# Plume Source Detection Using a Process Query System

Glenn T. Nofsinger and Keston W. Smith
Dartmouth College, 8000 Cummings Hall, Hanover, NH 03755 USA

## ABSTRACT

A Process Query System (PQS) has the capability of filtering large volumes of real time data originating from a field of networked Physical Sensors. Modern air quality monitoring techniques such as Fourier Transform Infra-Red (FTIR) spectroscopy will eventually provide massively distributed real time contamination data at high fidelity. As large networks of these sensors are deployed, improved techniques of data retrieval and assimilation will be required. The case of detecting a diffusion event such as a hazardous chemical plume is considered. In this scenario, a plume model based on an Ensemble Kalman Filter (EnKF)is submitted to the PQS which manages multiple hypotheses explaining the current observations. The feasibility of such an application is demonstrated and results from preliminary simulations are presented.

**Keywords:** Process query system, PQS, plume tracking, ensemble Kalman Filter

## 1. INTRODUCTION

### 1.1. Problem Description

Given a diffusion event in a planar space (assuming no wind and an ideal diffusion process) can a network of chemical sensors identify and locate the source of a distributed diffusion event? (Where distributed means large plume events originating from a single source being partially detected at multiple sensor locations). Using inverse diffusion modelling from a set of chemical sensors, a system is desired than can process queries requesting the origin of of chemical plume and return likely hypotheses. Most efforts in plume modelling consider the forward prediction problem - we attempt to examine the inverse problem. The desired system will report plume origin as numerous hypotheses each having likelihoods.

### 1.2. PQS Framework

In future sensor networks, chemical concentration data will generally be accessible in real time. Unlike modern-day data mining, future data retrieval will need to function on high volume real-time data streams. In the case of a chemical attack, latency introduced by data mining must approach seconds or minutes as opposed to hours or weeks. This creates a new generation of data mining problems. With the added data loads introduced by monitoring continuous real time data, an intermediate layer of message handling is required to filter and broker data without losing information of interest.

PQS is the presented as the next generation of information retrieval, able to identify and retrieve data based on process descriptions. The comparison of PQS to traditional data retrieval has been compared to the evolution from Aristotelian to Newtonian logic (Cybenko et al). As an example application of PQS, chemical plume tracking illustrates how to rapidly filter, assimilate, and fuse large amounts of real time sensor data. Such capabilities will be needed to monitor water supplies, chemical weapons, and terrorist attacks. The city of Washington DC has installed a hard wired system to monitor downtown and the metro for hazardous airborne materials - but the need exists for a regional and national network capable of real time information exchange and exploitation.

PQS, rather than being a single application, is designed as a framework to accommodate domain experts' development of process models. By abstracting data collection, filtering, subscription, and pre-processing - the expert can focus on building better models. PQS has thus far been demonstrated with vehicle target tracking, and Cyber worm detection. Plume tracking was selected as a problem set to demonstrate the extreme versatility of PQS.

Author contact information:
Glenn Nofsinger: E-mail: gtn@dartmouth.edu, Telephone: 1 603 646 3332
Keston Smith: E-mail: keston@dartmouth.edu, Telephone: 1 603 646 0263

### 1.3. Existing Approaches

The problem of data assimilation in sensor networks has recently been addressed in the communities of collaborative signal processing (Akbar Illinois), sensor network databases(Gehrke Cornell), as well as the TinyDB project at The University of California. The CASA project directed by the University of Massachusetts Amherst seeks to create a distributed collaborative adaptive senor network for much denser sampling of the atmosphere when are where needed. Holzhauer et al with the Air Force Research Laboratory (project CPAKS) have begun the integration of plume simulations into a command and control schema based on agent technology.

While PQS may eventually work in conjunction with these existing technologies, the main differentiation is the step to true process modelling and a new generation in the data retrieval. By allowing the end user to build custom process descriptions a new generation of efficiency is enabled.

## 2. INVERSE DIFFUSION PROBLEM

The heart of the plume source detection challenge is a classic inverse problem: Given a set of observations that may not correspond to a unique set of events, how can the best possible guess be made about the hidden states?

The classic diffusion equation (Ficks Law )[1] describes the phenomenon of particle migration from areas of high to low concentration. This differential equation, identical to the "Heat Equation" can be solved numerically, however the numerical approach yields discontinuities and becomes problematic for discontinuous media. For this reason monte carlo methods are often used. For example, the identical results can be derived using a particle random walk method. In the following section the relevant analytical solutions of the this 2nd order partial differential equation (PDE) are presented and used to derive useful geometric results for plume source inversion.

### 2.1. Relevant Solution of the Diffusion Equation

Ficks Law states:

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial c}{\partial x}\right) \tag{1}$$

Assuming $D$ is constant:

$$\frac{\partial c}{\partial t} = D\frac{\partial c^2}{\partial x^2} \tag{2}$$

Because this is a 2nd order PDE, we need at least 2 initial conditions: typically fixed end-point concentration, or fixed flux. Using the boundary conditions: $c(0,t) = c(1,t) = 0$ the concentration at the end points (0 and 1) of the normalized graph space are held at zero. This approximation works for the case of plume tracking assuming the selected grid space is large compared to the plume size, and that the plume does not fully migrate to the edges. The 1D Solution:

$$r(x,t) = \frac{M}{4\pi D(t)}\exp^{\left(-\frac{x^2}{4Dt}\right)} \tag{3}$$

Where $M$ is amount of diffusing matter released per distance, $D$ is the diffusion constant, and $x$ the spatial variable. The same solution in 2D:

$$r(x,y,t) = \frac{M}{4\pi D(t-t_0)}\exp^{\left[-\frac{(x-x_0)^2+(y-y_0)^2}{4D(t-t_0)}\right]} \tag{4}$$

where

$$\sigma^2 = 2D\,(t-t_0) \tag{5}$$

Which means the variance of the plume distribution is linear with respect to time, but $\sigma = \sqrt{2Dt}$, so the

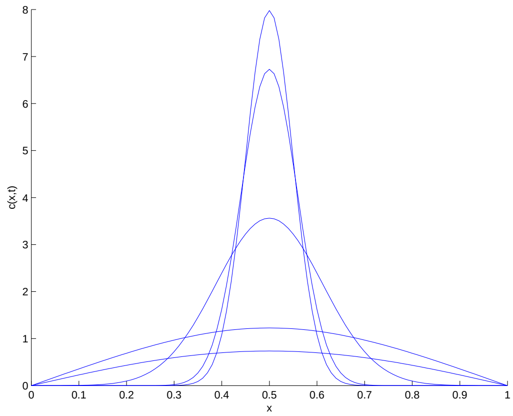


**Figure 1.** Ideal Particle concentrations which begin as a point source follow widening Gaussian distributions

spreading increases with time, never stopping but slowing down. The 95% perimeter of particles moves as $\sqrt{t}$ from the plume source. The characteristic solutions of 1D case for increasing times are plotted in Figure 1. Although this description of plume boundaries may be sufficient, very toxic pollutants may require the tracking beyond the bulk of the material ($3\sigma$). A more strict limit ($6\sigma$) or .13% on each side of the plume center may be required. Better, a random walk model can follow the precise edge of a distribution.

### 2.2. Inversion of Concentration Observations

Using the 1D diffusion equation solution and holding position constant provides the predicted observations of a chemical plume sensor as a function of time. This typical response curve is illustrated in Figure 2. Assuming the plume release is instantaneous and finite, the time of plume release may be described by the variable $T_0$. Depending on the distance of the chemical sensor to the plume source, a different concentration rise time will occur - until the concentration reaches a maximum at time $T_{max}$. Assuming a single plume source, we observe only a single local maximum in concentration, where $T_{max}$ occurs at the time identified by $\frac{\partial c}{\partial t} = 0$.

$T_{max}$ characterizes the distance from a sensor to the plume origin, assuming directionally uniform diffusion. Another interesting parameter is the ratio between $T_{max}$ and fall times, where fall time $T_{fall}$ is defined as the time required for concentration to be restored to a normal stationary level from the $T_{max}$ level. As distance from the source becomes large, this ratio $R$ changes characteristically with sensor-source separation distance.

$$R = \frac{T_{max}}{T_{fall}} \tag{6}$$

Inserting this value for $T_{max}$ into the diffusion equation solution along with position $x_n, y_n$ for sensor $n$ provides the maximum observed concentration at sensor $n$. This maximum concentration at sensor $n$ is defined as $Max_n$. In the 1D case an identified $T_{max}$ corresponds to a source probability distribution centered around two points equidistant from the sensor. When extending to 2D, an identified $T_{max}$ produces a source probability distribution around a circle of radius $r$ centered at $x_n, y_n$ for sensor $n$. In the non-ideal case this distribution would form a ring with an error term $\varepsilon_n$ for sensor $n$ describing the ring thickness. This simple inversion procedure allows for an estimation of distance between sensor and plume source in 2D. Goals for future work include accounting for non-uniform diffusion and external forces such as wind.

### 2.3. Inverse problem in 2 Dimensions

Assuming a 2D scenario of ideal Fickian diffusion a number of fundamental properties useful for inverse diffusion may be calculated. We also assume a single fixed plume source. We can estimate $T_{max}$ based on time delay between first observed increase and the time of maximum observed amplitude.

If more than one sensor is available, collaboration allows the elimination many possible source locations. The 2D approach is similar to the geometry used in the Global Positioning System (GPS) which leverages estimation

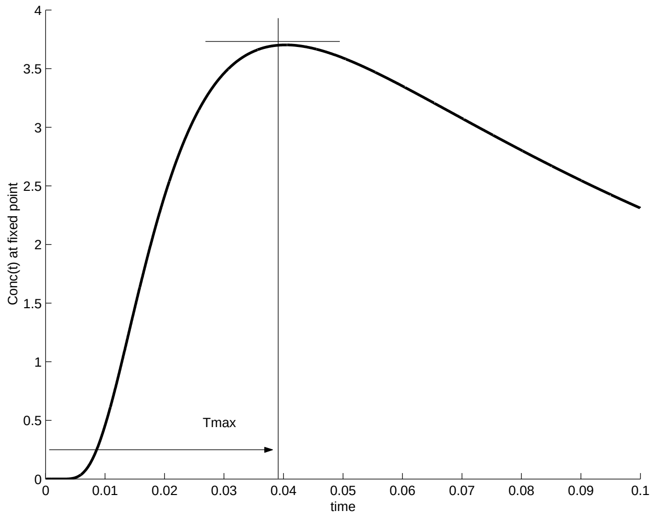


**Figure 2.** Characteristic Concentration as a Function of time observed from a fixed point

of distances to multiple satellites, and then finds regions of overlap between spheres. Based on concentration waveforms observed at individual nodes, path lengths to the plume sources are estimated. This pool of path lengths can then be exploited to enhance source location estimation.

We assume in these preliminary simulations that the time of initial release is known. (Which allows the calculation of $T_{max}$). Obviously in an application setting this assumption would not be valid, future work will investigate the estimation of $T_0$ based on the concentration rise time parameters. For example, the time between first observation above threshold and the time of the local maximum. In addition, agents within the sensor network may be able to share synchronization information providing more accurate $T_0$ knowledge throughout the network.

The general algorithm for finding a plume source in 2 dimensions requires:

- $M$ = Total plume mass
- find $\frac{\partial c}{\partial t} = 0$, solve for $T_{max}$
- Plug in this $T_{max}$ to solve for $Max_n = r\,(x, y, T_{max})$
- Calculate $Max_n = \frac{M}{L_n^2}$, solve for $L_n$
- Solve system of ratios iteratively
- $L_n$ = Length from source to each $n$ sensors results
- Combining solved values of $L_n$ can find $x_0, y_0$

With the above method, it is possible to solve for the 2D source location $x_0, y_0$ given 3 sensors of known location $x_n, yn$, where the length $L_n$ represents the distance between each sensor and the source. (See Figure 3). Using the ratio equation:

$$\left(\frac{Max_2}{Max_1}\right) = \left(\frac{L_1}{L_2}\right)^2 \tag{7}$$

we can iteratively find all the lengths in the sensor field that have detected an event originating from the same source. In $N$ dimensions, this approach requires knowing the location of $N + 1$ sensors for determination of the plume source.

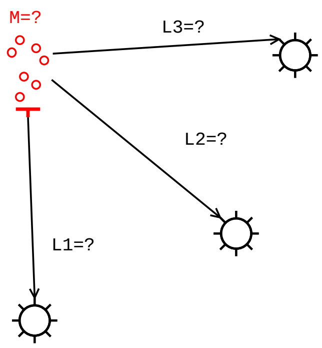


**Figure 3.** N-dimension plume source problem. Solving a simultaneous system of equations allows the calculation of L

Although these ideal geometric approaches may only allow crude approximations after assumptions are later removed - any regions that can be removed from the initial probability distribution will greatly enhance the performance of the EnKF.(refer next section)Once the geometric inverse algorithm has provided a diminished set of possible initial plume states, these constrained sets can be introduced as initial guesses (prior estimates) to a Monte Carlo Method known as EnKF.[3]

## 3. ENSEMBLE KALMAN FILTER IMPLEMENTATION

### 3.1. Introduction

We seek inverse methods whose implementations are forward model independent. One such family of schemes are Monte Carlo linear minimum variance methods, most widely used amongst these is the EnKF. We demonstrate the use of the EnKF and ensemble smoother (ES) with an application to chemical plume source detection. The advection setting is an urbanized circulation.

Data assimilation of large sets typically has the problem that plentiful observations (such as temperature) involve only a subset of the total variable space. Sensing units that measure a large number of atmospheric variables are uncommon and often separated by great distances. A recent[3] method introduced by Evensen to handle this situation is the Ensemble Kalman filter (EnKF). The EnKF applies a Markov Chain Monte Carlo method in which an ensemble of model states represent the probability density. A large cloud of model states, represented as points in state apace, corresponds to a specific probability density function. Integration of these model states in time according to the particular model dynamics can be used to estimate future probability densities of the system.

Leveraging the geometric diffusion analysis performed in section 2 reduces the total number of initial states involved in the forward integration. By reducing the total number of initial sates based on current observations (ruling out certain regions) the brute force forward calculation can be improved. For example, the concentration observations arriving at sensor $n$ require the plume source to be located within radius $r$ - creating a unique constraint disc centered at $n$. If a second sensor $m$ also has a unique constraint disc centered at $m$ - the constraint set produced as an intersection of $n$ and $m$ produces a greatly reduced number of initial states that must be integrated.

For algorithm testing, an EnKF simulation was developed in MATLAB which was able to predict plume source location based on distributed sensor observations.

### 3.2. EnKF Notation

- $\psi(x,t)$ model chemical concentration
- $\psi_0(x,0)$initial concentration
- $\psi^t$ true concentration

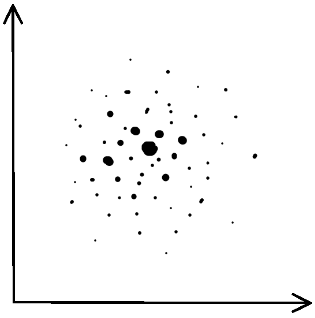

**Figure 4.** An ensemble of possible states can be represented in an n-dimensional phase space as a cloud of points, where each point represents an individual state.[3]

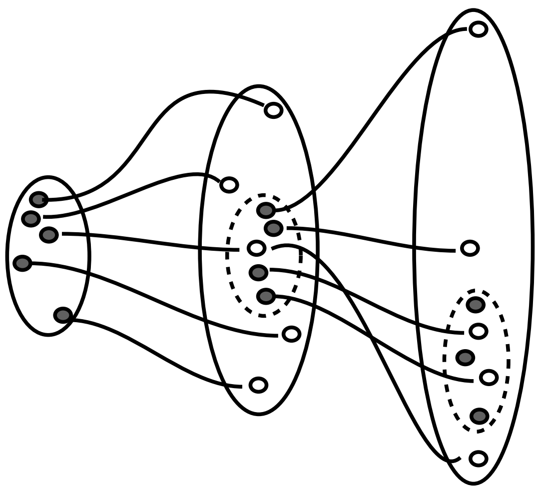

**Figure 5.** Two time steps evolve in an ensemble of initial phase space states. The EnKF integrates these individual points within an ensemble according to the particular process model. In this case the ensemble size (number of model realizations) remains fixed as the ensemble variance increases.

- $H$ measurement operator $d = H\psi^t + \xi$
- $d$ vector of data (concentration measurements)
- $V$ velocity field
- $D$ diffusion field

### 3.3. Ensemble Smoother

We seek to identify initial concentration from data recorded at fixed positions. The chemical concentration is assumed to evolve according to an advection diffusion equation:

$$\frac{\partial \psi}{\partial t} + V \cdot \nabla \psi - \nabla D \cdot \nabla \psi = 0 \tag{8}$$

We propose to solve the estimation problem with an ensemble smoother. The method is computationally expensive but provides useful error statistics and is simple to implement. (Alternatively one could solve the problem with as adjoint approach to the two dimensional advection diffusion equation). The ensemble smoother is a Monte Carlo approach to error variance minimizing estimation. The apriori model error distribution is defined implicitly through a stochastic perturbation model for the inital conditions and circulation. It is assumed that the distribution over intial conditions is Gaussian with some spatial structure.

A finite number, $n$, initial conditions are simulated from the prior model to create an ensemble of initial states $\{\psi_i^0\}$. The forward model is then run from each of the initial states to form the ensemble forecast:

$$\{\psi_i\} = \{f(\psi_i^0)\} \tag{9}$$

, where f represents the forward model operator.

Let $H$ represent the measurement operator for the insitu data cotemperous with the model run. We estimate a posterior distribution of inital conditions through the usual ensemble smoother method:

$$\psi_i^{0,posterior} = \psi_i^0 + PH^T(HPH^T + W)^{-1}(d - H\psi_i), \tag{10}$$

where P is the model error covariance matrix, $d$ is the vector of observations and $W$ is the observational error covariance matrix. We approximate the matrix $HPH^T$ with the ensemble covariance of the model measurements: $d_i = H\psi_i + \xi_i$. $\xi_i$ is a random variable simulated from the measurement error model (zero mean Gaussian with covariance $W$).

$$HPH^T = C^{d,d} = \left[c_{k,j}^{d,d}\right] = \left[\sum_i \frac{(d_i^k - \bar{d}^k)(d_i^j - \bar{d}^j)}{n-1}\right] \tag{11}$$

where $\bar{d} = \sum_i \frac{d_i}{n}$. Likewise the matrix $PH^T$ is approximated by it's ensemble analog:

$$PH^T = C^{\psi,d} = \left[C_{k,j}^{\psi,d}\right] = \left[\sum_i \frac{(\psi_i^k - \bar{\psi}^k)(d_i^j - \bar{d}^j)}{n-1}\right] \tag{12}$$

where $\bar{\psi} = \sum_i \frac{\psi_i}{n}$ It should be noted that $PH^T$ is only computed one row at a time so that memory use of the analysis step is independent of model size.

### 3.4. Prior model for initial conditions

Central to any smoother is specification of the apriori distribution over the model. In this case the distribution is defined implicitly through a model of perturbations to the initial and boundary conditions. Define the covariance function:

$$C(\{x1, y1\}, \{x2, y2\}) = e^{-(\frac{\Delta d}{\lambda_d} + \frac{\Delta b}{\lambda_b})} \tag{13}$$

where
$\Delta d = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$, and $\Delta b = |H(x1, y1) - (H(x2, y2)|$.

To generate the perturbations to the initial conditions we employ the following approach:

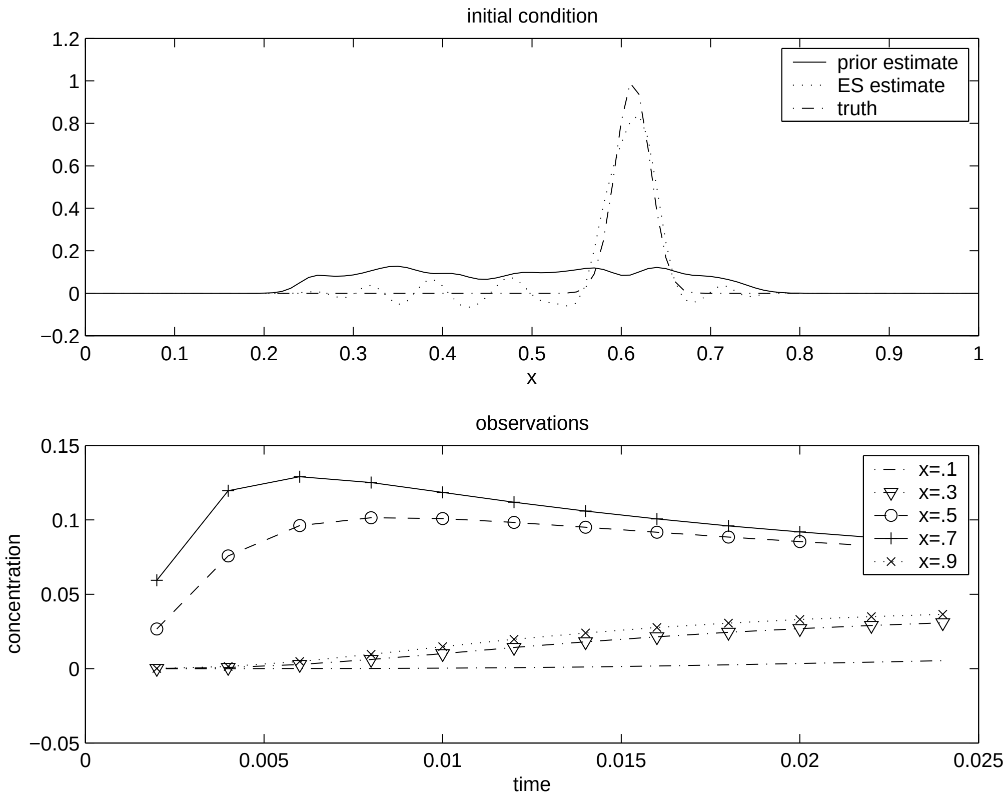


**Figure 6.** Simulation results of EnKF one dimensional plume source estimation.

1. construct the covariance matrix $C$ for all horizonal nodes.
2. compute B $= cholsky(C)$
3. to generate one initial perturbation, generate a random field $\xi = B^T\varsigma$, where $\varsigma$ are white standard normal random vectors of length $\eta_{nodes}$. Then let $T_i(x, y) = T_{mu}(x, y) + \nu\xi(x, y)$.

### 3.5. EnKF Plume Simulation

A One dimensional EnKF simulation developed in MATLAB illustrates the ability to invert sensor measurements and recover the source location. For programming simplicity a 1D vector space was used, normalized to the range (0,1). A random plume is generated (initial distribution) with center selected randomly in the range (0.2 - 0.8). An initial distribution corresponds to the plume particle density distribution at the time of plume release, $T_{max}$. This initial distribution ensemble is integrated forward in time, and generates sensor observations at regular intervals of x=(0.1, 0.3, 0.5, 0.7 and 0.9.) (Figure 7) Random initial conditions are then run through the same forward model to compute the ES estimate - which closely matches truth. The initial (prior) estimate distribution results from the possible starting locations of the plume. (Therefore 0 above 0.8 or below 0.2). The posterior result (ES estimate) in this case matches the truth extremely well. Positive and negative fluctuations in the posterior estimate are a byproduct of introduced noise and error in the estimator data.

## 4. INTEGRATION WITH PQS

We focus on diffusion events, especially chemical weapons, fires, or other aerosol producing events. These same methods could apply to other diffusion events such as heat or water pollutant tracking. The PQS has modular models, EnKF will serve as one of many. As the PQS system has the infrastructure readily available (Network

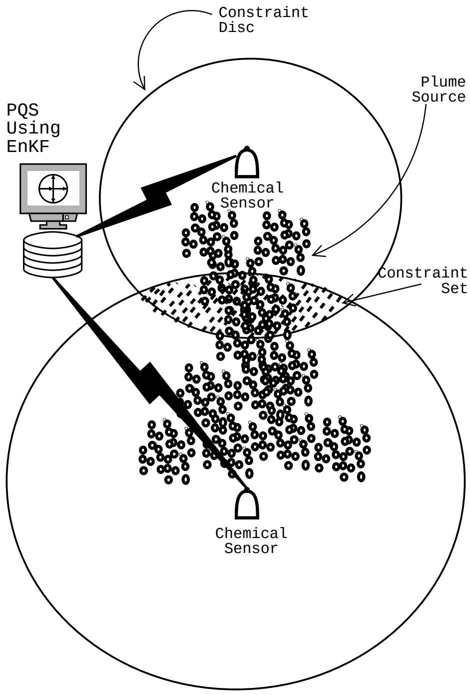


**Figure 7.** EnKF based plume source estimation with constraint sets used as initial guesses to reduce the total number possible source locations. Sensor observations are directed to the PQS system for filtering, assimilation, and hypotheses generation.

layers, model servers, message servers) to abstract the lower level system from the end user, more time and energy can be focused on the building of accurate models. Implementation of the EnKF model illustrates PQS can accommodate not only internet worms, HMM pattern recognition, and Track estimation - but also processes describing plume movement and origin. EnKF implementation further demonstrates the versatility of the PQS.

Submission of plume tracking PQS queries requires a plume query language (PQL) or graphical interface. This PQL defines a language or structure in which models are designed and prepared for submission. A proposed method for describing plume systems utilizes descriptive models of plumes. For example, a PQL must contain descriptions of plume shape, size, severity, and kinematics. Schemes for deconstruction of a complex system include these approaches: structural, ontological, and semantic.

One relevant industrial approach to flexible simulation design for complex logistical systems entails building descriptive models for all possible elements within the system.[4] The design of a reusable factory simulation package would require identifying three elements: objects, actions, and properties. Objects correspond to all possible entities capable of action. Actions describe all possible activities which the actors are capable of undertaking. Properties provide descriptive information about objects, actions, or other properties. The aggregate of these three elements provides a reusable generic meta-model in which queries can be constructed. With such a meta model, end users can design and submit high level queries descriptive of any possible scenario involving all plume elements. The following table illustrates a conceivable meta-model used to develop a PQL, including a few sample objects, actions, and property elements.

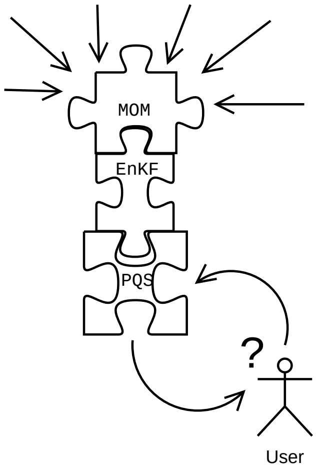


**Figure 8.** EnKF plugs into the PQS framework, receiving filtered observations from MOM (message oriented middleware)and passing query results to end users. The EnKF would post subscriptions for required plume data to MOM, and manage plume hypotheses with associated likelihoods.

| Objects | Actions | Properties |
|---|---|---|
| single static source | disperse | mass |
| single moving source | rise | volume |
| perimeter | fall | current location |
| center of mass | | origin location |
| | | direction of movement |
| | | predicted future location |
| | | temperature |
| | | turbidity |
| | | velocity |

Once a rigorous PQL has been developed, the next critical component of the EnKF module development will include hypothesis management. Conceivably even a well formed query could produce large numbers of matching ambiguous results. A ranking system must provide be able to track plume hypotheses as they evolve temporally. The most apparent techniques to apply derive from the literatures of multiple hypothesis tracking.[5] Methods such as Viterbi decoding may be implemented for managing exploding hypothesis sets. Historical records of plume hypotheses will be maintained as continuous data arrives from the sensor field. This will allow new hypothesis creation or irrelevant track elimination. Maintenance of unlikely hypotheses becomes increasedly important if detection of slow progressive events is desired.

## 5. FUTURE WORK

In the near future, a networked 2D simulation will create a 2D random walk model which generates a random plume field using existing PQS experimental networking hardware and software. Given sensor locations, the system will submit these observations to the MOM and PQS kernel. Using observations subscriptions PQS will answer queries from end users. This includes development of PQL for Plume queries.

Although the current focus is not on physical model precision, a greater variety of data sources would enhace our results. Currently all data used for model testing is simulated. Future versions may include historical data, or analysis of near real time data such as pollution in air or water.

Future models may also account for more complex flow regimes such as anisotropic diffusion (cross wind), high advection (wind or currents), as Diffusion rates differ in the vertical plane compared to the horizontal plane. This effect is evident by watching an emissions smokestack. Multi Dimensional models that include barriers and discontinuities such as geography and buildings will add additional realism.

## 6. ACKNOWLEDGMENTS

Many thanks to Prof. George Cybenko for inspiring this research, while also providing invaluable guidance and insight.


## REFERENCES


1. J. Crank, *The Mathematics of Diffusion*, Oxford University, London, 1975.
2. B. R. Roisin, , and E. Deleersnijder, *Introduction to Geophysical Fluid Dynamics - Physical and Numerical Aspects*, under contract with Academic Press, 2005.
3. G. Evensen, "Sequential data assimilation with a non-linear quasi-geostrophic model using monte carlo methods to forecast error statistics," *Journal of Geophysical Research* **99**(C5), 1994.
4. L. F. McGinnis and D. C. Gong, "Towards a manufacturing metamodel," *International Journal of Computer Integrated Manufacturing* , accepted for publication.
5. D. B. Reid, "An algorithm for tracking mutiple targets," *IEEE Transactions on Automatic Control* **AC-24**, pp. 843–854, December 1979.
6. G. Evensen, "The ensemble kalman filter: Theoretical formulation and practical implementation," *Ocean Dynamics* **53**, 2003.