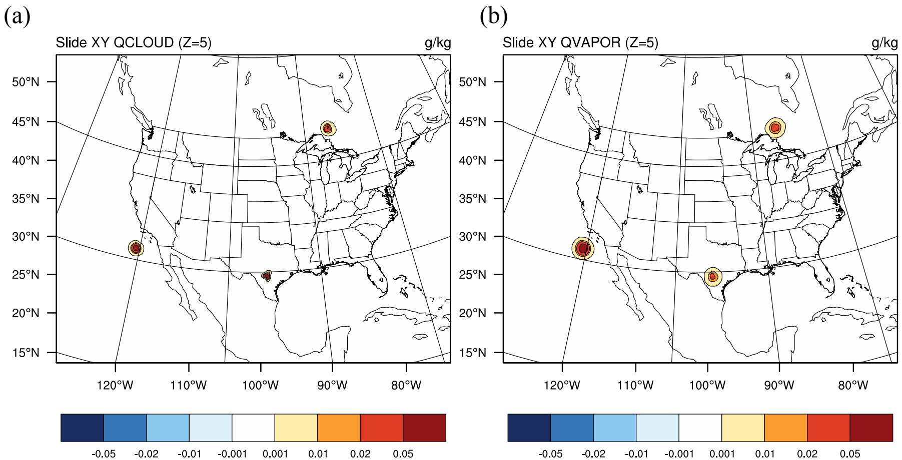

Figure 17. (a) Horizontal slide (vertical model level 5) of a pseudo observation test of cloud water condensate (0.1 $\mathrm{g\,kg^{-1}}$) in a multivariate approach using the 3-D variance, **(b)** as a consequence there is a positive increment on q_{vapor}.

GMDD
7, 4291–4352, 2014

Generalized Background Error covariance matrix model (GEN_BE v2.0)

G. Descombes et al.

Title Page
Abstract
Introduction
Conclusions
References
Tables
Figures
◀◀
▶▶
◀
▶
Back
Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

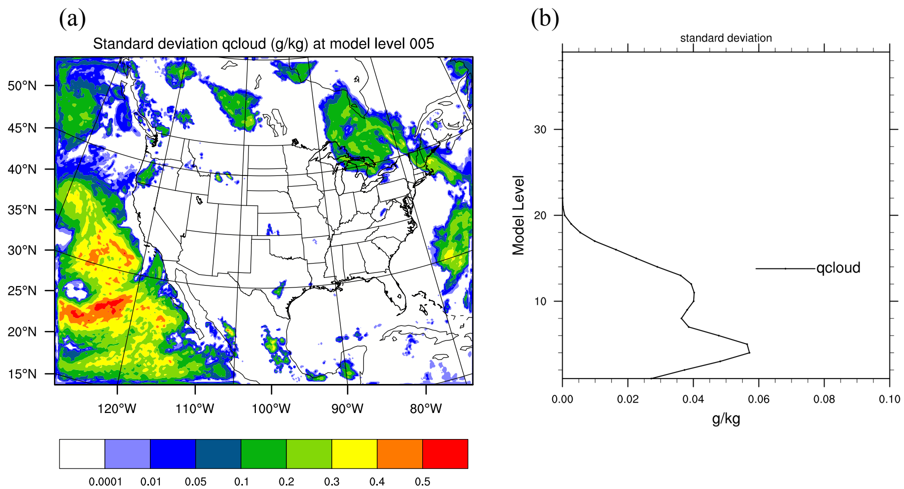

Figure 18. Standard deviation of liquid water condensate mixing ratio: **(a)** horizontal cross-section at the vertical model level 5 (950 hPa) and **(b)** profile of the average along the vertical. Both plots indicate the presence of low maritime cloud.

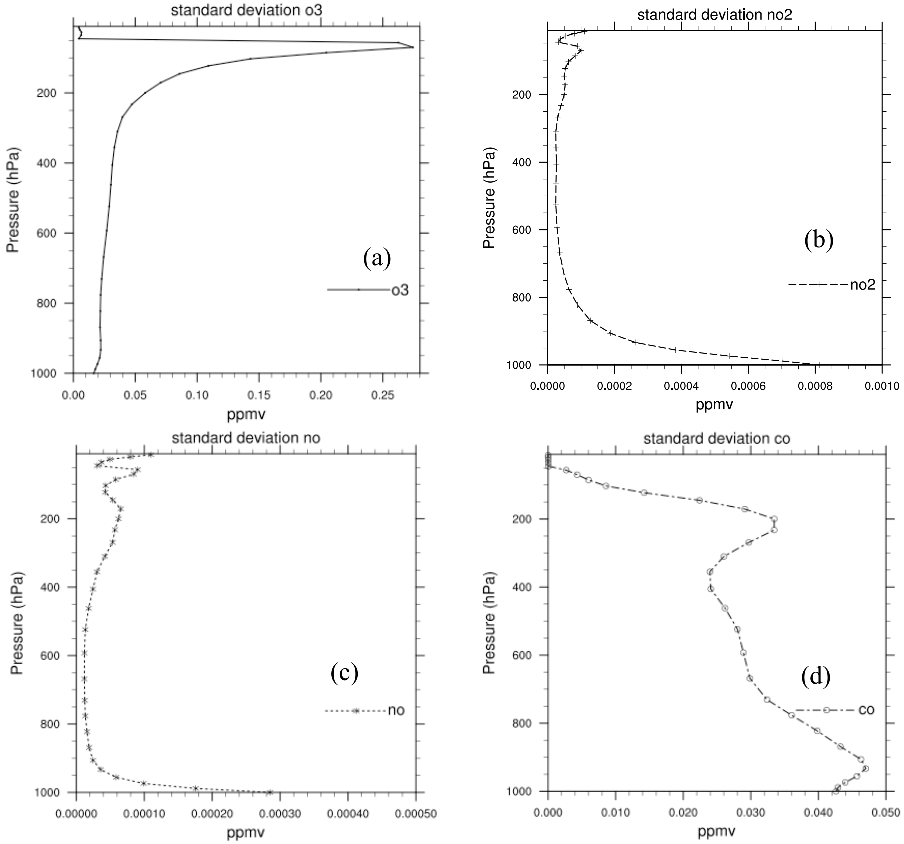

Figure A1. Vertical standard deviation of **(a)** O_3, **(b)** NO_2, **(c)** NO, and **(d)** CO.

GMDD
7, 4291–4352, 2014

Generalized Background Error covariance matrix model (GEN_BE v2.0)

G. Descombes et al.

Title Page
Abstract Introduction
Conclusions References
Tables Figures
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

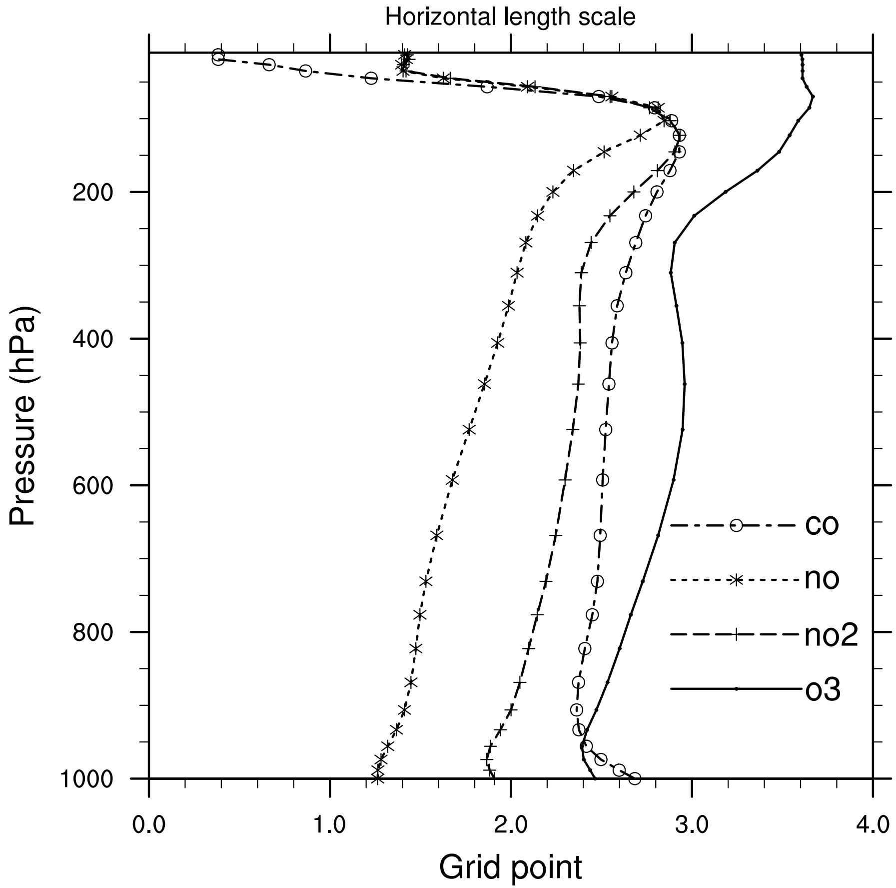

Figure A2. Horizontal length scale of O_3, NO_2, NO, and CO.

GMDD
7, 4291–4352, 2014

Generalized Background Error covariance matrix model (GEN_BE v2.0)

G. Descombes et al.

Title Page
Abstract
Introduction
Conclusions
References
Tables
Figures
◄
►
Back
Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

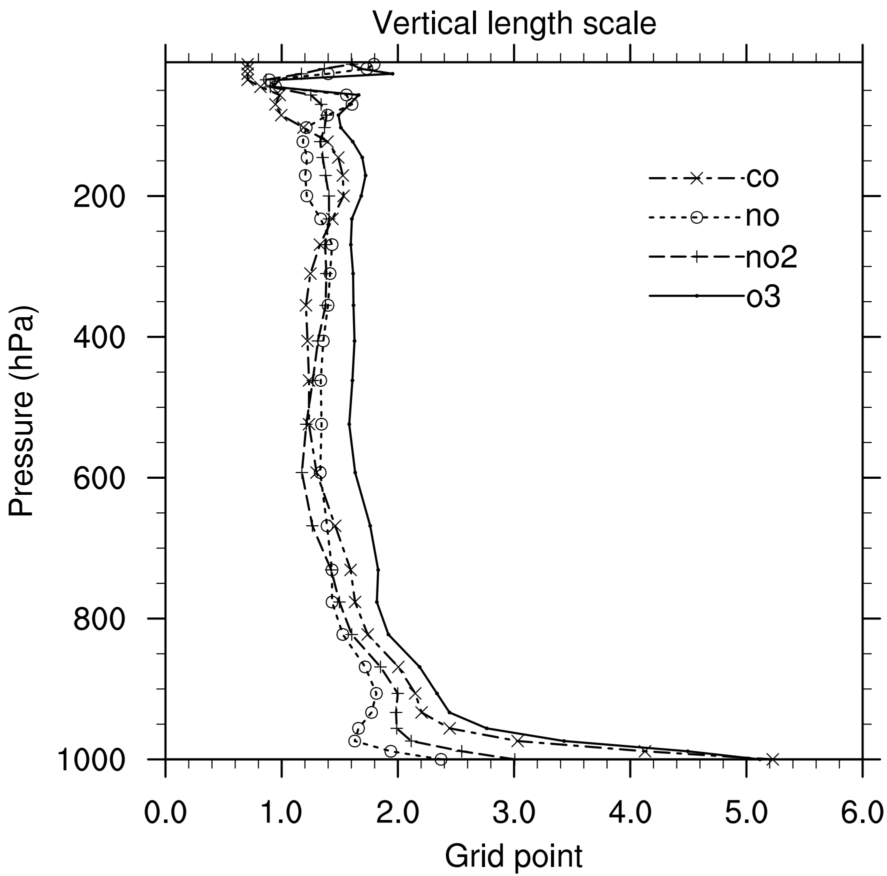

Figure A3. Vertical length scale of O_3, NO_2, NO, and CO.

GMDD
7, 4291–4352, 2014

Generalized Background Error covariance matrix model (GEN_BE v2.0)

G. Descombes et al.