J1.2

COMPARATIVE INVESTIGATION OF SOURCE TERM ESTIMATION ALGORITHMS USING FUSION FIELD TRIAL 2007 DATA

Nathan Platt * and Dennis Deriggi

Institute for Defense Analysis, Alexandria, VA

1. INTRODUCTION

Given a warning based on detection of hazardous materials at only a few sensors, it could be useful to rapidly provide an estimate of the source location, time of release, and amount of material released. Such an estimate can lead to refined predictions of the area affected by the release and can support near-term follow-on actions to investigate the cause and nature of the release. In some cases, refined predictions that could result from such source term estimation (STE) can support tactical decisions (e.g., which roads to travel on and which to avoid). For longer-range situations (tens of kilometers), accurate estimates of the source can allow for improved hazard-area predictions that could better support warnings and possible evacuation, efficient mission-oriented protective posture gear usage, or perhaps medical response.

In September 2007, a short-range (~500 meters), highly instrumented test was conducted at the U.S. Army's Dugway Proving Ground (DPG) (Storwold, 2007). This test, referred to as Fusing Sensor Information from Observing Networks (FUSION) Field Trial 2007 (FFT 07), was designed to collect data to support the further development of prototype algorithms. FFT 07 was sponsored by DTRA-Joint Science and Technology Office for Chemical and Biological Defense (JSTO-CBD) and was conceived of and planned within the Technical Panel 9 for Hazard Assessment (TP9) of The Technical Cooperation Program (TTCP) Chemical, Biological, and Radiological Defense (CBD) group, thus making this effort an international (in this case, U.S., U.K., Canada, and Australia) collaboration. Figure 1 illustrates the basic layout of a subset of FFT 07 instrumentation including 100 digiPIDs (digital photoionization detectors), used to continuously sample propylene concentration at 50 Hz, and the locations of various meteorological instruments. Not shown in this schematic are 20 UVIC (ultraviolet ion collector) sensors used to continuously sample propylene concentration at 50 Hz and positioned between the digiPIDs at lines 3 and 8.

The reasons for conducting FFT 07 were numerous. First, the experiment was meant to provide a set of data that STE model developers can use to improve their algorithms. Next, the collected information could be used to assist in identifying the strengths and weaknesses of the different modeling approaches chosen by the developers. Finally, assessment of STE algorithms using data collected during FFT 07, was meant to help the Department of Defense identify the current state of the STE algorithms in general (the "state of the art").

A comparative investigation of STE algorithms began in 2008. The general method of this investigation was to first provide participating developers with a subset of sensor data that was collected on selected FFT 07 trials. Next, developers provided "blind" predictions that were compared to parameters of the actual release. Phase I of this investigation consisted of 104 individual cases of simulated sensor data that were distributed in September 2008. Table 1 lists the composition of Phase I cases. These cases simulated continuous streams of concentration data for ingestion by STE algorithms. Each case selected for Phase I evaluation was created using available digiPID data. We notionally demonstrate this procedure using observed propylene concentration values at the array of digiPIDs during FFT 07 Release 6. First, the 50-Hz digiPID data are bin-averaged to 1 second (i.e., we assume that the simulated sensor produces a single value every 1 second). Second, the map depicting the locations where the concentration threshold is exceeded is constructed as shown in Figure 2. The colored circles denote digiPIDs whose concentration exceeds the color-coded concentration value (in ppm) shown in the legend for any 1-second time bin (i.e., "hit" sensors), while the black open circles denote the rest of the digiPIDs (i.e., "null" or malfunctioning sensors). The circles with 'X' inside them represent digiPIDs that malfunctioned and thus produced no useful data. The large purple circles denote locations that we arbitrarily selected to indicate four simulated sensors to be used to feed the sensor fusion algorithms. These locations roughly correspond

*Corresponding author address: Nathan Platt, Institute for Defense Analysis, 4850 Mark Center Drive, Alexandria, VA 22311-1882; e-mail: nplatt@ida.org.

to corners of the rectangular area. In this case, we chose three hits and a single null. To discourage any appearance of perceived data manipulation by algorithm developers and to improve the results of the evaluation, a simple data-obscuring mechanism was implemented as described in Platt, et al., 2008a. Figure 3 demonstrates the full procedure for the Stage I continuous simulated sensor output where the blue line depicts 10-minute, pre-release "padding" of the simulated sensor output. Further details on the composition of the phases of the evaluation and construction of the cases are available in Platt, et al., 2008a, 2008b.

Phase I consisted of cases that equally sampled the parameters that were potentially expected to most significantly affect the quality of STE predictions. These parameters included time of day of the tracer release (day or night), type of tracer release (continuous or instantaneous), and number of sensors reporting data (4 or 16). To provide some realism in the meteorological inputs to the STE algorithms, for some cases the developers were provided with surface wind velocity observations and a vertical wind velocity profile from sites up to 1-2 km removed from the tracer releases and sampler grid, instead of the more detailed meteorological observations made at the center location of the sampler grid itself. An additional sampled parameter that could affect the quality of STE predictions was the number of sources (single, double, triple, or quad). FFT 07 individual puff trials involved multiple (up to 10) realizations. These puffs were released by firing air cannons every few minutes resulting in "trains of puffs" periodically traversing the digiPID grid. Hence, for puff releases, some distributed cases included a single realization of the puff(s), while some of the distributed cases included multiple (up to 10) realizations. The full structure of the Phase I cases is given in Platt, et al., 2008a.

A total of eight different STE algorithm developers participated in this exercise. A total of 14 full and partial sets of predictions were received with some exercise participants providing multiple sets of predictions based on different algorithms that they have been developing. Table 2 depicts the organizations that participated in Phase I together with the composition of predicted cases that they provided. Red denotes that a full set of predictions was provided; blue denotes that predictions were provided for at least 50 percent of the distributed cases. Table 3 briefly lists some basic capabilities of each of the STE algorithms that include ability to predict number (e.g., single, double, triple, or quad source) and types of sources (e.g., continuous or instantaneous puff release).

2. STE ALGORITHM PERFORMANCE TRENDS

The goal of these evaluations is not to declare a "winning" algorithm, but rather to learn by examining the strengths and weaknesses of each of the proposed methodologies, because different approaches may best apply to different sets of tracer release scenarios (i.e., daytime versus nighttime, single- versus double- source release, richer information available from simulated sensors versus poorer) or for different specific applications (e.g., near real-time versus forensic). In this way, algorithm developers can learn from each other. The main motivation behind the evaluation matrix is an attempt to trade off the ability to cover the evaluation of a substantial number of potential variables that might influence algorithm performance with the desire to keep the sample sizes large enough to be able to arrive at reasonably robust conclusions. Therefore, we decided to start our analysis with the evaluation of algorithm performance trends instead of analyzing each individual algorithm. Two separate methodologies were pursued: linear regression analysis to discern similar trends among different algorithms and comparison of a selected global algorithm performance metrics.

2.1 Individual Case Metrics Used in the Analysis

As depicted in Table 3, the individual STE algorithms participating in Phase I evaluations have different capabilities with respect to predicting numbers and types of sources. In order to make a fair comparison among these algorithms, for each individual case for which a prediction was submitted, common metrics that are applicable to all of the algorithms needed to be defined. We selected two metrics: the distance between the averaged predicted and averaged observed source term locations, and the ratio of the observed and predicted release mass from all sources. Figure 4 illustrates the distance metric calculation. From the 14 sets of STE algorithm predictions, 12 algorithms provided enough information to calculate the mass ratio and all 14 algorithms provided enough data to calculate our distance metric.

2.2 Linear Regression Analysis Description and Results

This section describes the use of linear regression to the examination of source term estimation algorithms. In very general terms, this effort was an attempt to determine which of the underlying factors such as diurnal condition, number of release sources, type of

release, and several other independent variables, had the greatest effect on the estimation of the mass ratio (the ratio of reported to actual mass) or mean offset distance. Two regression techniques were used in this study: backward regression and stepwise regression.

Backward regression begins with all the independent variables in the regression equation, and then proceeds to eliminate those for which the associated sum of squares is not significant. In contrast, stepwise regression only allows independent variables into the regression equation if their associated sum of squares is significant, and eliminates previously admitted variables if their effect is substantially diminished by other variables in the equation.

Thus, roughly speaking, stepwise regression tests each independent variable to determine whether or not it should enter the regression equation, and again, if it should remain in the equation after others are admitted. Backward regression initially treats all variables as belonging to the equation, then eliminates those whose contribution is substandard. For reference, please see Draper and Smith (1966) and Seber (1977).

We chose the following independent regression variables: 1) "Diurnal," defined as either Night or Day release time; 2) "Met Num," defined as either "Close-in" met corresponding to meteorology obtained at the center of the digiPID grid, or "Operational" met, which corresponded to using meteorological stations approximately 1-2 km away; 3) "Sources" denoting the number of sources used in the definition of a case (single, double, triple, quad); 4) "Sensors" denoting the number of simulated sensors used in the definition of a case (4 or 16); 4) "Puff/Real," defined as: "-1" if case is constructed from a continuous trial, "0" if case is constructed using single realization of a puff trials, and "1" if case is constructed using multiple realizations of a puff trial. The "Puff/Real" independent variable is expected to succinctly represent two distinct parameters that could affect quality of STE predictions: continuous vs. instantaneous/puff releases and single vs. multiple releases from the same location.

The list of dependent regression variables included: "Mean," defined as the distance between averaged predicted and averaged observed source term locations for the individual case as shown in Figure 4, and "Mass Ratio," defined as a ratio of predicted to observed total mass of the material used to define a particular case.

The results for stepwise and backward regressions are summarized in Tables 4 and 5, respectively. To simplify viewing these tables, the colored background in the table entries is coded according to which independent variable is called by the particular significant factor.

The regression outcomes were ranked in decreasing order of their respective ("adjusted") coefficients of determination. The coefficient of determination is referred to as R^2 and is the square of the correlation between observations and values returned by the regression equation. It is equal to the proportion of the variance in observed data that can be "explained" by regression (G. Seber, 1977). The adjusted R^2, which determined the ordering, takes into account the number of variables in the model and is equal to $1 - (1 - R^2)(n-1)/(n - p - 1)$, where p is the number of independent variables in regression equation and n is the number of observations. The point of using the adjusted R^2 is to force models to be economical by penalizing excessive numbers of independent variables. This is in contrast to the (unadjusted) R^2, which increases with the number of independent variables.

Two types of regression coefficients are presented in this paper: standardized and unstandardized. The former refers to regression coefficients obtained after transforming all data so that the dependent variable and all independent variables have a mean of zero and a standard deviation of 1.0. In some sense, this treats all data as being on an equal footing. The unstandardized coefficients are the result of performing regression without this transformation. For each model listed in Tables 4 and 5, both types of coefficients appear in parentheses after each independent variable that was selected by the regression process. The level of significance – that is the probability of the same or more extreme observations under the null hypothesis that this coefficient was zero (loosely speaking, the probability that the coefficient is zero), also appears in parentheses after the coefficient.

Models with gray backgrounds in Tables 4 and 5 are those for which there were no data or which regression was not significant.

With respect to predicting miss distance between predicted and observed STE location, the regression analysis indicates:

1. "Diurnal" (Day/Night) is not a significant variable for both backward and stepwise regressions for almost all algorithms.

2. "Met Num" variable representing "Close-In" versus "Operational" met options is not a significant

variable for both backward and stepwise regression for almost all algorithms.

3. "Sources" variable representing number of sources used in the definition of a case is a significant predictor of algorithm performance for six algorithms. Six algorithms are called by stepwise regression and four algorithms are called by backward regression.

4. "Sensors" regression variable representing number of sensors (4 vs. 16) used in the definition of the case is a significant predictor of algorithms performance for only three algorithms. This indicates that most STE algorithms do not benefit from having larger number of sensors.

5. "Puff Real" regression variable is a significant predictor for algorithm performance for two algorithms using backward regression and one algorithm using stepwise regression.

With respect to mass ratio dependent variable, regression analysis indicates:

1. "Diurnal" (Day/Night), "Met Num" (Close-In/Operational met) and "Sensors" (4 vs. 16) are not significant variables for most algorithms for both backward and stepwise regression.

2. "Sources" independent regression variable representing number of sources used in the definition of a case is a significant predictor of algorithm performance for seven algorithms.

3. "Puff Real" regression variable is a significant predictor for algorithm performance for seven algorithms.

We would like to caution that regression analysis results should serve as a guide for further investigation of which algorithm/variable combinations are important. For instance, the regression analysis does not tell if the algorithm performed as expected with respect to a given variable.

2.3 Comparison of a Selected Global Algorithm Performance Metrics

In addition to using the linear regression methodology to discern trends among different sets of STE predictions, we devised metrics to capture some aspects of the global algorithm performance. As discussed earlier, for each individual case predicted by an STE algorithm, two measures were calculated: the distance between the averaged predicted and the average observed location of the source(s) that we will refer to as a "miss distance" and the ratio of the total predicted mass over the total released mass from all sources that we will refer to as the "mass ratio."

To compare various STE algorithm performances using the miss distance metric, we selected three levels of interest and then calculated a fraction of cases where the miss distance is less than the level of interest. These levels of interest include: 100 meters (i.e., miss distance is in the tens of meters), 250 meters (i.e., miss distance is less than half the size of the digiPID domain), and 500 meters (i.e., miss distance is less than approximate the size of the digiPID domain). We would like to note that even when a particular miss distance is less than some number *d*, it is quite possible that the individual distances between actual and predicted locations of the sources is greater than *d* as demonstrated in Figure 4. Figure 5 shows results for these calculations at the three levels of interest. For each set of STE predictions, the grouped colored bars denote the fraction of predictions that are less than the particular level of interest. Thick colored lines correspond to the medians of all fractions (i.e., half of the fractions are below/above this value) for various thresholds: 0.46 (blue line) for fraction of the miss distance less than 100 meters, 0.79 (brown line) for fraction of the miss distance less than 250 meters, and 0.94 (green line) for fraction of the miss distance less than 500 meters. With respect to predicting the miss distance, we observe the following: 1) when the miss distance is less than 100 meters, a wide spread is seen in algorithm performance; and 2) most algorithms seem to be capable of having more than 90 percent of the predictions with miss distances less than 500 meters (approximate size of the tracer measurement grid of FFT 07).

We examined the mass ratio metric for two types of statistics: 1) whether or not a particular algorithm has a tendency to over- or under-predict the total mass released from all sources, and 2) for any given set of predictions, what is the fraction of cases when the predicted and observed masses are within a factor of 2, 5, or 10 of each other. For each set of the 12 predictions that provided enough information to calculate the total predicted mass, Figure 6 shows the fraction of cases that were over-predicted. Thick brown lines (at 0.4 and 0.6) denote arbitrary limits that we used to distinguish different predictive behavior: 1) if the fraction of predictions is less than 0.4, then the particular algorithm is generally under-predicting the total release mass; 2) if the fraction of predictions lies between 0.4 and 0.6, then the particular algorithm has

about an equal amount of over- and under-prediction; and 3) if the fraction of predictions is greater than 0.6, then the particular algorithm has a tendency to over-predict the total released mass. For each set of predictions, Figure 7 shows the fractions of cases when the total observed and predicted masses are within factors of 2, 5, and 10. Thick colored lines correspond to the medians of all fractions (i.e., half of the fractions are below/above this value) for various factors: 0.34 (brown line) for a factor of 2, 0.62 for a factor of 5, and 0.77 for a factor of 10. With respect to predicting the total predicted over observed mass ratio metric, we observe the following: 1) there are wide variations in terms of algorithm performance with respect to over- or under-predicting masses of the releases with some algorithms exhibiting a large number of cases significantly over- or under-predicted; 2) with the exception of three algorithms, the fraction of mass within factors of 2, 5, 10 varies from 0.27 to 0.48, 0.59 to 0.81, and 0.69 to 0.92, respectively.

We would like to caution that these results are an attempt to capture global algorithm performance without any attempt to make sure that submitted predictions are intercomparable with each other. For instance, these results do not take into account that some algorithms provided only partial predictions (i.e., not a complete set of predictions for all Phase I cases). Some of these were preferentially selected (i.e., "Puff only" or "16 sensors only" predictions) by some of the algorithm developers.

3. SUMMARY

In September 2007, a short-range test – Fusing Sensor Information from Observing Networks (FUSION) Field Trial 2007 (FFT 07) – designed to collect data to support development of prototype STE algorithms was conducted. A comparative investigation of STE algorithms based on this field trial began in 2008. The general method of this investigation was to first provide the participating developers with a subset of sensor data that was collected on selected FFT 07 trials. Next, developers provided "blind" predictions that could then be compared to parameters of the actual release. Phase I of this investigation consisted of 104 individual cases of simulated sensor data that were distributed in September 2008. These cases simulated continuous streams of concentration data for ingestion by STE algorithms. A total of eight different STE algorithm developers participated in this exercise. A total of 14 full and partial sets of predictions were received with some exercise participants providing multiple sets of predictions based on different algorithms they have been developing.

The goal of these evaluations is not to declare a "winning" algorithm, but rather to learn by examining the strengths and weaknesses of each of the proposed methodologies, because different approaches may best apply to different sets of tracer release scenarios. Therefore, we decided to start our analysis with the evaluation of algorithm performance trends instead of analyzing each individual algorithm. Two separate methodologies were pursued: linear regression analysis to discern similar trends among different algorithms and comparison of selected global algorithm performance metrics.

In terms of the linear regression analysis results, we found that: 1) time of the release (night versus day), type of meteorology provided (detailed versus operational), and increased number of simulated sensors (4 versus 16) *did not* lead to improvements in the quality of predictions for most STE algorithms participating in the exercise; and 2) number of sources and type of release (continuous versus single realization of instantaneous puff(s) versus multiple realizations of instantaneous puff(s) *are significant variables* in terms of predicting algorithm performance for the majority of participating algorithms.

Two global performance metrics that compared STE algorithm performance were constructed and examined. With respect to our miss distance metric, all algorithms were able to predict "averaged" source term locations to within 500 meters (i.e., size comparable to the side of the tracer measurements grid of FFT 07), while a wide variation in the quality of the algorithm predictions was seen when the miss distance was on the order of tens of meters (i.e., less than 100 meters). With respect to predicting the total release mass, there was a wide variation in algorithm performance with respect to over- or under-predicting masses of the release with some algorithms showing large fractions of cases that were either under- or over-predicted. In addition, with the exception of a few algorithms, the fraction of the mass within factors of 2, 5, and 10 was within a reasonably tight range for most of the algorithms that were capable of predicting the total mass of the release.

Phase II of this exercise is being planned to start in FY 10 and incorporate: 1) lessons learned from Phase I, 2) the addition of "bar-sensor" input data stream, and 3) the use of a simulated environment to supplement the field trial data.

ACKNOWLEDGEMENTS

This effort was supported by the Defense Threat Reduction Agency with Dr. John Hannan as the project

monitor. The authors would like to thank Edward Argenta, Donald P. Storwold, Jr., and John White of the Meteorology Division, West Desert Test Center, U.S. Army Dugway Proving Ground for providing access to initial FFT 07 test results and for providing useful review of the evaluation plan. The authors also would like to thank Paul Bieringer and Jon Hurst from the National Center for Atmospheric Research for helping design and coordinate a common output format for the STE predictions. Finally, we would like to thank all STE participants who provided predictions. The views expressed in this paper are solely those of the authors.

REFERENCES

Draper, N. and H. Smith 1966: *Applied Regression Analysis*, Wiley.

Platt, N., Warner, S. and S.M. Nunes, 2008a: *Plan for initial comparative investigation of source term estimation algorithms using FUSION field trial 2007 (FFT 07)*. Institute for Defense Analyses Document D-3488.

Platt, N., Warner, S. and S.M. Nunes, 2008b: "Evaluation plan for comparative investigation of source term estimation algorithms using FUSION field trial 2007 data." *Croatian Meteorological Journal*, Proceedings of the 12th International Conference on Harmonization within Atmospheric Dispersion Modeling for Regulatory Purposes, Part 1: Oral Presentations, Vol. 43, p. 224-229.

Seber, G. 1977: *Linear Regression Analysis,* Wiley.

Storwold, D.P, 2007: *Detailed test plan for the Fusing Sensor Information from Observing Networks (FUSION) field trial 2007 (FFT 07)*. West Desert Test Center, U.S. Army Dugway Proving Ground, WDTC Document No. WDTC-TP-07-078.

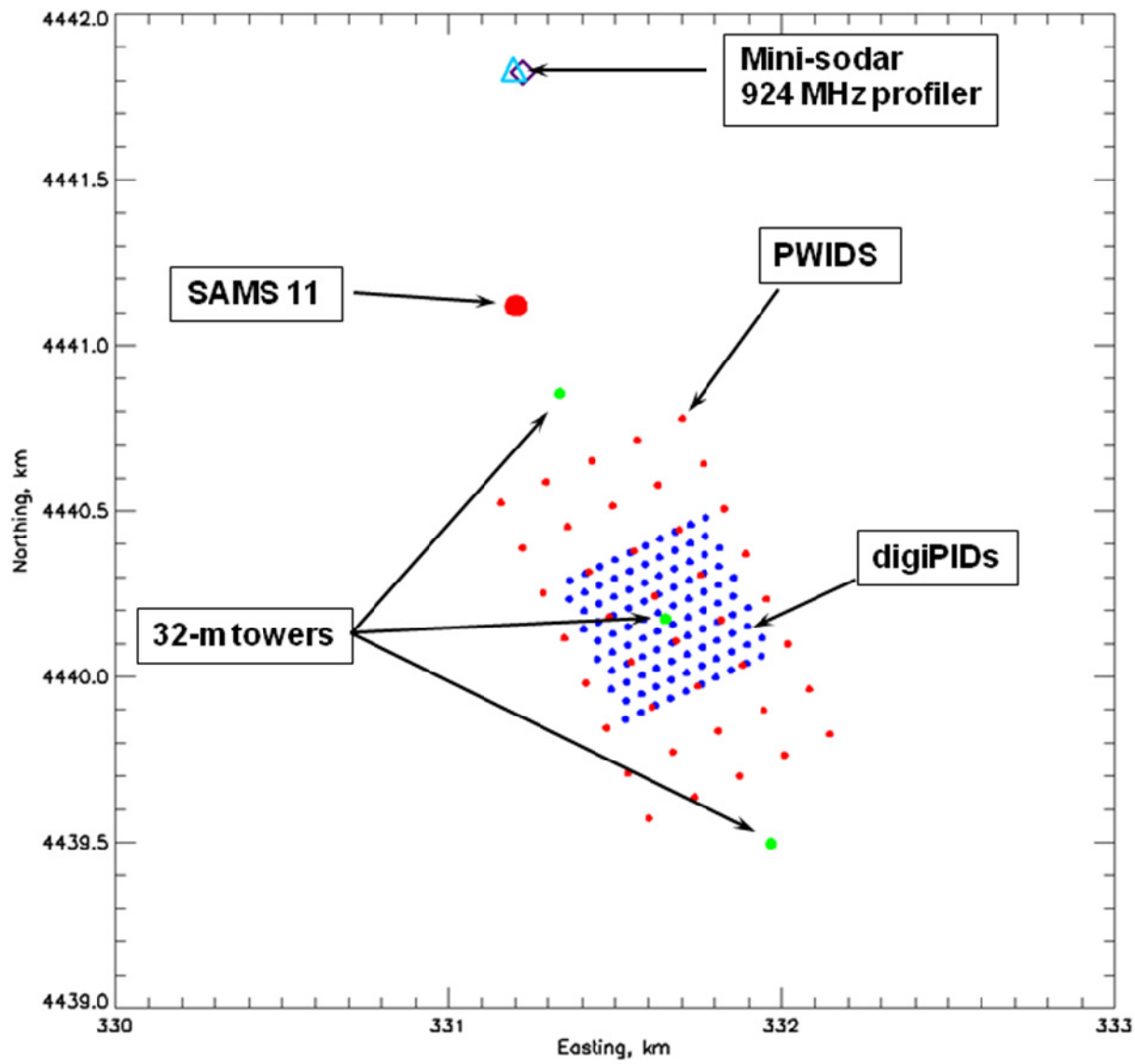

Figure 1. Subset of instrumentation utilized during FFT 07 field trials. The dense set of dots arranged in a square pattern denote locations of the 100 digiPIDs used to continuously sample propylene concentrations at 50 Hz; the larger dots arranged in a rectangular pattern denote locations of the 40 PWIDs (Portable Weather Information and Display Systems) used to collect detailed surface meteorology; the green dots denote locations of three 32-meter towers that carried additional meteorological instrumentation; the large red dot denotes the location of the SAMS (Surface Atmospheric Measurement System) 11 meteorological weather station; and the diamond and triangle at the top denote the location of a mini-sodar and a 924 MHz radar wind profiler, respectively.

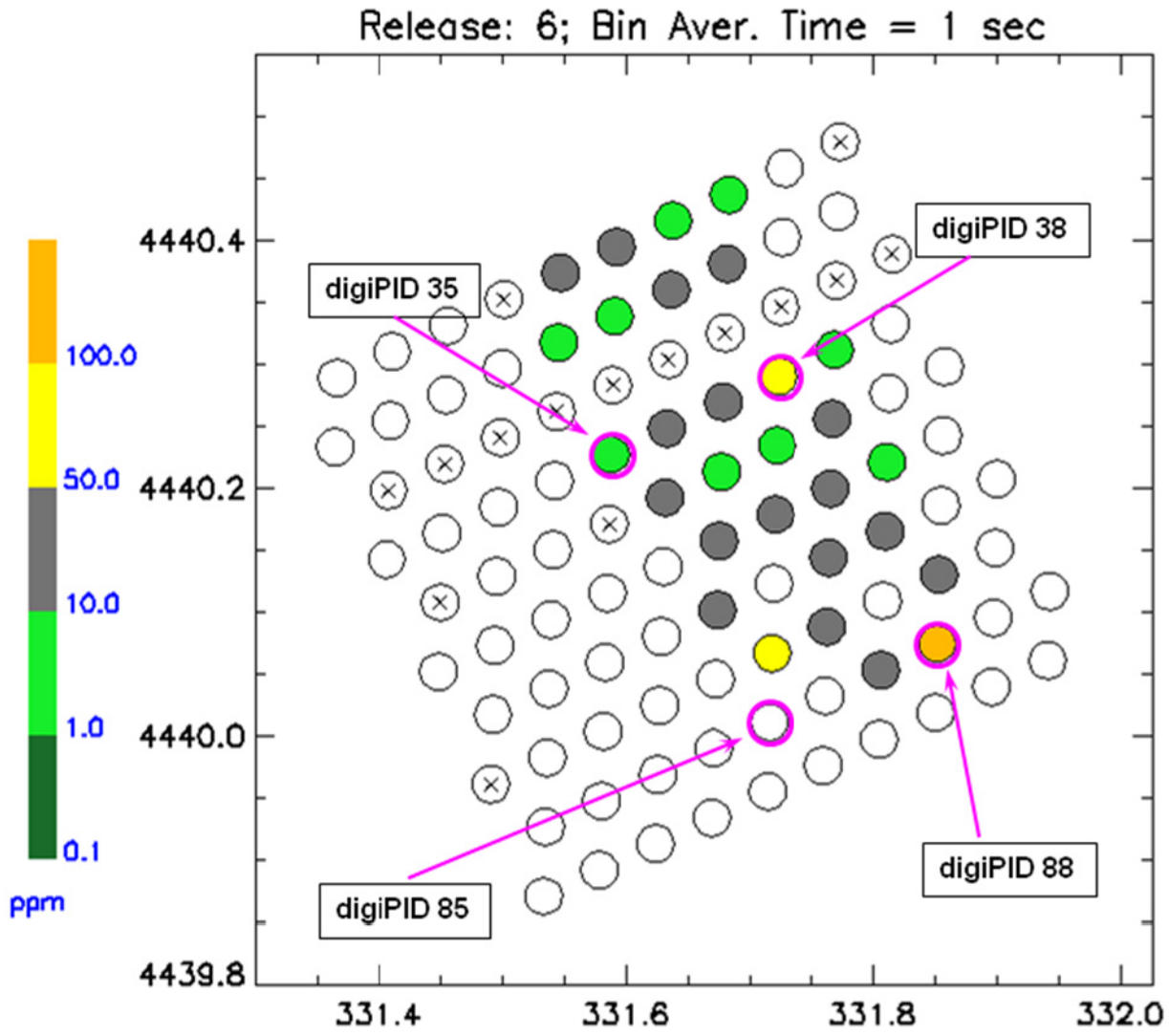

Figure 2. Notional demonstration of selecting digiPIDs to simulate sensor output. The color of the circles denotes the maximum 1-second averaged concentration observed at the particular digiPID in accordance with the legend. The "X" inside the circle denotes digiPIDs that were malfunctioning during this particular release and thus produced no data. Open circles denote digiPIDs that were not hit during the release. Purple circles denote the four digiPIDs that were selected for this case – three "hits" and one "null."

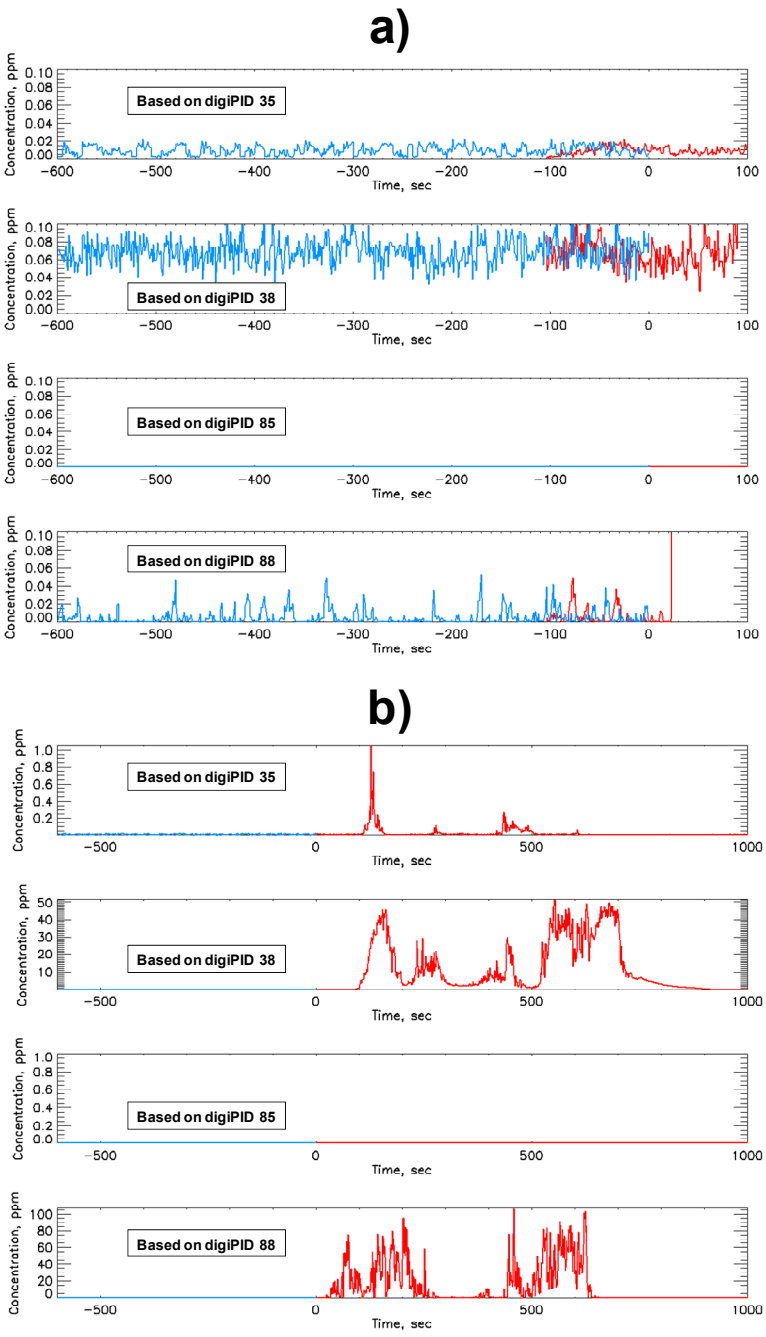

Figure 3. Demonstration of simulated 1-second averaged continuous stream STE input data for the notional case depicted in Figure 2. The blue line corresponds to 10-minute simulated pre-release padding added to the observed FFT 07 data. Part **a)** of the figure compares 10-minute simulated pre-release padding with observed pre-release data shown in red, while part **b)** of the figure shows full time series for the notional case depicted in Figure 2.

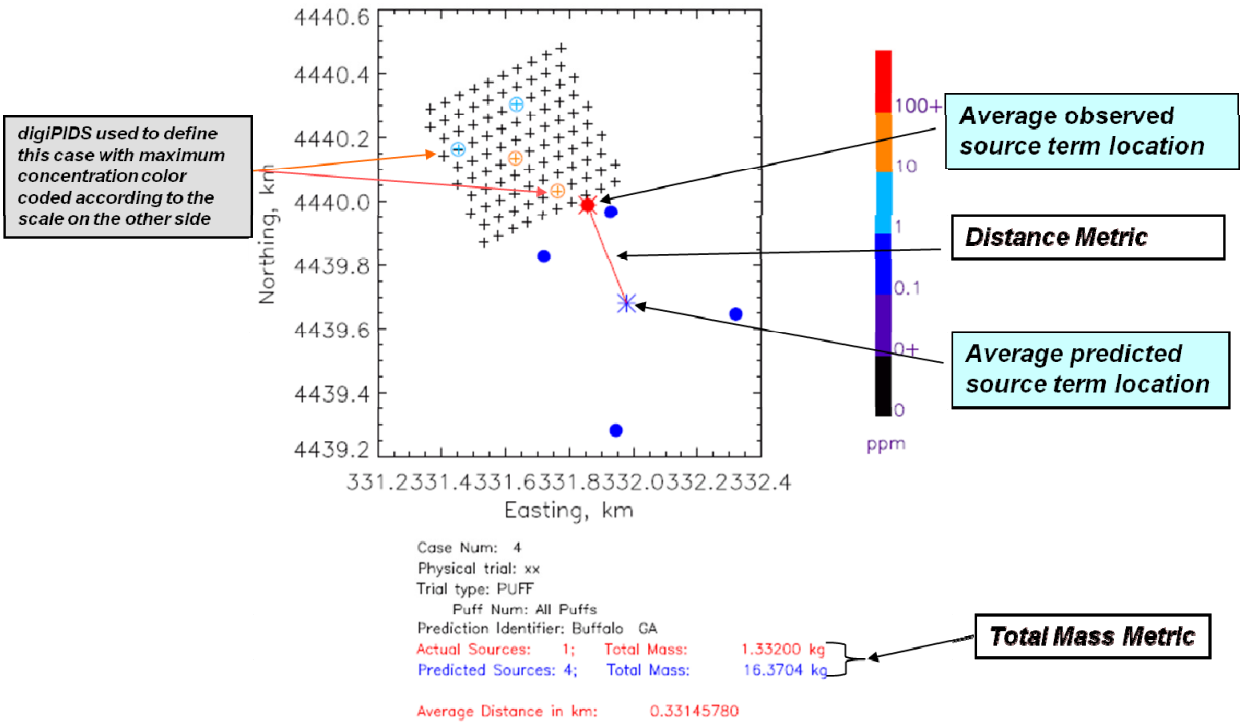

Figure 4. Example of the distance metric computation and total mass calculation used to compare algorithm performance for each individual case.

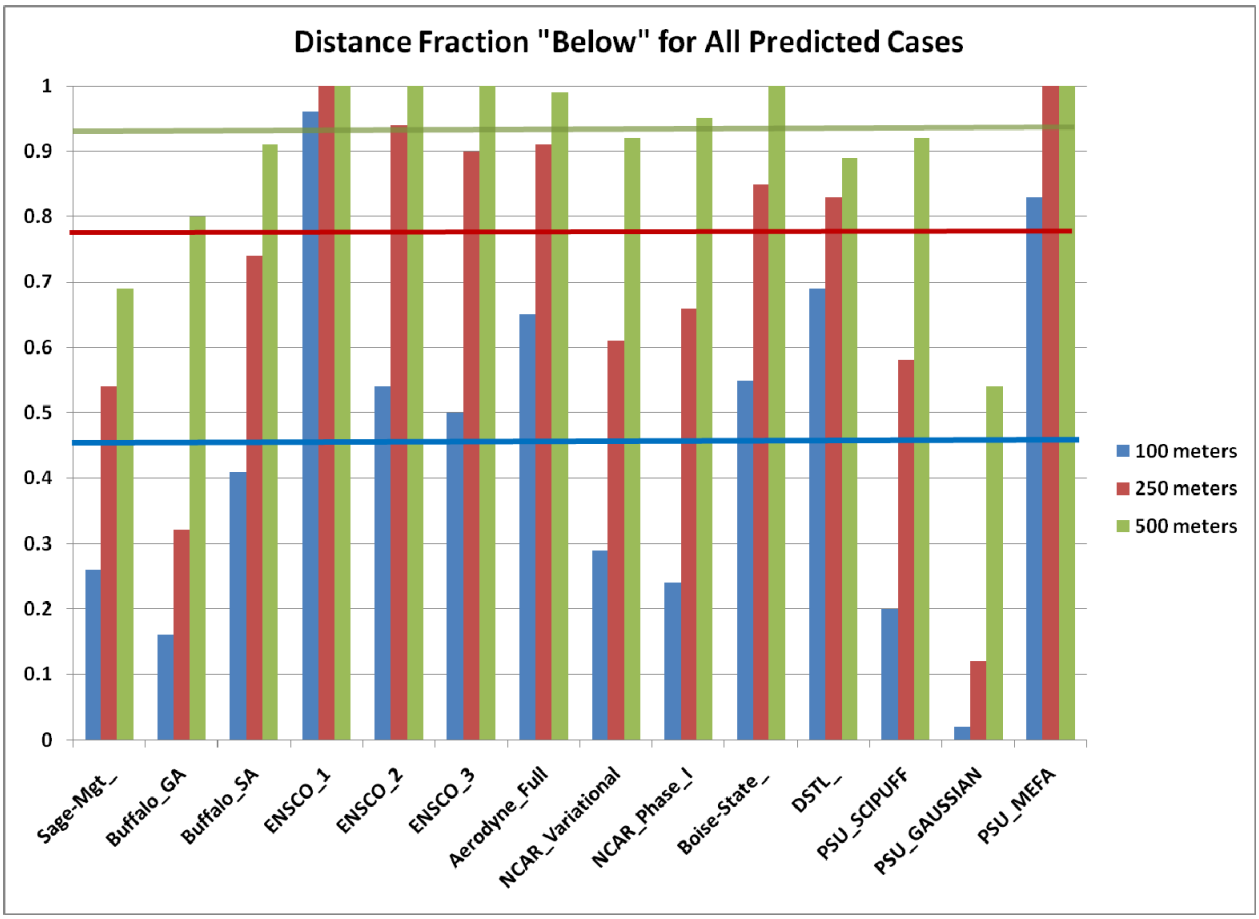

Figure 5. Algorithm inter-comparison using averaged miss distance fraction of cases below 100, 200, and 500 meters. Individual algorithm bars are color coded according to the legend. Thick colored lines correspond to the median for all the fractions (i.e., half of the fractions are below/above this value) for the various thresholds: 0.46 (blue line) for fraction of the miss distance less than 100 meters, 0.79 (brown line) for fraction of the miss distance less than 250 meters, and 0.94 (green line) for fraction of the miss distance less than 500 meters.

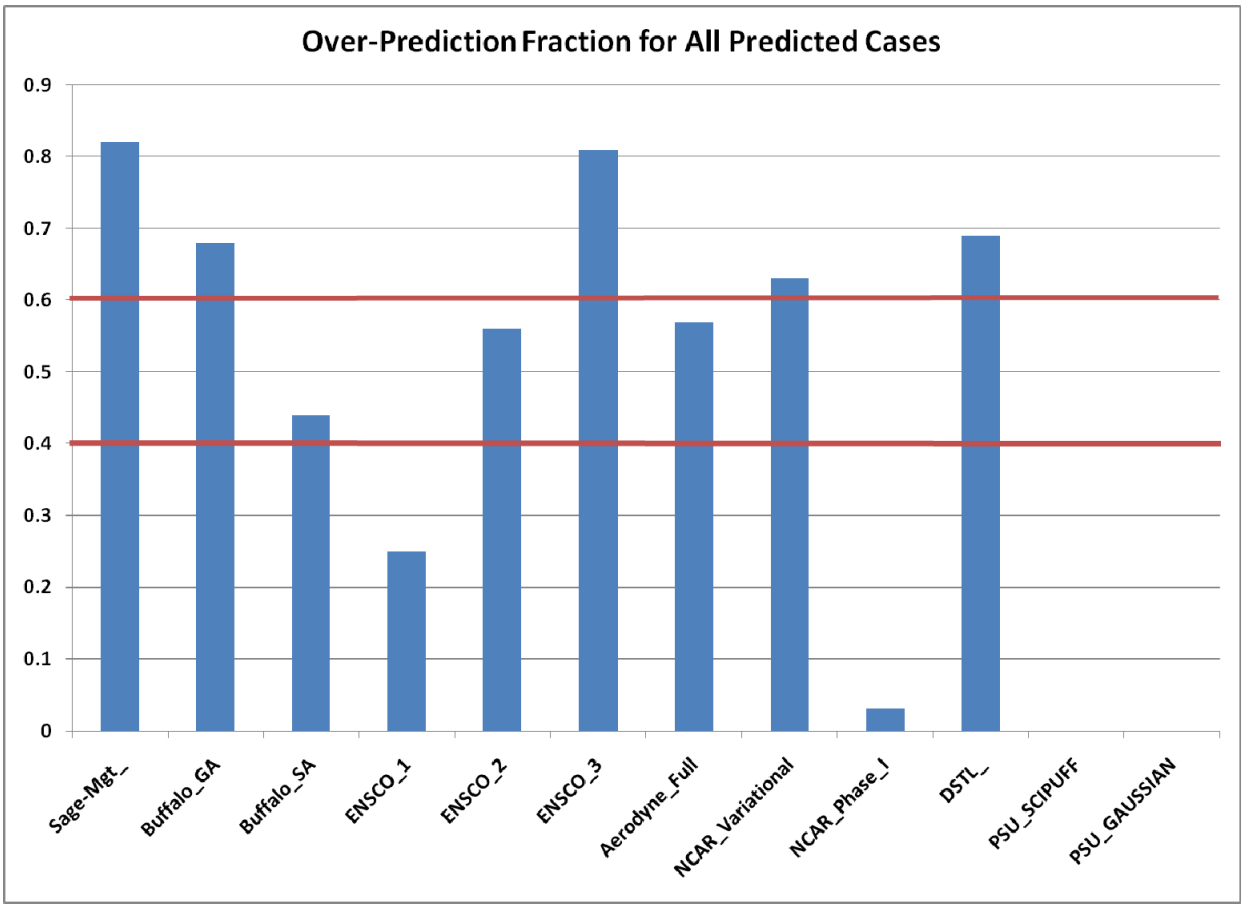

Figure 6. Total mass over-prediction fraction for the 12 STE algorithms that provided enough information to calculate total predicted release mass from all sources. Thick brown lines (at 0.4 and 0.6) denote limits that are used to distinguish different predictive behavior: a fraction below 0.4 implies an algorithm tendency to underpredict, a fraction in the range of 0.4 and 0.6 implies about an equal number of under- and overpredicted cases, and a fraction above 0.6 implies an algorithm tendency to overpredict.

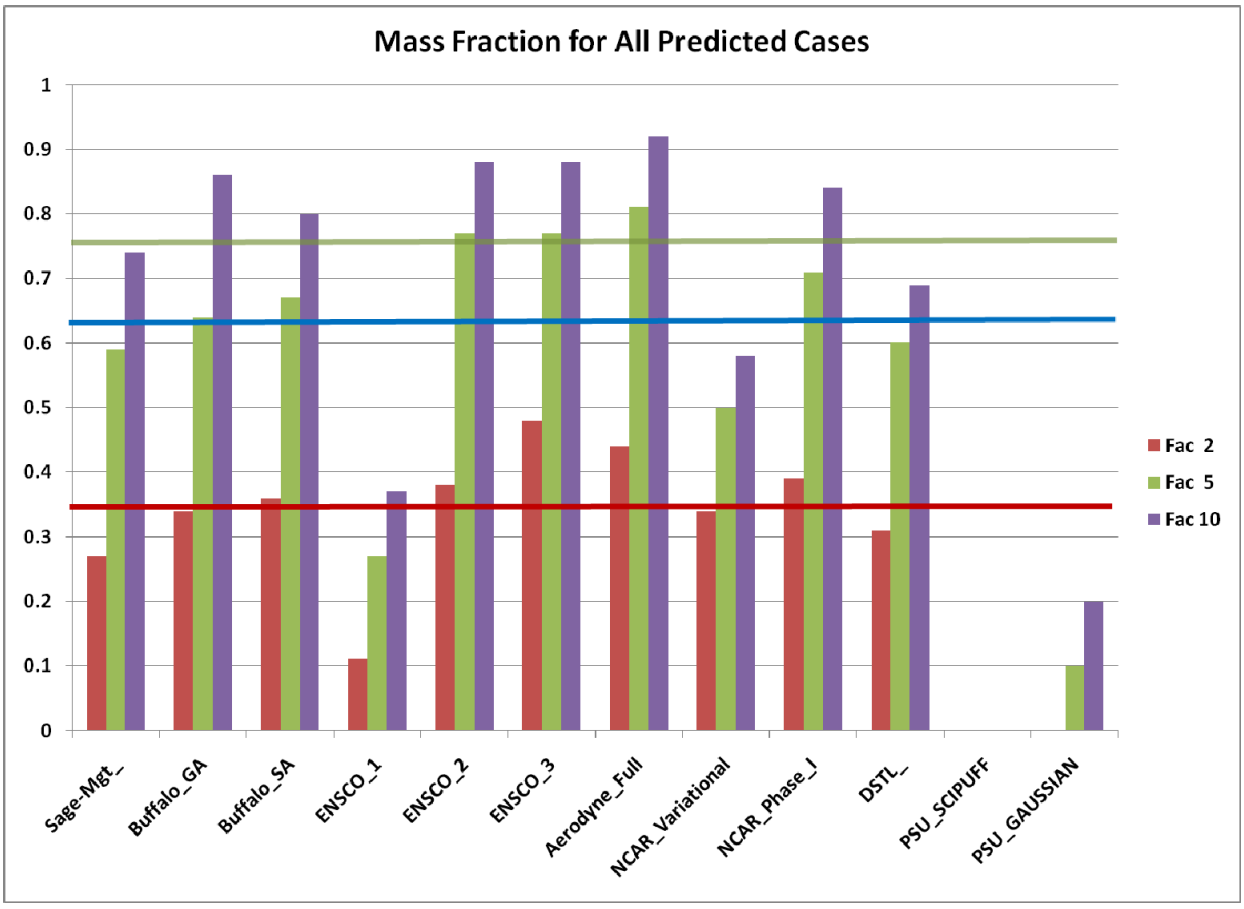

Figure 7. Algorithm inter-comparison using total observed and predicted mass fractions within factors of 2, 5, and 10. Thick colored lines correspond to the medians of all fractions (i.e., half of the fractions are below/above this value) for various factors: 0.34 (brown line) for factor of 2, 0.62 for factor of 5, and 0.77 for factor of 10.

Phase I Release Case Composition					
Condition	**All Trials**	**Single**	**Double**	**Triple**	**Quad**
none	**104**	**40**	**40**	**16**	**8**
Puff	**52**	**20**	**20**	**8**	**4**
Cont	**52**	**20**	**20**	**8**	**4**
Daytime	**52**	**20**	**20**	**8**	**4**
Nighttime	**52**	**20**	**20**	**8**	**4**
Daytime/Puff	**26**	**10**	**10**	**4**	**2**
Daytime/Cont	**26**	**10**	**10**	**4**	**2**
Nighttime/Puff	**26**	**10**	**10**	**4**	**2**
Nighttime/Cont	**26**	**10**	**10**	**4**	**2**

Table 1. Composition of Phase I cases that were distributed to STE algorithm developers to provide predictions.

Composition of the Prediction Sets Recieved									
Organization	Total	Cont	Puff	Daytime	Nighttime	Single	Double	Triple	Quad
Aerodyne	104	52	52	52	52	40	40	16	8
Boise-State	33	14	19	21	12	13	13	4	3
Buffalo / GA	104	52	52	52	52	40	40	16	8
Buffalo / SA	70	34	36	34	36	26	26	12	6
DSTL	35	5	30	20	15	12	14	7	2
ENSCO / Set 1	102	51	51	50	52	39	39	16	8
ENSCO / Set 2	104	52	52	52	52	40	40	16	8
ENSCO / Set 3	42	24	18	19	23	13	15	10	4
NCAR / Variational	38	3	35	20	18	16	14	4	4
NCAR / Phase I	38	3	35	20	18	16	14	4	4
Sage-Mgt	104	52	52	52	52	40	40	16	8
PSU / Gaussian	50	26	24	25	25	18	20	8	4
PSU / SCIPUFF	50	26	24	25	25	18	20	8	4
PSU / MEFA	35	19	16	17	18	13	16	5	1

Table 2. Organizations participating in Phase I together with the composition of predicted cases they provided broken down into several categories including release type, time of day, and number of sources. The red font values denote that a full set of predictions was provided; blue font values denote that the predictions were provided for at least 50 percent of the distributed cases.

Algorithm Capabilities		
Organization	**Number of Sources**	**Type**
Aerodyne	Multi	Cont/Puff
Boise-State	Single	Cont/Puff
Buffalo / GA	Multi	Cont/Puff
Buffalo / SA	Mostly Single	Cont/Puff
DSTL	Single	Puff
ENSCO / Set 1	Multi	Cont/Puff
ENSCO / Set 2	Single	Cont
ENSCO / Set 3	Single	Cont
NCAR / Variational	Single	Puff
NCAR / Phase I	Single	Puff
Sage-Mgt	Single	Cont/Puff
PSU / Gaussian	Single	Cont/Puff
PSU / SCIPUFF	Single	Cont/Puff
PSU / MEFA	Multi	Cont/Puff

Table 3. Basic capabilities of each of the STE algorithms that submitted predictions to the Phase I of the exercise.

model	dependent	R2	significant factor	significant factor	significant factor
ENSCO 3	Mass Ratio	0.379	Puff Real (0.51, 2.49. 0)	Sources (-0.447, -1.9, 0.001)	
Buffalo SA	Mass Ratio	0.273	Sources (-0.348, -0.723, 0.002)	Met Num (0.235, 0.632, 0.031)	Diurnal (0.231, 0.508, 0.029)
DSTL	Mass Ratio	0.254	Puff Real (-0.567, -287.1, 0.001)	Sources (-0.376, -75.9, 0.026)	
ENSCO 2	Mass Ratio	0.221	Puff Real (0.37, 1.3, 0.0)	Sources (-0.32, -0.93,0)	Sensors (0.17, 0.074, 0.06)
PSA Gaussian	Mass Ratio	0.209	Puff Real (0.46, 0.059, 0.01)	SourceS (-0.407, -0.037, 0.02)	
PSU SCIPUFF	Mass Ratio	0.203	Sources (-0.5, -0.011, 0.035)		
Buffalo GA	Mass Ratio	0.172	Sources (-0.365, -2.376, 0)	Puff Real (0.183, 1.417, 0.044)	Diurnal (0.177, 1.224, 0.051)
ENSCO 1	Mass Ratio	0.15	Puff Real (0.398, 14.64, 0)		
Aerodyne	Mass Ratio	0.096	Puff Real (0.262, 0.852, 0.006)	Sensors (-0.212, -0.089, 0.026)	
NCAR Phase I	Mass Ratio	0	constant		
NCAR Variation	Mass Ratio	0			
SAGE Mgt August	Mass Ratio	0			
Boise State	Mass Ratio	-1.00E-06	NO DATA		
PSU MEFA	Mass Ratio	-1.00E-06	NO DATA		
model	dependent	R2	significant factor	significant factor	significant factor
DSTL	Mean	0.67	Puff Real (-0.725, -1.105, 0)	Sources (0.212,0.129, 0.056)	
NCAR Phase I	Mean	0.266	Sources (0.534, 0.09, 0.001)		
NCAR Variation	Mean	0.204	Sources (0.475, 0.09, 0.003)		
ENSCO 3	Mean	0.148	Sources (-0.366, -0.031, 0.015)	Sensors (0.258, 0.003, 0.08)	
PSA Gaussian	Mean	0.102	Sources(0.306, 0.055, 0.029)	Puff Real (-0.254, -0.057, 0.069)	
SAGE Mgt August	Mean	0.083	Sources (0.303, 0.204, 0.002)		
ENSCO 1	Mean	0.043	Met Num (0.228, 0.009, 0.021)		
ENSCO 2	Mean	0.04	Sensors (-0.173, -0.002, 0.076)	Met Num (0.169, 0.017, 0.083)	
Aerodyne	Mean	0.033	Sensors (-0.206, -0.003, 0.036)		
Boise State	Mean	0	constant		
Buffalo GA	Mean	0	constant		
Buffalo SA	Mean	0			
PSU MEFA	Mean	0	constant		
PSU SCIPUFF	Mean	0	constant		

Table 4. Table of significant factors for backward regression. This table is divided into two sections, one for each dependent variable. Each section contains the proportion of variance explained by regression (adjusted R2), independent variables selected by backward regression (see Section 2.2 above), standard coefficient for that variable, unstandardized coefficient, and significance level. All computations were performed using SPSS 15.0 with a removal criterion of 10 percent significance as determined by the appropriate partial F-test. Background cell colors are used to designate and group individual variables that were considered significant by the regression analysis.

model	dependent	R2	significant factor	significant factor	significant factor
ENSCO 3	Mass Ratio	0.379	Puff Real (0.51, 2.49. 0)	Sources (-0.447, -1.9, 0.001)	
Buffalo SA	Mass Ratio	0.273	Sources (-0.348, -0.723, 0.002)	Met Num (0.235, 0.632, 0.031)	Diurnal (0.231, 0.508, 0.029)
DSTL	Mass Ratio	0.254	Puff Real (-0.567, -287.1, 0.001)	Sources (-0.376, -75.9, 0.026)	
PSU SCIPUFF	Mass Ratio	0.203	Sources (-0.5, -0.011, 0.035)		
ENSCO 2	Mass Ratio	0.201	Puff Real (0.37, 1.3, 0)	Sources (-0.32, -0.93, 0)	
ENSCO 1	Mass Ratio	0.15	Puff Real (0.398, 14.64, 0)		
Buffalo GA	Mass Ratio	0.125	Sources (-0.365, -2.376, 0)		
Aerodyne	Mass Ratio	0.096	Puff Real (0.262, 0.852, 0.006)	Sensors (-0.212, -0.089, 0.026)	
NCAR Phase I	Mass Ratio	0			
NCAR Variation	Mass Ratio	0			
PSU Gaussian	Mass Ratio	0			
SAGE Mgt August	Mass Ratio	0			
Boise State	Mass Ratio	-1	NO DATA		
PSU MEFA	Mass Ratio	-1	NO DATA		
model	dependent	R2	significant factor	significant factor	significant factor
DSTL	Mean	0.641	Puff Real (-0.807, -1.23, 0)		
NCAR Phase I	Mean	0.266	Sources (0.534, 0.09, 0.001)		
NCAR Variation	Mean	0.204	Sources (0.475, 0.09, 0.003)		
ENSCO 3	Mean	0.101	Sources (-0.35, -0.03, 0.023)		
SAGE Mgt August	Mean	0.083	Sources (0.303, 0.204, 0.002)		
ENSCO 1	Mean	0.043	Met Num (0.228, 0.009, 0.021)		
Aerodyne	Mean	0.033	Sensors (-0.206, -0.003, 0.036)		
Boise State	Mean	0			
Buffalo GA	Mean	0			
Buffalo SA	Mean	0			
ENSCO 2	Mean	0			
PSU Gaussian	Mean	0			
PSU MEFA	Mean	0			
PSU SCIPUFF	Mean	0			

Table 5. Table of significant factors for stepwise regression. As in the previous Table 4, Table 5 is divided into two sections, one for each dependent variable. Each section contains the proportion of variance explained by regression (adjusted R2), independent variables selected by stepwise regression (see Section 2.2 above), standard coefficient for that variable, unstandardized coefficient, and significance level. All computations were performed using SPSS 15.0 with an entry criterion of 5% and an elimination criterion of 10% significance as determined by the appropriate partial F-tests. Background cell colors are used to designate and group individual variables that were considered significant by the regression analysis.