Generalized Background Error Covariance Matrix Model (GEN_BE v2.0)

G. Descombes[1], T. Auligné[1], F. Vandenberghe[2], D. M. Barker[3] and J. Barré[4]

[1]{National Center for Atmospheric Research/MMM, Boulder, Colorado}

[2]{National Center for Atmospheric Research/RAL, Boulder, Colorado}

[3]{Met Office, Exeter, United Kingdom}

[4]{National Center for Atmospheric Research/ACD, Boulder, Colorado}

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

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. GEN_BE allows for a simpler, flexible, robust, and community-oriented framework that gathers methods used by some meteorological operational centers and researchers.

We present the advantages of this new design for the data assimilation community by performing benchmarks of different modeling of **B** 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 implement 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 applied to

other domains of science and be chosen to diagnose and to model **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 the best estimate of the background error covariances matrix (**B**) is a key component for data assimilation improvements, various operational meteorological centers such as the European Centre for Medium-Range Weather Forecasts (ECMWF), the National Centers for Environmental Prediction (NCEP), and the UK Met office, continue to develop new algorithms, techniques, and tools (Bannister, 2008a, b) to model **B** within a variational framework. The probability errors are supposed to be normally distributed and **B** is determined for a limited set of variables, called control variables. The dimensions of **B** are also reduced by diagnosing several parameters that drive a series of operators to model **B.** However, necessities to extend the capabilities of **B** subsist. For example, improving cloud (Auligné et al., 2011) and pollution forecast are major drivers of development of cloud and chemical data assimilation. In the meantime, as more and more observational datasets coming from radars, 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 may involve more variables, which are present in geophysical numerical models.

The opportunity has been taken 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) as a component of a three-dimentional variational data assimilation (3DVAR) method to estimate the background error of the fifth-generation Penn State/NCAR Mesoscale Model (MM5, Grell et al., 1994) for a limited-area system. Since this initial version, various branches of code have been developed at the National Center for Atmospheric Research (NCAR) and at the UK Met Office to address specific needs using different models such as the Weather Research Forecast (WRF, Skamarock et al., 2008) and the Unified Model (UM, Davies et al, 2005) on different data assimilation platforms such as the Weather Research Forecast Data Assimilation system (WRFDA, Barker et al., 2012) and the Grid point Statistical Interpolation system (GSI, Kleist et al., 2009). Different choices of control variables and their correlated errors used to mimic general physical balance (geostrophic, hydrostatic, etc.) in the atmosphere have been largely investigated by different operational centers and referenced in Banister (2008b). Since then, such multivariate

relationship approaches have been studied to characterize hererogeneous background errors in precipitating and nonprecipitating areas for regional applications (Fillon and al. 2010; Montmerle and Berre, 2010). Special emphasis is made in Michel et al. (2011) to include hydrometeors in the background error statistics as their direct analysis increment can come from data assimilation of radar reflectivity and satellites radiances. The framework of the GEN_BE code version 2.0 has been developed to merge these different efforts using linear regression to model the balance between variables, Empirical Orthogonal Functions (EOFs) decomposition techniques and diagnostic of length scales to apply recursive filters (RFs). It allows reading input from different models and providing output for different data assimilation platforms. This new flexibility associated with the possibility to define a set of control variables and their covariance errors as an input should potentially reduce further developments effort of the code and benefit the larger community of geophysical science in general.

This document describes the methods included in the GEN_BE code version 2.0 to investigate modeling of **B** for cloud and chemical data assimilation applications. Section 2.0 presents the role of the background error covariance and how a series of different operators (i.e. balance, vertical and horizontal transforms) can model **B**. The third section describes the general structure of the code, the methods to estimate the different parameters that model **B** and their role in the data assimilation processes. It explains how to modify and extend the control variables and to define multivariate background errors when correlated errors between variables are modeled by linear regression (i.e. balance transform Up). Section 4 presents results of a benchmark performed on two different systems of data assimilation (WRFDA and GSI) using different model of **B** based on WRF model forecast involving the same set of five control variables (referenced as CV5 hereafter) available in GSI (Kleist et al., 2009). Finally, Sect. 5 presents results of a multivariate cloud data assimilation approach that includes hydrometeors as control variables (referenced as CV9 hereafter) and their correlated error with humidity. In addition, the diagnostic of parameters such as standard deviation, vertical and horizontal length scales are discussed for the chemical species carbon monoxide (CO), nitrogen oxides (NO_x) and ozone (O_3) in a variational data assimilation framework.

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):

$$J(\boldsymbol{x}) = J_b(\boldsymbol{x}) + J_0(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{x})^{\mathrm{T}}\mathbf{B}(\boldsymbol{x}_b - \boldsymbol{x}) + \frac{1}{2}[\boldsymbol{y}_0 - H(\boldsymbol{x})]^{\mathrm{T}}\mathbf{R}^{-1}[\boldsymbol{x}_b - H(\boldsymbol{x})] \qquad (1)$$

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 and *H,* called the non-linear observation operator, is a map from the gridded model variables to the observation locations. The J_o term contains **R,** the observational error covariance matrix. The J_b term contains **B**, the background error covariance matrix defined in Eq (2):

$$\mathbf{B} = \overline{(\boldsymbol{x}_b - \boldsymbol{x}_t)(\boldsymbol{x}_b - \boldsymbol{x}_t)^{\mathrm{T}}} \qquad (2)$$

where $\boldsymbol{x}_\mathrm{t}$ is the true state vector and the overbare represent an average over a number of forecasts.

By definition, exact values of **R** and **B** would 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. Often, the **R** matrix is assumed to be diagonal, i.e. uncorrelated observation errors, 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, being comprised of nearly 10^8 x 10^8 = 10^{16} entries, is too large to be calculate explicitly and to be stored in present day computer memories. As a result, the **B** matrix needs to be modeled.

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:

$$\delta \boldsymbol{x} = (\boldsymbol{x}_b - \boldsymbol{x}) = \mathbf{B}^{1/2}\boldsymbol{u} \quad (3)$$

as it improves the conditioning (Courtier et al. 1994) 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:

$$J(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} + \frac{1}{2}(\boldsymbol{d} - \mathbf{H}\mathbf{B}^{1/2}\boldsymbol{u})^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{d} - \mathbf{H}\mathbf{B}^{1/2}\boldsymbol{u}) \quad (4)$$

Where d is the innovation vector defined as $\boldsymbol{d}=(\boldsymbol{y}_o-H(\boldsymbol{x}_b))$ and it represents the difference between observations and their modeled values using a non-linear observation operator H. $\mathbf{H}$ is the linearized observation operator, which makes the cost function quadratic and easier to minimize.

2.2.2 Background errors covariance matrix modelled by a succession of operators.

The square root of the $\mathbf{B}$ matrix as defined in Eq. (3) is decomposed to a series of sub-matrices, each corresponding to an elemental transform that can be individually modeled:

$$\mathbf{B}^{1/2} = \mathbf{U}_p\mathbf{S}\mathbf{U}_v\mathbf{U}_h \quad (5)$$

where:

- The $\mathbf{U}_\mathrm{p}$ matrix, called physical transform or balance operator, defines the set of control variables and their relationships. In practice, the control variables are calulated using the model variables and selected to minimize their cross-correlations. Also, the existing cross-correlations, called balanced part, can be reduced by appling statistical linear regressions (explained Sect. 3.2). The idea is that those new variables are less correlated with each other and so the corresponding off diagonal terms in the matrix vanish.

- The $\mathbf{S}$ matrix is diagonal and composed of the standard deviations of the background errors.

- The $\mathbf{U}_v$ matrix, called vertical transform, defines the vertical auto-correlations for each of the $\boldsymbol{u}$ control variables. It is modeled by either homogeneous Empirical Orthogonal Functions (EOFs) or application of a recursive iterative filter.

- The $\mathbf{U}_h$ matrix, called horizontal transform, 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).

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

3 Five stages to generate the background error covariance statistics (GEN_BE code version 2.0)

The general structure of the GEN_BE code version 2.0 has been designed to split the input, output, and algorithms in independent stages. The five steps, from stage 0 to 4, that model a background error covariance matrix, become independent of the choice of control variables and model input, which allows for more flexibility (Fig. 1). Stage 0 estimates the perturbations of the control variables based on variables coming from a Numerical Weather Prediction (NWP) model forecast. Stage 1 removes the mean of these perturbations and define the applied binning. Stage 2 defines the balance operator ($\mathbf{U}_p$) by estimating covariance errors between the control variables using linear regressions. Stage 3 determines the $\mathbf{S}$ operator by estimating the standard deviation that weighs the analysis increment for a given variable. It also computes the necessary parameters to spread out the information vertically ($\mathbf{U}_v$) in data assimilation processes. Stage 4 computes the horizontal length scale parameter used by the recursive filter to model correlated error on a two dimensional plane ($\mathbf{U}_h$). Technical details are presented in three Appendices. Appendix A describes the new features of the codes and should help to compute and to implement new modeling of $\mathbf{B}$. Appendix B presents the namelist options and Appendix C explains how to compile and run the code.

Here we present results obtained from a numerical experiment with the Advanced Research WRF (WRF-ARW, called WRF hereafter) model involving an ensemble of 50 members (D-ensemble) over the CONtiguous United States (CONUS) domain at 15 km resolution (Res. 15 km Fig. 2). Figure 3, shows the Pressure (hPa) against vertical model levels. Each member, is a six hour forecast valid at 12:00z on 3 June 2012. The Ensemble Adjustement Kalman Filter

(EAKF), coming from the community system Data Assimilation Research Testbed (DART, Anderson et al. 2009), was used by Romine et al. (2014) to generate the analysis ensemble. Table 2, shown in Sect. 4, contains detailed setup information of this data assimilation experiment.

3.1 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 hour minus 12 hour 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 perturbations valid at the same time (Fisher, 2003; Pereira and Berre, 2006). 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 and it may not provide automatically the optimum B for a particular system (Fisher 2003).

Pereira and Berre (2006) highlight the consequences of the evaluation of perturbations using the NMC method versus 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 B1 summarizes the general options to compute these raw perturbations.

Since the number of sample of perturbations can be limited, a strategy to model a static error covariance over an entire domain and filter the sampling noise is used. The statistics are spatially averaged by gathering grid points with similar characteristics. The different options available for this technique, referred as binning, are described in Table B2, and can be setup in the namelist input file (Table B3). The simplest way to compute statistics for a domain can be done by vertical levels (*bin_type=5*). Moreover, such formulation of B, which allows

modeling of homogeneous and isotropic covariance, may be inadequate to specify natural phenomena. Other binning option can be applied to the different transform $\mathbf{U}_p$, $\mathbf{U}_v$, $\mathbf{U}_h$ and **S** to have a heterogeneous formulation of B. For example, options *bin_type*=1, 2, 3, 4 compute statistics 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 (Wu et al., 2002). For example, the statistics of hydrometeors, as cloud liquid water, which are characterized by a high spatial and temporal variability can be skewed (Michel et al., 2011) 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, referred to "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 (see Appendix B). Stage 1 removes the mean of the perturbations and defines the binning which 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.

3.2 Balance through linear regressions (stage2)

Analysis increment for one variable may impact an another if they have correlated errors. The simplest way to model these multivariate error cross-covariances is to use linear regressions that mimic physical balance between variables. First, regression coefficient between variables can be estimated by solving Eq. (6) following the example of the regression of the temperature (t) by the stream function (psi):

$$\alpha_{\mathrm{psi},t}(b,k,l) \bullet VAR_{\mathrm{psi}}(b,k) = COVAR_{\mathrm{psi},t}(b,k,l) \qquad (6)$$

Where $\alpha_{psi,t}$ is the regression coefficient estimated, $COVAR_{psi,t}(b,k,l)$ represents the vertical cross-covariance between t and psi averaged over the vertical level k, l for the given binning class index b, and $VAR_{psi}(b,k)$ is the variance.

In practice, the regression coefficient can be directly calculated as the ratio of the inverted variance with the covariance or by performing a Cholesky decomposition (see Appendix B for more details). Then, linear regressions are performed to derive uncorrelated (i.e. unbalanced) perturbations by removing the balanced part from other perturbation variables.

Eq. (7) shows how the unbalanced part of the t perturbation (δt_u) is deduced by substrating its full perturbation (δt) to its balanced part coming from psi:

$$\delta t_u(i,j,k) = \delta t(i,j,k) - \sum_{j=1}^{N_k} \alpha_{psi\,t}(b,k,l)\delta \mathrm{psi}(i,j,l) \qquad (7)$$

where b is the index of the binning class according to the triplet indexes of the grid point position *(i,j,k)*. N_k is the total number of vertical model levels.

Note, that in variational data assimilation process, balance operator $\mathbf{U_p}$ is applied to the variable themselves. It models correlations between variables and allows to transform the **B** matrix as a block diagonal in the control (uncorrelated) space. The GEN_BE code version 2.0, has been developed to allow the use of a broad set of control variables (shown in Table 1) and to allow the definition of the $\mathbf{U_p}$ transform in a namelist input file. For example, Table B4 presents how to define the balance transform that involves five control variables (CV5) as it can be used in the GSI system developed at NCEP for analyses operational purpose (Kleist et al., 2009). The parameters *covar* equals 1 means the unbalanced part of the velocity potential ($\mathrm{chi_u}$), the temperature (t_u), and the pressure surface ($\mathrm{ps_u}$) are calculated by substracting their balanced part coming from the stream function (psi). Benchmark results of pseudo temperature test involving different modeling of **B** and the same $\mathbf{U_p}$ transform (CV5) are shown Sect 4.

Futhermore, Bannister (2008b) described the $\mathbf{U_p}$ transform used in different operational centers with special emphasis on the definition of the balance operator for humidity. To determine a balance operator, diagnostics of vertical cross-covariance or vertical cross-correlation are helpful to analyze the relationship between variables and can also be done through stage 2. For example, Fig. 4 shows the cross-correlation between humidity and temperature for all atmosphere conditions (mixing dry and wet conditions). The errors are mostly anti-correlated, and specific humidity (Fig. 4a) has weaker correlated errors with respect to temperature than relative humidity (Fig. 4b). Moreover, the errors between specific humidity and temperature become highly correlated close to saturation (Holm et al., 2002; Ménétrier and Montmerle, 2011). At saturation, these statistics likely rely on processes of condensation and precipitation when the released latent heat flux warms the atmosphere (Holm et al., 2002). These characteristics highlight how binning that differentiates background statistics in the presence of clouds can be important according to the choice of

control variables. Thus, various studies have been dedicated to better estimate the background error of humidity in cloudy areas (Carron and Fillon, 2010; Montmerle et Berre 2010; Ménétrier and Montmerle 2011). Carron and Fillon (2010) use the specific humidity (qs) and show benefit to characterize heterogeneous formulation of **B** defined for dry and precipitation areas. For a winter test-case where stratiform-type precipitation is predominant, they explain that geostrophic imbalance in precipitation areas can be characterized by the linear balance operator between the stream function and the mass fields (*t* and ps). Montmerle and Berre (2010) show potential improvements at convective scale by using a rainy mask in a multivariate approach for specific humidity that involves vorticity, divergence, temperature and surface pressure variables. While Ménétrier and Montmerle (2011) show the benefit of balancing the specific humidity only with the mass fields (*t* and ps) for fog data assimilation purposes. Dynamical variables such as vorticity and divergence are not included in the balance humidity operator since they do not drive fog formation processes.

Finally, results of an experiment that include hydrometeors and its correlated errors with humidity (CV9) are presented Sect. 5.1 and defined by the namelist input file Table B5.

3.3 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 ($\mathbf{U}_v$). The first method diagonalizes the VACM performing an EOF decomposition (i.e. computing eigenvectors 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 Daley's formula (1991, p110) for a one dimension homogeneous and isotropic case:

$$L_v = \sqrt{\frac{1}{\frac{\partial^2 \rho(0)}{\partial^2 z}}} \qquad (8a)$$

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

Approximating Eq (8a) with finite difference to the second order derivatives of $\rho(\delta z)$ and assuming ρ *is* symmetric around the origin results in:

$$L_{vp} = \frac{\delta z}{\sqrt{2[1-\rho(\delta z)]}} \tag{8b}$$

where L_{vp} represents the vertical length scale approximate by a parabolic function.

If the correlation is approximated at the origin by a Gaussian function as follows:

$$\rho(\delta z) = \exp(-\frac{\delta z}{2L_{vg}^2}) \tag{9a}$$

the length scale L_{vg} can be written:

$$L_{vg} = \frac{\delta z}{\sqrt{-2\ln\rho(\delta z)}} \tag{9b}$$

Pannekoucke et al. (2008) studied the sensitivity of sampling errors of these formulae and shows that the Gaussian and the parabolic approximation give similar results. Furthermore, the vertical length scale can be computed uniform by vertical model level or binned. Table B6 in Appendix B contains description of the namelist option to define the vertical length scale in stage 3 and the horizontal length scale in stage 4.

3.4 Estimation of the horizontal correlation (stage 4)

Horizontal auto-correlations can be computed for each control variable at each grid point. Figure 5 shows a diagnostic of correlation for a few selected points of the WRF computational domain around 500 m above the ground (model level 5). The stream function (5a) and velocity potential control variables have larger and more isotropic spatial correlations while the temperature (5b) and the humidity (5c) control variables show smaller and anisotropic correlations at different locations. The radius of the area where the correlation overpasses 0.9 is within a range of 100 km to 400 km for stream function while this radius reaches its maximum around 100 km for temperature and humidity. Hydrometeor mixing ratios show even more local structures due to their sparse location on the horizontal and the vertical (5d).

In stage 4, we estimate horizontal length scales averaged by vertical level or EOF mode for a field analysis in a 2-D plane. It represents the radius of influence, calculated in grid point space, around the position of an observation and is an input parameter for recursive filters to spread out horizontally the increment ($\mathbf{U}_h$). The different options available, as described below, are also contained in Table B6.

The first method (*ls_method=1*) employs a distribution function to fit the correlation for a 2-D field by vertical level or by EOF mode as explained in Sect 3.3. If a Gaussian function is chosen, the length scale is determined by solving Eq. (10a):

$$\rho(r) = \exp(-\frac{r^2}{2L}) \tag{10a}$$

where ρ(r) is the correlation calculated for a distance r between two grid points.

If a second order autoregressive (SOAR) correlation function is used, the length scale L is determined by solving Eq. (10b):

$$\rho(r) = (1 + \frac{r}{L}) \bullet \exp(-\frac{r^2}{L}) \tag{10b}$$

However, as this procedure is 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:

$$L = \left(\frac{8 \bullet Variance(\varphi)}{Variance(\nabla^2 \varphi)} \right)^{1/4} \tag{11}$$

Eq. (11) was used by Wu et al. (2002) and is similar to the diagnosic of Pereira and Berre (2006), which was analyzed in Pannekoucke et al. (2008).

The horizontal length scale can be uniformally calculated over a vertical model level or can be statistically binned. Homogeneous recursive filters are able to handle a unique length scale defined by model vertical level, or EOF mode. Inhomogeneous recursive filters (Purser et al., 2003b), as implemented in GSI, are able to handle heterogeneous length scale. In this case, the increment is spread out with a length scale according to the bin class of each grid point. Moreover, spatial filtering to smooth the length scale may be required because of recursive filters normalization issues (Michel and Auligné 2010).

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

We present a benchmark of different modeling of **B** perfomed on the GSI and WRFDA data assimilation platforms. Both systems can handle the set of five control variables (CV5) and their balance operator ($\mathbf{U}_p$) defined Table B4. By default, the GSI system allows using a **B** matrix statisitics ($\mathbf{B}_{nam}$), pre-computed over an enlarged CONUS domain, using the NMC method and NAM (North American Mesoscale) forecasts. $\mathbf{B}_{nam}$ is used with GSI (Wu, 2005) to produce daily forecasts with NDAS (NAM Data Assimilation System; Rogers et al., 2009). Based on the D-Ensemble dataset coming from the DART experiment (i.e. Sect 3. and Romine et al. 2014), we present in Sect. 4.1 the parameters that define the vertical transform $\mathbf{U}_v$ by using EOF decomposition for WRFDA ($\mathbf{B}_{eof}$) and by using recursive filters for GSI ($\mathbf{B}_{rcf}$). Table 2, gathers the general setup that leads to the modeling of these three **B** matrices ($\mathbf{B}_{eof}$, $\mathbf{B}_{rcf}$ and $\mathbf{B}_{nam}$) and additional information about the used datasets. The physics of the model can be found in Romine et al. (2014), Rogers et al. (2009). Sect. 4.2 compares the results of a pseudo single observation test experiment using $\mathbf{B}_{eof}$, $\mathbf{B}_{rcf}$ and $\mathbf{B}_{nam}$ on the WRFDA and GSI data assimilation system.

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 in Fig. 6 for the control variables (CV5) and Fig. 7 shows the associated eigenvalues. 99% of the variance of the stream function and the velocity potential are represented by the first ten and twenty modes respectively, while more than 30 modes are useful for temperature and relative humidity. Also, the EOF decomposition allows optionally some filtering as the largest

variances (i.e. eigen values) are associated with the first EOFs, the latest EOFs may be not taken into account if they mostly represent vertical noise in the system.

The horizontal length scales, estimated by Eq. (11), are presented in Figure 8. The stream function and the velocity potential have the largest length scale value reaching 600 km (39 grid points) for the first EOF mode. While, the unbalanced temperature length scale has a strong variation for the three first EOFs passing approximately from 135 km to 30 km (9 to 2 grid points) and from there, slightly decreases from 30 km to reach 15 km (2 to 1 point grid) for the last EOF mode. Relative humidity length scale remains small, decreasing from approximately 30 km to 15 km as a function of the EOF mode. The unbalanced temperature and the relative humidity have a relatively 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. As the horizontal length scale is associated to EOF mode and not directly related to a vertical model level and futher discussions on the association of length scale with physiscal event may be difficult.

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. (11), applied at every vertical model level for each variable, as shown in Fig. 9. The horizontal length scales diagnosed for each control variable by vertical level (Fig. 9) or by EOF mode (Fig. 8) have the same range of values. The length scales of the stream function and the velocity potential control variables have largest values above 150 km (10 grid points) for all the vertical model levels, while the length scales of temperature and relative humidity remain in a range of 30 km to 60 km (1 to 2 grid points) below 200 hPa level. 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 larger scale events. Direct comparison of these statistics with the $\mathbf{B}_{\text{nam}}$ horizontal length scale is difficult as they are performed with different methods, models, configurations, and physical options (i.e. Table 2). However, it can be noted than the horizontal length scale was approximately twice as small than those for $\mathbf{B}_{\text{nam}}$ (Wu 2005) performed by using the NMC method. Usually, sharper correlations are found in the D-ensemble compared to the NMC method (Fisher, 2003; Pereira and Berre, 2006). Furthermore, a factor contributing to this

difference may arise from the fact that we are comparing statistics from forecasts of different lengths.

The vertical correlation is modeled by the application of recursive filters based on the estimation of the vertical length scale coming from Eq. (8b). 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, decreasing 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 provide a benchmark as it allows visualizing the increment of an isolated observation and its impact on other variables. Thus, the following are pseudo observation tests of temperature with an innovation of 1 Kelvin and an obervation error of 1 Kelvin using different modeling of **B** ($\mathbf{B}_{eof}$, $\mathbf{B}_{rcf}$ and $\mathbf{B}_{nam}$). The position of the pseudo-observation is arbitrarily taken at the center of the domain and at 500 hPa pressure level. The series of plots (Figs 11-13) represent horizontal and vertical slices of the resulting increment for temperature and wind components.

As expected, the horizontal cross-section at the 500 hPa level for temperature shows an isotropic response to the innovation of 1 Kelvin. The maxima of intensity simulated depend on the standard deviation (diagonal matrix **S**) value coming from the **B** matrix.

On one hand, the operator ($\mathbf{U}_v$) employs EOF decomposition, the J_b term of the cost function is weighted by the standard deviation coming from the square root of the eigenvalues of $\mathbf{B}_{eof.}$ On the other hand, $\mathbf{U}_v$ is modeled by the estimation of a length scale and the recursive filters applied on the vertical ($\mathbf{B}_{rcf}$), the analysis is weighted by the standard deviation directly averaged on the vertical mesh grid. The increments of temperature are close for the three different tests and the increment from $\mathbf{B}_{nam}$ is slightly larger than that of $\mathbf{B}_{rcf}$ and $\mathbf{B}_{eof}$. In the case of $\mathbf{B}_{nam}$, recursive filters spread out the information in a larger area over a horizontal plane due to its larger length scales.

For the vertical cross-section (*XZ)*, vertical increments coming from $\mathbf{B}_{rcf}$ and $\mathbf{B}_{eof}$ spread out in the same range of altitude (~ between the 800 hPa and 450 hPa pressure levels). Based on the same D-ensemble datasets, the $\mathbf{U}_v$ operator using EOF decompostion and recursives filters gives similar results on different platforms, as expected. Moreover, the temperature increment

from $\mathbf{B}_{rcf}$ spreads out even more along the vertical compared to the $\mathbf{B}_{nam}$ experiment on the GSI system. This discrepancy can be associated with the computed vertical length scales from two different datasets. The length scales diagnosed over a D-ensemble are larger in this case for $\mathbf{B}_{rcf}$ than the one averaged over a long period of time in the NMC method (60 perturbations selected over a year). Also, statistics of $\mathbf{B}_{nam}$ are perfomed over an Eta grid of 60 vertical levels of WRF-NMM while the statistics of $\mathbf{B}_{rcf}$ and $\mathbf{B}_{eof}$ come from WRF defined on a hybrid-sigma grid of 39 vertical levels. Thus, the raw statistics of $\mathbf{B}_{nam}$ are interpolated on the WRF vertical grid in GSI before performing 3D-VAR data assimilation. Furthermore, differences in the definition of the physics of the model and the assimilated data may be contributing factors.

Finally, the multivariate approach, defined by CV5, induces increments in the wind components. The horizontal cross-section *(XY)* plotted for *U* and *V* showed dipole lobes, which can be explained by the geostrophic balance adjustment that the linear cross-covariances statistics reproduce. The vertical cross-section (*XZ)* follows the isocontour of 0 m s^{-1} for *U* while some differences can be observed on the slices of *V* for the $\mathbf{B}_{eof,}$ $\mathbf{B}_{rcf}$, and $\mathbf{B}_{nam}$ experiments. A larger spread of the V increment along pressure levels is observed for $\mathbf{B}_{eof}$ and $\mathbf{B}_{rcf}$ compared to experiment of $\mathbf{B}_{nam}$.

These ensemble based background error $\mathbf{B}_{eof}$ and $\mathbf{B}_{rcf}$ covariance matrices potentially have more skill in estimating error statistics related to the present meteorological event and using the same model configuration.

5 Cloud and chemistry variational data assimilation

5.1 Generation of a multivariate background error covariance for hydrometeors.

Code modifications have been done in the WRFDA code 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 of the hydrometeors is multivariate 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 3 provides the conversion from vertical model level to pressure level.

5.1.1 Definition of the Balance operator for hydrometeors (CV9)

The $\mathbf{U}_p$ transform CV5 (defined Table B4) is modified in the WRFDA code to include a multivariate analysis for humidity and hydrometeors (Eq. 12a-c). In a first approach, relative humidity (rh) is balanced in Eq. (12a) with the mass fields of unbalanced temperature (t_u), unbalanced surface pressure (ps_u) and does not include dynamic variables such as the stream function (psi) and unbalanced velocity potential (chi_u):

$$\text{rh}_u(i,j,k) = \text{rh}_u(i,j,k) - \sum_{l=1}^{N_k} \alpha_{\text{rh},t_u}(b,k,l) t_u(i,j,k) - \alpha_{\text{rh},\text{ps}_u}(b,k)\text{ps}_u(i,j) \qquad (12a)$$

The statitics coming from GEN_BE v2.0 code, i.e. regression coefficients and unbalance part of the variable can be estimated only by modifying the namelist file input. In this case, the line covar5 of Table B5 that describes the covariances between the fifth control variable, (relative humidity), with the third control variables t_u and the fourth ps_u is: *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. The regression coefficients can be computed without any

assumptions (Figs. 14a-b), or filtered to take into account the perturbations that represent the transition of a non-cloudy to a cloudy area only (Figs 14c-d). This latter choice is made to intensify the statistical relationship of the statistical balance to be able to remove misplaced clouds, or to create clouds. However, we may want to localize this balance 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 spread out by the recursive filters.

$$q_{\text{cloud}}(i,j,k) = q_{\text{cloud}}(i,j,k) - \sum_{l=1}^{N_k} \alpha_{q_{\text{cloud}},\text{rh}_\text{u}}(b,k,l)\text{rh}_\text{u}(i,j,l) \qquad (12b)$$

$$q_{\text{cloud}}(i,j,k) = q_{\text{cloud}}(i,j,k) - \alpha_{q_{\text{cloud}},\text{rh}_\text{u}}(b,k)\text{rh}_\text{u}(i,j,l) \qquad (12c)$$

Similar balance is applied to q_{ice}. q_{rain} and q_{snow} are defined univariate. Table B5 summarizes the definition of this balance operator called CV9

5.1.2 Statistics of the background error covariance matrix for hydrometeors.

The vertical and horizontal transforms retained are the recursive filters making the interpretation of the length scale parameter easier as they are directly associated to a vertical model level. The four main hydrometeors have been added in this study, as they could be useful for data assimilation in remote sensing such as satellite cloud radiances and radar reflectivity.

The horizontal length scale values of the different hydrometeors shown in Fig. 15a are smaller in comparison of other control variables (less than 30 km, 2 grid points). Significant values of length scale, that overpass 15 km (1 grid point), are related to the presence of hydrometeors: it occurs below 150 hPa pressure level for q_{ice} and q_{snow} and below 400 hPa pressure level for q_{cloud} and q_{ice}. The maximum of q_{cloud} length scale, located approximately at 950 hPa, can be associated to the presence of low maritime clouds above the Pacific ocean noted by the high standard deviation in Figs 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 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 low presence of snow hydrometeors in the first model levels, i.e. close to the ground, is characterized by small values of mixing ratio, averaged by vertical level, which tends to artificially reinforce vertical correlation, as well.

5.1.3 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 general information by vertical level and binning type without any assumption of horizontal location. It is mostly used when the perturbations come from the NMC method or when the variance is not diagnosed at the analysis time. In our test case, the increment is modulated by the 3-D variance computed from a 6-hour 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 greater than 10^{-3} g/kg where the variability of cloud presence exists (Fig. 17). The

strongest increment occurs over the Pacific Ocean for higher q_{cloud} standard deviation. A minimum value would likely need to be set to retain possibility of increments in the dry area.

The covariance between mixing ratio of cloud water condensate and relative humidity, described in Sect. 5.1.1 can reinforce the ability of adding clouds in the dry area or removing clouds in the cloudy area. The univariate version of the balance operator for hydrometeors may be beneficial at the analysis time as hydrometeors can be directly assimilated. 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 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 the D-ensemble may help to balance those high sensible variables. Furthermore, operational centers, such as Météo-France with the Application of Research to Operations at Mesoscale system (AROME, Seity et al., 2011) and the Met Office with the Met Office Global and Regional Ensemble Prediction System (MOGREPS, Bowler et al., 2008; Migliorini et al., 2011), 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 of CPU (Central Processing Unit) 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 (Pannekoucke et al., 2008; Ménétrier et al., 2014). When the ensemble size is small, methods that combine general statistics of the background errors and local balance are found to perform better (Hamill 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.

5.2 Background Error for Chemical Species

As a proof of concept, this last section shows the direct applicability of the GEN_BE v2.0 code as a diagnostic tool for other topics than meteorology. In recent decades, a large number 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; and 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.

Characterization of the background error covariance matrix **B** in chemistry is a very important aspect of a successful data assimilation system. During the last few years, different studies have used different techniques to characterize the **B** matrix. Barré et al. (2013) and Emili et al. (2014) estimated a quasi-constant **B** based on the Ménard and Chang (2000) and Desroziers et al. (2005) *a posteriori* statistics, for tropospheric and stratospheric ozone data assimialtion. Since the latter studies put their interests on large-scale events (global scale chemical assimilation and synoptic events) data assimilation perform reasonably well with those first order **B** matrix estimation. Depending on the region of the atmosphere that is analyzed **B** needs to be updated at different timescales. Massart et al. (2012) showed the importance of using a monthly **B** matrix ensemble estimate for stratospheric ozone data assimilation purposes. For surface ozone assimilation Jaumouillé et al. (2012) and Gaubert et al. (2014) showed that an hourly ensemble estimate of **B** that represent diurnal variations of model errors improves the data assimilation skills. The last few years, studies on aerosol data assimilation within WRF-Chem (Pagowski et al., 2010, 2014, Schwartz et al., 2012) showed the importance of having a detailed estimation of the **B** matrix.

Statistics were analyzed in detail to ensure that **B** reproduced relevant correlation structures during 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 its usage on different platforms. Several chemical trace gases such as CO (Carbon Monoxide), NO_x (Nitrogen Oxides) and O_3 (Ozone)

but also dust, sea salt and particulate matter (PM) have been already included as new possible control variables in the GEN_BE code. Results for CO, NO_x and O_3 are shown next.

The statistics are estimated using 20 members over the CONUS domain. Each member comes from a 12h forecast of WRF-CHEM (WRF model coupled with Chemistry, Grell et al., 2005), valid at 12:00z on 14 June 2008, at 36 km of horizontal resolution and 33 vertical levels. The lateral boundary conditions coming from MOZART (Model for OZone And Related chemical Tracers, Emmons et al., 2010) and emission factors coming from MEGAN (Model of Emissions of Gases and Aerosols from Nature, Guenther et al., 2006) are perturbed using a pseudo-normal random noise. In order to avoid unphysical or negative values of concentration and emissions and keep ensemble mean boundary conditions values close to the original values, we then perturb the boundary conditions (emissions and boundary conditions) by using a standard deviation (sigma) of 25% of the original boundary condition value and we limit the perturbation to be no more than 3 sigma (i.e. 75%).

Figure 19 present the standard deviations for the chemical species of interest. Standard deviation of the background error is directly related to the species concentrations. Most of the ozone variability takes place in the middle atmosphere (stratosphere) in the ozone layer around 100 hPa (Fig. 19a). Figures 19b and c highlight NO_x concentration fluctuations, due to photochemistry in the stratosphere and in the troposphere. Because the NO_x are also emitted from the ground with a short lifetime, a strong peak of standard deviation is observed. Carbon monoxide (Fig. 19d), which is also emitted at the surface and has relatively long life time (1-2 months), shows significant standard deviation values in all the troposphere with a maximum in the boundary layer.

Figure 20 displays the calculated horizontal chemical length scales. Ozone show horizontal length scales are around 100 km in the troposphere and around 125 km in the stratosphere. Pagaowski et al., 2010, used a NMC method and found that ozone horizontal length scale are around 100 km (150 km) in the troposphere (in the stratosphere). Concerning NO_2, GEN_BE v2.0 evaluates the tropospheric horizontal length scale between 70 km and 90 km. This range of values is consistent with the values found by Silver et al., 2013 that uses the NMC method. Horizontal length scales increase in the upper troposphere mostly due to the strong circulation (jets) and increased lifetime of species. Then strong advection of trace gases that are not short lived (with lifetimes that are more than a day) are likely to increase the horizontal correlations.

Concerning the vertical correlations (Fig. 21), all the 4 species diagnosed, present a maximum close to the surface where they are emitted (or secondarily produced for ozone). Correlation length scales sharply decrease between 1000 hPa and 850 hPa. Two main reasons explain this: (1) reactions with other short-lived species emitted near the surface create strong correlations in the lowest model levels, (2) increase of first model levels layer thickness from the surface to levels above creates stronger correlations in grid point. This strong decrease of correlation lenght scales is not fully understood and needs further investigations. Above the surface peak, vertical correlation also decrease around 800 hPa due to weaker vertical mixing above the planetary boundary layer. In the free troposphere where the vertical mixing is less significant, the evolution of the vertical length scale decreases slowly from approximately 70 km to 40 km. The vertical diffusion of possible data assimilation increments will be less significant than in the boundary layer. Compared to Pagowski et al., 2010, the ozone vertical length scale profile presents the same behavior. Strong vertical correlation close to the surface, followed by a strong decrease to the levels directly above. Resulting in lower values in the upper levels of the boundary layer.

Here we have shown that the GEN_BE v2.0 code is able to model a **B** matrix for chemical variables with features that are associated with physical processes i.e. ozone layer, tracer lifetime, emissions and planetary boundary layer mixing. 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 even if data assimilation does not involve the origin of the mismatch. Hybrid and ensemble methods may help to diagnose complex covariance structures in future work. In this paper, the chemical **B** matrix generated by GEN_BE v2.0 has not been extensively diagnosed. More investigations such as, the balance between chemical species, standard deviation and correlation length time and space variability could be investigated in further studies by the atmospheric chemistry modeling community using GEN_BE v2.0.

6 Summary and discussions

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 observations 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 first 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 based on WRF and WRF-CHEM forecasts. Parameters such as length scales, eigenvectors, eigenvalues, standard deviation and linear regression coefficients were first estimated for the control variables (CV5) described in Kleist et al. (2009) for the GSI system developed at NCEP.

Second, the GEN_BE v2.0 code has been validated through multivariate single observation tests of temperature using three different modeling of **B** ($\mathbf{B}_{eof}$, $\mathbf{B}_{rcf}$, and $\mathbf{B}_{nam}$) and on two different platforms. Based on the first dataset, D-ensemble, the single observation test performed with $\mathbf{B}_{eof}$ ($\mathbf{U}_v$, EOF decompostion) in WRFDA shows similar results than the single observation test of temperature performed with $\mathbf{B}_{rcf}$ ($\mathbf{U}_v$ recursive filters) in GSI. The increments were spread out in a larger area along the vertical than those coming from the test using the $\mathbf{B}_{nam}$ statititics calculated with the NMC method on a different vertical grid. While, the horizontal increments were spread out in a larger area using $\mathbf{B}_{nam}$.

Third, the GEN_BE code has been used to perform the statistics over an extended set of control variables that include mixing ratio of hydrometeors (CV9) for multivariate cloud data assimilation purpose. 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 hydrometeors of cloud and ice condensate water are also balanced with humidity to be potentially able to create or remove misplaced

clouds. The regression coefficients calculated, can be conserved for a next cycle analysis as they are averaged by bins or recalculated as they are not so expansive with regard to CPU time. In this paper, a pseudo observation test of cloud mixing ratio was performed using WRFDA and the next step is to test cloudy radiance data assimilation. Finally, statistics of background are estimated for chemical species such as carbon monoxide (CO), nitrogen oxides (NO_x) and ozone (O_3) coming from an ensemble of forecasts of WRF-CHEM, discussed and compared with existing studies. It has been shown that the statistics diagnosed are related to physical and chemical processes.

In these previous examples, GEN_BE code version 2.0 can handle input datasets coming from WRF, a model defined on a C-Arakawa grid, and the background error statistic outputs are computed on unstaggered A-Arakawa grid. Within minor modifications, the code would be able to handle other horizontal grids. Also, statistics could easily be done on models with different vertical grid definition. If we consider performing the background errors statistics on an unstructured grid, the structure of the code can remain the same but few mathematical operators, such as differential and laplacian, and estimation of the distance between two grid points, would need to be re-defined according to the grid. In fact, the $\mathbf{U_p}$ transform needs to be performed in the unstructured grid according to the user's choice of control variables. $\mathbf{U_v}$ transform will remain identical and $\mathbf{U_h}$ transform would be modified according to the mathematical operators. Another option would be to interpolate first the input dataset on a regular grid according to the data assimilation system used and then compute the statistics. Thus, implementation of models with different grid can be done in the GEN_BE v2.0 code based on its general framework and may be completed by adding new diagnostics.

The current trend is to model a more complex background error, expanding the control variables and correlated errors and using techniques to achieve more heterogeneity and anisotropy. The geographical binning and the 3-D variance available in the GEN_BE v2.0 code can be utilized with new data assimilation algorithms. For example, hybrid data assimilation that combines variational and ensemble methods may be helpful especially by adding flow dependence in the estimation of the background error while keeping a reduced ensemble size due to CPU time constraints (Hamill and Snyder, 2000). Wang et al. (2008a, 2008b) performed a study using a hybrid 3DVAR-ETKF (Ensemble Transform Kalman Filter) technique that combines static (modeled) with ensemble background error covariances. Better results were obtained over North America at a coarse resolution (200 km) especially in

data-sparse areas compared to those performed solely with 3DVAR. The extended control variable technique (Lorenc 2003) allows blending flow dependent errors with static covariance errors. Bannister et al. (2011b) investigated the benefit of a convection permitting prediction system ensemble (24 members) at a finer scale (i.e. 1.5 km of resolution) for nowcasting purposes based on MOGREPS (Migliorini et al. 2011a). Even though, the authors show how general balances that drive synoptic flow, in particular geostrophic balance, can diminish in convective situations at small scales, they highlight the necessity for a data assimilation system to better represent both the large scale and mesoscale components of the flow. In addition, Ménétrier et al. (2014) studied heterogeneous flow dependent background error covariances at a convective scale and showed that a small ensemble (6 members from AROME) contains relevant information together with sampling noise, which can be reduced through filtering. Finally, the GEN_BE v2.0 code may be a tool to diagnose inhomogeneous 3-D localization parameters in ensemble methods. The code has been tested in atmospheric science but the flexibility of the code may be useful in other geophysical applications.

Appendix A: FORTRAN code and input/output description.

New FORTRAN modules have been developed to generalize the calculation of the error covariance matrix from different input models and for new control variables. Table A1 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. 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. The NetCDF format has been chosen to improve robustness and flexibility in the input and output of the different stages as shown in Table A2. The final NetCDF output file be.nc contains all the information needed for a variational data assimilation system, as shown in Table A3. 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.

Appendix B: Description of the namelist options.

The "namelist.input" file drives the different stages 0 to 4 contains four different sections.

The namelist section "*&gen_be_info*", described in Table B1, defines the options to compute perturbations in stages 0 and 1 from input forecast model (e.g. WRF). Also, the data assimilation system can specified.

Table B2, presents eight binning available option and Table B3 explains how to set up the namelist section *"&gen_be_bin"*. In the GEN_BE code version 2.0, all the information that defines a binning option are encapsulated in the type bins_type. Since the algorithms of the different stages from 1 to 4 do not make any specific assumption on the binning option used, the implementation of new option is simplified as it needs to be defined just once in the *da_create_bins* FORTRAN routine of the module *io_input.f90*. In case of implementing geographical mask, developers have to introduce the method to update the mask in the *update_dynamical_mask* routine. All information related to binning is contained in the NetCDF file bin.nc created in stage 1.

The $\mathbf{U_p}$ transform is defined in section *"&gen_be_cv"* where the used control variables and balance operator are set up. Table B4 presents the CV5 control variable currently used in the GSI system (Kleist et al. 2009). In this example, the use of the relative humidity (rh) (line covar5) allows to performed statistics in GEN_BE for the normalized relative humidity described by Holm (2002) and implemented in GSI. Furthermore, when the regression coefficients are computed for a GSI regional application, a Cholesky decomposition is used and additional filtering is applied to the regression coefficients between stream function and temperature, and between stream function and pressure surface. This part of the code coming from the NCEP is flagged with *use_cholesky* variable in the gen_be_stage2.F FORTRAN program, and the called subroutines are contained in the *io_output_applications.f90* Fortran module. Table B5 shows the $\mathbf{U_p}$ transform, called CV9, which includes hydrometeors in a multivariate approach.

Table B6 contains the namelist section *"&gen_be_lenscale"* to diagnose parameters of the $\mathbf{U_v}$ and $\mathbf{U_h}$ transforms for stage 3 and stage 4 respectively. The vertical transform $\mathbf{U_v}$ can be performed by estimating a vertical length scale by model levels (*data_on_level=true*) or by a EOF decomposition (*data_on_level=false*). By default, statistics are binned with the same

option defined section *"&gen_be_bin"* of the namelist.input file. Otherwise, the statistics are averaged by vertical level if the flag *global_bin* is true (which is equivalent to the defininition of *bin_type=5*).

Appendix C: Installation, compilation, set up and visualization.

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 the Intel FORTRAN compiler, the Portland Group FORTRAN compiler, and the 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 in Appendix B. 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).

Acknowledgements

Funding for this work was provided by the U.S. Air Force Weather Agency. The authors benefited from numerous discussions with Yann Michel. Glen Romine is thanked for providing the ensemble over the CONUS domain. Syed Rizvi is thanked for discussions concerning the previous version of the code.

References

Anderson, J., Hoar ,T., Raeder, K., Liu, H., Collins, N., Torn, R., and Avellano A.: The data assimilation research testbed: A community facility, B. Am. Meteorol. Soc., 90, 1283-1296., doi: http://dx.doi.org/10.1175/2009BAMS2618.1, 2009.

Auligné, 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: http://dx.doi.org/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.

Bannister, R., Migliorini, S. and Dixon, M: Ensemble prediction with a convection-permitting model for nowcasting. Part II: Forecast error statistics. Tellus 63A, 497–51, DOI: 10.1111/j.1600-0870.2010.00500.x, 2011.

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: http://dx.doi.org/10.1175/BAMS-D-11-00167.1, 2012.

Barré, J., Peuch, V.-H., Lahoz, W. A., Attié, J.-L., Josse, B., Piacentini, A., Eremenko, M., Dufour, G., Nedelec, P., von Clarmann, T. and El Amraoui, L.: Combined data assimilation of ozone tropospheric columns and stratospheric profiles in a high-resolution CTM. Q. J. Roy. Meteorol. Soc., 140: 966–981. doi: 10.1002/qj.2176, 2014.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B. and Beare, S. E: The MOGREPS short-range ensemble prediction system. Q. J. Roy. Meteorol. Soc. 134, 703–722, DOI: 10.1002/qj.234, 2008.

Caron, J. F. and Fillion L.: An Examination of Background Error Correlations between Mass and Rotational Wind over Precipitation Regions. Mon. Weather Rev., 138 (2), 563–578, doi: http://dx.doi.org/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. (1994), 120, pp.1367-1387, 1994.

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

Davies, T., Cullen, M., J., P., Malcolm, A., Mawson, M., Staniforth., A., White, A. and Wood, N.:A new dynamical core for the Met Office's global and regional modelling of the atmosphere. Q. J. R. Meteorol. Soc. 131: 1759–1782, doi: 10.1256/qj.04.101, 2005.

Desroziers, G., Berre, L., Chapnik, B. and Poli, P. (2005), Diagnosis of observation, background and analysis-error statistics in observation space. Q. J. Roy. Meteorol. Soc., 131: 3385–3396. doi: 10.1256/qj.05.108

Elbern, H., Schimdt, H., and Ebel, A.: Variational data assimilation for tropospheric chemistry modeling, J. Geophys Res. Rev., 102, 15 967–15 985, 1997.

Emili, E., Barret, B., Massart, S., Le Flochmoen, E., Piacentini, A., El Amraoui, L., Pannekoucke, O., and Cariolle, D.: Combined assimilation of IASI and MLS observations to constrain tropospheric and stratospheric ozone in a global chemical transport model, Atmos. Chem. Phys., 14, 177-198, doi:10.5194/acp-14-177-2014, 2014.

Emmons, L. K., Walters, S., Hess, P. G., Lamarque, J.-F., Pfister, G. G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S. L., and Kloster, S.: Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4), Geosci. Model Dev., 3, 43-67, doi:10.5194/gmd-3-43-2010, 2010.

Fisher, M., 2003: Background error covariance modelling. Proceedings of the ECMWF Seminar on Recent developments in data assimilation for atmosphere and ocean, 45-63, 8-12 September 2003.

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: http://dx.doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2, 2000.

Holm, E., Andersson, E., Beljaars, A., Lopez, P., Mahfouf, J-F., Simmons, A.J. and Thepaut, J-N. Assimilation and Modelling of the Hydrological Cycle: ECMWF's Status and Plans, Technical Memoranda 383, 2002.

Gaubert, B., Coman, a., Foret, G., Meleux, F., Ung, A., Rouil, L., Ionescu, A., Candau, Y., and Beekmann, M.: Regional scale ozone data assimilation using an ensemble Kalman filter and the CHIMERE chemical transport model. Geoscientific Model Development, 7(1), 283–302. doi:10.5194/gmd-7-283-2014, 2014.

Grell, G.A., Dudhia J., and Stauffer D.: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Technical Note NCAR/TN-398+STR, DOI: 10.5065/D60Z716B, 1994.

Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G. J., Skamarock, W., and Eder, B.: Fully-coupled online chemistry within the WRF model, Atmos. Environ., 39, 6957–6975, 2005.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C.: Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), Atmos. Chem. Phys., 6, 3181-3210, doi:10.5194/acp-6-3181-2006, 2006.

Jaumouillé, E., Massart, S., Piacentini, A., Cariolle, D., & Peuch, V.-H.: Impact of a time-dependent background error covariance matrix on air quality analysis. Geoscientific Model Development, 5(5), 1075–1090. doi:10.5194/gmd-5-1075-2012, 2012.

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, http://journals.ametsoc.org/doi/pdf/10.1175/2009WAF2222201.1, 2009.

Lorenc, A., C.: The potential of the ensemble Kalman Filter for NWP-A comparison with 4-D VAR. Q. J. R. Meteor. Soc. 595: 3183-3203, DOI: 10.1256/qj.02.132, 2003.

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.

Ménard R, Chang LP. 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part II: χ2-validated results and analysis of variance and correlation dynamics. Mon. Weather Rev. 128: 2672–2686.

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.

Ménétrier, B., Montmerle, T., Berre, L. and Michel Y.: Estimation and diagnosis of heterogeneous flow-dependent background-error covariances at the convective scale using either large or small ensembles, Q. J. Roy. Meteor. Soc., 140, 2050–2061, DOI:10.1002/qj.2267, 2014.

Michel Y. and Auligné T.: Inhomogeneous Background Error Modeling Over Antarctica. Mon. Weather Rev., 138, 6, pp. 2229–2252, doi: http://dx.doi.org/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, 9, 2994-3015, doi: http://dx.doi.org/10.1175/2011MWR3632.1, 2011.

Migliorini, S., Dixon, M., A., G., Bannister, R., N., and Ballard, S., P.: Ensemble prediction for nowcasting with a convection-permitting model – Part I: description of the system and the impact of radar- derived surface precipitation rates. Tellus 63A, 468–496, DOI:10.1111/j.1600- 0870.2010.00503.x, 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.

Pagowski, M., Grell, G. A., McKeen, S. A., Peckham, S. E., and Devenyi, D.: Three-dimensional variational data assimilation of ozone and fine particulate matter observations: some results using the Weather Research and Forecasting – Chemistry model and Grid-point

Statistical Interpolation, Q. J. Roy. Meteorol. Soc., 136, 2013–2024, doi:10.1002/qj.700, 2010.

Pagowski, M., Liu, Z., Grell, G. A., Hu, M., Lin, H.-C., and Schwartz, C. S.: Implementation of aerosol assimilation in Gridpoint Statistical Interpolation (v. 3.2) and WRF-Chem (v. 3.4.1), Geosci. Model Dev., 7, 1621-1627, doi:10.5194/gmd-7-1621-2014, 2014.

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

Parrish, D. F., and J. C. Derber: 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: http://dx.doi.org/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: http://dx.doi.org/10.1175//1520-0493(2003)131<1524:NAOTAO>2.0.CO;2, 2003a.

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: http://dx.doi.org/10.1175//2543.1, 2003b.

Rogers, E., DiMego, G., Black, T., Ek, M., Ferrier, B., Gayno, G., Janjic, Z., Lin, Y., Pyle, M., Wong, V., Wu, W. S., and Carley, J.: The NCEP North American Mesoscale Modeling System: Recent changes and future plans, 23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction, https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154114.htm, 2009

Romine, G., S., Schwartz C., S., Berner J., Fossell, R., K., Snyder C., Anderson J. and Weisman M., L.: Representing forecast error in a convection-permitting ensemble system, Mon. Weather Rev., doi: http://dx.doi.org/10.1175/MWR-D-14-00100.1, 2014.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V., 2011: The AROME-France Convective-Scale Operational Model. Mon. Weather Rev., 139, 976–991, doi: http://dx.doi.org/10.1175/2010MWR3425.1, 2011.

Skamarock, W., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D., Duda, M. G., Huang, X. Y., and Wang, W., 2008: A Description of the Advanced Research WRF Version 3. NCAR Technical Note NCAR/TN-475+STR, doi: 10.5065/D68S4MVH, 2008.

Schwartz, C. S., Liu, Z., Lin, H. C., and McKeen, S. A.: Simulta- neous three-dimensional variational assimilation of surface fine particulate matter and MODIS aerosol optical depth, J. Geophys. Res., 117, D13202, doi:10.1029/2011JD017383, 2012.

Silver, J. D., Brandt, J., Hvidberg, M., Frydendall, J., and Christensen, J. H.: Assimilation of OMI NO2 retrievals into the limited-area chemistry-transport model DEHM (V2009.0) with a 3-D OI algorithm, Geosci. Model Dev., 6, 1-16, doi:10.5194/gmd-6-1-2013, 2013.

Wang, X., D., M., Barker, C. Snyder, Hamill, T., M.: A hybrid ETKF-3DVAR data assimilation scheme for the WRF model. Part I: observing system simulation experiment. Mon. Wea. Rev., 136, 5116-5131, doi: http://dx.doi.org/10.1175/2008MWR2444.1, 2008.

Wang, X., D., M., Barker, C. Snyder, Hamill, T., M.: A hybrid ETKF-3DVAR data assimilation scheme for the WRF model. Part II: real observation experiments. Mon. Wea. Rev., 136, 5132-5147. doi: http://dx.doi.org/10.1175/2008MWR2445.1, 2008.

Wu, W., S.: Background error for NCEP's GSI analysis in regional mode, Proc 4th WMO International Symposium on Assimilations of Observations in Meteorology and Oceanography, Prague, Czech Republic, 2005.

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: http://dx.doi.org/10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2, 2002.

Table 1: 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 water mixing ratio
q_{rain}	Rain water mixing ratio
q_{ice}	Ice mixing ratio
q_{snow}	Snow mixing ratio
sst	Sea Surface Temperature

Table 2: Description of the setup of the background error matrix modelling diagnosed over the CONUS Domain. $\mathbf{B}_{eof}$ and $\mathbf{B}_{rcf}$ are diagnosed using GEN_BE code version 2.0 and the D-Ensemble method while $\mathbf{B}_{nam}$ is performed by NCEP using the NMC method.

B Modeling		
	$\mathbf{B}_{eof}$ and $\mathbf{B}_{rcf}$	$\mathbf{B}_{nam}$
Model configuration	WRF model, resolution 15 km, 39 verticals levels on sigma hybrid grid	WRF-NMM model, resolution 12 km, 60 verticals levels on eta grid
Data assimilation setup	DART, EAKF with adaptative covariance inflation, cycling period of 6 h, Perturbated Boundary conditions from GFS, assimilation of conventional and cloud track winds observations	NDAS-GSI system, cycling period of 3 h, boundary conditions from GFS, assimilation of conventional and satellite radiances (clear-sky) observations
Method to compute the perturbations	D-ensemble method applied to 50 pertubations coming from 6 h forecasts of the different members of the ensemble	NMC method applied to 60 perturbations taken over a year, coming from time-lagged forecasts of 12 h and 24 h valid at the same time
B transforms	CV5 control variables	CV5 control variables
	$\mathbf{B}_{rcf}$: $\mathbf{U}_h$ and $\mathbf{U}_v$ transforms defined by recursives filters	$\mathbf{B}_{nam}$: $\mathbf{U}_h$ and $\mathbf{U}_v$ transforms defined by recursives filters
	$\mathbf{B}_{eof}$: $\mathbf{U}_v$ transform defined by EOF decomposition	Statistics $\mathbf{B}_{nam}$ binned by latitude band of 1 degree
	Statistics of $\mathbf{B}_{rcf}$ et $\mathbf{B}_{eof}$ averaged by vertical level.	

Table A1. 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

Table A2. 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 exists 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 split 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 split by vertical level.
gen_be_stage4.F	var(_u).ccyymmddhh.ennn.kkk	sl_print.blll.qcloud	Intermediate ASCII file format that contain the horizontal length scale.
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 readable by GSI.

Table A3. 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 length scale in EOFs space or physical space
vert_lenscale_cv1	Vertical length scale available only if the flag data_on_levels is true and the control variable number 1 is 3D.
vert_variance_cv1	Vertical variance of the control variable number 1 per bin
eigen_value_cv1	Eigenvalue of the control variable number 1 only available if the flag data_on_levels is false
eigen_vector_cv1	Eigenvector of the control variable number 1 only available if the flag data_on_levels is false
varce_cv1	Variance 3D
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 1D, 2D or 3D
vert_autocov_cv1	Vertical autocovariance of the control variable number 1
Binning parameters	
bin_type	Bin_type option selected
bin2d	Binning field 2D array
bins	Binning field 3D array

Table B1. General information defining the experiment in the namelist iput file (&gen_be_info part).

&gen_be_info	Namelist options	Description
model	'WRF'	Set up the acronym for the model input allows GEN_BE to read different input model in the stage 0.
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 data available, defined in hour (useful for the NMC method only).

Table B2. 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 define the width that splits the bins.
3	Binning by vertical model level 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 vertical model level and along the y direction.
5	Binning on vertical model level including all the horizontal point.
6	Average over all points.
7	Binning rain/no-rain by vertical model level and based on thresholds in the model background (Michel and al., 2011.).

Table B3. Parameters defining the binning options of the namelist input file (&gen_be_bin part).

&gen_be_bin	Namelist options	Description
bin_type	0-7	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)

Table B4. Information related to the control variables and their covariance errors in the namelist input file (&gen_be_cv part, example CV5). At present, the parameter covar can take three values: 0, 1, and 2, meaning "no regression", "full regression" and "diagonal only".

&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 does 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 cholesky decomposition (specific to GSI, CV5).

Table B5. 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_cv	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, 2, 0, 0, 0, 0, 0,
covar7	0, 0, 0, 0, 2, 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,

Table B6. 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) in stage 3
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 horizontal length scale (stage 4) See Sect. 3.4 for more details.
use_med_ls	'false'	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	'false'	The statistics can be binned (use_global_bin=false) or not in stages 3 and 4. Only inhomogeneous recursive filters can handle binned length scale.

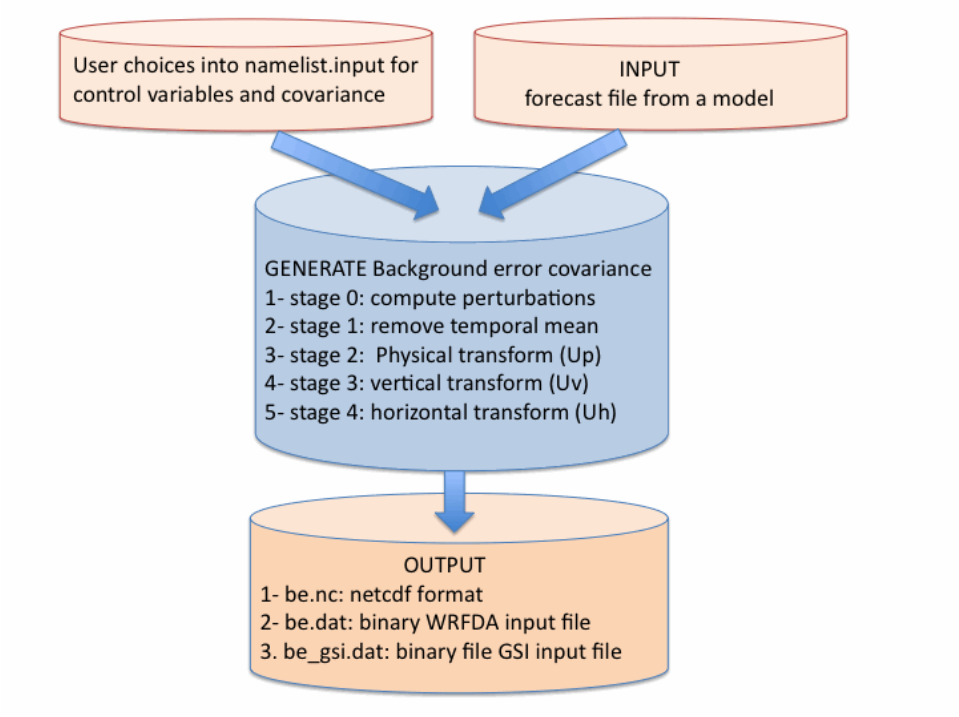


Figure 1. General structure of the code to generate a background error covariance matrix. The input and output are represented by the orange boxes and the five main stages that lead to model **B** are in blue.

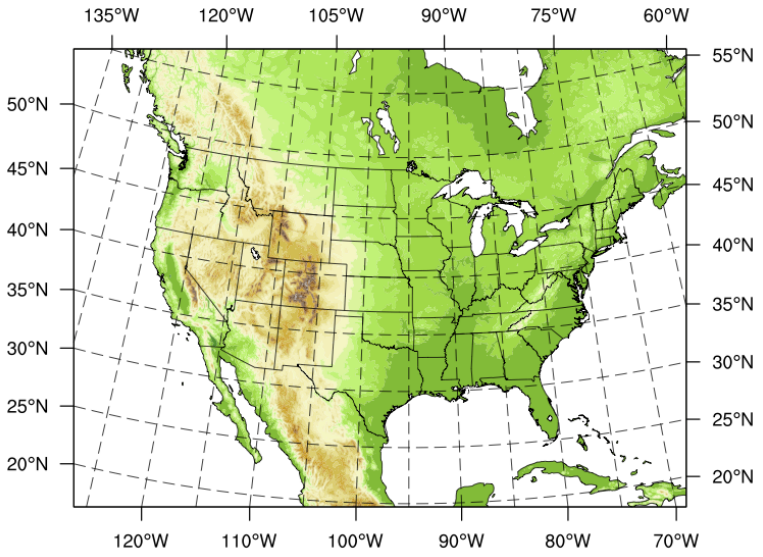


Figure 2. WRF domain over the conus area at the resolution of 15 km. Based on this configuration, the 50 members coming from a 6h forecast (DART experiment) are used to generate background error statistics.

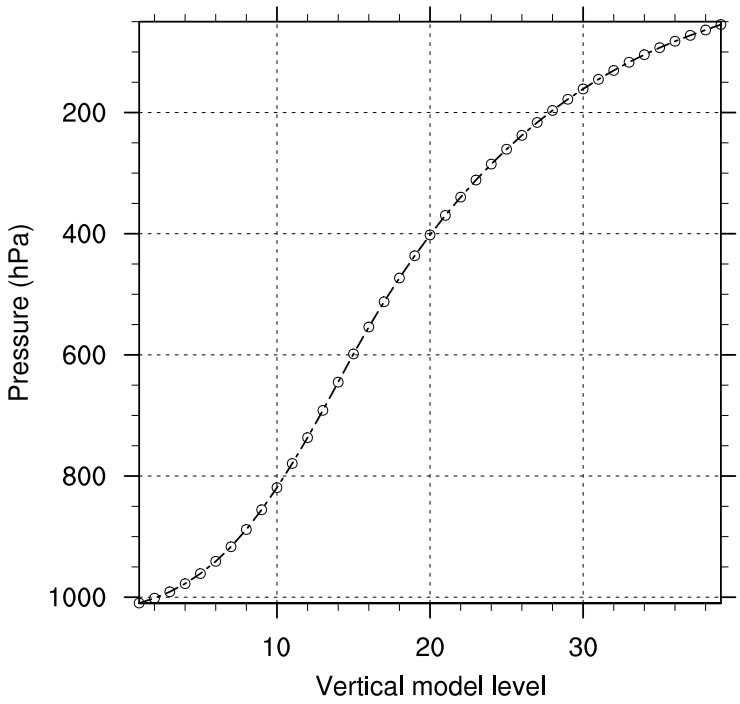


Figure 3. Plot of Pressure (hPa) against vertical model levels (WRF, Res. 15 km).

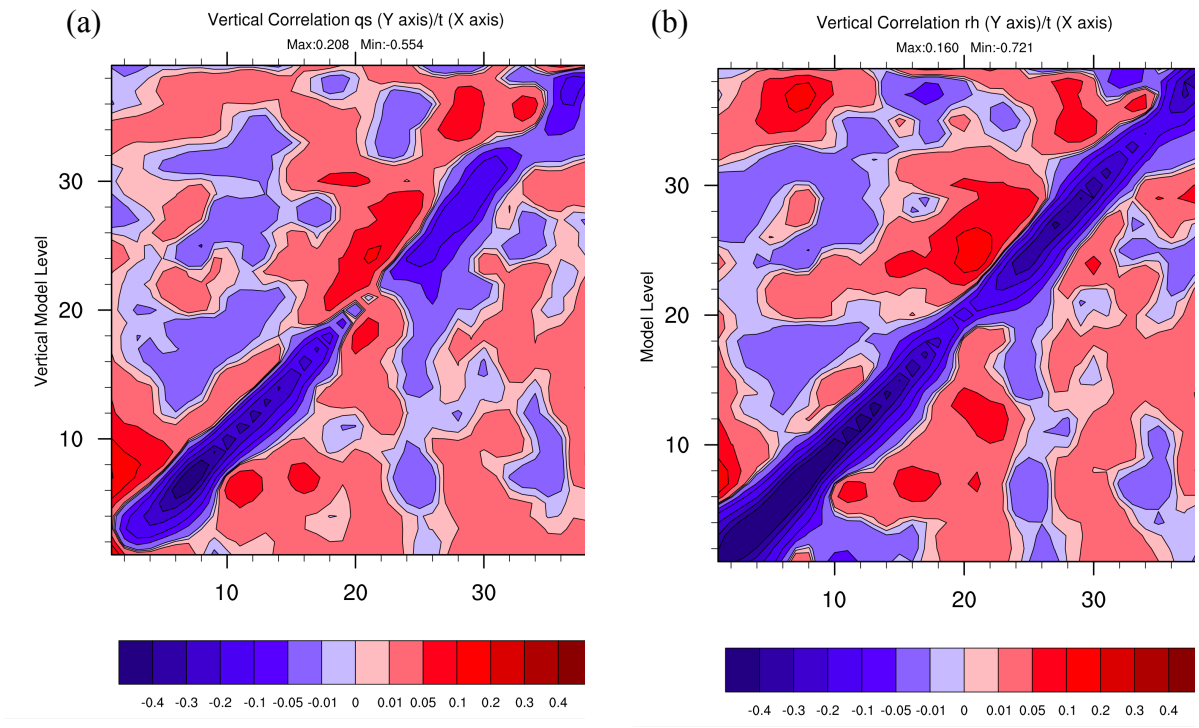


Figure 4. **(a)** Vertical cross-correlation between temperature (t) and specific humidity (qs), **(b)** vertical cross-correlation between temperature (t) and relative humidity (rh); (WRF, Res. 15 km, D-ensemble).

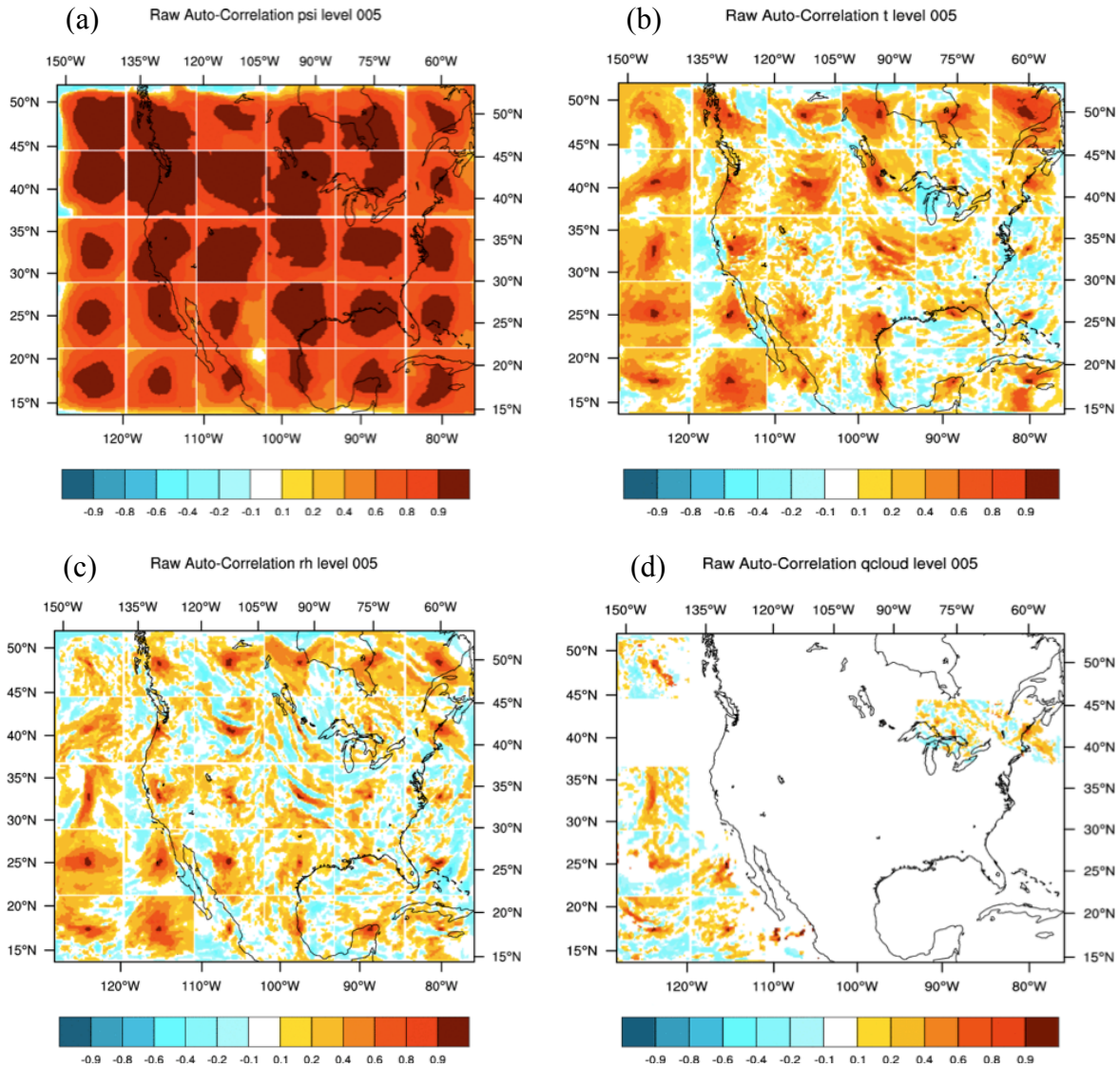


Figure 5. Horizontal autocorrelation performed at the center of each square grid over vertical model level 5, around 950 hPa, for the control variables **(a)** stream function (psi), **(b)** temperature (t), **(c)** relative humidity (rh), and **(d)** cloud mixing ratio (q_{cloud}). Larger correlations are observed for stream function compared to temperature and relative humidity. Cloud mixing ratio has the smallest correlation due to sparce location of hydrometeors (WRF, Res. 15 km, D-ensemble).

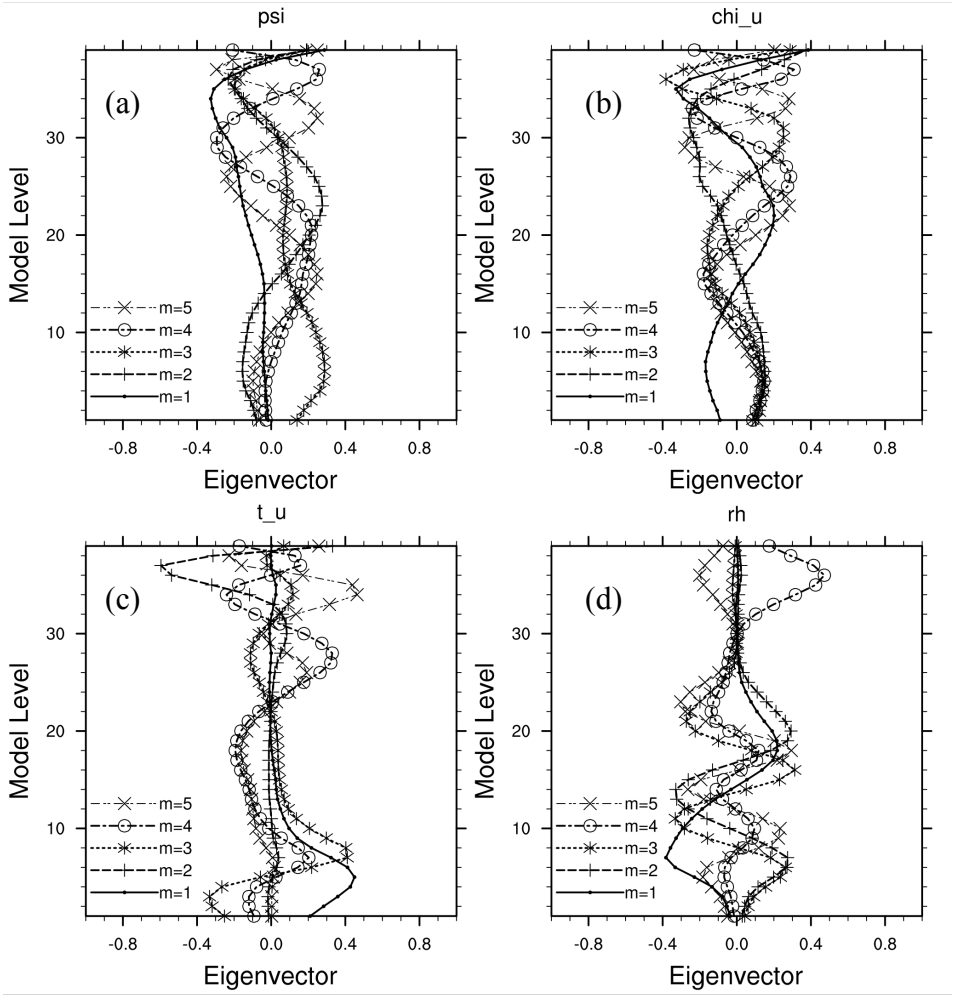


Figure 6. Representation of the first five eigenvectors resulting from the EOF decomposition of the vertical autocovariance matrix, eigenvectors of **(a)** psi, **(b)** chi_u, **(c)** t_u, and **(d)** rh. The eigenvectors are parameters that define the vertical transform ($\mathbf{U}_v$); (WRF, Res. 15 km, D-ensemble, EOFs).

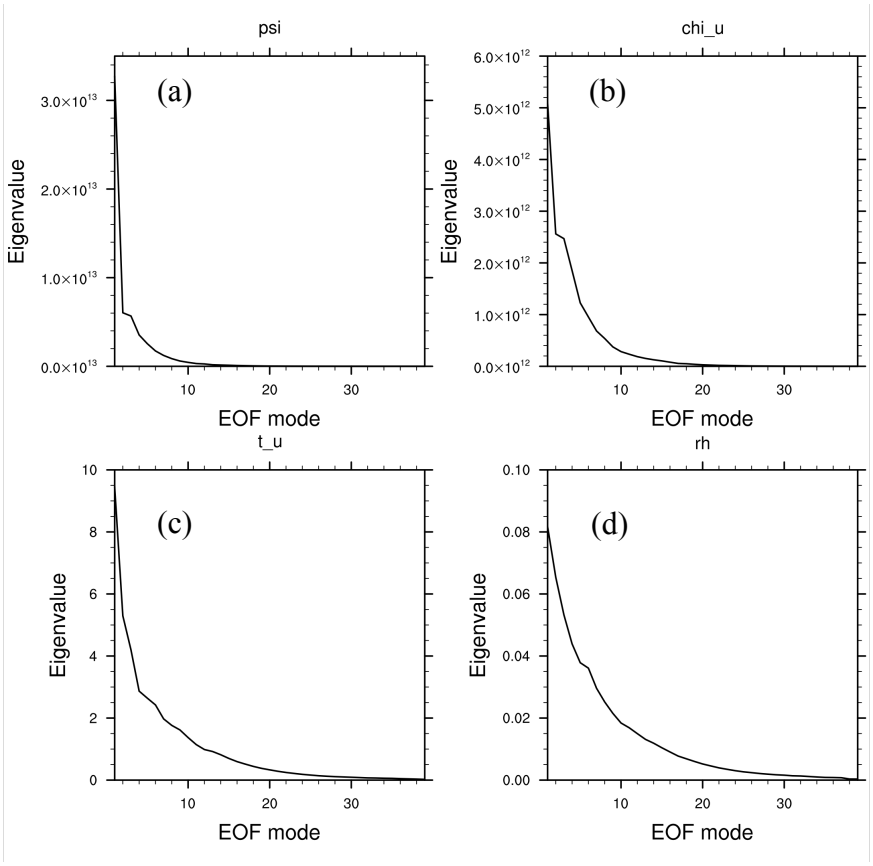


Figure 7. Eigenvalues computed by EOF mode for **(a)** psi, **(b)** chi_u, **(c)** t_u and **(d)** rh. They represent the variance of the control variables (WRF, Res. 15 km, D-ensemble, EOFs).

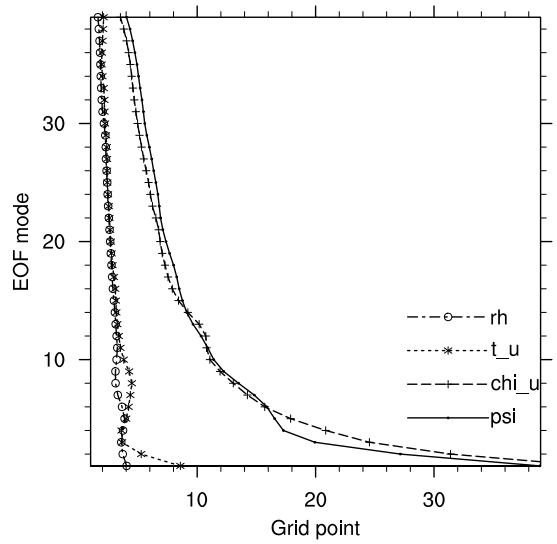


Figure 8. Length scales defined in grid point through EOF mode for CV5. The analysis control variables representating the dynamical variables, psi and chi_u, have longer length scales than t_u, and rh (WRF, Res. 15 km, D-ensemble, EOFs).

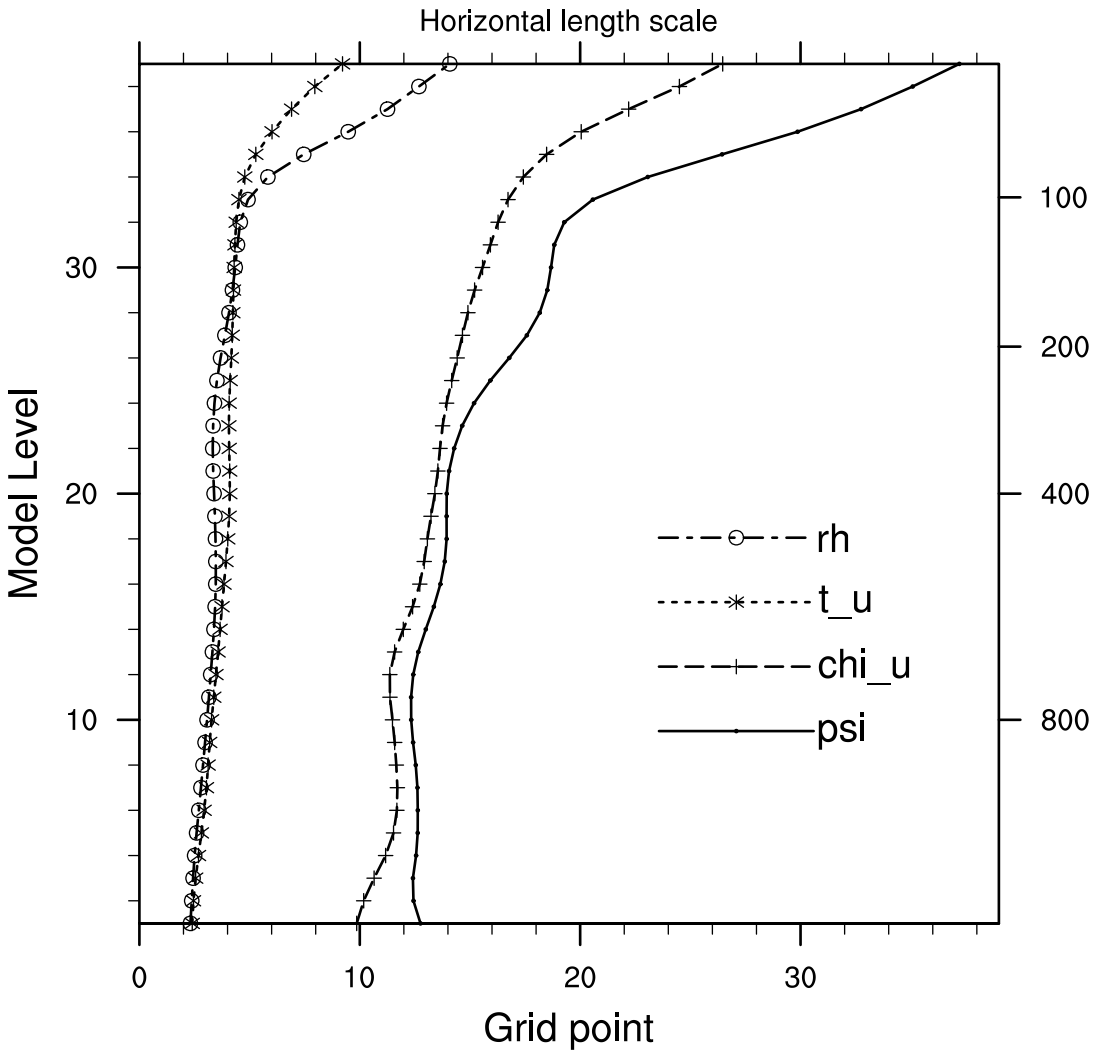


Figure 9. Horizontal length scales for CV5. t_u and rh, which have more local structures, are modeled by shorter length scales (WRF, Res. 15 km, D-ensemble, RFs).

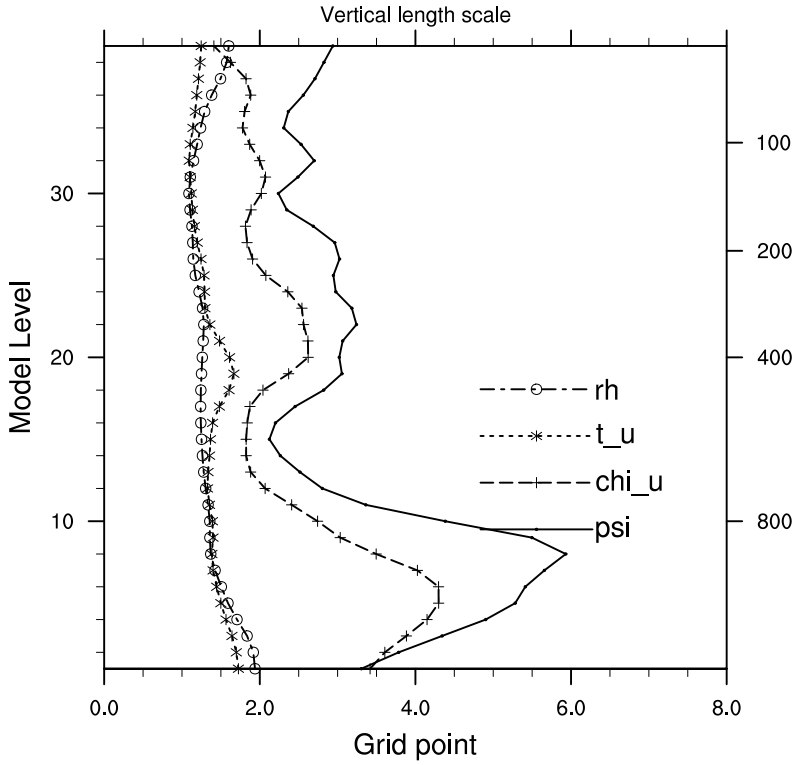


Figure 10. Vertical length scale for CV5 (WRF, Res. 15 km, D-ensemble, RFs).

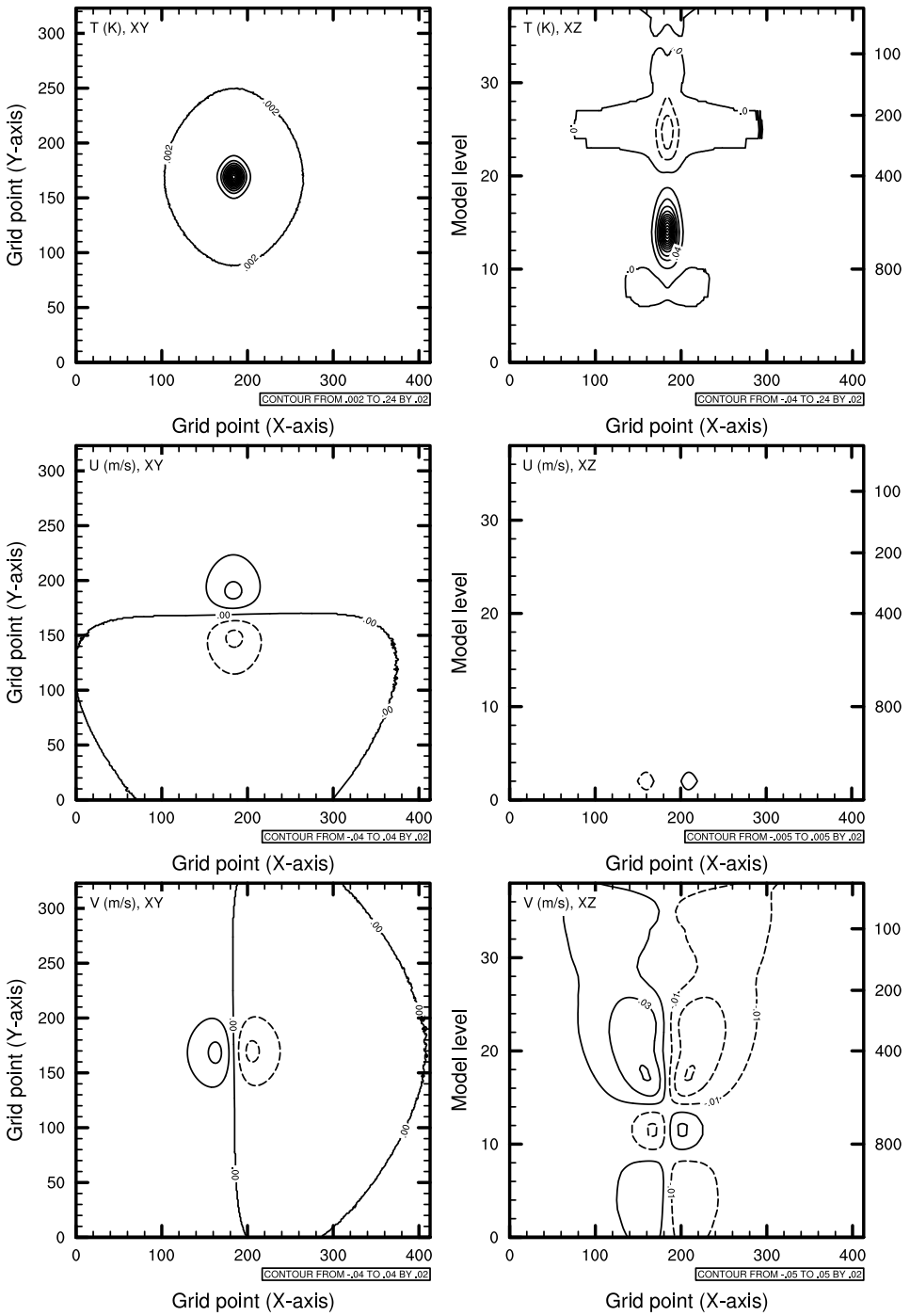


Figure 11. Pseudo observation test of temperature (innovation of +1 K) from the WRFDA application. The three plots on the left panel show, from top to bottom, horizontal cross-section *(XY)* of t (K), U and V wind component (m s^{-1}) respectively. Then, the right panel shows the corresponding cross-section (XZ) of the former variables ($\mathbf{B}_{eof}$: WRF Res. 15 km, D-ensemble, EOFs).

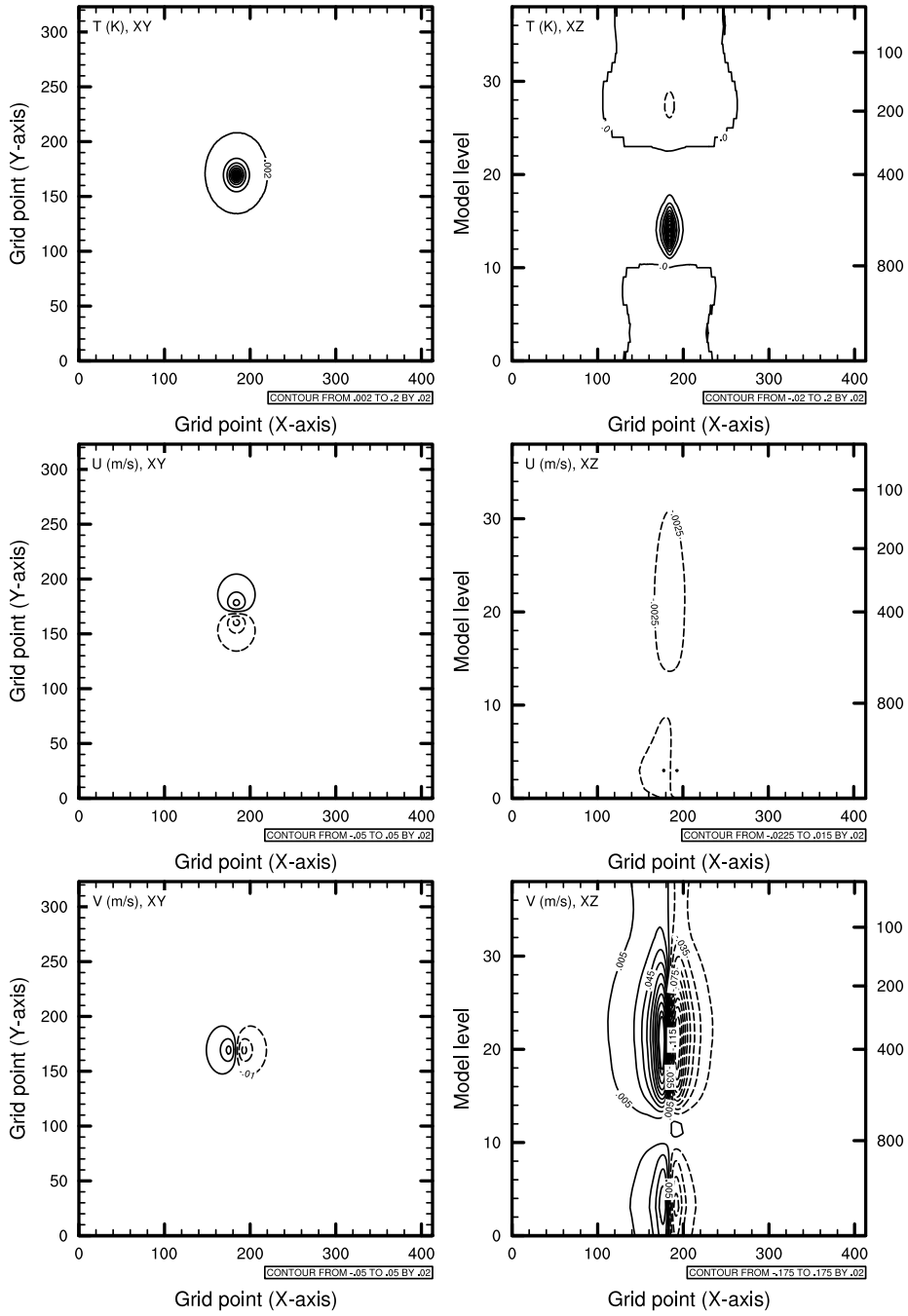


Figure 12. Pseudo observation test of temperature (innovation of +1 K) from the GSI application. The three plots on the left panel show, from top to bottom, horizontal cross-section *(XY)* of t (K), U and V wind component (ms^{-1}) respectively. Then, the right panel shows the corresponding cross-section (XZ) of the former variables. ($\mathbf{B}_{rcf}$: WRF Res. 15 km, D-ensemble, RFs).

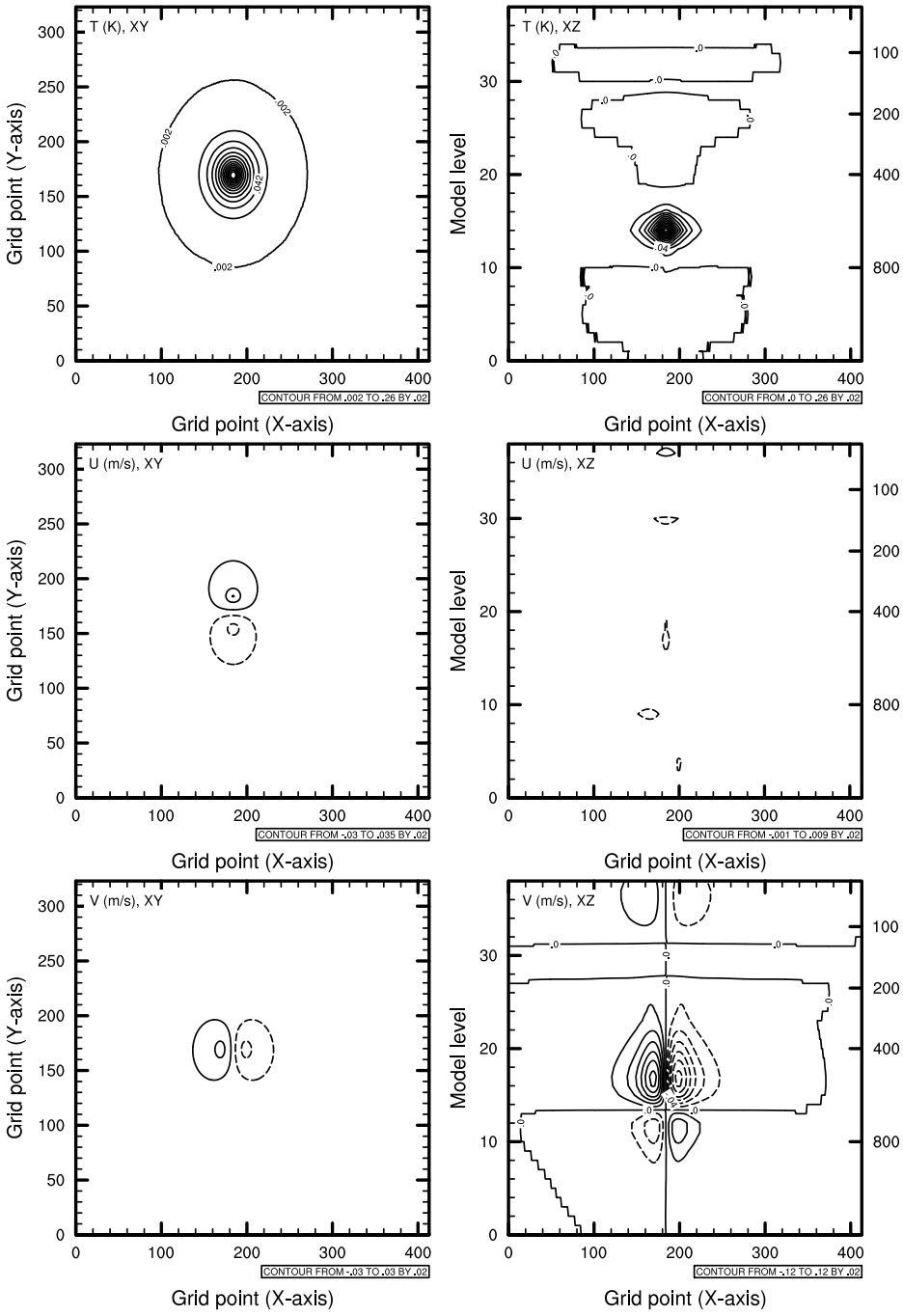


Figure 13. Pseudo observation test of temperature (innovation of +1 K) from the GSI application. The three plots on the left panel show, from top to bottom, horizontal cross-section *(XY)* of t (K), U and V wind component (ms^{-1}) respectively. Then, the right panel shows the corresponding cross-section (XZ) of the former variables. ($\mathbf{B}_{nam}$: WRF-NMM Res. 12 km, NMC, RFs).

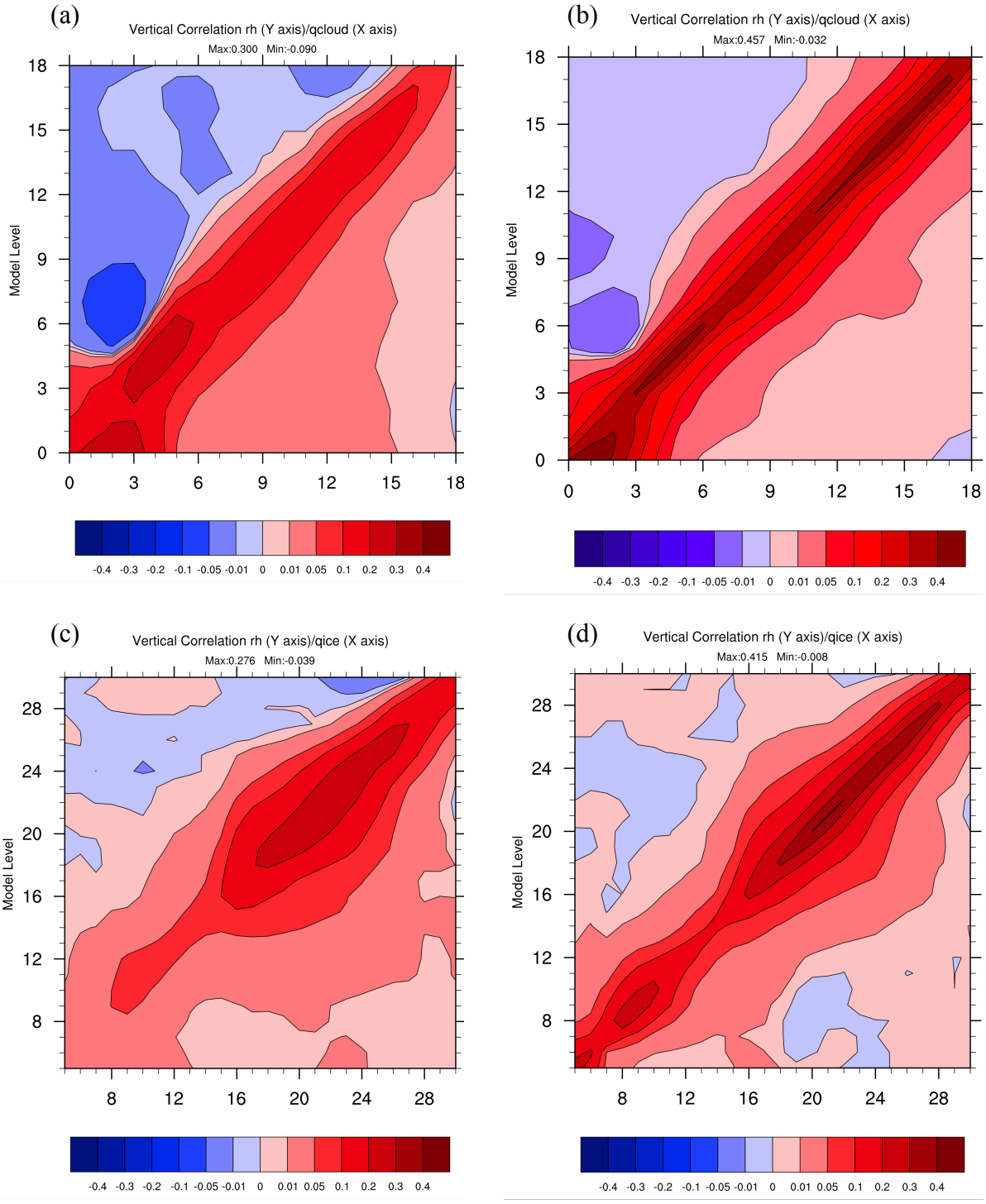
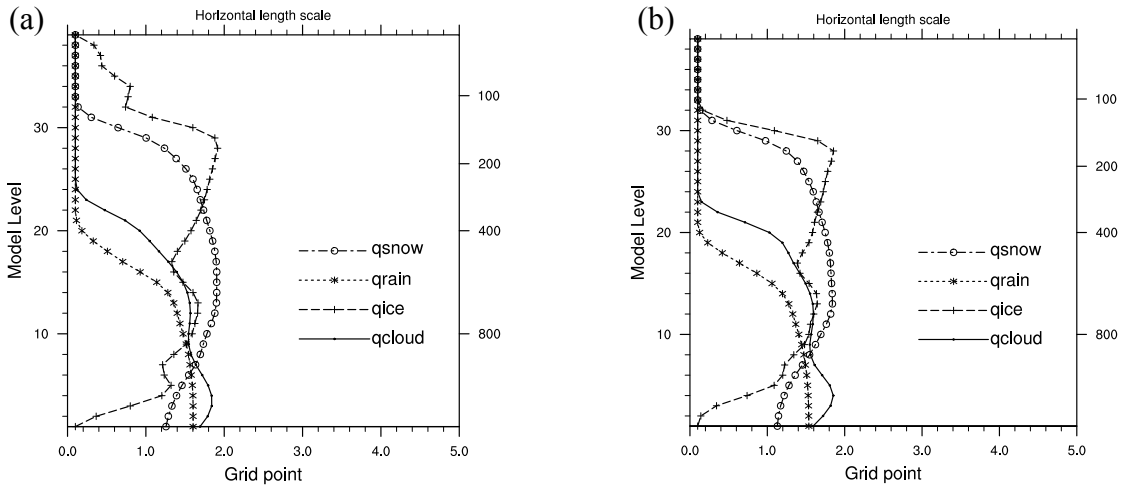


Figure 14. **(a)** Raw vertical cross-correlations between cloud mixing ratio (q_{cloud}) and relative humidity (rh), **(b)** filtered vertical cross-correlations between q_{cloud} and rh, **(c)** raw vertical cross-correlations between ice mixing ratio (q_{ice}) and rh, **(d)** filtered vertical cross-correlations between q_{ice} and rh. Taking into account the perturbations coming from the transition of a

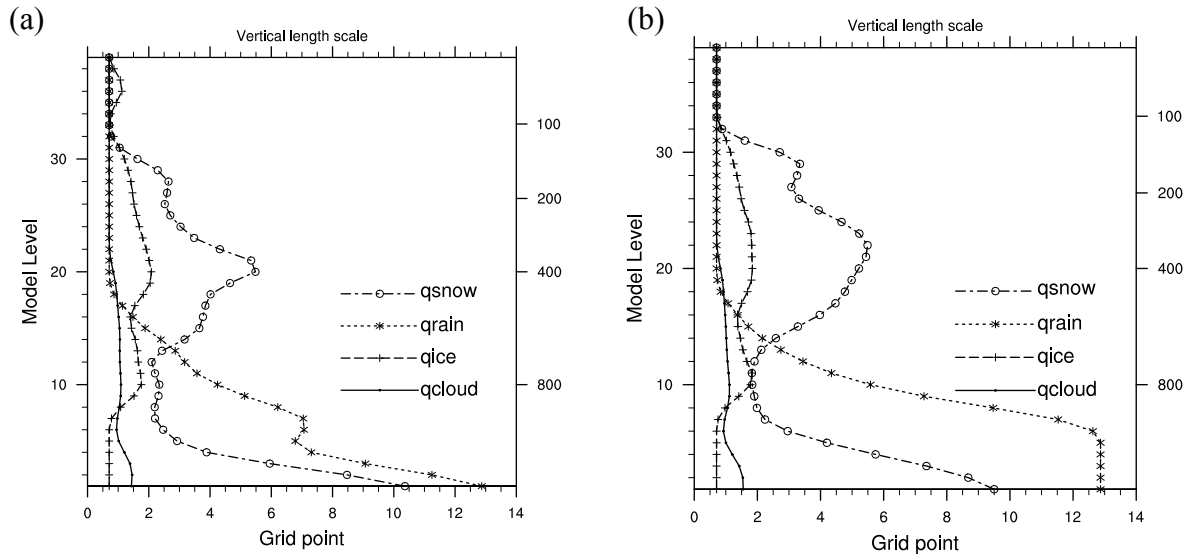
cloudy to a non-cloudy area only when reaching the threshold mixing ratio of 10^{-6} kg kg^{-1}, intensifies the vertical correlation (WRF, Res. 15 km, D-ensemble).

1



2 Figure 15. "Horizontal length scale for the hydrometeors using **(a)** 50 members and **(b)** using
3 5 members. The plots show similar characteristics regardless to the ensemble members (WRF,
4 Res. 15 km, D-ensemble).

5



6 Figure 16. Vertical length scale for the hydrometeors using (a) 50 members and (b) using 5
7 members. The plots show similar characteristics regardless to the ensemble members (WRF,
8 Res. 15 km, D-ensemble).

9

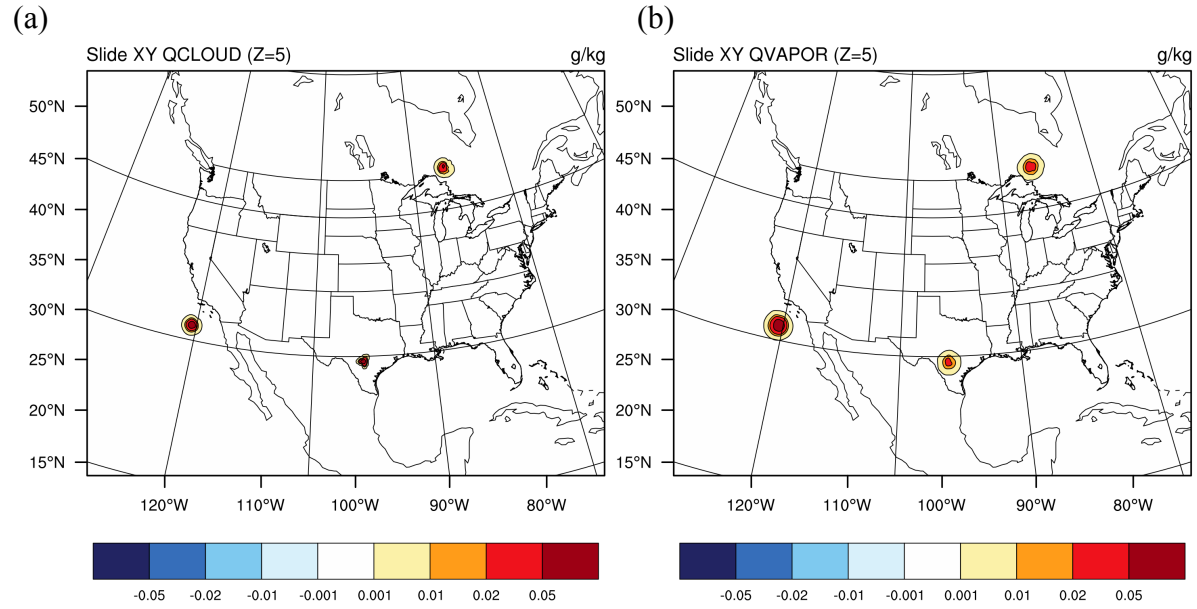


Figure 17. **(a)** Horizontal slide (vertical model level 5) of a pseudo observation test of cloud water condensate (Innovation and observation error of 0.1 g kg^{-1}) in a multivariate approach using the 3-D variance, **(b)** as a consequence there is a positive increment on qvapor (WRF, Res. 15 km, D-ensemble, RFs).

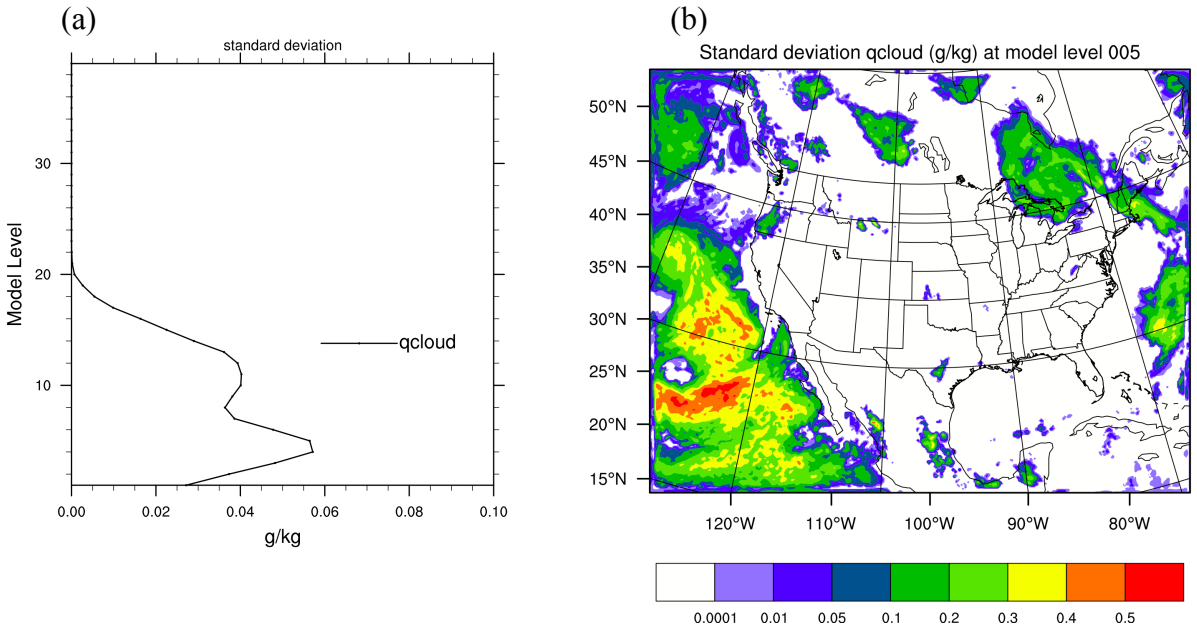


Figure 18. **(a)** Profile of standard deviation of liquid water condensate mixing ratio (q_{cloud} in g kg^{-1}) averaged along the vertical and **(b)** horizontal cross-section of standard deviation of

q_{cloud} at the vertical model level 5 (950 hPa). Both plots indicate the presence of low maritime clouds noted by high standard deviation (WRF, Res. 15 km, D-ensemble).

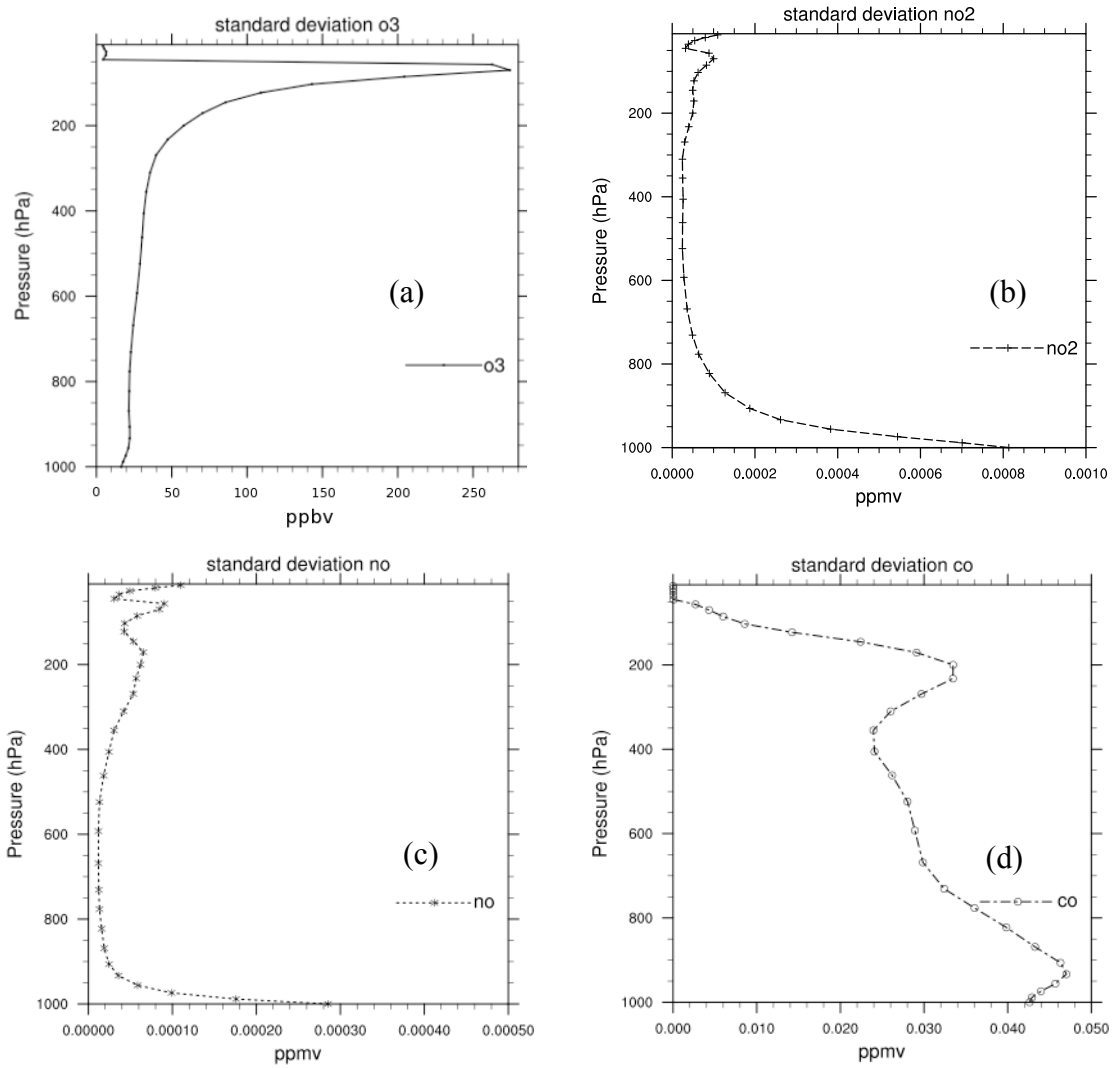


Figure 19. Vertical standard deviation in ppmv of (a) O_3, (b) NO_2, (c) NO, and (d) CO (WRF-CHEM, Res. 36 km, D-ensemble).

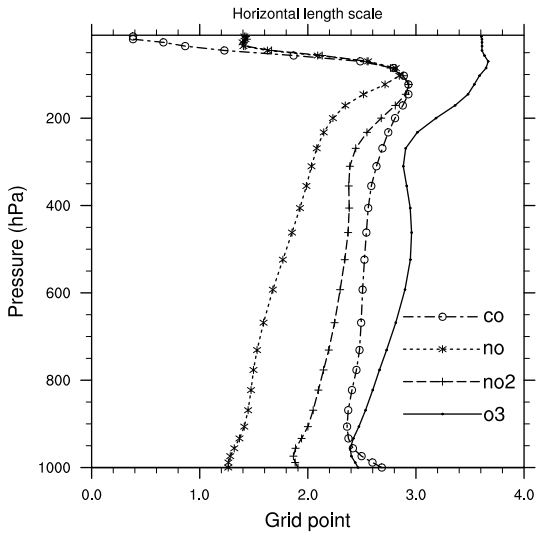


Figure 20. Horizontal length scale of O_3, NO_2, NO, and CO (WRF-CHEM, Res. 36 km, D-ensemble).

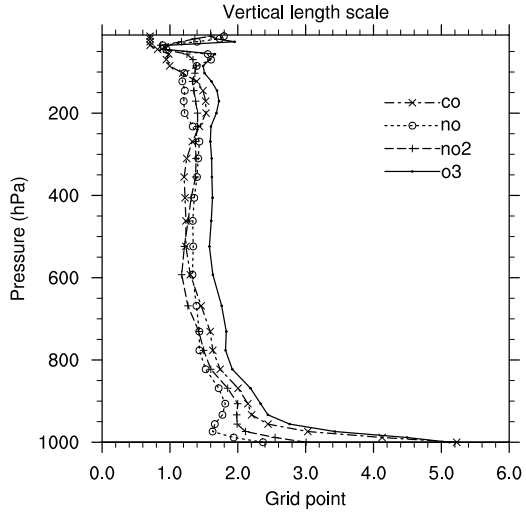


Figure 21. Vertical length scale of O_3, NO_2, NO, and CO (WRF-CHEM, Res. 36 km, D-ensemble).