A Practical Method to Estimate Information Content in the Context of 4D-Var Data Assimilation. II: Application to Global Ozone Assimilation

K. Singh[1], M. Jardak[1], A. Sandu[1], K. Bowman[2], and M. Lee[2]
[1]Computational Science Laboratory
Department of Computer Science,
Virginia Polytechnic Institute and State University
2201 Knowledgeworks II, 2202 Kraft Drive, Blacksburg, VA 24060, USA
[2]Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, CA 91109, USA

October 18, 2011

Abstract

Data assimilation obtains improved estimates of the state of a physical system by combining imperfect model results with sparse and noisy observations of reality. Not all observations used in data assimilation are equally valuable. The ability to characterize the usefulness of different data points is important for analyzing the effectiveness of the assimilation system, for data pruning, and for the design of future sensor systems.

In the companion paper [Sandu et al.(2011)] we derived an ensemble-based computational procedure to estimate the information content of various observations in the context of 4D-Var. Here we apply this methodology to quantify two information metrics (the signal and degrees of freedom for signal) for satellite observations used in a global chemical data assimilation problem with the GEOS-Chem chemical transport model. The assimilation of a subset of data points characterized by the highest information content, gives analyses that are comparable in quality with the one obtained using the entire data set.

Contents

1 Introduction

The information content of observations in the context of data assimilation is defined by their contribution to decreasing the uncertainty in the state estimate [Fisher(1922)]. This work employs several of the information theoretic metrics to quantify the observation impact on improving state estimates: the trace of the Fisher information matrix, the Shannon information, and the degrees of freedom for signal, which measure of the decrease in error variance, and the signal information. which measures the effects of data assimilation in terms of adjusting the mean.

In the companion paper [Sandu et al.(2011)] we have shown that the posterior statistics of the variational cost function and its gradient can be used to quantify the information content of observations in the context of four dimensional variational (4D-Var) data assimilation. An efficient computational approach was developed to estimate the information metrics using ensemble averages. Here we discuss how averages with respect to posterior probability density can be calculated effectively. One approach is based on weighted background ensemble averages, while other approaches construct samples drawn approximately from the analysis probability density.

While the information theoretic approach discussed in this work is general, our main application of interest is chemical data assimilation [Carmichael et al.(2008)] involving gas phase [Daescu et al.(2000), Carmichael et al.(2003), Constantinescu et al.(2007d), Liao et al.(2006)] and particulate phase [Sandu et al.(2005), Hakami et al.(2005), Henze et al.(2004)] atmospheric tracers. Examples of large scale applications are discussed in [Chai et al.(2006), Chai et al.(2007)]. Ensemble Kalman filters are an alternative based on estimation theory, and have been used in chemical data assimilation [Constantinescu et al.(2007a, Constantinescu et al.(20 Constantinescu et al.(2007c, Sandu et al.(2005)]. We consider the problem of global ozone estimation using the GEOS-Chem model, and assimilate satellite data from the tropospheric emission spectrometer. The information theoretic techniques are applied to this problem. The assimilation of a subset of data points characterized by the highest information content, gives analyses that are comparable in quality with the one obtained using the entire data set.

The paper is organized as follows. Section 2 reviews several metrics for information content and the computationally feasible estimation techniques developed in the companion paper [Sandu et al.(2011)]. All estimates require the ability to compute expected values with respect to the analysis probability distribution; obtaining such expected values is discussed in Section 3. Section 4 presents in detail the results of applying the proposed techniques to the data assimilation of global ozone. Section 5 summarizes the findings of this work and points to future research directions.

2 Information Metrics and Their Estimation

The 4D-Var analysis $\mathbf{x}_0^{\mathrm{A}}$ is the initial condition which minimizes the cost function

$$\mathcal{J}\left(\mathbf{x}_0\right) = \underbrace{\frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{B}}\right)^T \mathbf{B}_0^{-1} \left(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{B}}\right)}_{\mathcal{J}^{\mathrm{B}}(\mathbf{x}_0)} + \underbrace{\frac{1}{2}\sum_{i=1}^{N}\left(\mathcal{H}(\mathbf{x}_i) - \mathbf{y}_i\right)^T \mathbf{R}_i^{-1} \left(\mathcal{H}(\mathbf{x}_i) - \mathbf{y}_i\right)}_{\mathcal{J}^{\mathrm{obs}}(\mathbf{x}_0)}, \qquad (1)$$

subject to the model equation constraints. Here $\mathbf{x}_0^{\mathrm{B}} \in \mathbb{R}^n$ is the background value of the initial state, $\mathbf{B}_0$ is the covariance of the initial background errors $\mathbf{B}_0$, $\mathbf{y}_i \in \mathbb{R}^m$ are the observations at time t_i, $i = 1, \cdots, N$, and $\mathbf{R}_i$ are the corresponding observation error covariances.

The *information content* of the observations $\mathbf{y}$ quantifies the decrease in uncertainty from before data assimilation ($\mathcal{P}^{\mathrm{B}}$) to after data assimilation ($\mathcal{P}^{\mathrm{A}}$). The information content depends not only on the data ($\mathbf{y}_i$), but also on the data accuracy ($\mathbf{R}_i^{-1}$), on all other observations $\mathbf{y}_j, j \neq i$, on the background uncertainty ($\mathbf{B}_0^{-1}$), and on the model dynamics $\mathcal{M}$.

We are interested to rigorously quantify the information content of observations in 4D-Var. In the companion paper In the companion paper [Sandu et al.(2011)] we have discussed the following information theoretic metrics.

We seek to derive a computationally-easy way to estimate the information content of various observations in the context of 4D-Var. The proposed approach is based on an approximate sampling from the posterior error distribution in 4D-Var. We assume that we have the ability to compute expected values with respect to the posterior density $\mathbb{E}^{\mathrm{A}}\left[f\left(\mathbf{x}_0\right)\right]$.

2.1 Fisher information matrix

The Fisher information matrix (FIM) [Fisher(1922)] associated with the background probability density function $\mathcal{P}^{\mathrm{B}}(\mathbf{x})$ is

$$\begin{aligned} \mathcal{F}\left(\mathcal{P}^{\mathrm{B}}\right) &= \int_{\mathbb{R}^n} \left[\nabla_{x_0}\mathcal{J}^{\mathrm{B}}(\mathbf{x}_0)\right]\left[\nabla_{x_0}\mathcal{J}^{\mathrm{B}}(\mathbf{x}_0)\right]^T \mathcal{P}^{\mathrm{B}}(\mathbf{x}_0)\, d\mathbf{x}_0 = \mathbf{B}_0^{-1} \qquad (2) \\ \operatorname{trace}\mathcal{F}\left(\mathcal{P}^{\mathrm{B}}\right) &= \mathbb{E}^{\mathrm{B}}\left[\left\|\nabla_{x_0}\mathcal{J}^{\mathrm{B}}(\mathbf{x}_0)\right\|^2\right] \end{aligned}$$

where the last equality in the first relation holds when the background errors are normally distributed, with covariance $\mathbf{B}_0$). Similarly, the FIM associated with the analysis

probability density $\mathcal{P}^{\mathrm{B}}(\mathbf{x})$ is

$$\begin{aligned} \mathcal{F}\left(\mathcal{P}^{\mathrm{A}}\right) &= \int_{\mathbb{R}^n} \left[\nabla_{x_0}\mathcal{J}(\mathbf{x}_0)\right]\left[\nabla_{x_0}\mathcal{J}(\mathbf{x}_0)\right]^T \mathcal{P}^{\mathrm{A}}(\mathbf{x}_0)\, d\mathbf{x}_0 = \mathbf{A}_0^{-1} \\ \operatorname{trace}\mathcal{F}\left(\mathcal{P}^{\mathrm{A}}\right) &= \mathbb{E}^{\mathrm{A}}\left[\left\|\nabla_{x_0}\mathcal{J}(\mathbf{x}_0)\right\|^2\right] \end{aligned} \tag{3}$$

where the last equality in the first relation holds when the analysis errors are normally distributed, with covariance $\mathbf{A}_0$).

The information content of the observations used in data assimilation can be measured as the trace of the background FIM (total uncertainty in the background) minus the trace of the analysis FIM (total uncertainty in the analysis) [Rodgers(2000), Rodgers(1998)]. We obtain the following estimate for the FIM information content of all observations:

$$\mathcal{I}^{\mathrm{FIM}} = \operatorname{trace}\mathcal{F}\left(\mathcal{P}^{\mathrm{A}}\right) - \operatorname{trace}\mathcal{F}\left(\mathcal{P}^{\mathrm{B}}\right) = \mathbb{E}^{\mathrm{A}}\left[\left\|\nabla_{x_0}\mathcal{J}(\mathbf{x}_0)\right\|^2\right] - \operatorname{trace}\left(\mathbf{B}_0^{-1}\right) . \tag{4}$$

After the data assimilation has been performed, one runs the forward *and the adjoint* models $\mathrm{N_{ens}}$ times. The ensemble average of the norm of the adjoint gradients estimates the trace of the analysis FIM.

2.2 Degrees of freedom for signal

The Degrees of freedom for signal (DFS) metric for the information content has been previously employed in meteorological data assimilation [Rodgers(1996), Fisher(2003), Cardinali et al.(2004), Stewart et al.(2008), Zupanski(2009)].

The degrees of freedom for signal measures the total reduction in variance after assimilation. For Gaussian background and analysis distributions

$$\mathcal{I}^{\mathrm{DFS}} = n - \operatorname{trace}\left(\mathbf{B}_0^{-1}\,\mathbf{A}_0\right) . \tag{5}$$

The contribution of each observation $\mathbf{y}_\ell$ to the DFS information metric is estimated via [Sandu et al.(2011)]

$$\mathcal{I}^{\mathrm{DFS}}_{\mathbf{y}_\ell} = 2\,\mathbb{E}^{\mathrm{A}}\left[\mathcal{J}_\ell^{\mathrm{obs}}\left(\mathbf{x}_0\right)\right] - 2\,\mathcal{J}_\ell^{\mathrm{obs}}\left(\mathbf{x}_0^{\mathrm{A}}\right) . \tag{6}$$

After the data assimilation has been performed, one runs the forward model $\mathrm{N_{ens}}$ times. The ensemble average of the cost function, minus the cost function at the analysis, estimates the DFS information.

2.3 Signal information

The *signal part* of the relative entropy [Xu(2006)]

$$\mathcal{I}^{\text{Signal}} = \frac{1}{2}\left(\mathbf{x}_0^{\text{A}} - \mathbf{x}_0^{\text{B}}\right)^T \mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{A}} - \mathbf{x}_0^{\text{B}}\right) \tag{7}$$

measures the reduction of uncertainty due to the change in the best estimate from the background state to the analysis state. The contribution of the data point $\mathbf{y}_\ell$ to the signal information can be (coarsely) approximated as:

$$\mathcal{I}_{\mathbf{y}_\ell}^{\text{Signal}} \approx \left(\mathbf{y}_\ell - \mathcal{H}_\ell\left(\mathbf{x}_\ell^B\right)\right)^T \mathbf{R}_\ell^{-1}\left(\mathcal{H}_\ell\left(\mathbf{x}_\ell^A\right) - \mathcal{H}_\ell\left(\mathbf{x}_\ell^B\right)\right) . \tag{8}$$

The model is run from the analysis and the "synthetic observations" $\mathcal{H}_\ell\left(\mathbf{x}_\ell^A\right)$ are recorded. The model is run again starting from the background state, and (8) is evaluated for each data point $\mathbf{y}_\ell$.

3 Expected Values With Respect to the Analysis Probability Density

The information metric estimates discussed here require expected values with respect to the analysis probability distribution. Since 4D-Var does not provide immediately and approximation of the posterior density, a discussion of how to obtain these expected values is necessary.

The first approach, discussed in Section 3.1, is based on an ensembles drawn from background distribution, and expected values calculated as weighted sums. The second approach is to approximately sample initial conditions from the posterior distribution:

$$\mathbf{x}_0^r \in \mathcal{P}^{\text{A}}(\mathbf{x}_0) , \quad r = 1, \cdots, \text{N}_{\text{ens}} . \tag{9}$$

Based on it we can approximate expected values with respect to the posterior density by posterior ensemble averages as follows:

$$\mathbb{E}^{\text{A}}\left[f\left(\mathbf{x}_0\right)\right] \approx \left\langle f\left(\mathbf{x}_0\right)\right\rangle^{\text{A}} = \frac{1}{\text{N}_{\text{ens}}}\sum_{r=1}^{\text{N}_{\text{ens}}} f\left(\mathbf{x}_0^r\right) . \tag{10}$$

Various strategies to approximately sample the posterior distribution are discussed in Sections 3.2, 3.3, and 3.4

3.1 Expected values as weighted background ensemble averages

Consider the following sample from the background distribution:

$$\mathbf{x}_0^q \in \mathcal{P}^{\mathrm{A}}(\mathbf{x}_0) \ , \quad q = 1, \cdots, \mathrm{N_{ens}} \ . \tag{11}$$

The drawing is such that each sample has an equal weight $1/\mathrm{N_{ens}}$. Based on it we can approximate expected values with respect to the posterior density by weighted ensemble averages as follows [Wikle and Berliner(2007)]:

$$\begin{aligned} \mathbb{E}^{\mathrm{A}}\left[f\left(\mathbf{x}_0\right)\right] &\approx \sum_{q=1}^{\mathrm{N_{ens}}} \mathcal{P}^{\mathrm{A}}(\mathbf{x}_0^q)\, f\left(\mathbf{x}_0^q\right) = \sum_{q=1}^{\mathrm{N_{ens}}} \frac{\mathcal{P}^{\mathrm{A}}(\mathbf{x}_0^q)}{\mathcal{P}^{\mathrm{B}}(\mathbf{x}_0^q)}\, \mathcal{P}^{\mathrm{B}}(\mathbf{x}_0^q)\, f\left(\mathbf{x}_0^q\right) \\ &= \sum_{q=1}^{\mathrm{N_{ens}}} w^q\, f\left(\mathbf{x}_0^q\right) \ . \end{aligned} \tag{12}$$

The posterior average can be calculated as a weighted average of samples taken from the background distribution. Using the Bayes theorem, the new weights are:

$$w^q = \frac{\mathcal{P}^{\mathrm{A}}(\mathbf{x}_0^q)}{\mathcal{P}^{\mathrm{B}}(\mathbf{x}_0^q)} \frac{1}{\mathrm{N_{ens}}} = \frac{\mathcal{P}(\mathbf{y}|\mathbf{x}_0^q)}{\mathcal{P}(\mathbf{y})} \frac{1}{\mathrm{N_{ens}}} \, .$$

With the relationship that the observation part of the cost function is the logarithm of the observation likelihood, we can compute the weights as:

$$v^i = \exp\left(\mathcal{J}^{\mathrm{obs}}(\mathbf{x}_0^i)\right) \ , \quad w^q = \frac{v^q}{\sum_{i=1}^{\mathrm{N_{ens}}} v^i} \, .$$

The computational procedure is as follows. Start with an equally weighted sample $\mathbf{x}_0^q$ of the background probability density. For each sample run the model, and compute the observations part of the 4D-Var cost function $\mathcal{J}(x_0^q)$, as well as the metric of interest $\mathcal{J}(x_0^q)$. The analysis mean is a weighted average of the obtained values:

$$\mathbb{E}^{\mathrm{A}}\left[f\left(\mathbf{x}_0\right)\right] = \sum_{q=1}^{\mathrm{N_{ens}}} \left(\frac{\exp(\mathcal{J}^{\mathrm{obs}}(\mathbf{x}_0^q))}{\sum_{i=1}^{\mathrm{N_{ens}}} \exp(\mathcal{J}^{\mathrm{obs}}(\mathbf{x}_0^i))} \right) f\left(\mathbf{x}_0^q\right) \ .$$

3.2 Approximate posterior sampling by estimated posterior covariance

In this simple approach one assumes that the correlation structures of $\mathbf{B}_0$ and $\mathbf{A}_0$ are similar, and that the difference comes from changes in variances. The background and analysis variances can be estimated roughly by comparing the two solutions against data, and by measuring the model-data discrepancies. It is then assumed that the decrease in

the model-data discrepancy for each variable, vertical level, area, etc. is representative of the corresponding decrease in variance. The analysis variances are estimated by rescaling the background variances (for each variable, vertical level, area, etc.) The new variances, together with the specified correlation structure of $\mathbf{B}_0$, define the analysis covariance $\mathbf{A}_0$. Random draws are taken from the normal distribution with mean equal to the analysis, $\mathcal{N}(\mathbf{x}_0^{\mathrm{A}}, \mathbf{A}_0)$.

3.3 Approximate posterior sampling by estimated inverse Hessians

This approach uses the fact that the analysis covariance matrix is approximated by the inverse Hessian of the cost function, evaluated at the optimum [Thacker(1989), Gejadze et al.(2008)]

$$\mathbf{A}_0 \approx \left(\nabla^2_{\mathbf{x}0,\mathbf{x}0}\mathcal{J}\right)^{-1}.$$

Several eigenvectors corresponding to the smallest eigenvalues of the Hessian are computed. The inverses of these eigenvalues, together with their eigenvectors, approximate the principal components of the posterior error and can be used for approximate sampling from the posterior distribution. The computation of the smallest eigenpairs of the Hessian can be done using only Hessian vector products, for example obtained via a second order adjoint. Alternatively, if a quasi-Newton method is used in optimization (e.g., L-BFGS) the low rank quasi-Newton approximation of the inverse Hessian is constructed by the method and available for use in approximate sampling.

3.4 Approximate posterior sampling by subspace analysis

We use the hybrid of 4D-Var and ensemble approach discussed in [Cheng et al.(2010)] to generate our posteriori distribution. Suppose we are given the background state $\mathbf{x}_0^{\mathrm{B}} \in \mathbb{R}^n$ and the background error covariance matrix $\mathbf{B}_0 \in \mathbb{R}^{nxn}$, the $\mathrm{N_{ens}}$ normally distributed perturbation vectors with zero mean and covariance $\mathbf{B}$ can be generated as:

$$\Delta\mathbf{x}_0^{\mathrm{B}}(r) \in \mathcal{N}(\mathbf{0}, \mathbf{B}_0)\,, \quad r = 1, 2, \ldots, \mathrm{N_{ens}}\,. \tag{13}$$

Starting from $\mathbf{x}_0^{\mathrm{B}}$, we save the first k iterates $\mathbf{x}_0^{(j)}, j = 1, \ldots, k$, generated by the numerical optimization routine used in the 4D-Var assimilation. The value of k is chosen based on the rate of convergence of the optimization routine. Since the reduction in cost function is fastest during the initial iterations [Li et. al(1993), Navon et. al(1992), Sandu and Zhang(2008), Zou et. al(1993)], k is much smaller than the dimension of the state vector. Let $\mathbf{S}$ be the matrix with columns as normalized 4D-Var increments

$$\mathbf{S} = \left\{ \frac{\mathbf{x}_0^{(j)} - \mathbf{x}_0^{(j-1)}}{\|\mathbf{x}_0^{(j)} - \mathbf{x}_0^{(j-1)}\|} \right\}, \quad j = 1, 2, \ldots, k\,,$$

where $\mathbf{x}_0^{(0)} = \mathbf{x}_0^{\mathrm{B}}$. Using the singular value decomposition $\mathbf{S} = \mathbf{U}\,\Sigma\,\mathbf{V}^T$ we derive the orthogonal projector onto the orthogonal complement of range($\mathbf{U}$) as

$$\mathbf{P} = \mathbf{I}_{n\times n} - \mathbf{U}\,\mathbf{U}^T\,, \quad \mathbf{P}\mathbf{x} \perp \mathrm{range}(\mathbf{U})\,, \quad \forall \mathbf{x}\,.$$

Using $\mathbf{P}$, the ensemble perturbations $\Delta\mathbf{x}_0^{\mathrm{B}}$ are projected from forecast space onto the analysis space

$$\Delta\mathbf{x}_0^{\mathrm{A}}(r) = \mathbf{P}\cdot\Delta\mathbf{x}_0^{\mathrm{B}}(r)\,, \quad r = 1, 2, \ldots, \mathrm{N_{ens}}\,. \tag{14}$$

4 Application to Data Assimilation of Global Ozone

We apply the estimation methodology to a 4D-Var data assimilation study with a global chemical transport model. The data assimilation experiment focuses on ozone. Ozone is an important constituent of stratosphere which absorbs the high energy UV-B and UV-C rays, thus preventing the disintegration of DNA molecules and supporting the existence of life. However, ozone present in mid to low troposphere is a pollutant, a powerful oxidizing agent leading to destruction of tissues, damaging fibers and creating breathing problems.

The data are satellite ozone column retrievals. We estimate the information content of satellite observations taken at different times using different information theoretic metrics.

4.1 The model: GEOS-Chem

The model used for the numerical experiments in this paper is GEOS-Chem (`http://acmg.seas.harvard.edu/geos`), a global three-dimensional chemical transport model (CTM) driven by assimilated meteorological observations from Goddard Earth Observing System. A detailed description of the model is presented in [Bey et al.(2001)]. GEOS-Chem accounts in detail for emissions from both natural and anthropogenic sources, for gas phase chemistry, aerosol processes, long range transport of pollutants,troposphere-stratosphere exchanges, etc. GEOS-Chem is being widely used world-wide for global atmospheric chemistry studies.

The GEOS-Chem-Adjoint system (`http://wiki.seas.harvard.edu/geos-chem/index.php/GEOS-Chem_Adjoint`) has been developed through a joint effort of groups at Caltech, University of Colorado, Virginia Tech, Harvard, and Jet Propulsion Laboratory [Henze et al.(2007), Singh(2009a), Singh(2009b), Eller et al.(2009)]. The system can perform adjoint sensitivity analyses and 4D-Var chemical data assimilation. Inverse modelling studies with GEOS-Chem-Adjoint are exemplified in [Henze et al.(2009), Kopacz et al.(2007), Zhang et al.(2009)].

4.2 The data: Tropospheric Emission Spectrometer ozone column retrievals

We assimilate ozone profile retrievals from the Tropospheric Emission Spectrometer (TES), in order to obtain improved estimates of the ozone initial conditions. TES [Beer et al.(2001)], one of four science instruments aboard NASA's Aura satellite, measures the infrared-light energy (radiance) emitted by Earth's surface, and by the chemical tracers in the atmosphere (`http://tes.jpl.nasa.gov`). Vertical profiles of chemical concentrations are retrieved from the radiance measurements using an off-line inversion process.

A-priori information about the vertical concentration profile of the species of interest is needed to solve the retrieval inverse problem (the prior information does not come from the measurement). Let $\mathbf{x}^{\text{prior}}$ be the prior vertical ozone concentration profile (in volume mixing ratio units), and let $\mathbf{z}^{\text{prior}} = \ln \mathbf{x}^{\text{prior}}$. Let $\mathbf{x}^{\text{radiance}}$ be the atmospheric profile as resulting directly from the radiances and $\mathbf{z}^{\text{radiance}} = \ln \mathbf{x}^{\text{radiance}}$.

The vertical ozone profile is retrieved according to the formula [Parrington et al.(2009)]

$$\widehat{\mathbf{z}} = \mathbf{z}^{\text{prior}} + A_{\text{v}} \left(\mathbf{z}^{\text{radiance}} - \mathbf{z}^{\text{prior}} \right) + G\,\eta, \quad \widehat{\mathbf{x}} = \exp\left(\widehat{\mathbf{z}}\right). \tag{15}$$

Here A_{v} is the averaging kernel matrix, G is the gain matrix, and η is the spectral measurement error (assumed to have mean zero and covariance S_η). More details can be found in [Worden et al.(2004), Jones et al.(2003), Bowman et al.(2002)].

The corresponding TES observation operator is linear with respect to the logarithm of the concentrations, but nonlinear with respect to the concentration profile:

$$\mathcal{H}\left(\mathbf{x}\right) = \mathbf{z}^{\text{prior}} + A_{\text{v}} \left(\ln\left(L(\mathbf{x})\right) - \mathbf{z}^{\text{prior}} \right)$$

The ozone column $\mathbf{x}$ represented on the N_{lev} GEOS-Chem grid vertical levels is interpolated by the operator L to an ozone column $L(\mathbf{x})$ represented on the p TES profile retrieval levels.

For this reason several chemical data assimilation studies based on TES retrieved profiles [Jones et al.(2003), Bowman et al.(2006), Parrington et al.(2009)] have opted to perform the suboptimal Kalman filtering step in the logarithm of the concentrations:

$$\ln \mathbf{x}^{\text{A}} = \ln \mathbf{x}^{f} + K \left(\widehat{\mathbf{z}} - \mathcal{H}\left(\mathbf{x}^{f}\right) \right).$$

In case of 4D-Var data assimilation, the forcing calculation is carried out in the model state space. For this reason an adjoint of the observation operator needs to be derived

explicitly to update the gradients as described in equation (**??**).

$$
\begin{aligned}
(\mathcal{H}'(\mathbf{x}))^T \cdot v &= \left(\frac{\partial}{\partial \mathbf{x}} \left(A_{\mathrm{v}} \ln\big(L(\mathbf{x})\big) \right) \right)^T \cdot v \\
&= \left(\frac{\partial L}{\partial \mathbf{x}}\right)^T \cdot \begin{pmatrix} \frac{1}{[L(\mathbf{x})]_0} & 0 & \cdots & 0 \\ 0 & \frac{1}{[L(\mathbf{x})]_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{[L(\mathbf{x})]_p} \end{pmatrix} \cdot A_{\mathrm{v}}^T \cdot v \; .
\end{aligned}
$$

Here $(\mathcal{H}'(\mathbf{x}))^T$ is a matrix and $v = \mathbf{R}^{-1}\,(\mathcal{H}(\mathbf{x}) - \mathbf{y})$. The product $A_{\mathrm{v}} \cdot v$ is scaled by the diagonal matrix with the i-th diagonal entry $1/[L(\mathbf{x})]_i$. The result is fed to $(\partial L/\partial \mathbf{x})^T$, the adjoint of the interpolation operator, which entities from TES profile retrieval domain back to the GEOS-Chem model domain.

4.3 The validation data: INTEX ozonesonde profiles

In order to assess the quality of the data assimilation results, we compare the respective analyses against an independent data set. The independent data are the ozonesonde profiles measured during the INTEX Ozonesonde Network Study 2006 (IONS-6) (`http://croc.gsfc.nasa.gov/intexb/ions06.html`, [Thompson et al. (2007a, 2007b)]) for the month of August. There were 418 ozonesondes launched from 22 stations across North America. A detailed description of the number of ozonesondes launched per station with longitude and latitude information can be found in [Parrington et al.(2008)].

We use ozonesonde parameters such as launch time, longitude, latitude and pressure level to interpolate the concentration fields generated by the model. Differences between the ozone concentrations from ozonesonde observations, model forecasts, and model analyses are averaged individually over longitude, latitude and time to create vertical profiles of model errors. We report the vertical distribution of the mean and the standard deviation of model errors.

4.4 Experimental Setting

The GEOS-Chem simulations are carried out at a resolution of $4^\circ \times 5^\circ$. At this resolution, each latitude-longitude grid box on the ground level covers an area of about 400 Km $\times$ 500 Km. The chemical system accounts for 43 different chemical species. The dimension of the state space in our simulations is $n \approx 8$ million (72 longitude grid points, times 46 latitude grid points, times 55 vertical levels, times 43 chemical tracers).

The control variables are the initial concentrations of ozone throughout the simulation domain. While GEOS-Chem is capable of performing simulations up to 75 Km (55 vertical levels), the model error increases with height and the model bias is non-negligible in the upper troposphere and into the stratosphere. For this reason we perform data assimilation only up to 21 Km (the bottom 23 vertical levels). The dimension of the control vector for data assimilation is $n_c \approx 80,000$ (72 longitude grid points, times 46 latitude grid points, times 23 vertical levels, times 1 chemical tracer – ozone).

The assimilation time window has a length of 5 days, starting on August 1st, 2006 (00 GMT) and ending on August 6th, 2006 (00 GMT). The observation time window is 4 hours, i.e., the observation operator treats all retrievals available in a 4 hour window as a single data point. Specifically, the observation $\mathbf{y}_i$ at time t_i consists of all the data available for the time interval $[\, t_i - 2\,\text{hours}\,, t_i + 2\,\text{hours}\,]$.

We estimate the information content of ozone profile retrievals from TES when used to improve the ozone initial conditions in GEOS-Chem through 4D-Var data assimilation. The main computational costs come from: (1) the 4D-Var run, which requires 11 iterations of the optimization routine, with each iteration performing a forward and adjoint model run; and (2) an ensemble of 20 additional model runs, including adjoints, to gather the data needed for the estimation of different information content metrics. Concentrations and other time dependent variables are checkpointed during the forward runs, and are read during the adjoint runs. The adjoint forcing calculations are performed every observation window (4 hours). The numerical optimization method is the limited memory bound-constrained BFGS method [Zhu et al.(1997)], which has become the "gold standard" in solving large scale 4D-Var chemical data assimilation problems [Sandu et al.(2005)]. The total computational time is 14 minutes and 46 seconds per forward plus adjoint model runs. All the simulations are parallel and use eight cores; they were performed on a Dell Precision T5400 workstation with 2 quadcore Intel(R) Xeon(R) processors with clock speed 2.33GHz and a RAM of 16GB shared between the eight cores.

We consider a diagonal background error covariance matrix ($\mathbf{B}_0$) in all our experiments for simplicity. The setting can be easily extended to use a non-diagonal $\mathbf{B}_0$ that captures spatial error correlations[Singh et al.(2010)]. The initial variances (the diagonal entries of the $\mathbf{B}_0$ matrix) are constructed from the average background concentrations $\mathbf{x}_0^{\mathrm{B}}$ on each of the $\mathrm{N_{lev}}$ vertical layers

$$\mathbf{B}_0 = \begin{bmatrix} \mathbf{B}_0^{(0)} & 0 \dots & 0 \\ 0 & \mathbf{B}_0^{(1)} \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 \dots & \mathbf{B}_0^{(\mathrm{N_{lev}})} \end{bmatrix} \tag{16}$$

where

$$\mathbf{B}_0^{(\ell)} = \begin{bmatrix} \sigma_\ell^2 & 0 \dots & 0 \\ 0 & \sigma_\ell^2 \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 \dots & \sigma_\ell^2 \end{bmatrix}_{dim \times dim} , \quad dim = \mathrm{N_{lon}} \cdot \mathrm{N_{lat}} , \tag{17}$$

with

$$\sigma_\ell = \frac{rel}{dim} \sum_{i=1}^{\mathrm{N_{lon}}} \sum_{j=1}^{\mathrm{N_{lat}}} \mathbf{x}_0^{\mathrm{B}}(i, j, \ell, s_{O3}) , \quad \ell = 1, \cdots, \mathrm{N_{lev}} , \quad s_{O3} = \text{index of ozone} . \tag{18}$$

The relative uncertainty level in the background initial conditions is taken to be 50%, i.e., $rel = 0.5$.

The following simple technique is employed to approximately sample the analysis distribution. We perform data assimilation and compare the background and the analysis fields against the INTEX ozonesonde validation data set. This provides a vertical distribution of mean errors and of their variance. We make the following assumptions: the analysis covariance matrix is diagonal (the correlation length is smaller than one grid size); the relative error reduction realized through data assimilation is similar in all gridpoints at the same vertical level; and the relative error reduction is similar throughout the assimilation window. Under these assumptions the error reduction measured against the INTEX ozonesonde data is representative of the reduction in error at the initial time throughout the entire computational grid. Consequently, the analysis error standard deviation at a given grid point is obtained by scaling the background standard deviation. The scaling factor is the ratio of the standard deviation of the analysis against INTEX data over the standard deviation of the background against INTEX data; the same scaling factor is applied to all grids at the same vertical level. In summary, the analysis mean is provided by the result of the data assimilation. The analysis covariance matrix is diagonal, with the diagonal entries obtained by scaling the corresponding background variances. The scaling factors are obtained by comparing the background and the analysis against the validation data set. A more sophisticated method for sampling the posterior distribution is described in Appendix **??**.

4.5 Information content of TES ozone column retrievals

We exhibit four different sets of results that provide estimates of information content of aggregated and individual observation data sets in the context of 4D-Var data assimilation.

4.5.1 Aggregated information content of all available data

We first compute the aggregated information content of *all* the available data, i.e., of all the TES ozone profile retrievals available within the 5 days assimilation window. Since 4D-Var adjusts the initial conditions of ozone, the information content metrics describe the data impact on reducing the uncertainty at time t_0.

The estimate of the FIM information content (4) requires an ensemble of $\mathrm{N_{ens}}$ gradient values. Each gradient λ_0^r, $r = 1 \ldots \mathrm{N_{ens}}$, is calculated by running the forward and the adjoint models starting from one of the initial conditions $\mathbf{x}_0^r$ drawn from the posterior ensemble (12) . The ensemble average of the squared gradient entries is computed following (12)

$$\left\langle \lambda_0(i,j,\ell,s_{O3})^2 \right\rangle^{\mathrm{A}} = \frac{1}{\mathrm{N_{ens}}} \sum_{r=1}^{\mathrm{N_{ens}}} \left(\lambda_0^r(i,j,\ell,s_{O3})\right)^2 \ .$$

Using the average squared gradient values and the background error covariance matrix (16)–(17), the numerical approximation to Fisher information is calculated as

$$\begin{aligned}
\mathcal{I}^{\mathrm{FIM}} &= \left\langle \|\lambda_0\|^2 \right\rangle^{\mathrm{A}} - \mathrm{trace}\left(\mathbf{B}_0^{-1}\right) \\
&= \sum_{i=1}^{\mathrm{N_{lon}}} \sum_{j=1}^{\mathrm{N_{lat}}} \sum_{\ell=1}^{\mathrm{N_{lev}}} \left(\left\langle \lambda_0(i,j,\ell,s_{O3})^2 \right\rangle^{\mathrm{A}} - \frac{1}{\sigma_\ell^2} \right) \\
&= \sum_{\ell=1}^{\mathrm{N_{lev}}} \mathcal{I}_\ell^{\mathrm{FIM}} \\
\mathcal{I}_\ell^{\mathrm{FIM}} &= \sum_{i=1}^{\mathrm{N_{lon}}} \sum_{j=1}^{\mathrm{N_{lat}}} \left(\left\langle \lambda_0(i,j,\ell,s_{O3})^2 \right\rangle^{\mathrm{A}} - \frac{1}{\sigma_\ell^2} \right), \quad \ell = 1, 2, \cdots, \mathrm{N_{lev}} \ .
\end{aligned}$$

The first relation provides the scalar value for the FIM information content of all available observations. The last relation provides the Fisher information content relative to the vertical level ℓ of the model; this is a metric of how level ℓ benefits from the assimilation of the data. It is important to note that the breakdown of the information by vertical levels is possible only under the assumption that there is no correlation among errors at different levels. While this is not the case in general, the breakdown provides insight into how the uncertainty is reduced in models with varying pressure levels. The results are shown in Figure 1(a). The FIM information content is large between 400 hPa and 200 hPa, and is small for all other pressure levels. The uncertainty in the initial ozone field is reduced the most in the higher tropospheric area, according to the FIM metric; the levels between 400 hPa and 200 hPa benefit the most from the assimilation of TES ozone retrievals.

The signal information content of all the observations (7) is the background cost function evaluated at the optimal initial condition. Using the formula for background error covariance matrix (17), the level-wise signal contribution could be defined as

$$\begin{aligned}
\mathcal{I}^{\text{Signal}} &= \frac{1}{2}\left(\mathbf{x}_0^{\text{A}} - \mathbf{x}_0^{\text{B}}\right)^T \mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{A}} - \mathbf{x}_0^{\text{B}}\right) \\
&= \frac{1}{2}\sum_{i=1}^{\text{N}_{\text{lon}}}\sum_{j=1}^{\text{N}_{\text{lat}}}\sum_{\ell=1}^{\text{N}_{\text{lev}}}\left(\frac{\mathbf{x}_0^{\text{A}}(i,j,\ell,s_{O3}) - \mathbf{x}_0^{\text{B}}(i,j,\ell,s_{O3})}{\sigma_\ell}\right)^2 \\
&= \sum_{\ell=1}^{\text{N}_{\text{lev}}}\mathcal{I}_\ell^{\text{Signal}} \\
\mathcal{I}_\ell^{\text{Signal}} &= \frac{1}{2}\sum_{i=1}^{\text{N}_{\text{lon}}}\sum_{j=1}^{\text{N}_{\text{lat}}}\left(\frac{\mathbf{x}_0^{\text{A}}(i,j,\ell,s_{O3}) - \mathbf{x}_0^{\text{B}}(i,j,\ell,s_{O3})}{\sigma_\ell}\right)^2, \quad \ell = 1,2,\cdots,\text{N}_{\text{lev}}\,.
\end{aligned}$$

The results for the Signal information content of all observations are shown in Figure 1(b). The Signal information content is the largest between 400 hPa and 200 hPa, which correlates well with the distribution of the FIM information. The Signal information content decreases (almost) linearly for higher pressure levels, and approaches zero near the ground level. This indicates that the assimilation of TES ozone does little to reduce the uncertainty in ozone concentrations near ground level.

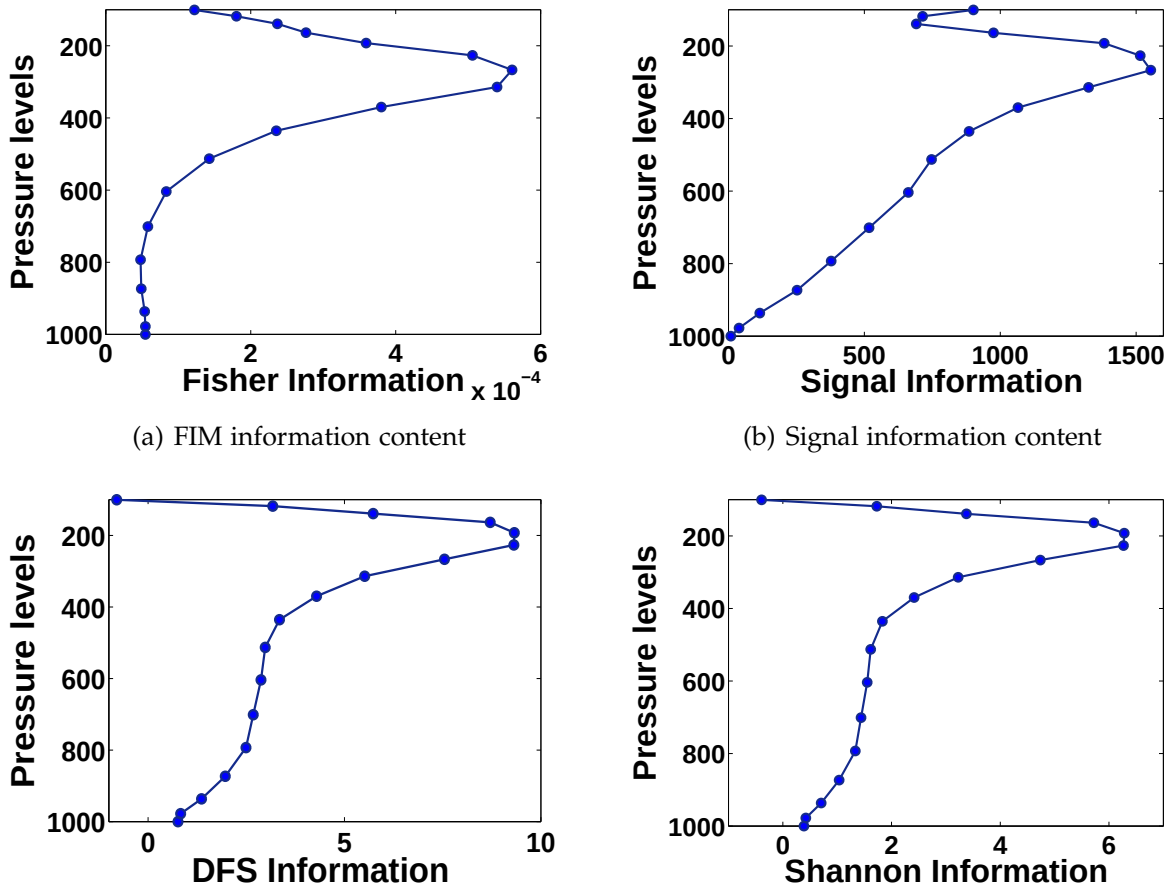
The DFS information (**??**) and the Shannon information content (**??**) are estimated from ensemble covariance eigenvalues using the formulas (**??**) and (**??**), respectively. The results for DFS are shown in Figure 1(c), and the results for Shannon information in Figure 1(d). The two information metrics have highest values between 400 hPa and 200 hPa indicating that a larger uncertainty reduction is obtained in the upper troposphere, and smaller reductions are obtained in the mid and lower troposphere.

4.5.2 The Signal information content

The signal information content of individual data points $\mathbf{y}_\ell$ is estimated using the formula (**??**). No gradient calculations are necessary. The estimate depends only on the innovation vectors associated with the background trajectory $\mathbf{d}_\ell^B = \mathbf{y}_\ell - \mathcal{H}(\mathbf{x}_\ell^B)$, and with the analysis trajectory $\mathbf{d}_\ell^A = \mathbf{y}_\ell - \mathcal{H}(\mathbf{x}_\ell^A)$. Equation (**??**) can be written as

$$\mathcal{I}_{\mathbf{y}_\ell}^{\text{Signal}} \approx \left(\mathbf{d}_\ell^B\right)^T \mathbf{R}_\ell^{-1}\left(\mathbf{d}_\ell^B - \mathbf{d}_\ell^A\right). \tag{19}$$

We first perform a forward model run starting with the optimal initial condition $\mathbf{x}_0^{\text{A}}$ and save the innovation vectors $\mathbf{d}_\ell^A$ for each observation location and for all observation windows. We then perform a second run starting with the background initial condition



(a) FIM information content

(b) Signal information content

(c) DFS information content

(d) Shannon information content

Figure 1: The aggregated information content *of all observations*, as measured by different information theoretic metrics. The breakdown of information content by vertical layers is possible only if the vertical error correlations are negligible.

$\mathbf{x}_0^B$. During this run we compute the innovation vectors $\mathbf{d}_\ell^B$, and, using the saved $\mathbf{d}_\ell^A$ values, we also compute the Signal information content (19).

The time series of the Signal information content per each observation window is shown in Figure 2. The difference between the contribution of observations taken earlier and taken later during the assimilation window is small. This difference is relatively large for the DFS information metric, as will be seen in Figure 5.

We next relate the signal information content of each observation with its location. This approach reveals the spatial distribution of observations that contribute more information to the data assimilation process.

Figure 3(a) presents the locations of observations with the highest Signal information content, specifically the observations within the top 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Signal}}$ averaged over all

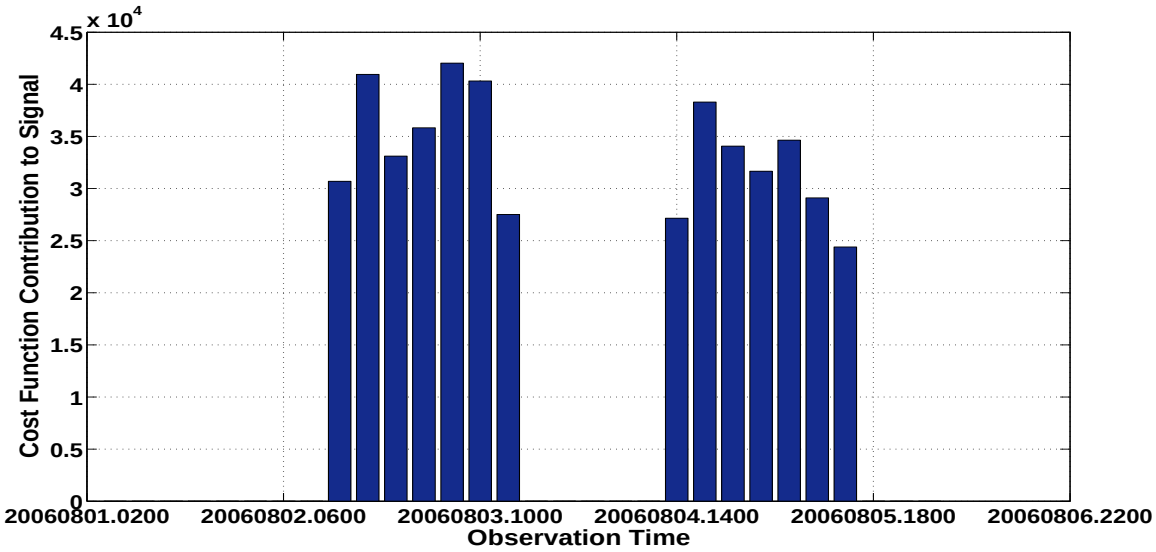
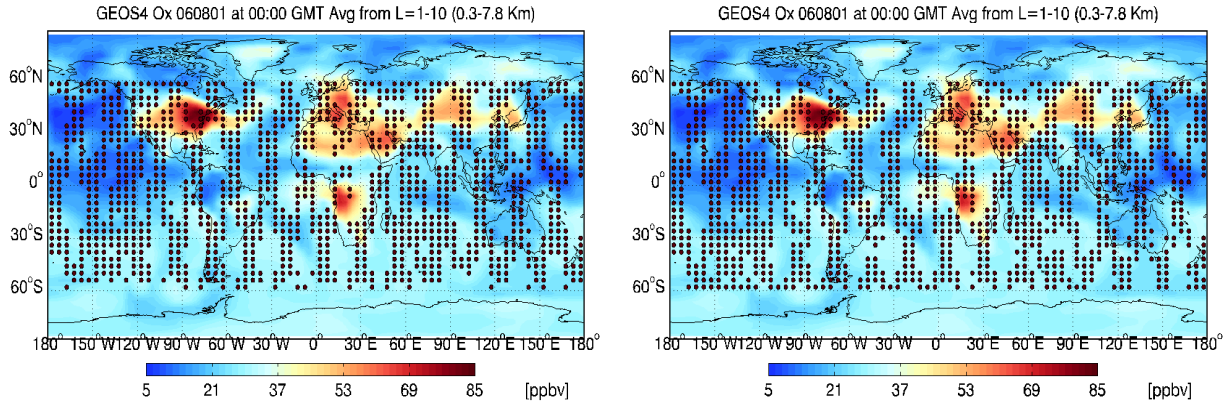


Figure 2: The Signal information content of observations taken at different times within the assimilation window.

vertical layers. Figure 3(b) shows the locations of observations within the bottom 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Signal}}$ averaged over all vertical layers. We see that the two plots are similar; many observations have a similar mean signal information content. Figures 3(c) and 3(d) add vertical information for the location of the top 20% and bottom 20% observations, respectively. The colorbar indicates the model vertical layer number corresponding to the height of the data point. It is evident from panels (c) and (d) that data points with higher Signal information are located in the low to mid troposphere (within 20 GEOS-Chem levels) while points with lower Signal information extend to upper tropospheric levels (45 GEOS-Chem levels).

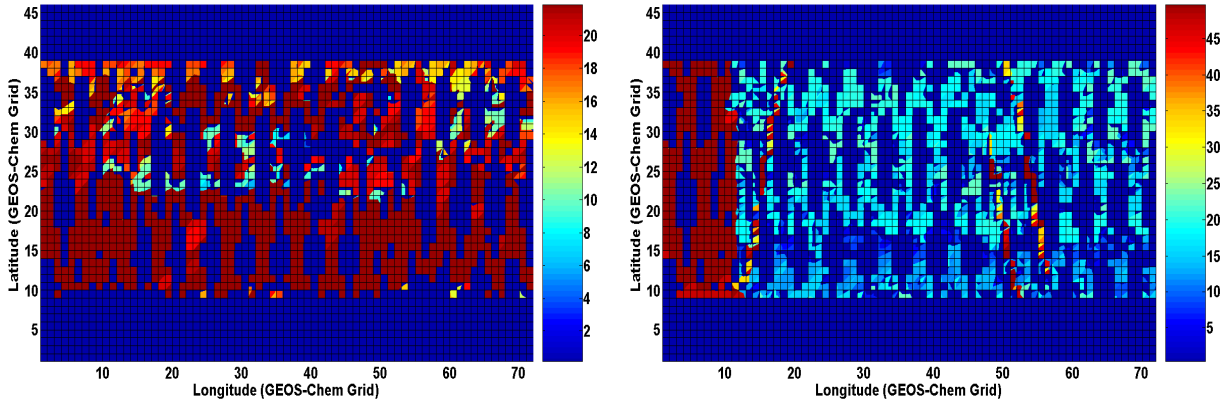
Assimilation of subsets of observations. We now investigate the relationship between the estimated Signal information content, and the benefit that the respective observations bring to the 4D-Var data assimilation process. Specifically, we perform a 4D-Var data assimilation using only the subset of observations within the top twentieth percentile, and another 4D-Var data assimilation using only the subset of observations within the bottom twentieth percentile when ranked by their Signal information content. All data assimilation experiments use the same covariance matrices and the same background field $\mathbf{x}_0^{\text{B}}$.

Figure 4 presents the results of the different data assimilation experiments. The errors are measured against the independent data set of INTEX Ozonesonde Network Study 2006 (IONS-6). The leftmost panel presents the mean ozone concentration vertical



(a) Location of observations within the top 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Signal}}$, averaged over all vertical levels

(b) Location of observations within the bottom 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Signal}}$, averaged over all vertical levels



(c) Location of observations within the top 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Signal}}$

(d) Location of observations within the bottom 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Signal}}$

Figure 3: The location of observations with the highest, and with the lowest signal information content. The colors represent vertical layer numbers of the model.

profiles. The central panel shows the mean errors, i.e., the relative difference between the mean model profiles and ozonesondes. The rightmost panel presents the corresponding error standard deviations. A detailed discussion of the 4D-Var data assimilation results using all the observations is provided in [Singh et al.(2010)].

The results in Figure 4 reveal that the observations with a higher signal information content contribute more to the 4D-Var analysis. The quality of the analysis using only the top 20% of observations is similar to the quality of the analysis using all observations. In contrast, the analysis based on the bottom 20% of the observations has considerably larger errors, albeit it still shows improvement compared to the background case.

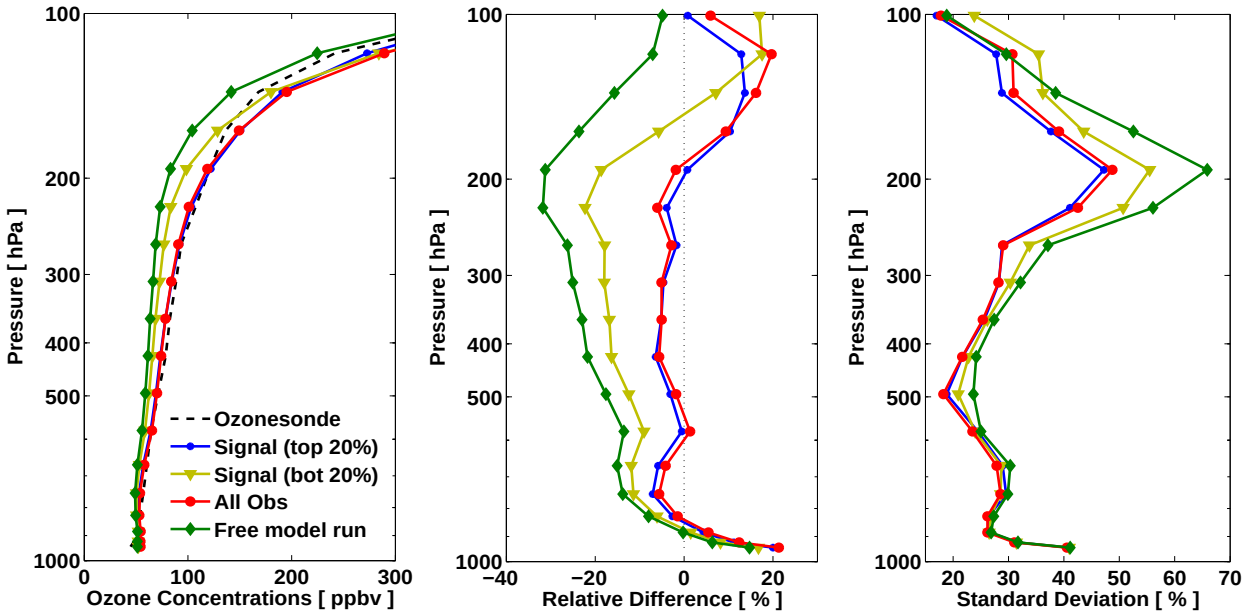


Figure 4: Plot of ozonesonde data, free model run, and 4D-Var analysis trajectories obtained using subsets of observation points. The subsets are selected according to their signal information content.

4.5.3 The DFS information content

Loosely speaking, the DFS metric (discussed in Section **??**) indicates the number of states that benefit from the assimilation of observations. The closer the $\mathcal{I}^{\mathrm{DFS}}$ is to the total number of model states n, the more information the observations have brought into the system through data assimilation. While the signal information content measures the change in the mean field obtained through assimilation, the DFS measures the relative decrease in the error (co-)variance through assimilation. Thus the two metrics measure different aspects of the data assimilation benefits.

The DFS information content for individual data points $\mathbf{y}_i$ is estimated using equation (6). Recall that in our simulations one data point $\mathbf{y}_i$ consists of all the ozone retrievals available in the 4 hours interval $[\, t_i - 2 \text{ hours}\,,\, t_i - 2 \text{ hours}\,]$. As the Aura satellite orbits the Earth the observations are taken over different locations and at different times of day. It is therefore expected that some data points will contain more information than other, i.e., are more useful in reducing uncertainty when assimilated. We utilize the data from the ensemble of $\mathrm{N}_{\mathrm{ens}} + 1$ model runs initialized with states drawn from the analysis distribution (this is the same set of runs used for the aggregated information content calculations). During each of the runs the cost function contribution of each data point

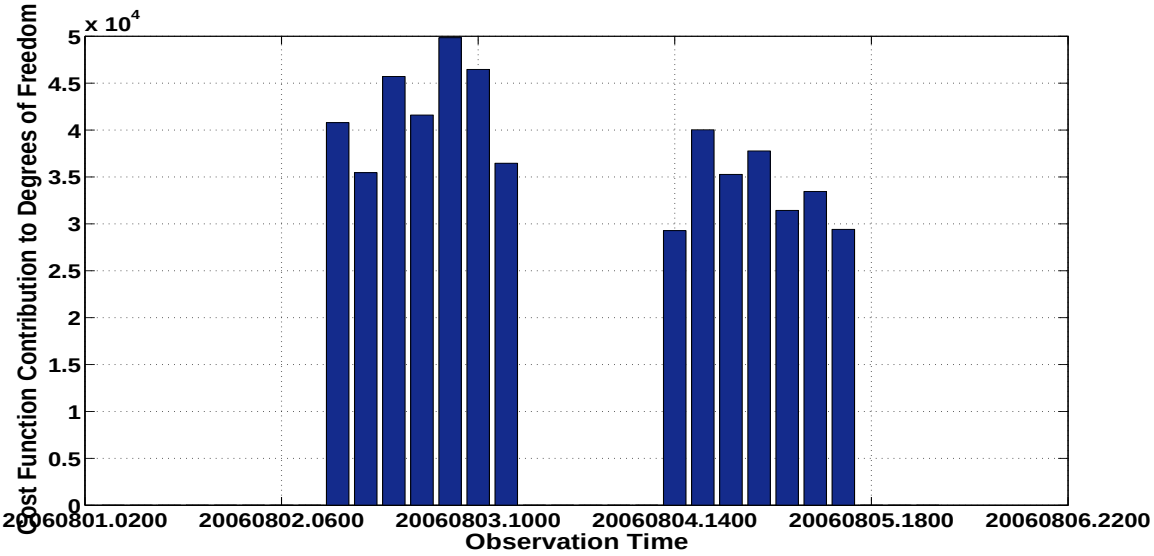
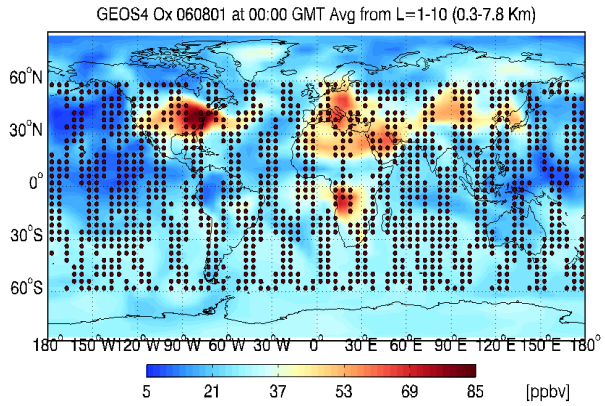


Figure 5: The DFS information content of observations taken at different times within the assimilation window.

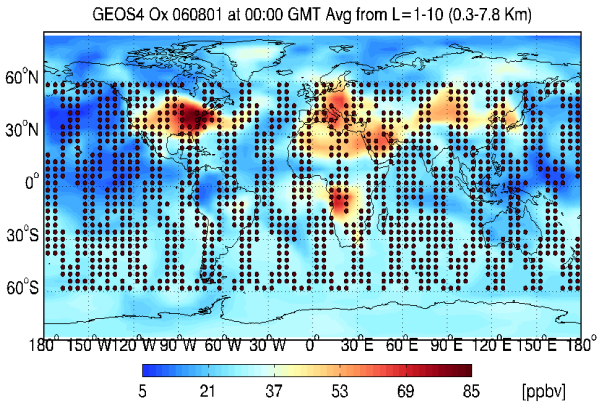
is saved. These results are used to estimate $\mathcal{I}^{\text{DFS}}$ via (6).

Figure 5 presents the DFS information content of the data in each observation window. The data in the observation window 16:00 GMT - 20:00 GMT, August 3rd, 2006 has the highest DFS information content. The DFS information content decreases with time, and the impact of the observations taken later in the assimilation window is smaller. The decreasing trend is more pronounced than the case of the signal information content.

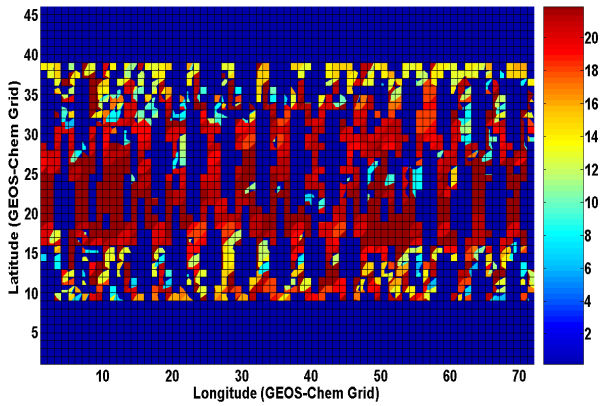
We next study the DFS contribution of each observation point to the assimilation results. Specifically, the data points are classified into subsets according to their estimated DFS information values. Figure 6 shows the location of different observation subsets plotted over the global ozone distribution (averaged over the first 23 levels on August 1st, 2006, 00 GMT). First, all columns of observations are ranked according to their $\mathcal{I}^{\text{DFS}}_{\mathbf{y}_i}$ averaged over all vertical layers. Figures 6(a) and 6(b) represent the locations of the columns within the top and and within the bottom 20-th percentile. The distribution is rather uniform. Next, we rank individual observation points according to their $\mathcal{I}^{\text{DFS}}_{\mathbf{y}_i}$. Figures 6(c) and 6(d) show the locations of the top and of the bottom twentieth percentiles, with the color coordinate representing the height of the data point, in model level units. Similar to Signal information case, data points with higher DFS information content are located in the low to mid troposphere while points with lower DFS information content are extended to upper tropospheric levels. However, there is a clear difference between Figures 3(d) and 6(d) in that points with lower DFS information content are distributed evenly over the globe.



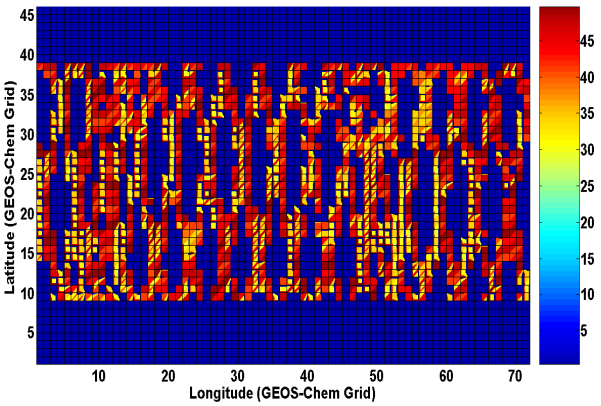
(a) Observation lon-lat coordinates with top 20% $\mathcal{I}^{\text{DFS}}_{\mathbf{y}_i}$



(b) Observation lon-lat coordinates with bottom 20% $\mathcal{I}^{\text{DFS}}_{\mathbf{y}_i}$



(c) Observation lon-lat-lev coordinates with top 20% $\mathcal{I}^{\text{DFS}}_{\mathbf{y}_i}$



(d) Observation lon-lat-lev coordinates with bottom 20% $\mathcal{I}^{\text{DFS}}_{\mathbf{y}_i}$

Figure 6: The location of the most important observations, filtered by their DFS information content.

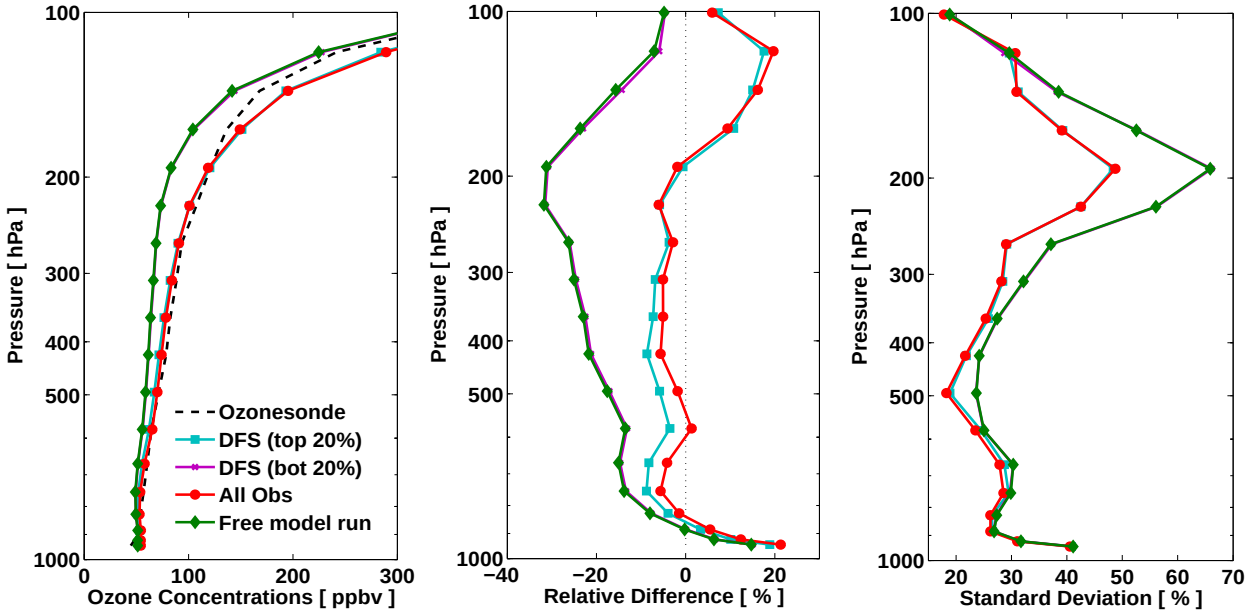
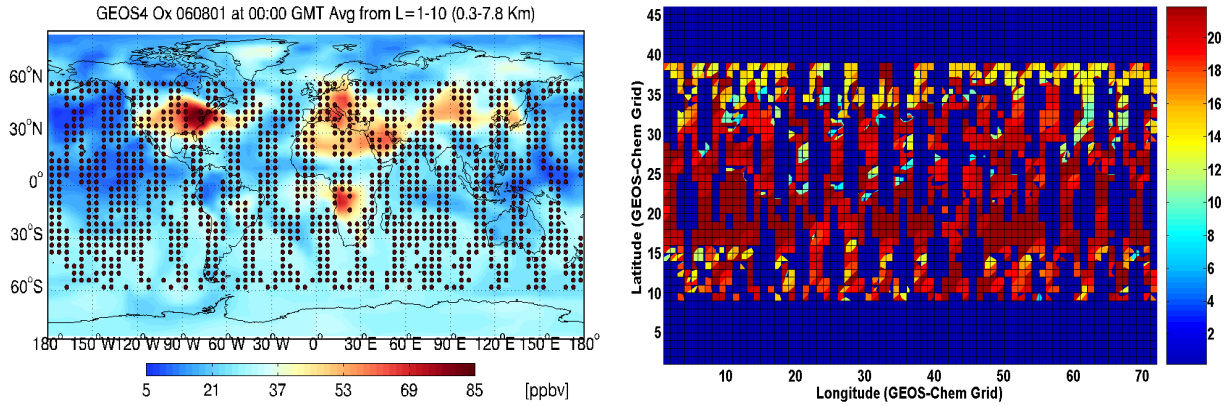


Figure 7: Plot of ozonesonde data, free model run, and 4D-Var analysis trajectories obtained using subsets of observation points. The subsets are selected according to their DFS information content.

Assimilation of subsets of observations. We perform several data assimilation experiments using only subsets of observations, filtered by their estimated DFS information content. The results are presented in Figure 7. The assimilation results using only the top 20% of observation data points (according to the DFS) are almost as accurate as the results using all observation points. The quality of analysis obtained using only the bottom 20% observation points (according to the DFS) is similar to that of free model run. The fact that almost all information is captured by the top 20%, and almost no information is captured by the bottom 20%, suggests that the DFS provides a sharp diagnostic criterion to distinguish between the most and the least important observation data points.

4.5.4 Common to DFS and Signal information content

As described in the previous section, DFS and signal provide complementary measures of the information content. Therefore, It would be of interest to consider observation points that have high DFS as well as high signal information content. We choose the top 20% of all observation points that rank high on both DFS and signal metrics. In order to come up with such a selection, we arranged the complete set of observation data points in two different three-dimensional arrays, first array with descending DFS information



(a) Observation lon-lat coordinates with top 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Common}}$

(b) Observation lon-lat-lev coordinates with top 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Common}}$

Figure 8: The location of the most important observations, filtered by their information content common to DFS and Signal.

content and second with ascending coordinate points (longitude, then latitude) with their signal information content in the third column. We also calculated a threshold $\mathcal{I}_{\mathbf{y}_i,50\%}^{\text{Signal}}$ that determines whether an observation point belongs to the top or the bottom 50% of the signal information content of all observation points. Since DFS provides a clear distinction between points with higher and lower information content, we start with index [0][0][0] and go up to the first half of the first array to collect observation data points that have signal information greater than $\mathcal{I}_{\mathbf{y}_i,50\%}^{\text{Signal}}$ using the second array. In our case, we were able to find 20% of all observation points that meet this criteria. If not, the next step would have been to compare first half of the first array and collect observation points that have signal information less than $\mathcal{I}_{\mathbf{y}_i,50\%}^{\text{Signal}}$. Proceeding until we find the required number of points, next would have been to compare second half of the first array and second array of points with higher signal content, and lastly second half of the first array and second array of points with lower signal content.

Figure 8 represents their longitude-latitude coordinate locations in panel (a), while panel (b) provides the number of levels associated with each location. The color bar in panel (b) indicates that top 20% of observation points which rank high on both DFS and signal fall within 20 vertical model levels.

Assimilation of subsets of observations. We perform data assimilation using only the top 20% of all observations chosen according to the combined DFS and signal criteria. Figure 9 compares the quality of the vertical ozone profiles generated by the free model run, and by the 4D-Var analysis using all observations, the top 20% signal, the top 20%

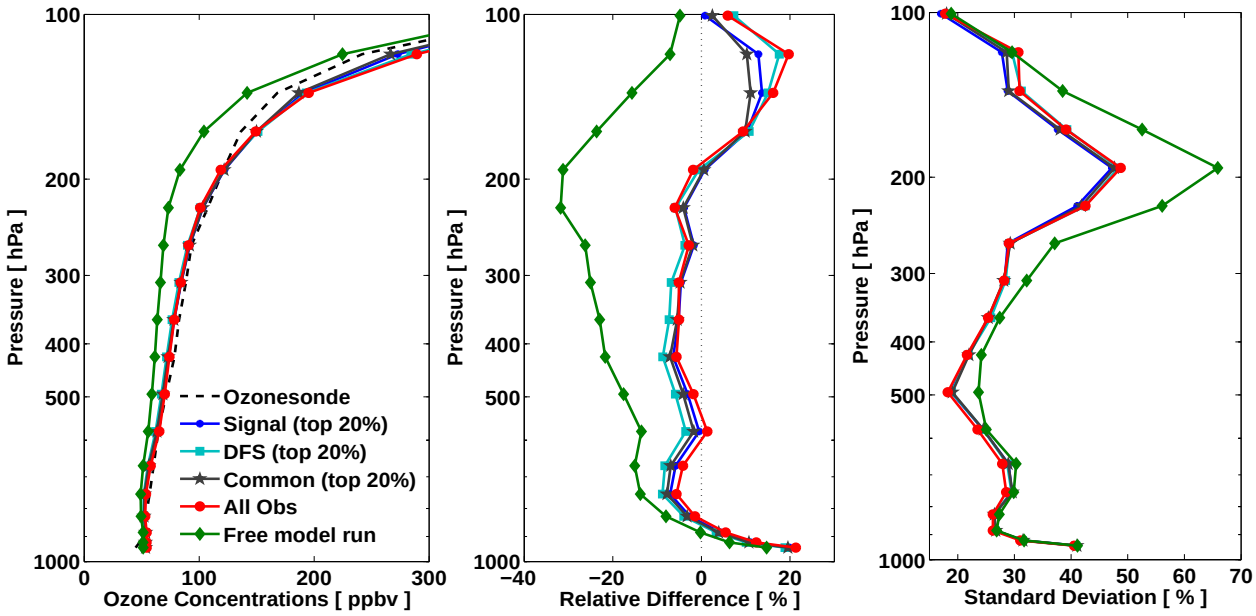


Figure 9: Plot of ozonesonde data, free model run, and 4D-Var analysis trajectories obtained using subsets of observation points. The subsets are selected according to their information content common to DFS and Signal.

DFS, and the top 20% data points common to both DFS and signal. Results indicate that the data points satisfying combined signal and DFS criterion provide the most accurate analysis overall. The analysis generated using these points follows closely the analysis generated by using full observation data set from ground level up to 300 hPa and is better than other observation data sets in the 100–300 hPa vertical region. This indicates that pruning the least informative data points may actually improve the quality of the overall analysis.

A direct comparison of different assimilation results is provided in Figure 10. Specifically, we plot the differences in global ozone concentrations at the beginning of the assimilation window (00:00 GMT on August 6, 2006) averaged over the first 10 GEOS-Chem vertical levels. Panels (a)-(c) show differences between the 4D-Var analysis fields and the model forecast (solution without data assimilation); the analyses use observation data points with top 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{DFS}}$, top 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Signal}}$, and top 20% signal and DFS, respectively. Panels (d)-(f) show absolute differences between 4D-Var analyses using all observation data and using only the data within the top 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{DFS}}$, top 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Signal}}$ and top 20% $\mathcal{I}_{\mathbf{y}_i}^{\text{Common}}$.

Since Figure 10 does not provide any comparison of 4D-Var analysis with real ozone

observations, we use Figure 9 as a baseline to assess the results. There is a limitation to this assessment however which is the fact that the IONS-06 ozonesondes data used in Figure 9 are available only for North America (mainly United States). Comparing Figures 9 and 10 reflect that in lower to mid troposphere (up to 10 GEOS-Chem levels, 400 hPa), 4D-Var analysis using all observations is slightly different from analysis using observations with highest DFS information while it is closer to analysis using observations with highest signal information and observations which rank high on both DFS and signal metrics.

4.5.5 Virtual ground-level observations

So far we have analyzed the information content of real data: the ozone profile retrievals from TES. We next illustrate the use of the proposed methodology to asses the potential impact of *virtual* observations. This is useful for planning new field campaigns, and for guiding the design of new observing networks.

Here we focus on virtual observations taken at ground level. The concentrations of the analysis field $\mathbf{x}^{\mathrm{A}}$ provide the virtual observations. We perform a forward model run starting from $\mathbf{x}_0^{\mathrm{B}}$ and compute the following approximation of the signal information content at hourly intervals

$$\mathcal{I}_{\mathrm{ground}}^{\mathrm{Signal}}\left(\mathbf{x}^{\mathrm{B}}\right) = \frac{1}{2}\left(\mathbf{x}^{\mathrm{B}}{}_{\mathrm{ground}} - \mathbf{x}^{\mathrm{A}}{}_{\mathrm{ground}}\right)^{T} \mathbf{G}^{-1} \left(\mathbf{x}^{\mathrm{B}}{}_{\mathrm{ground}} - \mathbf{x}^{\mathrm{A}}{}_{\mathrm{ground}}\right) . \qquad (20)$$

Note that equation (20) is derived from (19) with the observation data replaced by the analysis field, and with an observation operator that selects the ground level ozone concentrations. The error covariance matrix $\mathbf{G}$ of the virtual observations is diagonal; the standard deviation of each virtual observation is chosen to be 10% of the analysis field. Figure 11 presents the time series of the signal information content of the virtual ground observations. The total signal information initially increases, reaches a peak on August 2nd, 2006, 18:00 GMT, and then decreases. Note that the peak information time for virtual ground level observations is the same as the peak DFS information time for TES ozone column retrievals. This indicates that the ground level observations (possibly) taken on August 2nd at 18:00 GMT are most useful for the assimilation scenario under consideration.

Figure 12 plots the locations of the most important virtual ground level observations, ranked based on their signal information content. These locations are overlaid on top of the global ozone distribution on August 1st, 2006, 00:00 GMT. Figures 12(a) and 12(b) indicate that larger signal information is associated with the region between $60^{\circ}\,N$ and $30^{\circ}\,S$. The reason for this scattering in ground observation case could be attributed to the northern and southern hemisphere subtropical jet streams. The virtual observations

with the highest signal information are located around the Equator and at about $45^\circ N$, as seen in Figures 12(c) and 12(d).

5 Conclusions and Future Work

This paper discusses a characterization of the information content of observations in the context of four dimensional variational (4D-Var) data assimilation framework. The ability to characterize the usefulness of different data points is important for analyzing the effectiveness of the assimilation system, for data pruning, and for the design of future sensor systems.

Several metrics from information theory are used to quantify the information content of data, including the trace of the Fisher information matrix, the signal information, and the degrees of freedom for signal. The companion paper [Sandu et al.(2011)] shows how these metrics can be computed from expected values of the 4D-Var cost function and its gradient. The expected values

The estimates require a sampling from the posterior distribution, which is not readily available in 4D-Var data assimilation. Different approximate methods are possible to obtain analysis samples. Here we use a normal distribution with the mean given by the assimilation result, a diagonal covariance matrix, and the analysis variances obtain by properly scaling the background variances. The error ratios obtained by comparing the model results against an independent data set are used to determine the scaling factors. More sophisticated methods for sampling the posterior distribution are possible, as discussed in Section 3.

The information content estimation approach is applied to a global ozone data assimilation problem using TES satellite observations and the GEOS-Chem chemical transport model. The quality of the assimilation is assessed by comparing the results against an independent data set (INTEX ozonesonde measurements). The assimilation of a subset of 20% of the data points characterized by the highest signal, DFS, and combined information content, gives analyses that are comparable in quality with the one obtained using the entire data set. This results are very encouraging since they indicate the effectiveness of the proposed approach as a diagnosis tool for the value of observations used during the assimilation. Moreover, pruning the least informative observations seems to improve the quality of the analysis in the upper atmosphere. This point deserves future investigation.

Acknowledgements

This work has been supported in part by NASA through the ROSES-2005 AIST project, by NSF through the awards NSF CCF–0635194, NSF OCI–0904397, NSF CCF–0916493, and NSF DMS–0915047, and by the Computational Science Laboratory at Virginia Tech.

References

[Abramov(2004)] Abramov, R. V. and Majda, A. J., Quantifying uncertainty for non-Gaussian ensemble in complex systems. *SIAM Journal on Scientific Computing*, 2004; **26(2)**, 411-447.

[Bartlett(1962)] Bartlett, M. S., An introduction to stochastic processes, with special reference to methods and applications. *Cambridge University Press*, 1962.

[Beer et al.(2001)] Beer, R., Glavich, T. A., and Rider, D. M., Tropospheric emission spectrometer for the Earth Observing System's Aura satellite. *Applied Optics*, 2001; **40(15)**, 2356-2367.

[Bernardo(1994)] Bernardo, J. M. and Smith, A. F. M., Bayesian theory. *Wiley, Chichester, UK*, 1994.

[Bey et al.(2001)] Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B., Fiore, A. M., Li, Q., Liu, H., Mickley, L. J., and Schultz, M., Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research*, 2001; 106, 23, 073-23,095.

[Bowman et al.(2008)] Bowman, K. W., Jones, D. B. A., Logan, J. A., Worden, H., Boersma, F., Kulawik, S. S., Osterman, G., Worden, J., and Chang, R., Impact of surface emissions to the zonal variability of tropical ozone and carbon monoxide for November 2004. *Atmospheric Chemistry and Physics*, 2008; **8**, 1505-1548.

[Bowman et al.(2006)] Bowman, K. W., Rodgers, C. D., Kulawik, S. S., Worden, J., Sarkissian, E., Osterman, G., Steck, T., Luo, M., Eldering, A., Shepherd, M., Worden, H., Lampel, M., Clough, S., Brown, P., Rinsland, C., Gunson, M., and Beer, R., Tropospheric Emission Spectrometer: Retrieval method and error analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 2006; **44(5)**, 1297-1307.

[Bowman et al.(2002)] Bowman, K. W., Worden, J., Steck, T., Worden, H. M., Clough, S., and Rodgers, C., Capturing time and vertical variability of tropospheric ozone: A study using TES nadir retrievals. *Journal of Geophysical Research*, 2007; **107**, D23.

[Cardinali et al.(2004)] Cardinali, C., Pezzulli, S., Andersson, E., Influence-matrix diagnostic of data assimilation system. *Quarterly Journal of the Royal Meteorological Societ*, 2004; **130**, 2767-2786.

[Carmichael et al.(2003)] Carmichael, G. R., Daescu, D. N., Sandu, A., Chai, T., Computational aspects of chemical data assimilation into atmospheric models. *Computational Science - ICCS*, 2003, PT IV, Book series title: Lecture Notes in Computer Science , 2660: 269-278 , 2003.

[Carmichael et al.(2008)] Carmichael, G. R., Sandu, A., Chai, T., Daescu, D. N., Constantinescu, E. M., Tang, Y., Predicting air quality: Improvements through advanced methods to integrate models and measurements. *Journal of Computational Physics*, 227 (7): 3540-3571 , 2008

[Chai et al.(2006)] Chai, TF;Carmichael, GR; Sandu, A; Tang, Y., Daescu, D. N., Chemical data assimilation of Transport and Chemical Evolution over the Pacific (TRACE-P) aircraft measurements. *Journal of Geophysical Research – Atmospheres*, 111 (D2): Art. No. D02301, 2006.

[Chai et al.(2007)] Chai, TF; Carmichael, GR; Tang, YH; Sandu, A., Hardesty, M., Pilewskie, P., Whitlow, S., Browell, E. V., Avery, M. A., Nédélec, P., Merrill, J. T., Thompson, A. M., Williams, E., Four-dimensional data assimilation experiments with International Consortium for Atmospheric Research on Transport and Transformation ozone measurements. *Journal of Geophysical Research – Atmospheres*, 112 (D12): Art. No. D12S15, 2007.

[Cheng et al.(2010)] Cheng, H., Jardak, M., Alexe, M. and Sandu, A., A hybrid approach to estimating error covariances in variational data assimilation. *Tellus A.* **Vol.** 62, Number 3, May 2010 , pp. 288-297(10).

[Constantinescu et al.(2007a] Constantinescu, E. M., Sandu, A., Chai, T. F., Ensemble-based chemical data assimilation. I: General approach. *Quarterly Journal of the Royal Meteorological Society*, 133 (626): 1229-1243 Part A , 2007

[Constantinescu et al.(2007b] Constantinescu, E. M., Sandu, A., Chai, T. F., Carmichael, G. R., Ensemble-based chemical data assimilation. II: Covariance localization. *Quarterly Journal of the Royal Meteorological Society*, 133 (626): 1245-1256 Part A , 2007

[Constantinescu et al.(2007c] Constantinescu, E. M., Sandu, A., Chai, T. F., Carmichael, G. R., Assessment of ensemble-based chemical data assimilation in an idealized setting. *Atmospheric Environment*, 41 (1): 18-36 , 2007.

[Constantinescu et al.(2007d)] Constantinescu, E. M., Chai, T. F., Sandu, A., Carmichael, G. R., Autoregressive models of background errors for chemical data assimilation. *Journal of Geophysical Research – Atmospheres*, 112 (D12): Art. No. D12309 , 2007.

[Courtier and Talagrand(1987)] Courtier, P., and Talagrand, O., Variational assimilation of meteorological observations with the adjoint vorticity equations Part 2: Numerical results. *Quarterly Journal of the Royal Meteorological Society*, 1987; **113**, 1329-1347.

[Daescu et al.(2000)] Daescu, D., Carmichael, G. R., Sandu, A., Adjoint implementation of Rosenbrock methods applied to variational data assimilation problems. *Journal of Computational Physics*, 165 (2): 496-510, 2000.

[Eller et al.(2009)] Eller, P., Singh, K., Sandu, A., Bowman, K., Henze, D. K., and Lee, M., Implementation and evaluation of an array of chemical solvers in a global chemical transport model. *Geophysical Model Development*, 2009; **Vol. 2**, 1-7.

[Fisher(1922)] Fisher, R. A., On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, 1922; **Series A, 222**, 309-368, URL: http://www.jstor.org/stable/91208.

[Fisher(2003)] Fisher, M., Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems. *ECMWF Technical Memoranda*, 2003; **397**.

[Gejadze et al.(2008)] Gejadze, I. Y., Le Dimet, F. X., and Shutyaev, V., On analysis error covariances in variational data assimilation. *SIAM Journal on Scientific Computing*, 2008; **30(4)**, 1847-1874.

[Hakami et al.(2005)] Hakami, A., Henze, D. K., Seinfeld, J. H., Chai, T. F., Tang, Y., Carmichael, G. R., and Sandu, A., Adjoint inverse modeling of black carbon during the Asian Pacific Regional Aerosol Characterization Experiment. *Journal of Geophysical Research – Atmospheres*, 110, (D14): Art. No. D14301, 2005.

[Heinkenschloss(2008)] Heinkenschloss, M., Numerical Solution of Implicitly Constrained Optimization Problems. CAAM Technical Report TR08-05, Department of Computational and Applied Mathematics, Rice University.`http://www.caam.rice.edu/~heinken/software/matlab_impl_constr`.

[Henze et al.(2004)] Henze, D. K., Seinfeld, J. H., Liao, W., Sandu, A., Carmichael, G. R., Inverse modeling of aerosol dynamics: Condensational growth. *Journal of Geophysical Research – Atmospheres*, 109 (D14): Art. No. D14201 , 2004.

[Henze et al.(2007)] Henze, D. K., Hakami, A., and Seinfeld, J. H., Development of the adjoint of GEOS-Chem. *Atmospheric Chemistry and Physics*, 2007; **7**, 2413-2433.

[Henze et al.(2009)] Henze, D. K., Seinfeld, J. H., and Shindell, D. T., Inverse modeling and mapping U.S. air quality influences of inorganic PM2.5 precursor emissions with the adjoint of GEOS-Chem. *Atmospheric Chemistry and Physics*, 2009; **9**, 5877-5903.

[Houtekamer and Mitchell (1998)] Houtekamer, P. L. and Mitchell, H. L., Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review* **126**, 796–811, 1998.

[Jazwinski(1970)] Jazwinski, A. H., Stochastic processes and filtering theory. *Academic Press, New York*, 1970.

[Jones et al.(2003)] Jones, D. B. A., Bowman, K. W., Palmer, P. I., Worden, J. R., Jacob, D. J., Hoffman, R. N., Bey, I., and Yantosca, R. M., Potential of observations from the Tropospheric Emission Spectrometer to constrain continental sources of carbon monoxide. *Journal of Geophysical Research*, 2003; **108**, D24.

[Kalman(1960)] Kalman, R. E., A new approach to linear filtering and prediction problems. *Transaction of the ASME - Journal of basic Engineering*, 1960; **Series D(82)**, 35-45.

[Kopacz et al.(2007)] Kopacz, M., Jacob, D. J., Henze, D. K., Heald, C. L., Streets, D. G., and Zhang, Q., A comparison of analytical and adjoint Bayesian inversion methods for constraining Asian sources of CO using satellite (MOPITT) measurements of CO columns. *Journal of Geophysical Research*, 2009; **114**, D04305.

[Kullback(1968)] Kullback, S., Information theory and statistics. *Wiley, New York*, 1968.

[Li et. al(1993)] Li, Y., Navon, I. M., Courtier, P., and Gauthier, P., Variational data assimilation with a semi-Lagrangian semi-implicit global shallow water equation model and its adjoint. *Monthly Weather Review*, 1993; **121(6)**, 1759–1769.

[Le Dimet(1986)] Le Dimet, F. X. and Talagrand, O., Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, **v.38**, 97–110, 1986.

[Liao et al.(2006)] Liao, W. Y., Sandu, A., Carmichael, G. R., Chai, T. F., Singular vector analysis for atmospheric chemical transport models. *Monthly Weather Review*, 134 (9): 2443-2465 SEP 2006.

[Lions(1971)] Lions, J. L., Optimal control of systems governed by partial differential equations. *Springer-Verlag*, 1971.

[Lorenc(1986)] Lorenc, A. C., Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society* 1986;**112**, 1177-1194.

[Lorenz(1996)] Lorenz, E., Predictability: A problem partly solved. *Proceedings of the Seminar on Predictability*, 1996; Shinfield Park, Reading, UK, ECMWF.

[Majda(2006)] Majda, A. J. and Wang, X., Nonlinear dynamics and statistical theories for basic geophysical flows. *Cambridge University Press*, 2006.

[Navon et. al(1992)] Navon, I. M., Zou, X., Derber, J., Sela, J., Variational data assimilation with an adiabatic version of the NMC Spectral Model. *Monthly Weather Review*, 1992; **120(7)**, 1433-1446.

[Palmer et al.(2003)] Palmer, P. I., Jacob, D. J., Jones, D. B. A., Heald, C. L., Yantosca, R. M., Logan, J. A., Sachse, G. W., and Streets, D. G., Inverting for emissions of carbon monoxide from Asia using aircraft observations over the western Pacific. *Journal of Geophysical Research*, 2003; **108**, 8828.

[Parrington et al.(2008)] Parrington, M., Jones, D. B. A., Bowman, K. W., Horowitz, L. W., Thompson, A. M., Tarasick, D. W., Witte, J. C., Estimating the summertime tropospheric ozone distribution over North America through assimilation of observations from the Tropospheric Emission Spectrometer. *Journal of Geophysical Research*, 2008; **Vol 113**, D18307.

[Parrington et al.(2009)] Parrington, M., Jones, D. B. A., Bowman, K. W., Thompson, A. M., Tarasick, D. W., Merrill, J., Oltmans, S. J., Leblanc, T., Witte, J. C., Millet, D. B., Impact of the assimilation of ozone from the Tropospheric Emission Spectrometer on surface ozone across North America. *Geophysical Research Letters*, 2009; **36(4)**, L04802.

[Rabier et al.(2002)] Rabier, F., Fourrie, N., Chafa, D., and Prunet, P., Channel selection methods for Infrared Atmospheric Sounding Interferometer radiances. *Quarterly Journal of the Royal Meteorological Society*, 2002; **128**, 1011-1027.

[Rodgers(1976)] Rodgers, C. D., Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation. *Reviews of Geophysics and Space Physics*, 1976; **14**, 609-624.

[Rodgers(1996)] Rodgers, C. D., Information content and optimization of high spectral resolution measurements. *Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research*, **SPIE Volume 2830**, 136-147.

[Rodgers(1998)] Rodgers, C. D., Information content and optimization of high spectral resolution measurements. *Advances in Space Research*, 1998; **21**, 361-367.

[Rodgers(2000)] Rodgers, C. D., Inverse methods for atmospheric sounding: Theory and Practice. *World Scientific: Singapore*, 2000.

[Sasaki (1958)] Sasaki, Y. K., An objective analysis based on the variational method. *Journal of the Meteorological Society of Japan*, 1958; **II(36)**, 77-88.

[Sandu et al.(2005)] Sandu, A., Liao, W., Carmichael, G. R., Henze, D. K., and Seinfeld, J. H., Inverse modeling of aerosol dynamics using adjoints: Theoretical and numerical considerations. *Aerosol Science and Technology*, 2005, 39 (8): 677-694.

[Sandu et al.(2005)] Sandu, A., Constantinescu, E. M., Liao, W. Y., Carmichael, G. R., Chai, T., Seinfeld, J. H., and Daescu, D. N., Ensemble-based data assimilation for atmospheric chemical transport models. *Computational Science - ICCS*, 2005, PT 2 Book series title: Lecture Notes in Computer Science , 3515: 648-655.

[Sandu et al.(2005)] Sandu, A., Daescu, D. N., Carmichael, G. R., and Chai, T., Adjoint sensitivity analysis of regional air quality models. *Journal of Computational Physics*, 2005; **Volume 204**, 222-252.

[Sandu and Zhang(2008)] Sandu, A. and Zhang, L., Discrete second order adjoints in atmospheric chemical transport modeling. *Journal of Computational Physics*, 2008; **227(12)**, 5949-5983.

[Sandu et al.(2011)] Sandu, A., Singh, K., Jardak, M., Bowman, K., Lee, M., A Practical Method to Estimate Information Content in the Context of 4D-Var Data Assimilation. I: Methodology. *Journal of Geophysical Research*, 2011; submitted.

[Shannon and Weaver(1949)] Shannon, C. E. and Weaver, W., The mathematical theory of communication. *University of Illinois Press, Urbana, IL.*, 1949.

[Singh(2009a)] Singh, K., Eller, P., Sandu, A., Bowman, K. W., Jones, D., Lee, M., Improving GEOS-Chem model forecasts through profile retrievals from Tropospheric Emission Spectrometer. *International Conference on Computational Science*, 2009, Lecture Notes on Computational Science **volume 5545**, 302-311.

[Singh(2009b)] Singh, K., Eller, P., Sandu, A., Henze, D., Bowman, K., Kopacz, M. and Lee, M., Towards the construction of a standard adjoint GEOS-Chem model. *High Performance Computing Symposium at Spring Simulation Multiconference*, 2009.

[Singh et al.(2010)] Singh, K., Jardak, M., Sandu, A., Bowman, K. W., Lee, M., Jones, D., Construction of non-diagonal background error covariance matrices in global chemical data assimilation. *Geophysical Model Development*, 4, 299-316, doi:10.5194/gmd-4-299-2011, 2011.

[Stewart et al.(2008)] Stewart, L. M., Dance, S. L., Nichols, N. K., Correlated observation errors in data assimilation. *International Journal for Numerical Methods in Fluids*, 2008; **56**, 1521-1527.

[Thacker(1989)] Thacker, W. C., The role of the Hessian matrix in fitting models to measurements. *Journal of Geophysical Research*, 1989; **94(C5)**, 6177-6196.

[Thompson et al. (2007a, 2007b)] Thompson, A.M., et al. (2007a), "Intercontinental chemical transport experiment ozonesonde network study (IONS) 2004: 1. Summertime upper troposphere/lower stratosphere ozone over northeastern North America". *Journal of Geophysical Research*, **112**, D12S12.
Thompson, A.M., et al. (2007b), "Intercontinental chemical transport experiment ozonesonde network study (IONS) 2004: 2. Tropospheric ozone budgets and variability over northeastern North America". *Journal of Geophysical Research*, **112**, D12S13.

[Worden et al.(2004)] Worden, J. R., Bowman, K. W., and Jones D. B. A., Characterization of atmospheric profile retrievals from Limb Sounding Observations of an inhomogeneous atmosphere. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 2004; **86**,(03)00274-7.

[Xu(2006)] Xu, Q., Measuring information content from observations for data assimilation: relative entropy versus Shannon entropy difference. *Tellus, A.* 2006, 198-209.

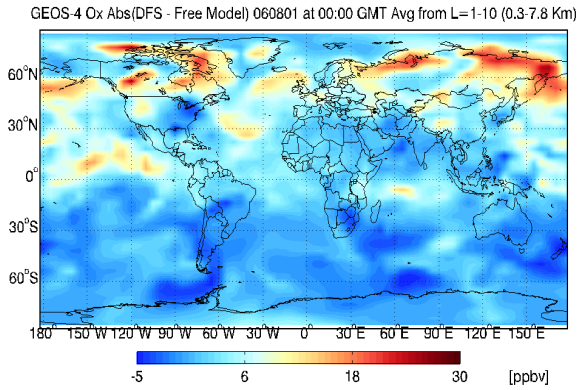
[Zhang et al.(2009)] Zhang, L., Jacob, D. J.,Kopacz, M., Henze, D. K., Singh, K., Jaffe, D. A., Intercontinental source attribution of ozone pollution at western U.S. sites using an adjoint method. *Geophysical Research Letters*, 2009; **Volume 36**, L11810.

[Zhu et al.(1997)] Zhu, C., Byrd, R. H., and Nocedal., J., L-BFGS-B: Algorithm 778, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 1997; **Vol 23, Num. 4**,550-560.

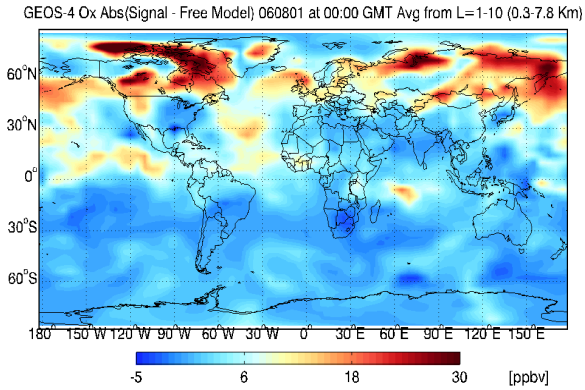
[Zou et. al(1993)] Zou, X., Navon, I. M., Sela, J., Variational data assimilation with moist threshold processes using the NMC spectral model. *Tellus A.*, 1993; **45A**, 370–387.

[Zupanski(2009)] Zupanski, D., Information measures in ensemble data assimilation. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, 2009.

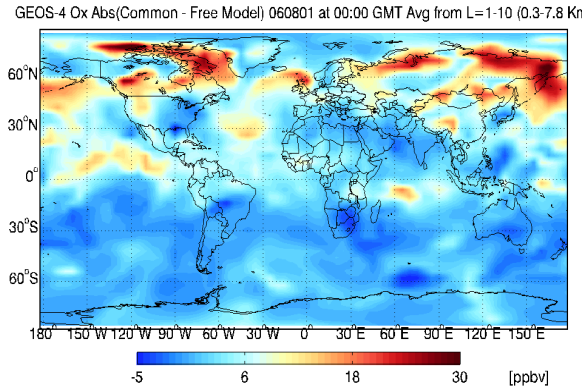
[Wikle and Berliner(2007)] Wikle, C.K., and Berliner, L.M. A Bayesian tutorial for data assimilation. *Physica D*, 2007; **230**, 1–16.



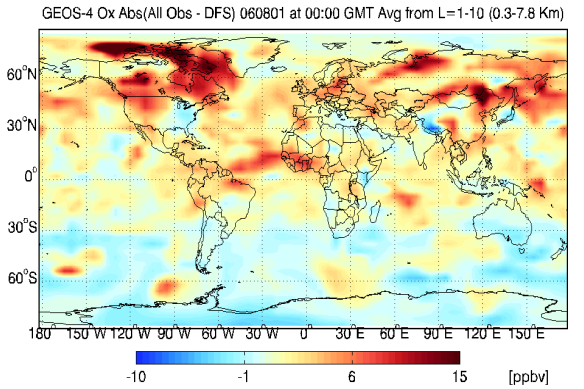
(a) Absolute difference between the 4D-Var analysis using data points with top 20% DFS and the free model run



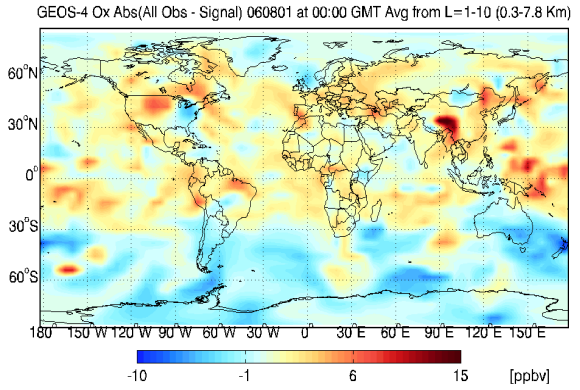
(b) Absolute difference between the 4D-Var analysis using data points with top 20% signal and the free model run



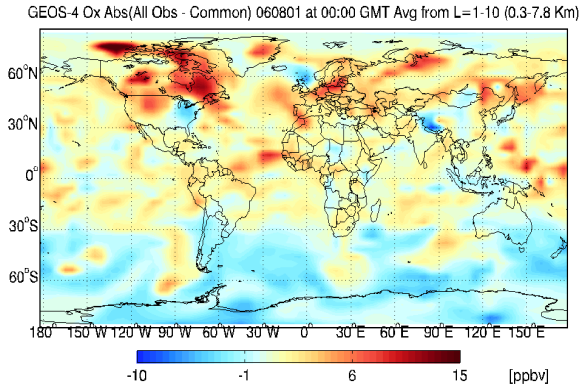
(c) Absolute difference between the 4D-Var analysis using data points with top 20% common signal and DFS and the free model run



(d) Absolute difference between the 4D-Var analysis using all observations and data points with top 20% DFS



(e) Absolute difference between the 4D-Var analysis using all observations and data points with top 20% signal



(f) Absolute difference between the 4D-Var analysis using all observations and data points with top 20% common signal and DFS

Figure 10: Direct comparison of different assimilation results using various subsets of the data. Differences in global ozone concentrations are shown at 00:00 GMT on August 6, 2006 and averaged over the first 10 GEOS-Chem vertical levels.

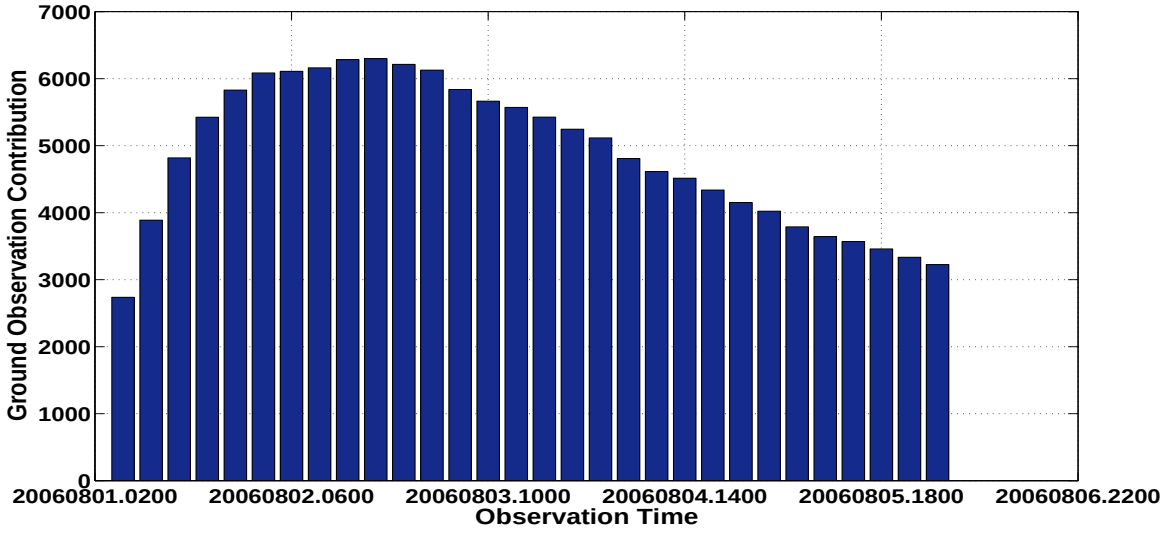
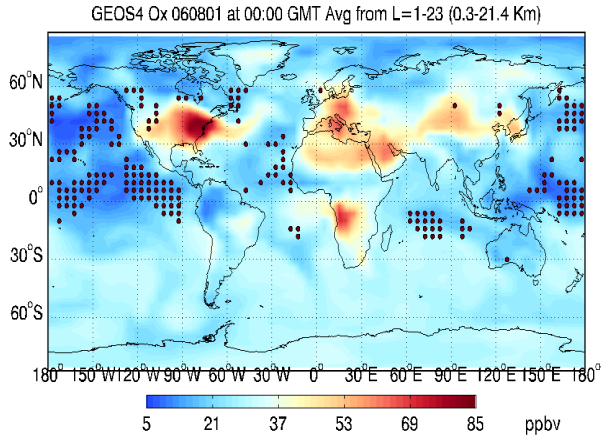
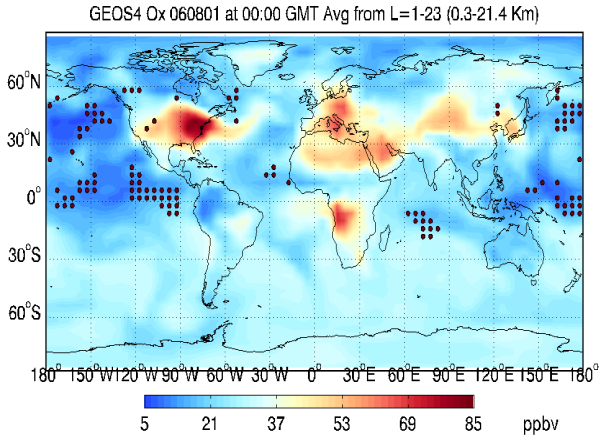


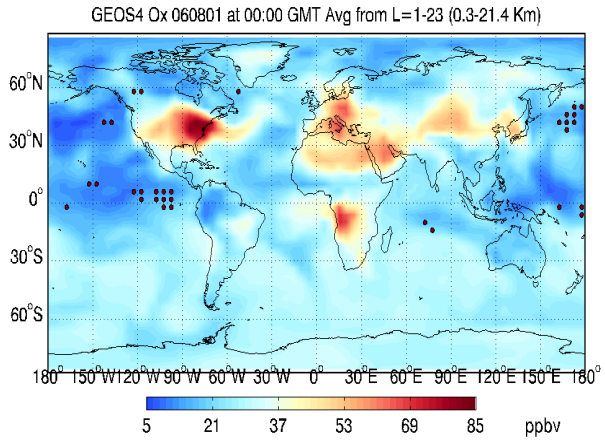
Figure 11: Time evolution of the total signal information content of virtual ground level observations during the assimilation window.



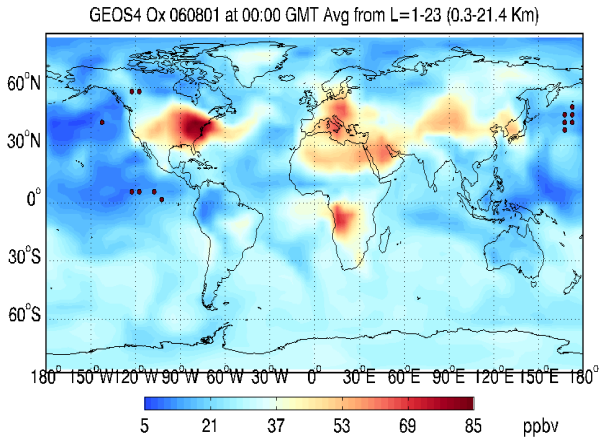
(a) Virtual ground observations with $\mathcal{I}_{\text{ground}}^{\text{Signal}} \geq 3$



(b) Virtual ground observations with $\mathcal{I}_{\text{ground}}^{\text{Signal}} \geq 5$



(c) Virtual ground observations with $\mathcal{I}_{\text{ground}}^{\text{Signal}} \geq 10$



(d) Virtual ground observations with $\mathcal{I}_{\text{ground}}^{\text{Signal}} \geq 15$

Figure 12: The location of virtual ground level observations with the largest signal information content.