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

Nathan Platt, Steve Warner and Steve M. Nunes

Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311, USA

Abstract: Given a warning based on detection of hazardous materials at just a few sensors, it could be useful to rapidly (minutes) 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 impacted by the hazardous release, and can support near-term follow-on actions to investigate the cause and nature of the hazardous release. In September 2007, a short-range, highly-instrumented test was conducted at the U. S. Army's Dugway Proving Ground. This test, referred to as Fusing Sensor Information from Observing Networks (FUSION) Field Trial 2007, or simply FFT 07, was designed to collect data to support the further development of source term estimation algorithms. This presentation describes how the field trial data collected during FFT 07 is being used for investigating several prototype algorithms including the goals of the comparisons, the comparison protocol, and the design of the comparative evaluation matrix.

Key words: *sensor data fusion, source term estimation, short distance transport and dispersion*

1. INTRODUCTION

Given a warning based on detection of hazardous materials at only a few sensors, it could be useful to rapidly (minutes) 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 hazardous release and can support near-term follow-on actions to investigate the cause and nature of the hazardous 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 roads to avoid). For longer range situations (tens of kilometres), 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). This test, referred to as FFT 07, was designed to collect data to support the further development of prototype algorithms (Storwold, 2007). 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 50Hz, 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 50Hz that were positioned between the digiPIDs at lines 3 and 8.

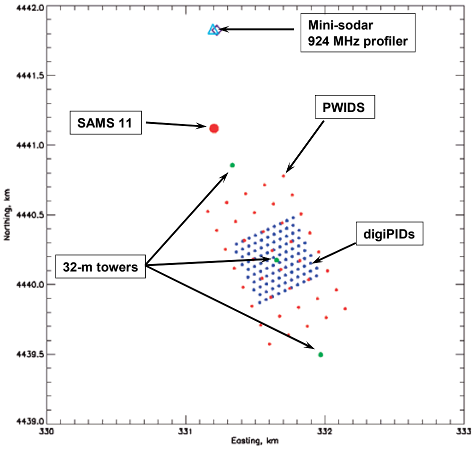

Figure 1. Illustration of a subset of instrumentation utilized during FFT 07 field trials. Blue dots denote locations of 100 digiPIDs used to continuously sample propylene concentrations at 50 Hz, small red dots denote locations of 40 PWIDs (Portable Weather Information and Display Systems) used to collect detailed surface meteorology, green dots denote locations of three 32-meter towers that carried additional meteorological instrumentation, the large red dot denotes the location of 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.

There were several reasons for conducting FFT 07. First, the experiment was meant to provide a set of data that the STE model developers can use to improve their algorithms. A second reason for FFT 07 was to use the collected information to assist in identifying the strengths and weaknesses of the different modelling 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").

2. PURPOSE

An evaluation plan is needed to allow for credible and fair comparisons among the prototypes and to best use the data collected during FFT 07 to further the development of STE algorithms, identify strengths and weaknesses of different approaches, and assess the "state-of the-art." A well-described and agreed-to plan will help avoid perceived intentional, or more likely unintentional, model parameter tweaking to fit the unique data and observations of FFT 07 in a way that might be considered unfair. Most importantly, an evaluation plan is meant to support the most credible assessment of the state of the art, and ultimately, to allow for the identification of scientific insights via careful comparative analyses. Comparative analyses are important, in part, because specific detailed STE accuracy and timeliness requirements do not exist. Therefore, assessment against clearly defined evaluation criteria is not possible. Rather, one seeks to identify which approaches work best and why and to define expectations for current algorithm performance.

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 be best applicable to different sets of tracer release scenarios (i.e., daytime versus nighttime, single- versus double- versus triple- versus quadruple-source releases, 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 design described later is an attempt to trade off the ability to cover the evaluation of a substantial number of potential variables that might affect algorithm performance with the desire to keep the sample sizes large enough to be able to arrive at reasonably robust conclusions.

This paper describes the plan designed to provide for comparative evaluations of a variety of prototype STE algorithms. Further details can be found in (Platt, et. al., 2008). This plan should be considered as detailed guidance, but only guidance. That is, it is understood that changes to the protocol, depending on findings, may be required over the course of this study. For instance, time and budget permitting, an additional stage (or stages) of the evaluation could be added. This additional stage could examine the effects of greatly increasing the numbers of meteorological observations relative to the size of the area and its effect on the quality of the STE and/or could help answer additional questions that might arise during the earlier evaluations.

3. OVERVIEW OF THE EVALUATION PLAN

This section provides an overview of the proposed evaluation plan. Three considerations are briefly highlighted to help focus the overview: Relevance, Data Obscuration, and Design of the Evaluation.

Relevance

In order to provide some relevance to the evaluation, at least some of the data provided to the STE algorithm developers must be representative of the data that would be available to drive STE algorithms in reality. For that reason, in addition to data from 16 sensors (potentially considered too much information for such a small area), data from 4 sensors will also be provided per tracer release for STE algorithm evaluation. Furthermore, the time resolution of the data will be no faster than 1 sample per second. Finally, the developers will not only be provided with time series data from sensors that record continuum concentration values, but also with data sets that simulate data from more realistic, near-term sensors that are only capable of recording where discrete concentration thresholds are exceeded. To provide some realism in the meteorological inputs to the STE algorithms, the developers will be 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, in addition to more detailed meteorological observations made at the centre location of the sampler grid itself.

Data Obscuration

To provide the community with the perception of a fair evaluation of the STE algorithms, a semi-blind protocol is required in which data from FFT 07 have been obscured in order to hide information that could be used to identify major source term characteristics ahead of the calculation or to tweak model parameters to fit the FFT 07 data. This requires the date and time stamps of the recorded sensor data to be altered, as that information can be used to exactly identify source term information from the FFT 07 web site. Furthermore, since an important part of STE is identifying the time of the release and sensor recording began near the start of the tracer gas release, false concentration data representing background noise must be added to the beginning of the sensor concentration time series in order to somewhat obscure the starting time of the release.

Design of Evaluation

The STE algorithm developers must be provided with data that not only sufficiently spans the space of the parameters that may affect algorithm performance, but also includes enough sample realizations of each parameter value so that statistically robust conclusions can be drawn. Furthermore, the total amount of data released to the developers must be manageable enough so that the proposed evaluation timeline can be met. The compromise between competing objectives is described as follows.

First, the evaluation is divided into two or more stages of manageable size. Each stage represents an evaluation of the STE algorithms using sensor data simulated to be of different fidelities: continuous concentration data, multiple-threshold concentration data, and, optionally, binary-threshold concentration data. During each stage of the evaluation, developers will be given sets of simulated sensor output data based on observed tracer data at selected sampling locations and meteorological input data designed to span the parameter space appropriately. It is desirable to repeat the data sets from stage to stage in order to provide fair comparisons of algorithm performance as a function of sensor fidelity; unfortunately, full repetition of all the released data from stage to stage would not provide enough sampling of each set of parameter values without significantly increasing the amount of data released in each stage to an unreasonable number of cases for which model developers would need to provide predictions. As a compromise, only a subset of the released data (the "control subset") is repeated from stage to stage.

Full sampling of the parameter space is ensured by a structured release of the data. During each stage of the evaluation, simulated sensor output data along with either the aforementioned "good quality" meteorological data or the "realistically degraded" meteorological data that were available for the accompanying tracer release will be provided. A set of tracer releases will be chosen that appropriately sample the number of tracer sources that were used during FFT 07 releases. For every combination of meteorological input type (good quality or more realistically degraded) and number of sources, one or more sets of sensor output data representing "scenarios" will be released. Each scenario consists of 8 "cases" representing the 8 permutations of the time of day of the tracer release (day or night), the type of tracer release (continuous or instantaneous), and the number of sensors reporting data (4 or 16). Each case is constructed from an actual FFT 07 release, and an STE algorithm prediction will have to be provided for each case. For non-control "scenarios," the 8 cases are chosen randomly to be independent from each other: they may represent the same FFT 07 tracer release or different releases, or they may be built using the same set of sensors or different sets. The cases chosen for "control" scenarios will not change throughout the stages of the evaluation with the exception of the quality of the simulated sensor outputs. Additionally, each 4-sensor case within the "control" scenario will use sensors that are a subset from the corresponding 16-sensor case. Figures 2 and 3 provide flowchart schematic definitions of "scenario" (in terms of individual cases) and "Stages" (in terms of individual "scenarios") of evaluations.

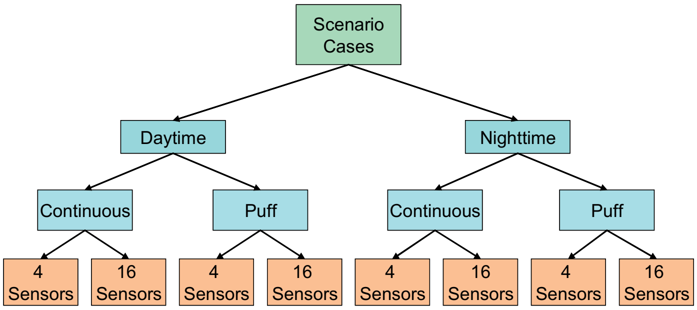

Figure 2. Flowchart diagram depicting the definition of a scenario.

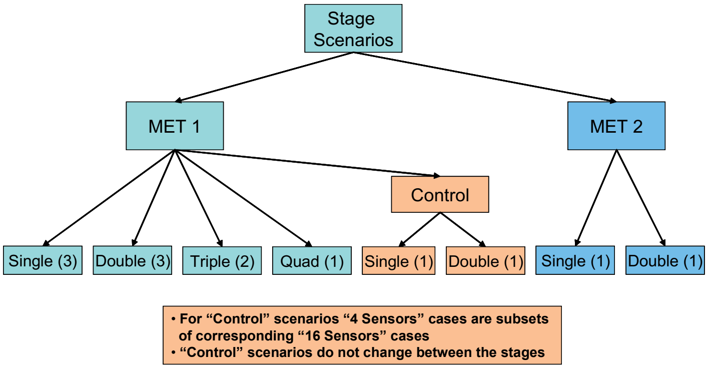
Stage Scenarios
MET 1
MET 2
Control
Single (3)
Double (3)
Triple (2)
Quad (1)
Single (1)
Double (1)
Single (1)
Double (1)
• For "Control" scenarios "4 Sensors" cases are subsets of corresponding "16 Sensors" cases
• "Control" scenarios do not change between the stages

Figure 3. Flowchart diagram depicting definition of a stage of the evaluation. Nomenclature "single", "double", "triple" and "quad" at the bottom level denotes cases with the named number of sources used in the definition.

To provide algorithm developers with data sufficient to test and develop their models without biasing the evaluation, a "raw" subset of the full tracer and meteorological data collected for selected subset of FFT 07 releases will be released to the developers and is meant to be used outside of this comparative protocol in any way that the developers choose. These data will not be altered to conform to the evaluation protocol described here. This fully released subset of data is expected to be small enough to not permit the tweaking of model parameters to fit the FFT 07 data.

4. DEMONSTRATION OF CREATION OF SIMULATED SENSOR OUTPUT

Each case selected for algorithm evaluation will contain simulated sensor outputs that will be created using available digiPID observations. Each case will contain at least two simulated sensors that were "hit," and at least one "null" sensor. 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 locations where the concentration threshold is exceeded is constructed as shown in Figure 4. The coloured circles denote digiPIDs whose concentration exceeds the colour-coded concentration value (in ppm) shown in the legend on the left for any 1-second bin (i.e., "hit" sensors), while the black open circles denote the rest of the digiPIDs (i.e., "null" or "broken/not working" 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 to corners of the rectangular area. In this case, we chose three hits and a single null.

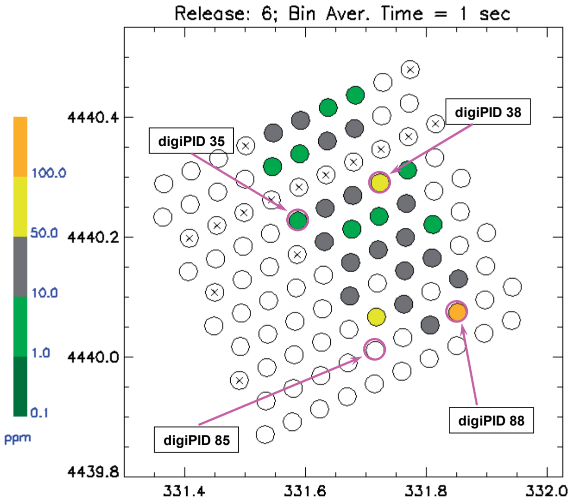

Figure 4. Notional demonstration of selecting digiPIDs to simulate sensor output.

Given 50-Hz propylene concentration time-series for a selected subset of digiPIDs to simulate sensors at each case, there is a need to translate this information into the sensor output expected from different types of sensors. We assume a somewhat idealized and perhaps futuristic situation in which the sensor sends a continuous stream of data. Due to practical limitations on available data frequency bandwidth (e.g., on the battlefield), it is quite unlikely that raw 50-Hz data would be fed directly to some central location. Thus, the raw digiPID data needs to be reduced to some lower frequency, for example, a single data point transmitted each second (1 Hz). Next we show how digiPID data could be translated into simulated sensor output to be used with sensor fusion algorithms. As shown in the previous section, four digiPIDs (35, 38, 85, and 88) are used to simulate four operational sensors. Figure 5 depicts 1-second bin-averaged concentration data for the four selected digiPIDs. For the second and optionally third stage of the exercise, given the high frequency of the available digiPID data, it will be straightforward to translate continuous 1-second backward bin-averaged data into eight-level "bar"-like simulated sensor output by using eight separate thresholds corresponding to each bar.

To discourage any appearance of perceived data manipulation by algorithm developers and to improve the results of the evaluation, a data-obscuring mechanism is proposed. This obscuration mechanism should be simple, in terms of the referee's data manipulation and adequate for this purpose. We implemented the following procedure to simulate up to 10 minutes of pre-release concentration data for each case. For each digiPID selected to represent a sensor, a minute or so of the pre-release 50-Hz raw data is randomly "block" re-sampled to create longer term simulated 50-Hz data. Random positions are used to select the beginning of each block and then these fixed-length blocks are appended together to create a longer data sequence. Using the trial and error method, we found that a block size of 250-500 data points (5-10 seconds) produces visually reasonable approximations of the original pre-release data when backward bin-averaged to 1 Hz. Figure 6 demonstrates this procedure to create 10 minutes of simulated pre-release data when applied to the digiPIDs selected in Figure 5. Finally, Figure 7 demonstrates the full procedure for the stage 1 continuous simulated sensor output. The blue line depicts 10-minute, pre-release “padding” of the simulated sensor output shown in Figure 5.

5. OUTPUTS TO BE PROVIDED BY MODELERS

Estimated source terms from the participating models are to be compared (e.g., differences are computed) to actual source terms and to each other. These comparisons will focus on identifying test conditions (scenarios and cases) where relative model performance was improved or degraded. For each individual case, the minimal information provided by any participating algorithm developer should include: 1) best estimate for the source location(s) (x and y) at the concluding time of each case, 2) source type, strength, and number of sources, 3) release start time, 4) for continuous releases, release end time (or duration) should also be provided. Additionally, it is highly desirable that algorithm developers provide the following information, if available: 1) uncertainties in the source estimation, especially for location, 2) for algorithms that are based on a continuous estimate of the source term as a function of the simulated sensors’ time history, a time history of the STEs is requested (especially location and strength), 3) for algorithms that use some form of transport and dispersion code to simulate backward and forward propagation of the tracer gas, a concentration time-history at every digiPID/UVIC location using their own estimated source terms as input is requested.

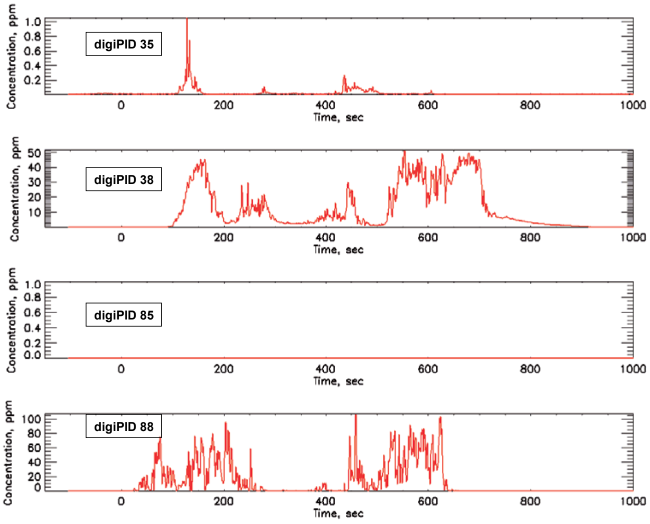

Figure 5. 1-Second averaged concentration time series for selected digiPIDs.

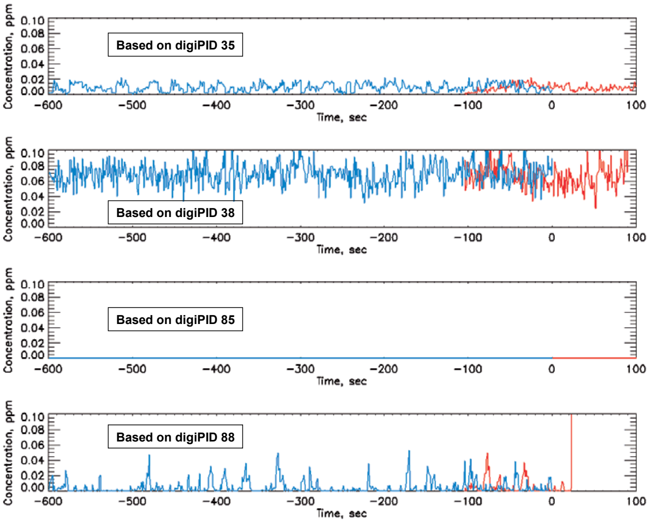

Figure 6. Demonstration of 10-minute simulated pre-release data (red lines correspond to 1-second, bin-averaged observed concentrations and blue lines correspond to 1-second, bin-averaged, 10-minute, simulated pre-release data).

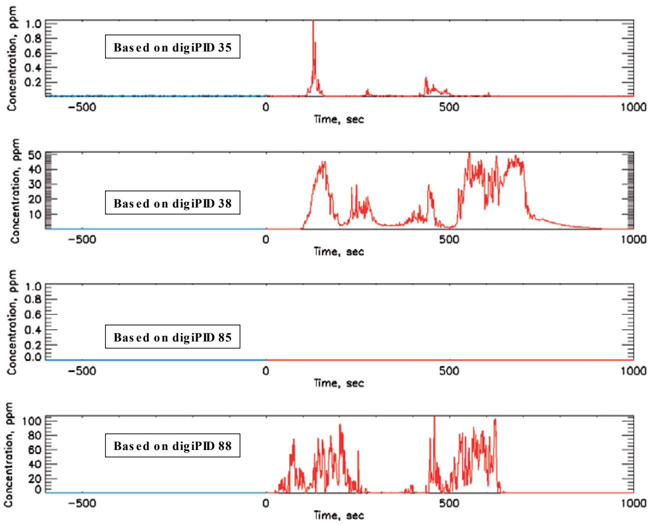

Figure 7. Demonstration of 10-minute initial padding of the simulated sensor output.

Acknowledgments: *This effort was supported by the Defense Threat Reduction Agency with Dr. John Hannan as the project monitor and the Central Research Program of the Institute for Defense Analyses. The authors 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 this plan. The authors also acknowledge the comments and discussions of the FFT 07 Management/Science Team, especially Dr. Ray Hosker (Hosker Science Services) for providing a careful review of this plan. The views expressed in this paper are solely those of the authors.*

REFERENCES

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

Platt, N., Warner, S. and S.M. Nunes, 2008: Plan for Initial Comparative Investigation of Source Term Estimation Algorithms Using FUSION Field Trial 2007 (FFT 07), Institute for Defense Analyses Document D-3488.