2009 IEEE International Conference on Data Mining Workshops

Spatiotemporal Modeling and Monitoring of Atmospheric Hazardous Emissions using Sensor Networks

Guido Cervone
*Dept. of Geography
and Geoinformation Science
George Mason University
Fairfax, VA, USA
gcervone@gmu.edu*

Anthony Stefanidis
*Dept. of Geography
and Geoinformation Science
George Mason University
Fairfax, VA, USA
astefani@gmu.edu*

Pasquale Franzese
*Center for Earth Observing
and Space Research
George Mason University
Fairfax, VA, USA
pfranzes@gmu.edu*

Peggy Agouris
*Dept. of Geography
and Geoinformation Science
George Mason University
Fairfax, VA, USA
pagouris@gmu.edu*

***Abstract*—A spatiotemporal methodology is presented for the analysis and visualization of atmospheric emissions in a metropolitan area. Numerical transport and dispersion models are used to build a library of time-dependent emissions of hazardous gases under various atmospheric conditions and from multiple potential sources in Washington DC. This library comprises representative emergency events that may involve natural or man-made hazardous emissions. To represent and analyze the events of this library we use the model of the spatiotemporal helix, which provides concise summaries of complex spatiotemporal events. We demonstrate the ability to compare emerging situations to library entries in order to predict their future evolution, thus recognizing potentially hazardous conditions early in their development.**

I. Introduction

Accidental or intentional releases of chemical, biological and nuclear agents in the atmosphere can have a a serious impact on health. A fundamental problem associated with the analysis of atmospheric emission data is the identication of the characteristics of the source such as, e.g., its location and emission rate [1]. This application requires the spatiotemporal analysis and visualization of data which are often rapidly changing, incomplete, and very noisy. The identication of the source characteristics has become a primary interest of national security. The accuracy of a forward transport and dispersion simulation (from source to sensor) of a toxic gas depends on a number of factors, such as the scale of the phenomenon, the accuracy of the source term, the availability and representativeness of meteorological data, the coverage of the sensor network and the averaging times of its measurements, and the approximations inherent in the numerical model in order to perform the simulation in a real- istic time frame and with a realistic data storage capability. Even in controlled eld experiments, model simulations can at best give only an approximate representation of the evo- lution of an atmospheric contaminant. Backward transport and dispersion simulations (from sensors to source) are even more uncertain.

To complement the capabilities and overcome some of the limitations, numerical models are ever more frequently paired with methods from statistics, computer science and machine learning. This trend is fueled by the increasing availability of fast and scalable algorithms from different disciplines. Bayesian methods aim at an efficient execution of an ensemble of forward simulations, where statistical comparisons with observed data are used to improve the estimates of the unknown source location [2]. These methods consist of forward dispersion simulations from each candidate source, where the goal of the algorithms is to minimize the error between simulated and measured concentrations. These methods are independent of the type of model used, the type and amount of data, and can be applied to non-linear processes as well. Monte Carlo algorithms also have the advantage of working well with real-valued attributes, which are the most common type in a source detection problem.

Powerful methodologies based on Bayesian inference coupled with Monte Carlo based stochastic sampling were employed to reconstruct atmospheric contaminant dispersion [3], [4], [5]. A similar line of research aimed at using directly Monte Carlo simulations to minimize the error between simulated and measured concentrations was proposed by [6].

A similar approach was followed by [7], [8], [9], who use an iterative process based on genetic and evolutionary algorithms (GAs) (e.g., [10], [11]) to find the characteristics of unknown sources. They perform multiple forward simulations from tentative source locations, and use error functions to quantify the agreement between simulated and measured concentrations. An extension of evolutionary algorithms was used by [1], where an evolutionary process guided by machine learning was employed to identify the source characteristics of atmospheric emissions.

Although these models proved to work well, there is a lack of an efcient spatiotemporal representation for the analysis and visualization of the emission information, namely, they do not take full advantage of the spatiotemporal information associated with emissions. In this paper we investigate the use of our spatiotemporal helix structure [12], [13] to represent and analyze atmospheric emissions. We are particularly interested in instantaneous release of contaminants, as they

978-0-7695-3902-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ICDMW.2009.67

IEEE computer society

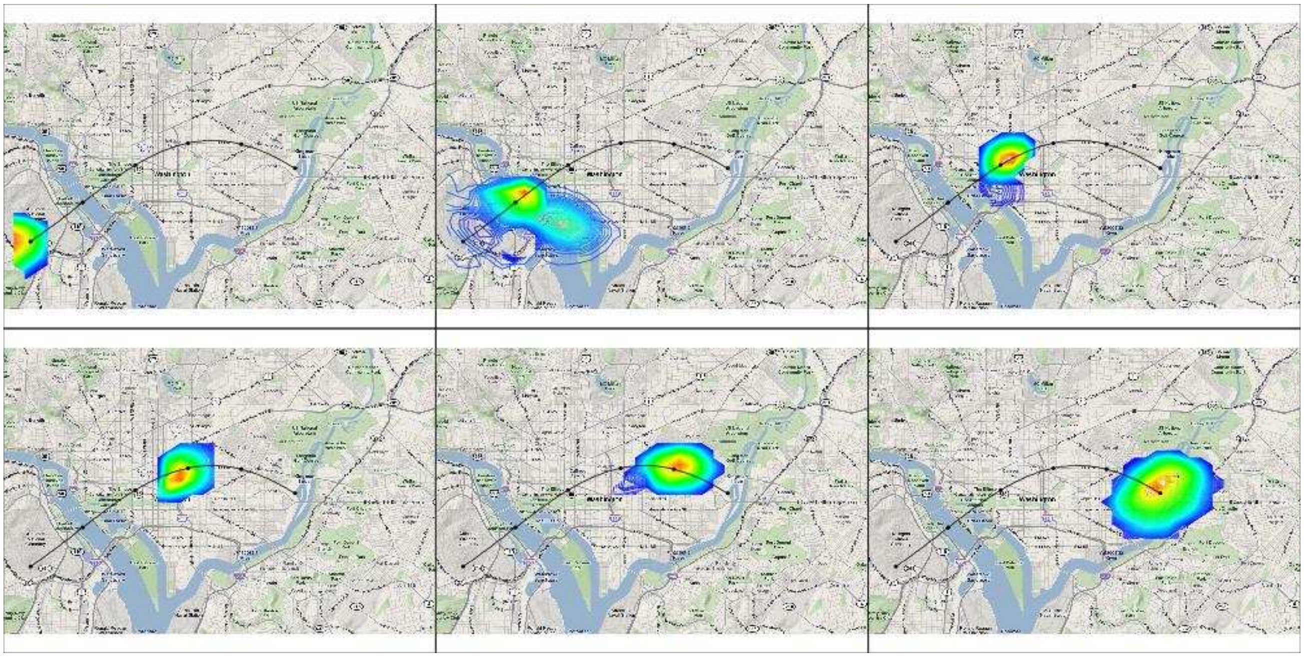


Figure 1. A spatiotemporal event: distribution of a pollutant in the atmosphere.

may be monitored by sensors distributed in a metropolitan area.

The spatiotemporal helix is a framework for describing and summarizing spatiotemporal phenomena. Designed to allow efficient querying of data, the strength of the spatiotemporal helix lies in its ability to simultaneously describe both an objects movement and deformation through time. The helix representation was extend to characterize the avarage concentration of pollutants in the toxic cloud. The helix 3D structure was applied to discover correlations between 500 simulated emissions in Washington DC. The goal is to find the maximum correlation between the emissions in the library of pre-computed cases and a new event. In this manner, developing events are compared to a library, and provide early warning for emerging situations of interest. This forms the basis for a broader integrated framework for situation awareness and monitoring using atmospheric transport and dispersion models, networks of ground sensors, and spatiotemporal modeling and analysis through helixes.

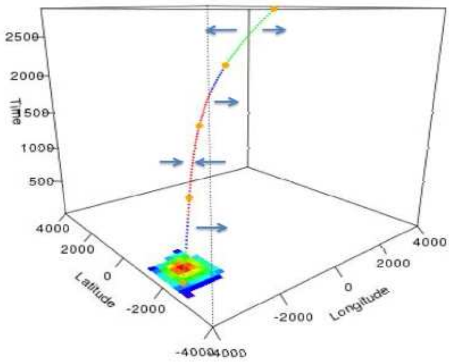


Figure 2. 3D Helix representation of a release. The concentration field represents the cumulative ground deposition of contaminant particles. The helix color indicates the acceleration of the contaminant cloud as increasing (blue) or decreasing (red). The yellow circles show the break-points, the locations of significant changes in the helix structure.

II. METHODOLOGY

The first task consists of building a library of simulated or observed emissions for a specific area, under different meteorological conditions. The scope of the library is to sample as extensively as possible all potential outcomes of an accidental release, which can occur anywhere within the domain boundaries. In some cases, intelligence analysis (e.g., of potential terrorist attacks) can reduce the search domain by identifying potential locations, such as, for example, power plants, chemical complexes, or sensitive buildings.

Each emission consists of a time-dependent cloud which travels and disperses three dimensionally through the atmosphere. Each cloud is represented by an helix structure, which is defined in Section II-B. The advantage of the helix is to provide a compact representation which preserves both spatial and temporal information, but at the same time provides an ideal way to analyze and visualize the emissions. Each contaminant cloud is also associated with a threat map, depending on the sensitivity of the area covered.

When a new emission is detected by ground sensors a concentration field is quickly reconstructed, and converted to helix structure. A correlation coefficient is computed between the new concentration field and those stored in the library. When the correlation coefficient is above the acceptance threshold, a classification is made.

A. *Transport and Dispersion Simulations*

The numerical simulation are performed using a reflected Gaussian plume dispersion model. Each simulation requires 8 input variables: x, y, z, θ, U, Q, S and ψ. x, y, and z are the coordinates of the release in m; θ and U are respectively the wind direction and speed in degrees and ms^{-1} and are time-dependent variables; Q is the source strength in gs^{-1}; S is proportional to the area of the release in m^{2}; and ψ describes the atmospheric stability according to Pasquill's stability classes [14], [15].

The concentration at the sensors C_s are simulated using a 3D Gaussian dispersion model:

$$C_s = P_1 P_2 P_3 (P_4 + P_5) \tag{1}$$

where

$$P_1 = \frac{Q}{U\sqrt{(2\pi)^3 (S+\sigma_x^2)(S+\sigma_y^2)(S+\sigma_z^2)}} \tag{2}$$

$$P_2 = \exp\left[-\frac{(x-x_0)^2}{2(S+\sigma_x^2)}\right] \tag{3}$$

$$P_3 = \exp\left[-\frac{(y-y_0)^2}{2(S+\sigma_y^2)}\right] \tag{4}$$

$$P_4 = \exp\left[-\frac{(z-z_0)^2}{2(S+\sigma_z^2)}\right] \tag{5}$$

$$P_5 = \exp\left[-\frac{(z+z_0)^2}{2(S+\sigma_z^2)}\right] \tag{6}$$

where $\sigma_x(x, x_0; \psi)$, $\sigma_y(x, x_0; \psi)$, and $\sigma_z(x, x_0; \psi)$ are the dispersion coefficients, which were computed from the tabulated equations of Briggs [16].

B. *Helix Spatio-Temporal Representation*

Lets consider a spatiotemporal phenomenon, like the dispersion of a pollutant in the atmosphere, moving from its point of origin in Arlington, VA, towards DC, as it was captured by a network of ground sensors, overlaid on a corresponding map (Figure 1). It can be easily understood that this could also reflect a similar spatiotemporal phenomenon captured e.g. by satellite remote sensing, as the proposed methodology is not application- or data-dependent, but instead has broad usage potential.

Figure 2 shows a summarization of this phenomenon in the form of its spatiotemporal helix. This spatiotemporal helix comprises two components [12]. The first is a central spine, which depicts the trajectory of the entitys center of gravity, and variations in the entitys attributes (density for this example). It is defined by a series of nodes, $s_i = \{x, y, t, q_m, q_a\}$ where x, y, and y are the nodes spatiotemporal coordinates, q_m is a qualifier that is characterizing the nodes dominant type of movement (acceleration, deceleration, or rotation), and q_a is a qualifier that is characterizing the nodes type of attribute variations (e.g. the event becomes more or less dense). The second component of the helix are the *prongs* protruding from the spine (represented by the arrows in Figure 2). Prongs, $p_i = \{t, r, q_1, a_2\}$, describe an objects deformation, and are defined in terms of the time coordinate t, the magnitude of outline change r (with positive values for r indicating expansion and negative values indicating contraction), and a_1 and a_2 denoting the azimuth range of the deformation. Prongs exist independent of nodes, and appear at any time t along a helixs spine with outward facing arrows indicating expansion and inward facing arrows indicating contraction. A helix comprises spine and prong information to describe an events spatiotemporal behavior. As such, any given helix can be expressed as an aggregate of nodes and prongs as: and can be stored in a geospatial or standard open-source databases (e.g. PostgreSQL).

Combined, spine and prong information provide a concise signature of an events spatiotemporal behavior, capturing for example the changing shape of the pollutant cloud in our example. The movement of a car will be represented through a spine only (as the car does not change its shape).

C. *Helix Correlation and Classification*

The degree of correlation between helixes is computed by mean squared difference (MSD) of the normalized internal representation of the helixes (Equation (7)).

$$MSD = \sum_{i=0}^{n} (x_i^a - x_i^b)^2 + (y_i^a - y_i^b)^2 + (q_{ai}^a - q_{ai}^b)^2 + (q_{mi}^a - q_{mi}^b)^2 \tag{7}$$

where n is the length of the helixes, x, y, q_a, q_m are the dimensions of the helix, and a, b are the two helixes being compared.

First, the internal representation of the helix is normalized to weight each variable equally. Then the correlation is computed by aggregating the MSD of all corresponding tuples in the helixes. If the helixes have different lengths, due to a different travel time of the respective releases, multiple similarities are computed by considering all possible combinations, where the smallest helix is compared with the different segments of the longer helix, until all possible combinations are considered. The result correlation is the smallest of all possible correlations.

Comparing helixes of different lengths occurs frequently when new releases need to be classified within the smallest

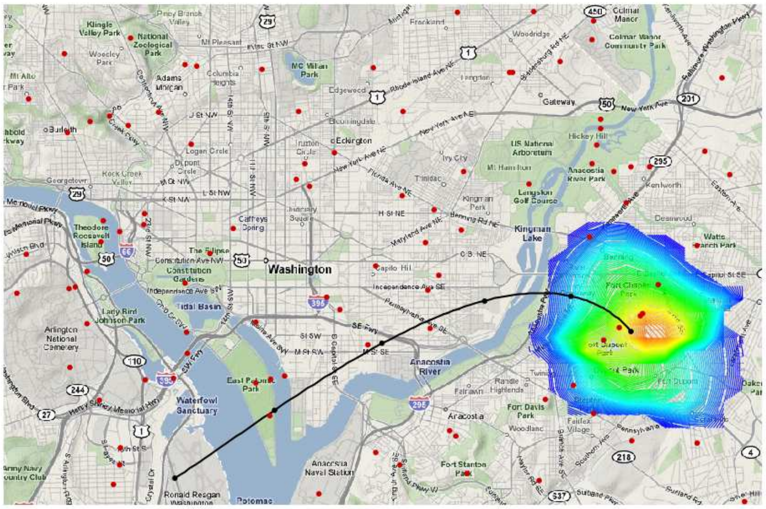


Figure 3. Terrain map for the domain of the study, and sample contaminant cloud and detected trajectory. The location of the sensor receptors is also shown.

possible time. Therefore although the emissions in the library can be simulated for a long time, up to hours or days, information on new releases is assumed to be available only for few minutes. Furthermore it is not safe to assume that the first temporal observation of the release corresponds to its origin, since it might have avoided previous detection due to wind patterns or lack of proper sensor measurements. Because such assumption cannot be made, the shorter temporal observation, which corresponds to the new accidental release, is compared to all helixes in the library at all times, and not just to their origins.

Once a degree of correlation between the helixes is calculated, the classification calculations are based on threshold (T) and tolerance parameters (τ). The tolerance range, specified by the τ parameter, is the percentage by which the degree of match can fall below the top degree of match threshold T and be selected as an alternative decision. If a helix matches more than one in the database within the tolerance range, then corresponding decisions are presented as alternative decisions. The acceptability threshold is defined as the minimum degree of match for which a specific classification decision will be made. If all degrees of match fall below the acceptability threshold, “no decision” (or “don’t know”) is returned.

The classification mechanism used is designed to go beyond the typically used predictive accuracy measure that assumes that for each testing event, the system assigns a single classification decision that is either correct or incorrect. This single-decision schema makes the evaluation of a method simple, but is not adequate for many real-life problems. The helix approach assumes that it is better to give a few alternative answers (“multiple decisions”) that include the correct decision, or to give no definite answer (“no decision”), than to be forced to give just one answer and be incorrect.

The idea of producing “multiple decisions” or “no-decisions” is implemented by computing degrees of match between a new helix, and the library of helixes, and, depending on the distribution of these degrees, computing the final classification decision. Consequently, an evaluation of the performance consists of not just one number, but several numbers, each with a simple cognitive interpretation. Specifically, the output includes the following measures:

- Predictive Accuracy-S ($\mathrm{P^S}$)
- Predictive Accuracy-M ($\mathrm{P^M}$)
- Ambiguity-S ($\mathrm{A^S}$)
- Ambiguity-M ($\mathrm{A^M}$)

The predictive accuracy scores are success metrics, expressing our success in finding the correct response. The ambiguity parameters are scores of the uniqueness of the responses, thus describing how easy it is for an analyst to evaluate the results presented to him/her.

Predictive Accuracy-Single refers to the percentage of correct classifications which receive the highest degree of match. If more than one candidates have tied for the highest matching score and the right response in included in this group, the answer is considered correct. Thus this metric is a variant of the standard measure of predictive accuracy when multiple top matches are returned. In testing situations, when a helix is compared to a set of candidates that includes itself (e.g. during cross validation), it is expected that $\mathrm{P^S}$ will always be 100%, assuming that the matching algorithm is fundamentally correct.

Predictive Accuracy-Multiple applies to situations where the response to a classification returns not only the highest score, but all results that exceed threshold T but are within tolerance τ of the best match. In these situations, P^{M} denotes the percentage of having the correct classification within the returned matches. Thus, the Predictive Accuracy-M reduces to Predictive Accuracy-S when both T and τ are 0. All helixes that return degrees of match both above the threshold and within the tolerance of the highest degree of match attained by the new helix are returned as possible classifications.

The ambiguity metrics account for cases of multiple decisions. Ambiguity is calculated both for P^{S} and P^{M}. It is possible that even in the PS case the answer may include more than one choice, namely if two or more decisions tie for the highest degree of match. The ambiguity metric is defined as log(1/e) where e is the number of events that are returned. Accordingly,ambiguity is equal to 0 when only one solution is chosen, and its absolute value increases with the number of returned alternative solutions. A^{M} reduces to A^{S} when τ = 0.

III. Results

We simulate a network of ground sensors distributed over our area of interest (See Figure 3) and record concentration measurements at these locations over different time instances. In this manner we simulate the measurements of a sensor network distributed in the DC metropolitan area. The concentration field is then reconstructed using only the measurements at these locations. In addition to the simulated sensor positions, Figure 3 shows the study domain, a sample contaminant cloud at the end of its 60-minute path, reconstructed from sensor measurements, and the track of its center during this 60 minute period. At various time instances we can reconstruct the instantaneous record of the concentration fields from the sensor measurements. By analyzing these records we generate the trajectory, average concentration and speed of each release and encode this information into spatiotemporal helixes as presented in Section II-B. These 500 simulations form a library of atmospheric releases for our area of interest.

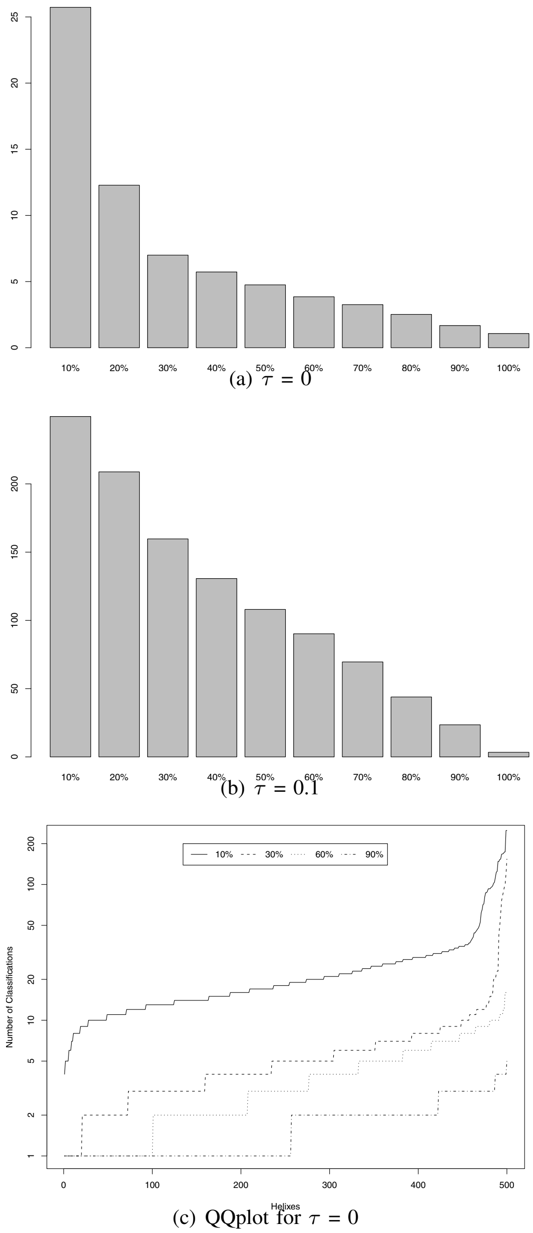
To assess the performance of the proposed methodology, cross correlation values were computed between all helixes in the libraries. Therefore each helix is compared to itself, and to all other 499 helixes. Therefore, in order to achieve high accuracy and low ambiguity, each helix must obtain the highest correlation when matched only with itself. Our particular interest is to assess how well we can predict the future evolution of a specific emission by comparing it during its initial stages to our library of 500 events. Accordingly we performed tests using various helix sample sizes, ranging from 10% (corresponding to using sensor data collected during the first 6 minutes only of an emission event) to 100% (corresponding to using sensor data collected during the complete 60-minute emission event). The tests were designed to test for how long an emission must be observed in order to be correctly classified. Ideally, a release is unambiguously recognized in the shortest possible time.

The classification parameters used are T=0 and τ=0.1. In all cases the accuracy is 100%, as the correct answer is always included among the solutions. However, the ambiguity varies depending on both the size of helix and by the size of the threshold τ. Figure **??**a,b summarizes the average number of solutions for τ=0 (top) and τ=0.1 (middle) for different helix sizes. Ambiguity metrics (A^{S} and A^{M}) are computed as log(1/e) where e is the number of events that are returned. For τ = 0, when only a small part of the helix (105) is used, we have on the average 25 matches returned as events that are potentially similar to our observed samples. Of these 25, one is the correct answer, and the other 24 are incorrect solutions. Therefore, although the correct answer is among the solutions, the result has an ambiguity of log(1/e)= -1.398. Similarly, for τ=0.1, when only 10% of the helix is used, about 250 helixes are identified as potentially similar to our sample, yielding to an ambiguity of -2.40. In both cases, the ambiguity is drastically reduced with longer sample measurements. Figure **??**c shows τ= 0 as a function of helix size. The plot represents the distribution of classifications for the 10%, 30%, 60% and 90%. When longer helixes are used, the distribution of alternative solutions also varies greatly. In particular, there is always about 5% of the helixes that perform considerably worse than the average. This 5% of events correspond to helixes which occurs with very variable wind, lack very distinct features that can uniquely help the classification.

IV. Conclusions and future work

In this paper we presented a spatiotemporal methodology for the analysis and visualization of atmospheric emissions in a metropolitan area. The objective was to devise an approach based on spatiotemporal analysis for the early detection of potentially hazardous situations. We built a library of 500 emission events using numerical transport and dispersion models, under various atmospheric conditions and from multiple potential emission sources in the Washington DC metropolitan area. This library comprises representative emergency events that may represent natural or man-made hazardous emissions. To represent and analyze the events of this library we used the model of the spatiotemporal helix, which provides concise summaries of complex spatiotemporal events. We demonstrated successfully the ability to compare emerging situations to our library of events, and thus predict their future evolution of evolving events.

In a practical application we expect analysts to go over a library of events like the one we generated, and identify in it the ones they consider semantically important (e.g. due to their path crossing points of interest, or due to their high density levels). Developing events can then be compared to that library using the methodology we presented in this paper, and provide early warning for events of interest. Our



(a) τ = 0

(b) τ = 0.1

(c) QQplot for τ = 0

experiments addressed the accuracy and ambiguity of the matching technique (and thus of the corresponding early warning system) and the results are very promising. Even though we considered a specific type of application in this paper, namely atmospheric emissions, the approach is application-independent and thus can be used to model and analyze any other event that has an evolving spatiotemporal footprint (e.g. flooding, wildfire), especially as they are captured in a geosensor network.

Although the method was presented in the framework of an urban environment, the same method can apply to atmospheric emissions at different scales, including mesoscale and synoptic scale.

V. Acknowledgments

Work performed under this project has been partially supported by the NSF through Awards 0429644 and 0849191, and by NGA through a NURI grant.

References

[1] G. Cervone, P. Franzese, and A. P. Keesee, “AQ learning for the source detection of atmospheric releases,” *WIREs: Computational Statistics*, vol. in press, 2009.

[2] F. Chow, B. Kosović, and T. Chan, “Source inversion for contaminant plume dispersion in urban environments using building-resolving simulations,” in *Proceedings of the 86th American Meteorological Society Annual Meeting*, January 2006, pp. 12–22.

[3] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*. Chapman & Hall/CRC, 2003, 668 pp.

[4] L. Delle Monache, J. Lundquistand, B. Kosović, G. Johannesson, K. Dyer, R. Aines, F. Chow, R. Belles, W. Hanley, S. Larsen, G. Loosmore, J. Nitao, G. Sugiyama, and P. Vogt, “Bayesian inference and markov chain monte carlo sampling to reconstruct a contaminant source on a continental scale,” *J. Appl. Meteor. Climatol.*, vol. 47, pp. 2600–2613, 2008.

[5] I. Senocak, N. Hengartner, M. Short, and W. Daniel, “Stochastic event reconstruction of atmospheric contaminant dispersion using Bayesian inference,” *Atmos. Environ.*, vol. 42, no. 33, pp. 7718–7727, 2008.

[6] G. Cervone and P. Franzese, “Stochastic source detection of atmospheric releases,” *Computers & Geosciences*, no. submitted, 2009.

[7] S. E. Haupt, “A demonstration of coupled receptor/dispersion modeling with a genetic algorithm,” *Atmos. Environ.*, vol. 39, no. 37, pp. 7181–7189, Dec. 2005.

[8] S. E. Haupt, G. S. Young, and C. T. Allen, “A genetic algorithm method to assimilate sensor data for a toxic contaminant release,” *Journal of Computers*, vol. 2, no. 6, pp. 85–93, August 2007.

[9] C. T. Allen, G. S. Young, and S. E. Haupt, “Improving pollutant source characterization by better estimating wind direction with a genetic algorithm,” *Atmos. Environ.*, vol. 41, no. 11, pp. 2283–2289, 2007.

[10] J. Holland, *Adaptation in Natural and Artificial Systems*. Cambridge, MA: The MIT Press, 1975.

[11] K. De Jong, “Evolutionary computation: a unified approach,” in *Proceedings of the 2008 GECCO Conference on Genetic and Evolutionary Computation*. ACM New York, NY, USA, 2008, pp. 2245–2258.

[12] P. Agouris and A. Stefanidis, “Efficient summarization of spatiotemporal events,” *Communications of the ACM,*, vol. 46, no. 1, pp. 65–66, 2003.

[13] A. Croitoru, K. Eickhorst, A. Stefanidis, and P. Agouris, “Spatiotemporal event detection and analysis over multiple granularities,” *patial Data Handling*, pp. 229–245, 2006.

[14] F. Pasquill, “The estimation of the dispersion of windborne material,” *Meteorol. Mag*, vol. 90, no. 1063, pp. 33–49, 1961.

[15] F. Pasquill and F. Smith, *Atmospheric Diffusion*. Ellis Horwood, 1983.

[16] P. S. Arya, *Air pollution meteorology and dispersion*. Oxford University Press, 1999.