J1.1 ASSIMILATING CONCENTRATION DATA INTO DISPERSION MODELS WITH A GENETIC ALGORITHM

Sue Ellen Haupt
Kerrie J. Long
Anke Beyer-Lout
George S. Young
The Pennsylvania State University, University Park, Pennsylvania

1. INTRODUCTION

In the event of a hazardous release of a toxic contaminant, atmospheric transport and dispersion (AT&D) models are critical to quickly and accurately predicting hazard areas that might be affected. AT&D models require accurate source term information as well as meteorological input data such as wind speed and direction. While meteorological data is routinely available from land surface stations, it is likely to be sparse, may not be representative of specific locale or current time, and can also contain errors. Such inaccuracies can produce a large impact on the accuracy of concentration predictions. Data assimilation provides a strategy to make the most effective use of the available data in an efficient manner, thus providing emergency managers with timely, accurate information.

Figure 1 illustrates the problem. In the left pane, a toxic contaminant event has occurred. The wind direction is uncertain; therefore, the hazard area (light grey) initially predicted is relatively wide. However, the concentration measurements from the surrounding sensors provide information to narrow the prediction. Note that the null sensor data provides useful information to fine-tune the hazard area. Not only has the concentration prediction been narrowed, conserving sparse resources, but the sensor information has refined the estimate of the most probable wind direction, so that the next time's prediction is more accurate (right pane). As that time is realized, further sensor data again helps to fine-tune the direction and the extent of the hazard region.

For dispersion problems, forecasting the transport and dispersion of a contaminant requires knowledge of a coupled system: the wind field providing the transport and a concentration equation predicting the dispersion. For this problem the coupling is one way: the wind field drives the concentration field but the concentration field has no direct impact on the resulting wind field. Both models are described in the following sections.

The goal of this paper is to demonstrate a technique to assimilate sensor data to improve the concentration predictions. This method uses a genetic algorithm (GA) to identify the characteristics of the dispersion realization that is occurring (including wind data), then to back-calculate the best modeling parameters so that the predicted concentration field best matches the observed concentrations. It is akin to the variational methods of data assimilation because the method optimizes a minimization problem. Thus, we refer to this method as GA-Var.

*Corresponding author address: Sue Ellen Haupt, Applied Research Laboratory, P.O. Box 30, The Pennsylvania State University, State College, PA, 16804-0030; e-mail: seh19@psu.edu

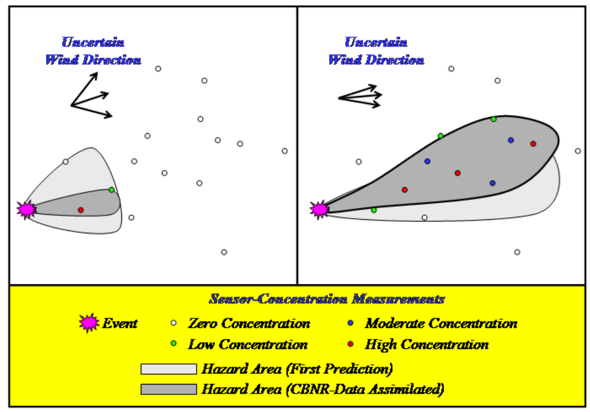

Figure 1. Schematic indicating the usefulness of assimilating sensor data to refine hazard area predictions.

2. ASSIMILATION FOR DISPERSION

2.1 Mathematical Formalism

The assimilation problem can be posed as a dynamical system as:

$$\frac{\partial \mathbf{x}}{\partial t} = \mathbf{M}\mathbf{x} + \boldsymbol{\eta} \qquad (1)$$

where **x** represents the predictands (concentration and wind), $\frac{\partial \mathbf{x}}{\partial t}$ is their time rate of change or tendency, **M** is a linearized operator based on the potentially nonlinear dynamics, and $\boldsymbol{\eta}$ is a stochastic noise term incorporating the errors in the model and the unresolved subgrid processes.

The assimilation process can be expressed as adding a term to the system that acts to force the system toward the observations:

$$\frac{\partial \mathbf{x}}{\partial t} = \mathbf{M}\mathbf{x} + \boldsymbol{\eta} + G(\mathbf{x}^{\mathbf{0}}, \mathbf{x}^{\mathbf{f}}) \qquad (2)$$

where $\mathbf{x}^{\mathbf{0}}$ is the observed value of the field and $\mathbf{x}^{\mathbf{f}}$ is the forecast field. Here *G* is the adjustment function.

Forecasting the transport and dispersion of a contaminant requires a set of coupled systems: a wind equation and a concentration equation. This system is coupled one-way: the transport equations influence the dispersion equation but the concentration equation has no direct impact on meteorological transport.

Equation (2) can be written for the transport and dispersion problem by separating it into one equation for the wind field and another for the concentration equation:

$$\frac{\partial \vec{\mathbf{v}}}{\partial t} = \mathbf{M}_{\mathrm{v}}(\vec{\mathbf{v}})\vec{\mathbf{v}} + \boldsymbol{\eta}_{\mathrm{v}} + G_v(\vec{\mathbf{v}}^{\mathbf{o}}, \vec{\mathbf{v}}^{\mathbf{f}}, C^o, C^f) \tag{3}$$

$$\frac{\partial C}{\partial t} = \mathbf{M}_{\mathrm{C}}(\vec{\mathbf{v}})C + \boldsymbol{\eta}_{\mathrm{C}} + G_C(\vec{\mathbf{v}}^{\mathbf{o}}, \vec{\mathbf{v}}^{\mathbf{f}}, C^o, C^f) \tag{4}$$

where $\vec{\mathbf{v}}$ denotes a continuous two- or three-dimensional wind field, and C is the two- or three-dimensional concentration field (a single nonreactive species is assumed). Subscripts v and C on the two dynamics operators, $\mathbf{M}_{\mathrm{v}}$ and $\mathbf{M}_C$, denote separation into a wind operator and a concentration operator. Both adjustment functions depend on the meteorological data forecasts and observations and also on the concentration forecasts and observations. The wind equation (3) depends on the previous state of the wind field and the concentration equation (4) depends on the previous state of the concentration field. The indirect impact is through the adjustment function $G_C(\vec{\mathbf{v}}^{\mathbf{o}}, \vec{\mathbf{v}}^{\mathbf{f}}, C^o, C^f)$. Both adjustment functions in (3) and (4) depend on the forecast (superscript *f*) and the observations (superscript *o*) of both the wind and the concentration fields. The wind field is altered according to the innovation between the observed and forecast wind fields, plus includes information on the innovation of the concentration field from that forecast. Similarly, the concentration adjustments depend on wind field innovations as well as concentration innovations. This system is dynamic, coupled, and nonlinear. Thus the coupled systems interact through their adjustment functions.

2.2 Assimilation with a Genetic Algorithm

Using the GA-Var assimilation method, we seek to directly optimize the wind direction to produce a predicted concentration field closest to that monitored. This method is most akin to the variational approaches to assimilation (for example see Sasaki 1970 or Kalnay 2003), but rather than using the variational formalism, it directly optimizes the match. We choose the continuous genetic algorithm (GA) as our optimization tool. The GA is an artificial intelligence optimization method inspired by the biological processes of genetic recombination and evolution. It begins with a population of potential solutions and evolves them closer to the correct solution through the implementation of mating and mutation operators. The GA in general, and the continuous parameter version used here, is described in detail by Haupt and Haupt (2004).

Figure 2 demonstrates the concentration assimilation procedure using a GA. An AT&D model makes an initial prediction of the concentration. This prediction is compared to an observed concentration field. The difference is the innovation vector that can be used to evolve a better guess for the modeling variables. The process to improve those variables used here is the GA-Var method, which uses a genetic algorithm to directly optimize the modeling variables. The genetic algorithm works by initially “guessing” model variables in the prescribed range, applying the operations of mating and mutation to produce new variables that combine the information from the previous generation while continuing to generate new “guesses”, then uses a cost function to determine whether the new variables have improved the solution. The process iterates through a number of generations until converging on an improved solution. The updated modeling variables then provide a better prediction for the next time. The process iterates dynamically as new data becomes available for the assimilation. The genetic algorithm has been applied to assimilation-type problems in various publications (Haupt 2005; Haupt et al. 2006, 2007, 2008, 2009 a,b; Allen et al. 2007a,b; Long et al. 2008; Rodriguez et al. 2008).

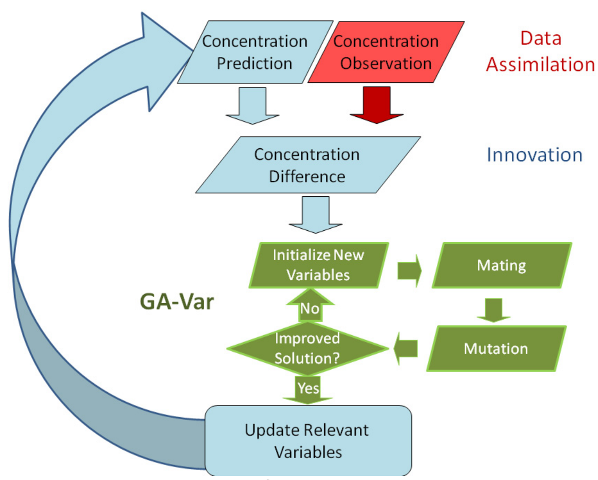

Figure 2. Flowchart of concentration assimilation using the GA-Var technique.

3. MEANDERING PLUME EXAMPLE

3.1 Problem Description

Gaussian dispersion in a meandering wind field is used as a first testbed for our assimilation methods here. Such a configuration is simple, it varies smoothly in time and space, and it represents an important realizable state of the atmosphere. Meandering wind conditions are particularly common during nocturnal stable boundary layer conditions (Hanna 1983, Mahrt 1999, among others).

We concentrate on an instantaneous release of contaminant in a neutrally buoyant atmosphere, which can be modeled with a Gaussian puff equation:

$$C_r = \frac{Q\Delta t}{(2\pi)^{1.5}\sigma_x\sigma_y\sigma_z}\exp\left(\frac{-(x_r - Ut)^2}{2\sigma_x^2}\right)\exp\left(\frac{-y_r^2}{2\sigma_y^2}\right)\times \left[\exp\left(\frac{-(z_r - H_e)^2}{2\sigma_z^2}\right)+\exp\left(\frac{-(z_r + H_e)^2}{2\sigma_z^2}\right)\right] \qquad (5)$$

where C_r is the concentration at receptor r, (x_r, y_r, z_r) are the Cartesian coordinates downwind of the puff, Q is the emission rate, Δt is the length of time of the release itself, t is the elapsed time since the release, U is the wind speed, H_e is the effective height of the puff centerline, and $(\sigma_x, \sigma_y, \sigma_z)$ are the standard deviations in the concentration distribution in the x-, y-, and z-directions, respectively. The standard deviations of the model are computed according to Beychok (1994):

$$\sigma = \exp\left[I + J\ln(x) + K(\ln(x))^2\right] \qquad (6)$$

where: x is the downwind distance in km, and I, J, and K are coefficients dependent on Pasquill stability class for both σ_y, and σ_z and can are tabulated in Beychok (1994). We use $\sigma_x = \sigma_y$ here.

For our test scenario, the puff transport and dispersion occurs in a sinusoidally varying wind field with a constant wind speed of 5 m s^{-1} and direction, θ, defined as:

$$\theta = \theta_0 \sin(2\pi\omega t) \qquad (7)$$

where θ_0 is the maximum amplitude (set at 20°), ω is the oscillation frequency (set at 1/600 s^{-1}), and t is the time variable, measured from the time of release (for a total time period of 1000 s), consistent with its use in (5).

For an instantaneous release, it is equivalent to view this sinusoidal wind variation as either varying in time (the entire field with a single wind that varies in time) or in space (meandering wind field, as would be the case where local terrain, gravity waves, or inactive turbulence affect the flow). The goal is to assimilate only the concentration data to adjust the wind field by defining the adjustment function in (3), $G_v(\vec{\mathbf{v}}^{\mathbf{f}}, C^o, C^f)$, and thus improving on the concentration forecast.

The domain of the identical twin experiment is 5 km × 5 km with grid points sited every 125 m, yielding a full spatial resolution of 41 x 41 grid points. The synthetic data are produced with an integration time interval of 20 s for a total integration time of 1000 s, the time required for the puff to traverse the domain. We assume that the source characteristics are known and seek to compute the time evolving wind direction. Figure 3a indicates the domain and contours the "true" puff concentration calculated every 20 s but plotted only every 100 s.

The GA-Var approach to wind assimilation is applied dynamically. We assume that we have already modeled the past dispersion history and seek to assimilate the most recent concentration observations into the transport and dispersion model. This is a two step process: 1) we use the current concentration measurements to compute the optimal wind direction and 2) we use that wind direction to forecast the location of the plume at the next observation time.

3.1 Meandering Plume Results

Puff concentration values predicted with the GA-Var technique appear in Figure 3b. The concentration values predicted using this field-based technique is nearly indistinguishable from the original data (Fig. 3a).

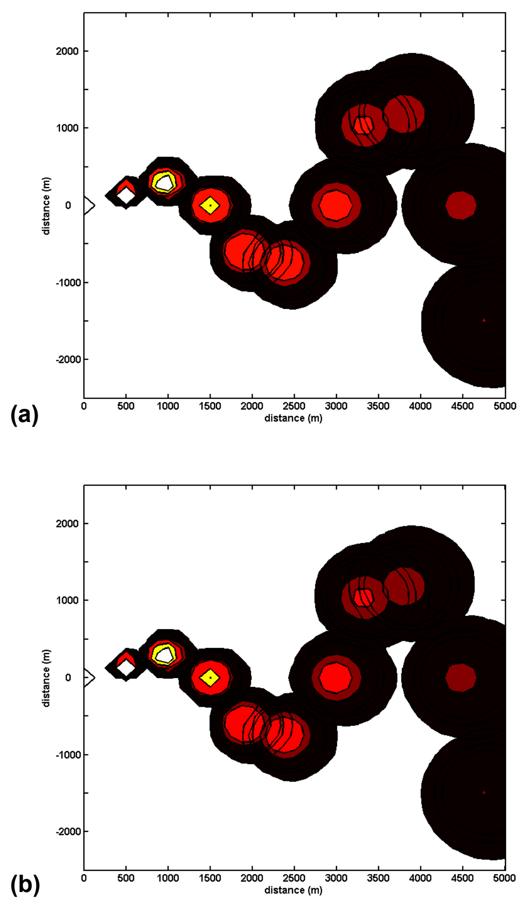

Figure 3. Series of concentration puff locations in meandering wind field calculated with a time step of 20 s, plotted every 100 s for visualization for a total time of 1000 s. a) original data created for comparison and b) puffs assimilated with the GA-Var technique.

Figure 4a shows the true (solid black line) and GA-Var computed wind direction (blue dashed line) at each time step for the same assimilation setup. The GA-Var approach produces smooth time trajectories because the wind direction is computed directly from all observations. The solution is ill-posed at the first sample time because there is not yet sufficient puff spread to reach the sensor grid; hence, there is a large wind direction error at this time. After the first time step, the GA curve is indistinguishable from the truth curve. Figure 4b shows the locations of the puff centroid plotted as a trajectory in time. The GA curve is smooth and follows the truth quite accurately. In addition, the GA method does not rely on identifying features, but instead compares entire fields of concentration data.

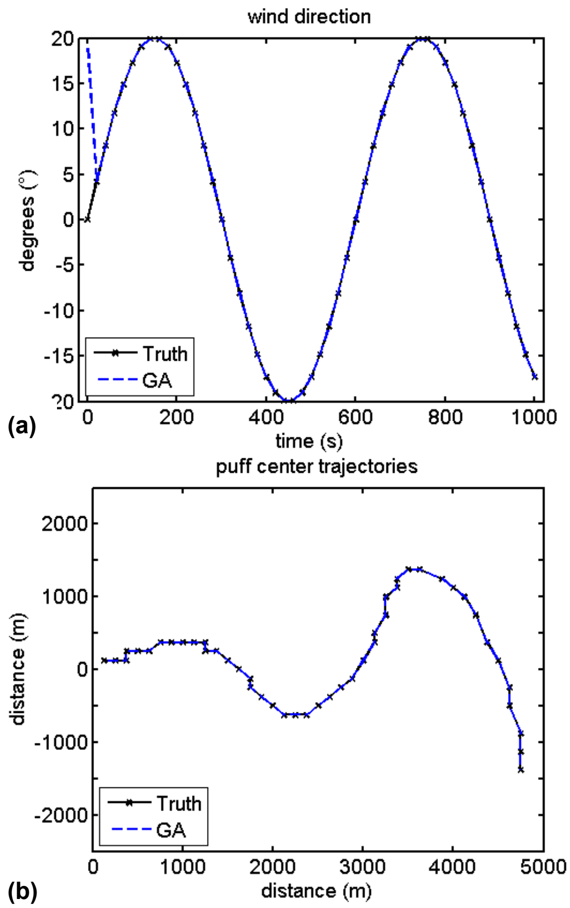

Figure 4. a) Wind direction as a function of time for full spatial and temporal resolution for the truth (solid black) and as computed by the GA-Var technique (dashed blue) and b) location of the puff centroid plotted as a trajectory in time for full spatial and temporal resolution comparing the three (same markings).

The results above have demonstrated success at using concentration observations to infer a meteorological variable, in this case wind direction, to assimilate into the wind equation (3). That assimilated variable then forces the prognostic concentration equation (4). Haupt et al. (2009a) also addresses issues of how many concentration observations are necessary to correctly infer the wind direction in terms of both spatial and temporal data denial. Figure 5 contours the root mean squared errors for the puff centroid location depending on the observational grid and the model time step used for the GA–Var technique. It is interesting to note that the centroid location RMSE increases steadily as fewer observations in space are incorporated into the model, yet shows less sensitivity to temporal data denial since computations at each time step are independent.

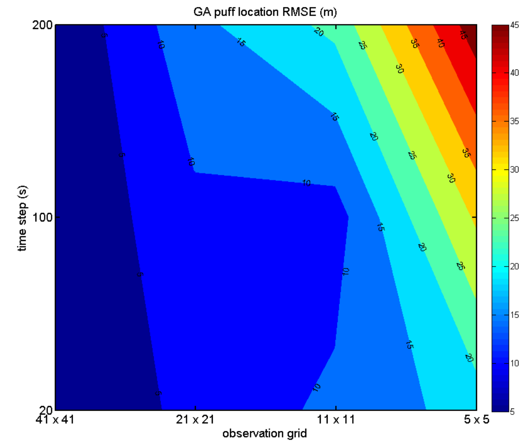

Figure 5. Puff centroid location root mean squared errors contoured dependent on the model time step and the observational grid for the GA-Var technique.

These results show promise for predicting the transport and dispersion of a contaminant in a time varying wind field. They also provide hope for being able to model elements of a specific realization of concentration dispersion if sufficient concentration observations are available. The data make it possible to determine the realization, allowing a more accurate prediction in spite of the one-way coupling of the wind and concentration models.

4. TUSSEYPUFF EXAMPLE

4.1 The TusseyPUFF System

The second example tests the GA-Var technique in a more complex Shallow Water wind field coupled to a basic Gaussian puff dispersion model – the TusseyPuff system. TusseyPuff's Gaussian puff model (see (5) above) requires a minimum of input information, is easily implemented, and treats the release as a computationally efficient single entity. To drive the Gaussian puff model a reduced gravity two dimensional Shallow Water model (Holton 1992) is chosen. TusseyPuff couples these two model elements and provides a basic but comprehensive test environment and is described in detail in Beyer-Lout (2007). The Shallow Water equations are:

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=-g'\frac{\partial h}{\partial x}-\frac{1}{\rho h}\tau_{Bx} \qquad (8)$$

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}=-g'\frac{\partial h}{\partial y}-\frac{1}{\rho h}\tau_{By} \qquad (9)$$

$$\frac{\partial h}{\partial t}+\frac{\partial}{\partial x}(uh)+\frac{\partial}{\partial y}(vh)=0 \quad (10)$$

where u is the zonal velocity component, v is the meridional velocity component, g' is the reduced gravitational acceleration, $g' = g\Delta T / T$, where *g* is the gravitational acceleration, *T* is the temperature of the fluid and *ΔT* is the temperature discontinuity above the fluid layer. Furthermore, *ρ* denotes the density and *h* is the depth of the fluid layer being modeled. The stress at the bottom of the layer τ_B is parameterized as:

$$\tau_{Bx} = \rho c_D u\sqrt{u^2+v^2}$$
$$\tau_{By} = \rho c_D v\sqrt{u^2+v^2} \quad (11)$$

where c_D is the drag coefficient, taken here to be 0.02. The model domain spans 9 km x 9 km, with 30 x 30 grid points, yielding a grid spacing of 300 m in both the x- and the y-direction. The mean fluid depth of the layer is chosen to be 500 m, the temperature of the layer *T* is 300 K and the temperature discontinuity above the layer Δ*T* is chosen to be 12 K, resulting in a gravity wave speed $U = \sqrt{g'h}$ of approximately 14 ms^{-1}. Periodic boundary conditions are imposed in both the x and y directions.

4.2 Test Case Set-Up

Cone-like topography is added to the Shallow Water model in order to create a more realistic wind field. The topography is centered in the domain and reaches a height of 400 m and extends approximately 4 km in the horizontal. In addition the model is initialized with a small fluid height perturbation. For more information about the specific shallow water model used here see Beyer-Lout 2007.

The same coupled model is used to generate observations for use in the data assimilation tests. This identical twin approach is advantageous for a test environment because it allows us to compare the performance of the different data assimilation schemes with a known 'truth'. The 'truth' is created by first initializing TusseyPuff with a uniform wind field and a small fluid-height perturbation (see Figure 6). After a startup time during which the model is run forward to achieve dynamically consistent velocity and height fields, the contaminant puff is released. The release location was chosen such that the puff trajectory passes through the terrain driven flow (Figure 7). The resulting data fields are then used to derive the observations needed for the data assimilation tests. The observation stations are located at model grid points, for simplicity, so no interpolation is necessary. The sensors record wind velocity and direction every ten model time steps.

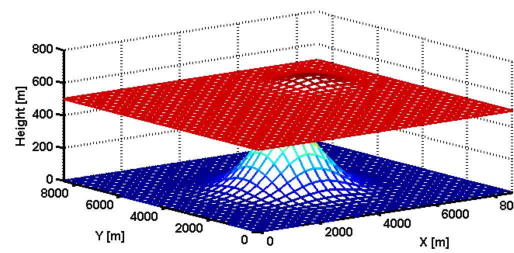

Figure 6. Topography (blue) and initial fluid height (red) of the two dimensional shallow water model.

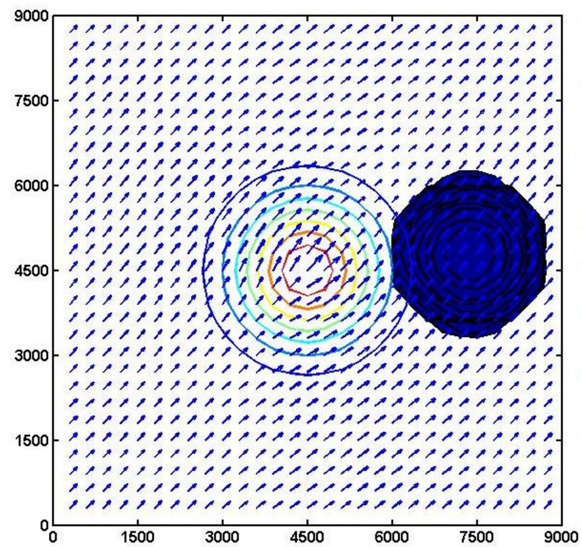

Figure 7. Wind field with topography and concentration puff in blue

4.3 The Assimilation Process

This study does not assimilate the continuous concentration fields, but rather the four puff parameters, the puff center $\bar{x}$ and $\bar{y}$, the puff spread *σ*, and the puff concentration amplitude, *q*. These puff parameters are derived from the continuous concentration field by fitting a Gaussian distribution to the observations. While computationally more efficient than assimilating the continuous concentration observations, this approach adds uncertainty when the puff is still small and only a few sensors detect concentration. The model is initially spun-up to assure that the transients have dissipated. Then a dynamic assimilation process is initiated for each time.

4.4 Results for TusseyPUFF

Model results carried out with the TusseyPuff model in a neutral atmosphere are generally good. Figure 8 indicates the root mean squared error in three source variables as a function of time. (Note the x and y puff parameters are collapsed into a single distance metric called location.) For all three, there is some initial oscillation as the puff grows large enough to influence the computations. The magnitude of the location error stabilizes to very low values most quickly. It is interesting to note the spike at a time of about 1500 s: that is the point in time when the puff first encounters the topographic bump. Thereafter, the RMSE is quite low, indicating the success of the GA-Var procedure.

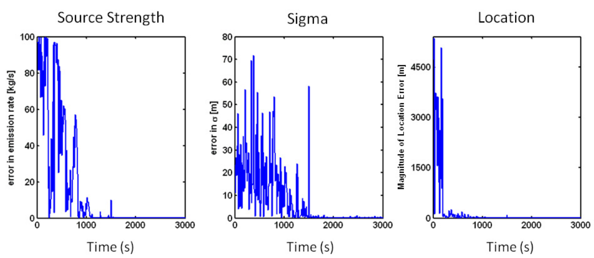

Figure 8. Results of estimating source strength, spread, and location for TusseyPUFF.

5. CONCLUSIONS

The GA-Var technique is shown to be effective for assimilating concentration data into a model to successfully recover information necessary to predict future atmospheric transport and dispersion despite the issue of one-way coupling in the system. Not only does the genetic algorithm successfully identify the realization, but it also computes modeling variables necessary to provide a better prediction. By assimilating observation data we are able to more closely predict a specific realization of a dispersion event, thus providing emergency responders with more accurate predictions.

Acknowledgements: This research is supported by the Defense Threat Reduction Agency (contract numbers W911NF-06-C-0162 and DTRA01-03-D-0010-0012). We particularly thank John Hannan and Christopher Kiley for valuable feedback throughout this project. Joel Peltier drafted Figure 1. McKenzie McNeal assisted with some coding.

References

Allen, C.T., S.E. Haupt, and G.S. Young, 2007a: Source Characterization with a Receptor/Dispersion Model Coupled With A Genetic Algorithm, *Journal of Applied Meteorology and Climatology,* **46**, 273-287.

Allen, C.T., G.S. Young, and S.E. Haupt, 2007b: Improving Pollutant Source Characterization by Optimizing Meteorological Data with a Genetic Algorithm, *Atmospheric Environment,* **41**, 2283-2289.

Beyer-Lout, A, 2007: *Concentration assimilation into wind field models for dispersion modeling.* Master's Thesis, The Pennsylvania State University, University Park, PA.

Beychok, M.R., 1995: *Fundamentals of stack gas dispersion*. 4rd Edition, California, USA, p. 50 – 56.

Hanna S.R., 1983. Lateral turbulence intensity and plume meandering during stable conditions. Journal of Climate and Applied Meteorology. 22, 1424-1430.

Haupt R. L., & Haupt S. E., 2004. Practical Genetic Algorithms, 2nd edition with CD. John Wiley & Sons, New York, NY.

Haupt, S.E., 2005: A Demonstration of Coupled Receptor/Dispersion Modeling with a Genetic Algorithm, *Atmospheric Environment*, **39**, 7181-7189.

Haupt, S.E. G.S. Young, and C.T. Allen, 2006: Validation Of A Receptor/Dispersion Model Coupled With A Genetic Algorithm, *Journal of Applied Meteorology,* **45**, 476–490.

Haupt, S.E., G.S. Young, and C.T. Allen, 2007: A Genetic Algorithm Method to Assimilate Sensor Data for a Toxic Contaminant Release, *Journal of Computers*, **2**, 85-93.

Haupt, S.E., R.L. Haupt, and G.S. Young, 2008: A Mixed Integer Genetic Algorithm used in Chem-Bio Defense Applications, accepted by *Journal of Soft Computing*.

Haupt, S.E., A. Beyer-Lout, K.J. Long, and G.S. Young, 2009a: Assimilating Concentration Observations for Transport and Dispersion Modeling in a Meandering Wind Field, *Atmospheric Environment,* in press.

Haupt, S.E., C.T. Allen, and G.S. Young, 2009b: Addressing Air Quality Problems with Genetic Algorithms, in *Artificial Intelligence Methods in the Environmental Sciences*, Haupt, S.E., A. Pasini, and C. Marzban, eds., Springer, pp. 269-296.

Hoke, J.E., and R.A. Anthes, 1976: The Initialization of Numerical Models by a Dynamic Initialization Technique. *Mon. Wea. Rev.*, **104**, 1551-1556.

Holton, J.R. 1992: *An introduction to dynamic meteorology*. Volume 48 of International Geophysics Series. 3rd edition, Academic Press.

Kalnay, Eugenia, 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge, 136-204.

Long, K.J., S.E. Haupt, and G.S. Young, 2008: Optimizing Source Characteristics and Meteorological Data Describing a Contaminant Release with a Genetic Algorithm. Submitted to *Optimization and Engineering.*

Mahrt L., 1999. Stratified atmospheric boundary layers. Boundary Layer Meteorology 90, 375-396.

Rodriguez, L.M., S.E. Haupt, and G.S. Young, 2008: Impact of Sensor Characteristics on Source Characterization for Dispersion Modeling, submitted to *Measurement*.

Sasaki, Y., 1970: Some basic formalisms in numerical variational analysis. *Mon. Wea. Rev.*, **98**, 875-883.