CHARACTERIZING CONTAMINANT SOURCE AND METEOROLOGICAL FORCING USING DATA ASSIMILATION WITH A GENETIC ALGORITHM

Kerrie J. Long*, Sue Ellen Haupt, George S. Young, and Chris T. Allen
The Pennsylvania State University, Department of Meteorology, University Park, Pennsylvania

1. INTRODUCTION

The release of harmful contaminants is a potentially devastating threat to homeland security. Accurate identification of the source strength and location is essential to minimize the impact. Insufficient spatial and temporal resolution as well as inherent uncertainty in wind field data makes characterizing the source and predicting subsequent transport and dispersion extremely difficult. The solution requires a robust technique such as a genetic algorithm (GA) in order to precisely characterize the source and obtain the required wind information. The method uses a GA to find the combination of source location, source strength, and surface wind direction that best matches the monitored receptor data with the forecast pollutant dispersion model output. The approach is validated with an identical twin experiment that generates the observation data using the same model embedded in the solution method (Daley 1991). Such an experiment allows the validation of the solution methodology in a controlled environment.

Researchers have characterized the source of a release using a Bayesian probability density algorithm to determine the mass of material released to within an order of magnitude as well as the location, type, and time of release (Robins, et al. 2005a). Robins, et al. (2005b) use a probabilistic dispersion model to better simulate the small scale effects of a meandering plume. In the current study, we use the Gaussian puff equation as the dispersion model and add wind direction to the list of parameters sought.

2. EXPERIMENTAL DESIGN

A release of a toxic contaminant is more likely to be an instantaneous release rather than a continuous emission. Therefore, the Gaussian puff equation is used as the forecast pollutant dispersion model rather than the Gaussian plume equation as was used in previous studies (Allen, et Al 2006). The Gaussian puff model is defined as:

$$C_r = \frac{Q\Delta t}{(2\pi)^{1.5}\sigma_x\sigma_y\sigma_z}\exp\left(\frac{-(x_r-Ut)^2}{2\sigma_x^{\,2}}\right)\exp\left(\frac{-y_r^{\,2}}{2\sigma_y^{\,2}}\right)\left[\exp\left(\frac{-(z_r-H_e)^2}{2\sigma_z^{\,2}}\right)+\exp\left(\frac{-(z_r+H_e)^2}{2\sigma_x^{\,2}}\right)\right] \quad (1)$$

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 time since the release, U is the wind speed, H_e is the height of the puff centerline, and $(\sigma_x, \sigma_y, \sigma_z)$ are the standard deviations of the concentration distribution in the x-, y-, and z-directions, respectively. As the puff traverses a regular gridded domain, the concentrations monitored at many receptors will be extremely small or zero (Figure 1). This analysis determines the receptor density necessary to obtain an accurate solution. In order to test the method's robustness, the receptor data is generated by an identical twin experiment.

*Corresponding author address: Kerrie J. Long, Department of Meteorology, The Pennsylvania State University, University Park, PA 16802; kjl203@psu.edu

2.1 Data Characteristics

The simulated source release is located at the center of a 16,000 m by 16,000 m equally spaced grid. The receptors are located at the intersection and corners of the grid (Figure 2). We examine two different wind directions, 225º and 180º. In the 225º case, the centerline of the puff falls directly over a number of receptors. In the 180º case, the centerline of the puff falls directly between the receptors. Currently, sensors can retrieve data up to once a second but the concentrations are likely to be highly correlated (Robins, et al. 2005a). The grid size and grid-spacing is constructed such that the puff remains on the domain for the specified time period. Based on a 16,000 m domain size and a wind speed transporting the puff at 5 ms^{-1}, the best intervals for data retrieval are 6, 12, 18, 24, and 30 minutes following the release. The spacing configurations investigated are listed in table 1. The height of release is 10 m above the surface. We assume neutral Pasquil stability D.

2.2 Procedures

The GA technique is based on initializing solutions using a random number generator. Therefore, each run will yield somewhat different results. We run each configuration of the model ten times in order to gain statistics on the model's performance. In order to keep the same size domain throughout the study, we quadrupled the number of receptors and halved the grid-spacing with each configuration. As a result, we are only examining the effects of increased resolution. The puff may not grow appreciably in the first time step and may be partially off the domain by the last time step, thus leaving an overwhelming number of zero data points at those times. For example, figure 2 illustrates that the puff does not pass over many of the receptors in the first and last time steps. Thus, for those times we may not have enough non-zero reporting sensors to obtain an accurate solution. Because of this, we expect the exclusion of certain time steps may aid in our analysis.

2.3 Modeling

A GA is used to determine the best combination of source strength, source location, and wind speed that matches the dispersion model output with the receptor data. A chromosome comprised of four parameters, wind direction, source strength, and source location (x,y) is fed into the GA as a vector. Each chromosome is initialized with a different random number for each of the four variables sought for a population size of 1200 chromosomes. After the initial population is generated, the fitness of each chromosome is evaluated based on the cost function below:

$$\text{cost function} = \frac{\sum_{t=1}^{5} \sqrt{\sum_{r=1}^{TR} \left(\log_{10}\left(aC_r + \varepsilon\right) - \log_{10}\left(aR_r + \varepsilon\right)\right)^2}}{\sum_{t=1}^{5} \sqrt{\sum_{r=1}^{TR} \left(\log_{10}\left(aR_r + \varepsilon\right)\right)^2}} \qquad (2)$$

where C_r is the concentration as predicted by the dispersion model given by (1), R_r is the receptor data value at receptor r, TR is the total number of receptors, a and ε are constants, and the cost function is summed over all five time steps. The value of a depends on the maximum values of C_r and R_r and is determined by dividing the sum of all the receptor data by one. The concentrations given by the puff equations are often very small. To avoid taking the logarithm of zero, ε is added to aC_r and aR_r quantities. The value of ε is based on the 10% largest concentrations of C_r and R_r. If ε approaches one, then it will dwarf the concentration values, aC_r and aR_r , thus rendering the cost function meaningless.

The population is then sorted by the cost function value. A lower cost function represents a combination of parameter values that most closely matches the truth. Haupt (2005) found that large populations coupled with smaller mutation rates or small populations coupled with high mutation rates allow the cost function to converge towards zero in the fewest number of function calls. Thus, for a population size of 1200 the best mutation rate was determined to be 0.015 (Allen et al. 2006). The GA used here is elitist and retains the best solution found while the remainder of the population is subject to mating and mutation. The GA uses a mating procedure in which all parameters of the original two chromosomes, or parents, are blended to create two new unique chromosomes, or children. In order to encourage a complete search of the solution space, a number of the population is subject to mutation. This portion of the population is further modified by the multiplication of a random number. After each iteration the cost function is recalculated for the new chromosomes and the population re-sorted. The best candidate solution found by the GA after 100 iterations is used as the first guess for a Nelder-Mead simplex algorithm (Nelder and Mead 1965). The GA is generally good at finding the correct solution basin and the Nelder-Mead simplex performs a local search to find the minimum of that basin.

Skill scores are also used to evaluate the closeness of each solution to the exact value. The skill score is evaluated based on a linear composite of three component equations, one each to quantify the accuracy of strength, source location, and wind direction as determined by the GA and/or Nelder-Mead simplex method. The minimum and most desirable skill score is zero; the maximum and least desirable skill score is three. If the found source location in either x or y is greater than 4000 m, the skill score is assigned a maximum value of one for that parameter. If the strength is found to be more than five times the actual value, then the skill score for the strength parameter is assigned a maximum value of one. Finally, if the wind direction is determined to be more than 180º off from the known value, then it is assigned a maximum value of one. A solution that exactly matches the known solution is assigned a skill score of zero. All other solutions are assigned a skill score based on the equations described in Allen (2006).

3. RESULTS

The model is run ten times for each of the four receptor configurations for both southerly and southwesterly wind directions, yielding the eight total configurations. Table 1 displays the averages for each of the set-ups both before and after the Nelder-Mead simplex optimization.

The GA solution alone is within 0.01º of the wind direction, 0.01 of the strength, and 1 m of the location for every configuration for the 180º wind direction case. Thus, the GA alone is sufficient for accurately identifying the source in all of the θ=180º configurations. This is not true, however, for the 225º wind direction case. The 4x4, θ=225º configuration yielded an average wind direction of 223.58º, a strength of 1.09, and a location of (82,-73) for a total skill score of 0.3768. As the resolution of the grid increases, the average solution for the 225º wind direction improves by approximately an order of magnitude. The 32x32, θ=225º configuration yields solutions as good as θ=180º.

Even for the 225º wind direction where accuracy was less than perfect for small grid sizes, the GA is very good at finding the deepest trough when multiple optima exist. Once in the correct basin, the GA solution is used to initialize the Nelder-Mead simplex method which then performs a local search to fine tune the minimum. Table 2 demonstrates that this hybrid GA with a traditional gradient descent method is extremely effective at optimizing the solution. In all eight configurations, the post-Nelder-Mead search finds the solution to within 0.01º of the wind direction, 0.01 of the strength, and 1 m of the location. The worst solution generated by the GA is the 4x4, θ=225º case discussed above and the Nelder-Mead search still finds the correct solution.

When the grid size is reduced to 2x2, the model fails (results not shown). With only four receptors, the 180º wind direction has only two non-zero data points. Lacking sufficient information, the GA is unable

to determine the solution to the degree of accuracy shown above. However, the Nelder-Mead simplex method is still able to optimize the wind direction to within 0.2º, the strength to within 11%, and the location to within 30m of the actual value. When the wind direction is 225º, only one receptor receives data. In this case, the average wind direction found is 12º off, the strength is 50% off, and the location is on the order of 1000 m off. Clearly a grid size of 2x2 receptors does not provide enough information to produce meaningful results and higher resolution configurations are necessary.

In every configuration the cost function is reported as zero in Table 1. A closer analysis reveals that prior to the application of the Nelder-Mead simplex method, the cost function becomes progressively smaller as the resolution of the configuration increases. For example, the 4x4, θ=225º configuration has an average cost function of 4.0e-4, whereas the 32x32, θ=225º configuration has an average cost function of 6.0e-6.

A second method for evaluating the accuracy of each solution is the skill score which measures how close the solution comes to the correct known solution. In Table 1, the skill score prior to applying the Nelder-Mead simplex method for all grid sizes when θ =180º is zero. Because every solution the GA found for these configurations was correct, θ=180º, strength=1.0, and location (0,0), it makes sense that the skill score would also be zero. The GA did not find perfect solutions for the θ=225º configurations and thus the skill scores start relatively high, 0.3768 for a 4x4 grid, but progressively improve towards 0.0004 for the 32x32 grid. In all cases when the best candidate chromosome is fed into the Nelder-Mead simplex algorithm, the model is able to find the optimal solution and this is corroborated by a zero skill score.

4. SUMMARY & DISCUSSION

The Gaussian puff equation was used as the dispersion model in conjunction with a genetic algorithm to best characterize a pollutant emission and the transport wind direction from receptor observations of the time-evolving concentration field. The approach was validated by using an identical twin experiment to create receptor data. The GA successfully identified the source location, strength, and wind direction for the higher resolution grid sizes at both a southerly and southwesterly wind direction. Using a Nelder-Mead simplex algorithm initialized with the best candidate chromosome generated by the GA, the model was able to correctly identify source location, strength, and wind direction in every case evaluated here.

One motivation for using the Gaussian puff equation is to be able to identify wind speed; something we were unable to do with the time-independent Gaussian plume equation (Allen, et al. 2006). When the Gaussian plume equation was used (Allen, et al. 2006) the source strength and wind speed are to first approximation inversely proportional. In contrast, with a puff for a given concentration field we have information as the puff evolves in time so that speed can be inferred independent. Thus, it should be possible to identify wind speed in addition to the wind direction found here, thereby providing the basic weather data required by a transport and dispersion model. Adding this additional parameter to the model may increase the data requirements beyond those documented in table 1. It will also increase CPU time.

The model used here assumes neutral stability atmospheric conditions and completely accurate sensors. In reality, sensors monitoring pollution are fraught with error and uncertainty. In order to simulate the inherent uncertainty associated with the data collection process and the turbulent nature of the atmosphere, we will add noise to the model's input data. We will incorporate both additive and multiplicative noise as was done in Allen, et al. (2006). As part of this study, we will determine where the model fails and how much information is necessary to obtain good solutions.

The model works extremely well when given sufficient information. For this study we had receptor data for five different time steps which amounts to 5120 pieces of data for the 32x32 grid. Clearly, this configuration provides enough information to yield good results. In real world applications it may not be feasible to have sensors in place to monitor pollutant every six minutes. We will consider what happens to the model if we feed in only the last two or three sets of data. Using the last several sets of information allows the puff time to expand, and in doing so, more receptors are able to pick up information. Three sets of data may be enough information to yield an accurate solution while saving CPU time. In future studies, we will examine how much information is necessary to produce accurate solutions in a realistic scenario, particularly when observational noise is introduced.

ACKNOWLEDGEMENTS

This work was supported by DTRA under grant number W911NF-06-C-0162 and Penn State's Applied Research Laboratory as internal research and development. We would like to extend a special thanks to Anke Beyer for frequent discussions on this project.

REFERENCES

Allen, C. T., G. S. Young, and S. E. Haupt, 2006: Improving Pollutant Source Characterization by Optimizing Meteorological Data with a Genetic Algorithm, Submitted to Atmos. Environ.

Allen, C. T., 2006: Source Characterization and Meteorological Data Optimization with a Genetic Algorithm-Coupled Dispersion/Backward Model. M.S. Thesis, Dept. of Meteorology, The Pennsylvania State University, 79 pp.

Daley, R., 1991: Atmospheric Data Analysis. Cambridge University Press, New York, NY, 457 pp.

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

Nelder, J. A. and R. Mead, 1965: A Simplex Method for Function Minimization. *Computer Journal*, 7, 308-313.

Robins, P., V. Rapley, and P. Thomas, 2005a: Non-Linear Bayesian CBRN Source Term Estimation, Dstl, Salisbury, UK.

Robins, P., V. Rapley, and P. Thomas, 2005b: A Probabilistic Chemical Sensor Model for Data Fusion, Dstl, Salisbury, UK.

FIGURES

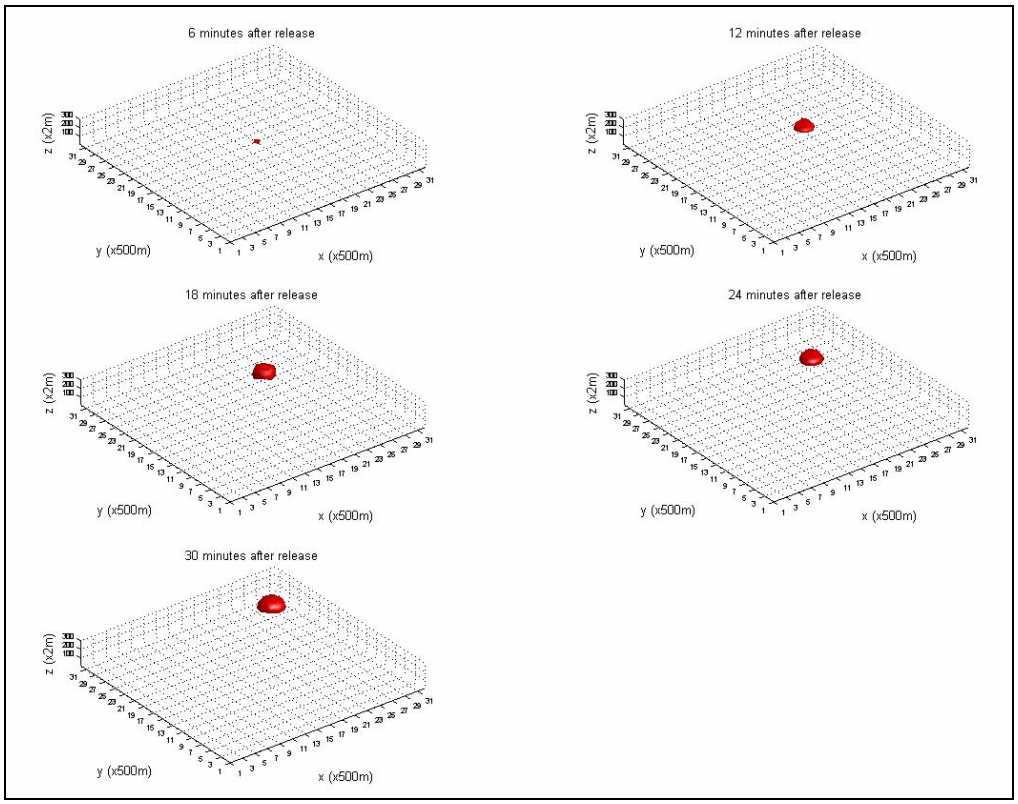

Figure 1. The evolution of the Gaussian puff over a 32x32 grid with a wind direction of θ=225º.

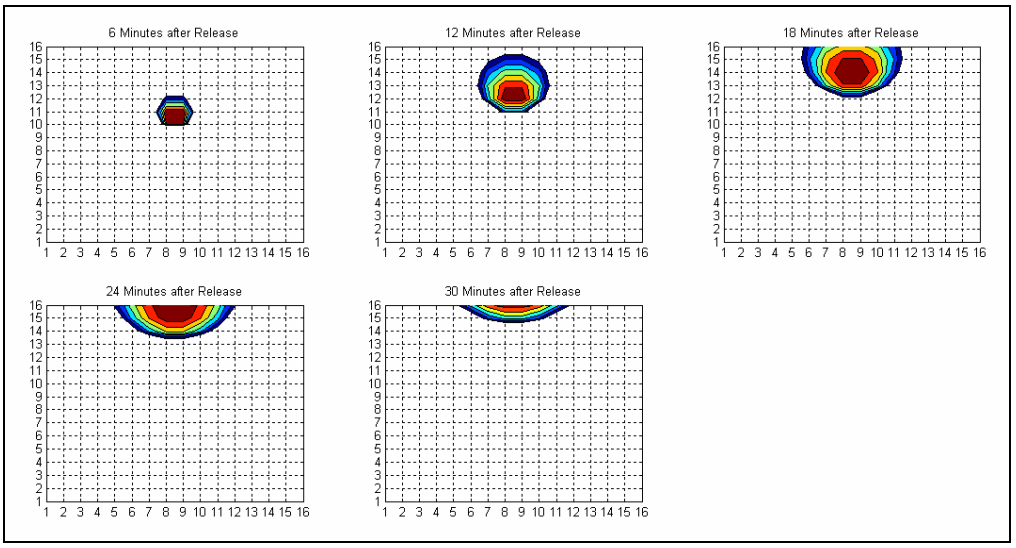

Figure 2. A 180º wind direction release on a 16x16 grid with a grid-spacing of 1000 m. The source is located at (8.5,8.5) and receptors are located at all intersections.

TABLE

Table 1. The average values over ten runs for each of the parameters are listed below. The actual wind direction, θ, is given below, the actual strength is 1.00, the actual source location is (0,0). The cost function value as well as the skill score should be close to zero.

	Grid Size	Grid Spacing (m)	Actual θ (°)	Found θ (°)	Strength	(x,y) (m,m)	Cost Function	Skill Score
GA Alone	4x4	4000	180.00	180.00	1.00	(0,0)	0.000	0.0000
post-Nelder-Mead	4x4	4000	180.00	180.00	1.00	(0,0)	0.000	0.0000
GA Alone	4x4	4000	225.00	223.58	1.09	(82,-73)	0.000	0.3768
post-Nelder-Mead	4x4	4000	225.00	225.00	1.00	(0,0)	0.000	0.0000
GA Alone	8x8	2000	180.00	180.00	1.00	(0,0)	0.000	0.0000
post-Nelder-Mead	8x8	2000	180.00	180.00	1.00	(0,0)	0.000	0.0000
GA Alone	8x8	2000	225.00	224.89	1.00	(10,-10)	0.000	0.0499
post-Nelder-Mead	8x8	2000	225.00	225.00	1.00	(0,0)	0.000	0.0000
GA Alone	16x16	1000	180.00	180.00	1.00	(0,0)	0.000	0.0000
post-Nelder-Mead	16x16	1000	180.00	180.00	1.00	(0,0)	0.000	0.0000
GA Alone	16x16	1000	225.00	224.99	1.00	(1,-1)	0.000	0.0032
post-Nelder-Mead	16x16	1000	225.00	225.00	1.00	(0,0)	0.000	0.0000
GA Alone	32x32	500	180.00	180.00	1.00	(0,0)	0.000	0.0000
post-Nelder-Mead	32x32	500	180.00	180.00	1.00	(0,0)	0.000	0.0000
GA Alone	32x32	500	225.00	225.00	1.00	(0,0)	0.000	0.0004
post-Nelder-Mead	32x32	500	225.00	225.00	1.00	(0,0)	0.000	0.0000