EAKF-CMAQ: DEVELOPMENT AND INITIAL EVALUATION OF AN ENSEMBLE ADJUSTMENT KALMAN FILTER BASED DATA ASSIMILATION FOR CO

Alexis Zubrow *
CISES, The University of Chicago, Chicago, IL, USA

Li Chen
CISES, The University of Chicago, Chicago, IL, USA

V. R. Kotamarthi
Argonne National Laboratory, Argonne, IL, USA

Michael L. Stein
Statistics Department, The University of Chicago, Chicago, IL, USA

1 INTRODUCTION

An integrated approach to modeling atmospheric chemistry with trace gas data assimilation is a relatively new focus of the atmospheric chemistry modeling community. It is expected that the predictive capability of CTMs can be significantly improved by assimilating measurements of key trace gases from satellite-based platforms and surface monitors. Ensemble adjustment Kalman filter (EAKF) methods are simple to implement, don't need adjoints and backward integration, and are capable of handling non-Gaussian model errors. These factors have led to the adoption of EAKF methods for weather and climate simulations. Additionally, EAKF provides a measure of error resulting from the assimilation. We have combined EAKF data assimilation with a single-tracer version of CMAQ. The Data Assimilation Research Testbed (DART), developed by NCAR, was used to create an EAKF enabled CMAQ for assimilating CO. DART provides a modular environment that can integrate dynamical models with various assimilation techniques. Specifically, we ran CMAQ in ensemble adjustment Kalman filter mode to assimilate both synthetic and real observations of CO for the period of June 2001. We argue that it is a viable approach for further data assimilation experiments and potentially for air quality forecasting.

*Corresponding author: Alexis Zubrow, the Center for Integrating Statistical and Environmental Science (CISES), the University of Chicago, 5734 S. Ellis Ave. Chicago, IL 60637; email: azubrow@uchicago.edu.

2 MODIFICATIONS TO CMAQ

A single tracer version of CMAQ (Im et al. 2005) was modified to model Carbon Monoxide (CO). CO has a relatively simple chemistry given OH; therefore our numerical experiments could focus on the data assimilation of a single trace and not have to consider multiple chemical reactions in evaluating the results. The relatively long life of CO means that assimilations would have influence beyond one diurnal cycle. Because of it's long memory, CO experiments may help determine the potential of data assimilations to improve forecasts.

Our modified CMAQ model, from here out CCTM_CO, retained the ingestion of emissions and advection of CO from the full CMAQ model. The model transports only one tracer (CO) and performs the following chemistry:

$$\frac{\partial[CO]}{\partial t} = K_{form}[OH][HCOH] - K_{oh}[CO][OH] \quad (1)$$

where K_{oh} and K_{form} are the reactions rates for OH and formaldehyde, respectively. The concentrations of OH and formaldehyde are fixed and updated hourly from a full offline CMAQ run. The OH and formaldehyde concentrations are read into CCTM_CO from a separate file (OH_CONC) during the run time of each ensemble member.

A second reason for creating CCTM_CO is the computational cost of running multiple ensemble members. For our relatively small beowulf cluster (16 processors), the full CMAQ model would make ensemble runs of 10 or more realizations impractical. The full CMAQ model takes approximately 50 minutes per model day on 4 processors. The CCTM_CO

runs in 3 to 4 minutes for the same domain on 1 processor.

3 DART

The Data Assimilation Research Testbed (DART) was developed at the National Center for Atmospheric Research (NCAR). It provides a modular approach for testing various data assimilation schemes, based on ensemble simulations [1]. It uses a Bayesian approach for data assimilation. In DART, the prior is the model output, in other words it is our best estimation of the state of the system (the model) before we assimilate any observations. The spread of the prior is represented by the spread of the ensemble members, each a unique realization of the model. In other words, the value of the ensemble members at a particular location is the sample from the prior at that location. The observation likelihood is the distribution of each observation. By combining the prior and the likelihood, a posterior estimate can be generated. The posterior is our updated state of the model after the data has been assimilated; our best understanding of the system given the model and the data. In the parlance of DART, the process of combining the prior with the likelihood is called "filtering". The model values, either in the prior or in the posterior, is called the "state variable".

DART provides multiple types of filters to use for data assimilation. We chose the Ensemble Adjustment Kalman Filter (EAKF) because of a series of advantages (Anderson 2001). First, EAKF reduces the extent of the problem by only looking at the correlation between observations and nearby state variables. In other words, one can limit the calculation updates to only model values within a certain spatial region (a cutoff). Second, the relative relationship between the prior ensembles is maintained in the posterior ensembles. For example if ensemble 1 was predicting a lower value of CO in a region than ensemble 2 in the prior, then the posterior would still have ensemble 1 with a lower prediction than ensemble 2. The EAKF first shifts the prior ensembles to match the posterior mean, then adjusts their spread to the posterior variance. In this way, individual ensemble trajectories are more physically relevant. Third, EAKF is relatively computationally efficient. The model does need to be run N times, which is a fixed cost for any ensemble method. The cost of filtering is on the order O(mnN), where m is the number of observations, n is the size of the state variable (the model), and N is the number of ensembles.

[1] See http://www.image.ucar.edu/DAReS/DART/ (Sept 2006) for more details.

Creating an interface between DART and CMAQ had two stages. First, the development of a DART module in FORTRAN 90 provides necessary information to DART about translating back and forth between the model state and the observation state. DART does all of it's assimilation in observation space:

$$Y = H(x) + \epsilon \tag{2}$$

where x is the model state, Y is the expected observation, H is a function that gives the expected value of the observation given the model state, and ϵ is the error. The DART module needs to define this function for each type of observation. In the case of surface observations, this may be as simple as nearest neighbor or bilinear interpolation (the latter in our case). For other observation types, for example with MOPITT satellite data , this H function may be much more complicated (Emmons et al. 2004). The module also creates functions for determining which state variables are "close" given a certain distance from an observation and provides additional information about the model.

The second stage is coordinating DART's filter with the CCTM_CO ensembles. We developed a series of python programs using the ioapiTools [2] module that could run CCTM_CO (and CCTM for that matter) directly from python, translate the prior (the CMAQ CONC file) to DART format, then take the DART output and translate it back to CMAQ format.

The flow of data through the overall assimilation system is: (a) start the ensemble of CCTM_CO models from the perturbed initial conditions (see next section for details), (b) run the ensembles forward to some date; (c) take the last hour of the CONC file and produce the priors for DART; (d) DART applies the EAKF filter to the observations prepared for that period; (e) take the posterior from DART and translate the data into initial condition files for the next stage of the CCTM_CO run; repeat continuing from (b).

4 PRELIMINARY RESULTS

4.1 Perturbing the Ensembles

Developing the ensembles is a key component to the success or failure of the data assimilation. Our initial experiments have only modified the initial conditions (IC). We tested various schemes for

[2] See http://www-pcmdi.llnl.gov/software-portal/Members/azubrow/ioapiTools/index_html (Sept 2006) for more details.

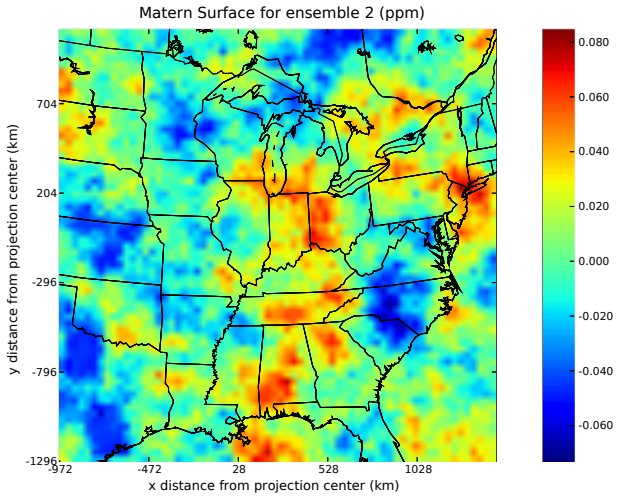

Figure 1: The simulated surface from the Matérn function, which is used to perturb the CO IC

perturbing the IC. We started with applying Gaussian noise to the CO IC. This had unintended consequences, namely that the CO values would jump dramatically. A large value of CO could advect through a region, creating non-physical, dramatic shifts in CO concentrations at particular points in the model.

A second approach was to apply a constant perturbation. In other words, multiply the whole CO field by a single value. This had the consequence of smoothing out the previous, excessive spatial variation. But, the constant perturbation was too smooth, in fact it created no variation in the spatial structure of the IC.

The third approach was to apply Gaussian noise, which included spatial structure. In our case, we have used a Matérn function (Handcock and Stein 1993) with a spatial range of 200 km and smoothness parameter of 0.5 (an exponential decay function) to create spatially correlated noise. This Matérn function can be used to simulate a series of surfaces that have mean 0 and a σ of our choosing. For example, Figure 1 is the surface for the second ensemble member. The resulting surfaces are then added to CO IC to create N IC files. The mean of all the N perturbed IC files is CO IC, the standard deviation is σ. Each ensemble has been perturbed from CO IC, but the perturbation is not excessively "patchy"; it retains some spatial correlation.

A related issue to perturbing the ensembles is how to maintain the ensemble spread. Our ensembles collapse over time due to diffusion and identical emissions and meteorology. If the ensembles collapse, then data does not impact it during the filtering stage. In other words, a collapsed ensemble indicates that the prior is very certain (all the ensemble members have the same value), therefore the prior and posterior become more and more identical.

Data assimilation tends to accelerate this collapse. There are two main methods for counteracting this collapse: (a) Prior inflation, take the covariance between the ensemble members before the assimilation (the prior) and multiply it by a constant. (b) Add perturbation to another part of the system. For example, we have increased the ensemble spread by perturbing the OH in the OH_CONC file. Future experiments should investigate the perturbation of the emissions and the meteorology data.

4.2 Cutoff

A second factor that greatly impacted the data assimilation was the cutoff. The cutoff determines the spatial distance around a particular observation that should be considered for data assimilation. It defines a smooth function that decreases the correlation between an observation and the state variable as the distance increases. In our assimilations, the cutoff value determines the half-width of the Gaspari-Cohn 5th order polynomial (Gaspari and Cohn 1999).

In the horizontal, the cutoff may dramatically change the assimilation results. We found that if the cutoff was too large, then observations could impact state variables that were spuriously correlated but spatially remote. The resulting assimilation tended to smooth out the expected spatial variability in each ensemble member. If the cutoff was too small, only state variables very near the observations would be updated. The result was a "bullseye" pattern of regions just around the observation having potentially dramatic changes while the rest of the domain was unchanged.

In the vertical, we experimented with multiple schemes for limiting the impact of a surface observation on the state variables. Initially, we limited the update to the first vertical level of the state variable. Because of diffusion, the updated CO value quickly returned to it's original non-assimilated position, often within a model hour.

A second approach was to modify the DART module to consider both the cutoff and the model PBL height. If the state variable was within the cutoff function distance and below the PBL height, it would be potentially updated. If it was outside

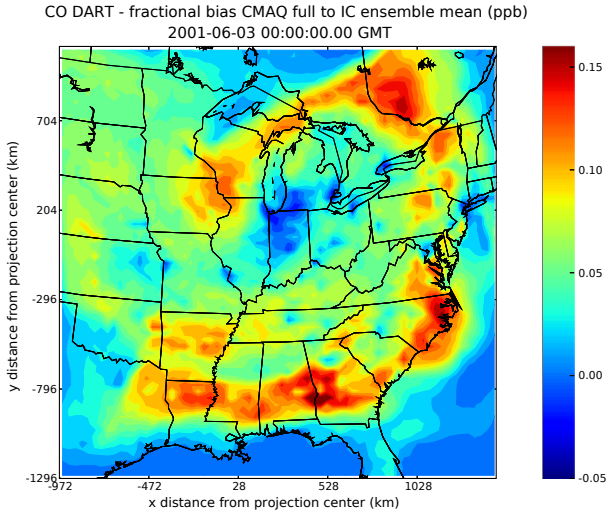

Figure 2: Fractional bias between the mean of the ensembles without data assimilation and the original full CMAQ run

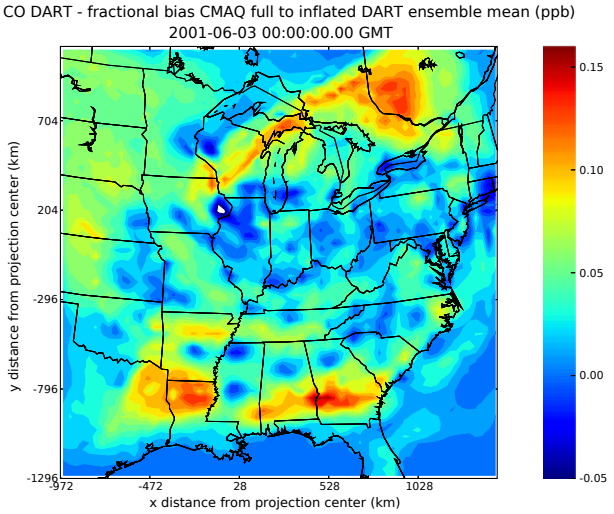

Figure 3: Fractional bias between the mean of the ensembles after data assimilation and the original full CMAQ run

of the cutoff function or above the PBL, it would not be updated. This improved our results, but created unintended consequences. For reasonable cutoffs, the PBL test would sometimes create a strong vertical gradient. In these cases the posterior values below PBL height would be significantly changed by the update, while the values above the PBL height would remain the same.

Our present solution is to disregard the PBL height and allow all levels to be considered for update within the cutoff function distance. This smoothes the vertical gradient, while not discarding state variables that may be correlated despite being above the PBL. The determination of the ideal horizontal and vertical cutoff is an area for further research.

4.3 Synthetic Results

In the synthetic experiment, synthetic observations were drawn from the full CMAQ run and assigned a σ of 2.5 ppbv. The observations were choosen to match the physical monitor locations. The ensembles were run from the perturbed IC. At first, they were run without any data assimilation. A second run used the EAKF filter to assimilate the synthetic observations every three hours. The mean of the ensembles were compared against the full CMAQ run using fractional bias:

$$b = \frac{\mathrm{CO}_{full} - \mathrm{CO}_{mean}}{\left[\frac{\mathrm{CO}_{full} + \mathrm{CO}_{mean}}{2}\right]} \quad (3)$$

where b is the fractional bias, CO_{full} is CO from the full CMAQ run, and CO_{mean} is the mean of the CCTM_CO ensembles. In comparing Figure 2, no data assimilation, to Figure 3, data assimilation, it is clear that the EAKF filter drew the mean of the ensembles closer to the "truth" (in this case the full CMAQ run). The ensembles were adjusted to more closely match the "truth" not only at the monitor locations, but over large sections of the domain.

Future work will include experimenting with different perturbation schemes, cutoff lengths, and the assimilation of MOPITT data.

References

Anderson, J. L. (2001). An ensemble adjustment kalman filter for data assimilation. *Monthly Weather Review 129*, 2884–2903.

Emmons, L., M. Deeter, J. Gille, D. Edwards, J. Attie, J. Warner, D. Ziskin, G. Francis, B. Khattatov, V. Yudin, J. Lamarque, S. Ho, D. Mao, J. Chen, J. Drummond, P. Novelli, G. Sachse, M. Coffey, J. Hannigan, C. Gerbig, S. Kawakami, Y. Kondo, N. Takegawa,

H. Schlager, J. Baehr, and H. Ziereis (2004). Validation of measurements of pollution in the troposphere (MOPITT) CO retrievals with aircraft in situ profiles. *Journal of Geophysical Research-Atmospheres*.

Gaspari, G. and S. E. Cohn (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society 125*, 723–757.

Handcock, M. S. and M. L. Stein (1993). A bayesian analysis of kriging. *Technometrics 35*, 403–410.

Im, H. K., M. L. Stein, and V. R. Kotamarthi (2005). A new approach to scenario analysis using simplified chemical transport models. *Journal of Geophysical Research 110*.