Lawrence Berkeley National Laboratory
Lawrence Berkeley National Laboratory

Title

Rapid Data Assimilation in the Indoor Environment: Theory and Examples from Real-Time Interpretation of Indoor Plumes of Airborne Chemical

Permalink

https://escholarship.org/uc/item/4kb4k9f7

Author

Gadgil, Ashok

Publication Date

2007-08-01

Rapid Data Assimilation in the Indoor Environment: Theory and Examples from Real-Time Interpretation of Indoor Plumes of Airborne Chemical

Ashok Gadgil, Michael Sohn, and Priya Sreedharan

Abstract Releases of acutely toxic airborne contaminants in or near a building can lead to significant human exposures unless prompt response measures are identified and implemented. Possible responses include conflicting options, such as shutting the ventilation system off versus running it in a purge (100% outside air) mode, or having occupants evacuate versus sheltering in place. The right choice depends in part on quickly identifying the source location, the amount released, and the likely future dispersion of the pollutant. This paper summarizes recent developments to provide such estimates in real time using an approach called Bayesian Monte Carlo updating. This approach rapidly interprets measurements of airborne pollutant concentrations from multiple sensors placed in the building, and computes best estimates and uncertainties of the release conditions. The algorithm is fast, and can continuously update the estimates as measurements stream in from sensors. As an illustration, two specific applications of the approach are described.

Keywords: Bayesian Updating, Data Fusion, Real-Time Source Reconstruction

1. Introduction and Motivation

Airborne releases of acutely toxic contaminants in or near a building can lead to significant human exposures unless prompt response measures are taken. Unfortunately, possible response strategies can include conflicting options, such as shutting the ventilation system off versus running it in a purge mode, or having occupants evacuate versus sheltering in place. The right choice depends in part on knowing the source locations, the amounts released, and the likely future dispersion routes of the pollutants.

Determining this information is complicated by the uncertain and variable airflows typically found in multi-room, multi-story buildings. Merely detecting an airborne pollutant may not reveal the location or strength of the source. The sensor

measurements must be interpreted to estimate the source characteristics and quantify the uncertainties. For effective decision making, the measurements must also be interpreted quickly and continuously as data stream in from the sensors.

Traditional data interpretation algorithms generally use an inverse modeling approach (e.g., optimization and Gibbs sampling) to fit an indoor airflow and pollutant transport model to measurements of airborne pollutants. The fit is usually achieved by iteratively adjusting model input parameters until they reasonably predict the data. For on-line, real-time, sensor data interpretation, these approaches may be too slow. They (i) wait to execute many computationally intensive fate and transport models until data are first obtained; (ii) execute the models repeatedly as new data become available; and (iii) require a considerable amount of data before the algorithm finds a unique solution or estimates the uncertainty in the calibrated parameters.

Many of these problems can be solved using Kalman filtering [Grewal and Andrews, 2001]. However, Kalman filters work best on linear systems with well-conditioned input-to-output parameter covariance matrices and strong observability between the internal state variables and the model outputs. For these and other reasons (skipped here for brevity), Kalman filtering is not well suited to the current problem.

We describe an alternative algorithm called Bayes Monte Carlo updating, based on Bayesian statistics. This approach succeeds where traditional methods fail because it (i) decouples the simulation of predictive fate and transport models from the interpretation of measurements; and (ii) integrates uncertainty analysis into the calculations. Thus, we can pre-compute the time-consuming airflow and pollutant transport predictions and uncertainty estimates, before an event occurs, and then interpret sensor data in real time during an event. The technique works even with highly nonlinear transport physics, and may be used to estimate the location, magnitude, and duration of the release, to characterize any unknown or variable building or weather conditions, and to predict future pollutant transport in the building. Initial estimates are provided as soon as a sensor detects a pollutant, and can be updated as each new measurement arrives.

Bayesian updating approaches have been applied to assess environmental health risk [Brand and Small 1995, Pinsky and Lorber 1998], analyze groundwater data [Wolfson et al. 1996, Sohn et al. 2000], and conduct environmental value-of-information analyses [Dakins et al. 1996]. However, these papers describe applications to interpret data well after they were collected, where interpretation and response are not time-critical. In the present work, we exploit a feature of Bayes Monte Carlo updating not previously recognized: modeling and data analysis can be decoupled, which allows for data to be interpreted while they stream in during a pollutant release event. This is a significant advance over previous applications. Furthermore, our research group has been the leader in application of this general approach to indoor air pollutant source characterization and airborne pollutant transport predictions.

In this keynote paper, we (i) elucidate the Bayesian algorithm for interpreting sensor data in real time; and (ii) demonstrate the approach with two selected examples. In the first example, we characterize a pollutant release in a hypothetical five-room building, comparing concurrent and sequential sampling of the sensor data. We also examine degradation of the predictive results with increasingly noisy

data. In the second example, we characterize a tracer gas release in a three-story building. In this case study, not only are the sensors noisy, but they are trigger-type sensors, i.e., they only report a "yes" or a "no" regarding whether the local tracer gas concentration exceeds the pre-set trigger level.

2. Bayesian Updating

Since this is a diverse audience, we assume that not all readers are familiar with Bayesian Monte Carlo updating.

Our Bayesian Monte Carlo (BMC) data interpretation approach is divided into two stages. First, in the *pre-event* or *simulation* stage, the practitioner builds a fate and transport model of the building, characterizes uncertainties in the model inputs, and simulates many hypothetical airflow and pollutant transport scenarios. These time-consuming tasks are completed before a pollutant release occurs. In the second stage, during a pollutant release event, the agreement between each of the model simulations and sensor data is evaluated using a technique called Bayesian updating [Brand and Small 1995, Sohn et al. 2000]. This real-time stage may be conducted as data stream in from the sensors.

2.1 Pre-Event Computations

Before a release event, the practitioner develops a model of the building's indoor airflow and pollutant transport. Best estimates for model inputs are generated from, for example, previous building characterization exercises, tracer gas flow experiments and modeling, published literature, and professional judgment. Any uncertain model parameters (e.g., effective leakage areas) or variable inputs (e.g., outdoor temperature and width of door openings) is assigned a probabilistic distribution of possible values. Release characteristics (e.g., the location, duration, and amount of pollutant released in an incident) also are assigned uncertainty distributions.

The practitioner next generates a library of model simulations, by repeatedly sampling the distributions of the model inputs, and predicting the airflow and pollutant transport for each resulting model. The final library may comprise many thousands of simulations or *realizations*. Each realization represents a single possible combination of building configuration, weather conditions, and pollutant release scenario, and at this pre-event stage, each is deemed equally likely.

Because this library will be used to assess sensor data during an actual event, it is important to (i) characterize the uncertainty and variability in the model inputs properly; and (ii) draw sufficient samples from the distributions. Artificially narrow uncertainty distributions may not cover an actual release (in which case the algorithm may fail to assign a high likelihood to any library realizations). Similarly, insufficient sampling may miss a combination inherent in the original

distributions. One method for ensuring sufficient sampling is to increase the sample size until summary statistics (e.g., means, variances, coefficients of variation) of the model predictions no longer change.

2.2 During-Event Data Interpretation

During an actual release, the BMC algorithm compares data streaming in from sensors to each realization in the library of model simulations.

To summarize the process, each realization in the library is compared against the data to *quantitatively* assess the likelihood that the realization describes the event in progress. A high likelihood indicates a realization (i.e., the scenario described by the model simulation) fits the sensor data well. This in turn suggests that the model inputs used to generate that realization have high probability of describing the event in progress. By evaluating the relative fits for each realization, the Bayesian method estimates the best-fitting model inputs and the associated uncertainties.

The Bayesian updating process updates the probability associated with each of the library realizations, as a result of comparing the freshly-obtained data (say, because a sensor produced a new value) with the corresponding prediction from each realization. Suppose the BMC library contains N realizations. Then the kth realization predicts values Y_k, which are to be compared against the actual observations, O. Before the updating step, the *prior* probability, $p(Y_k)$, represents the assessed likelihood that the kth realization represents the actual event (recall that, before the onset of data comparison, each of the model realizations is equally likely, $p(Y_k) = 1/N$). After the update, that realization has *posterior* probability $p(Y_k/O)$, read as the probability of Y_k given O.

Using Bayes' rule, $p(Y_k/O)$ is calculated as [Brand and Small 1995]:

$$p(Y_k \mid O) = \frac{p(O \mid Y_k)p(Y_k)}{\Sigma_{i=1}^{N} p(O \mid Y_i)p(Y_i)} \tag{1}$$

The posterior probability, $p(Y_k/O)$, describes the probability of all of the model assumptions and predictions associated with the kth realization.

The posterior probability (Bayes Factor) of each simulation is applied to each parameter associated with that simulation, and thus a re-estimation is obtained for all input uncertainties. The posterior mean is calculated for each input parameter and a weighted estimate of the value of the parameter is obtained [Brand and Small 1995]:

$$Mean: \quad \mu_V = \sum_{i=1}^{N} V_i * p_i \tag{2}$$

The likelihood function, $p(O/Y_k)$, in eq. 1 quantifies the error structure of the data, i.e., the differences between the data and the model predictions resulting from

measurement error, spatial, and temporal averaging or correlations, and imperfect model representation. If S independent measurements are considered, for example following sequential concentration measurements returned from sensors or from concurrent measurements sampled in several locations, the likelihood of observing all of the measurements is the product of all of the individual likelihoods:

$$p(O \mid Y_k) = \prod_{j=1}^{S} p(O_j \mid Y_k) \tag{3}$$

For unbiased measurements with a normally distributed error, the likelihood of observing a sensor measurement, O_s, given a model prediction, $Y_{s,k}$, is given as:

$$p(O_s \mid Y_{s,k}) = \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{O_s - Y_{s,k}}{\sigma_\varepsilon}\right]^2\right) \tag{4}$$

where O_s is the concentration measured by a sensor in a room at $t=t_s$; $Y_{s,k}$ is the airborne concentration predicted from the kth Monte Carlo realization that corresponds to O_s; and σ_ε^2 is the error variance of the measurements. The error variance describes not only the error in the sensor instruments, but also the error associated with comparing model predictions with sensor measurements having different spatial and temporal averaging. Alternative likelihood functions can be used, as appropriate, without affecting the overall approach.

This second stage of the approach is mathematically simple and can be executed very quickly, much quicker than the rate at which new data are likely to arrive from sensors.

3. First Illustrative Application

This first illustration is to locate and characterize a hypothetical pollutant release in a five-room building, selected from the full description and results presented in Sohn et al. [2002]. Uncertainties in source location, duration, and amount, and in some building characteristics, were estimated using synthetic data. We also compare concurrent and sequential sampling, and examine how noise in the sensor data degrade the algorithm's predictions.

3.1 Building Description

The study building is a single story building comprising three rooms, a common area (CA), and a bathroom (Figure 1). Each of the partitioned areas are treated as well-mixed zones. The zones connect to the outside via windows and doors, and interconnect via doors. The building does not have a ventilation system.

The status of one of the CA zone windows and the door between the CA zone and Room 3 is "unknown" to the simulations (e.g., owing to failed position sensors at these locations). These are denoted in Figure 1 with question marks. Wind blows at a steady 3 m/s from the North. The temperatures of the rooms are indicated in Figure 1, and the outside temperature is unknown.
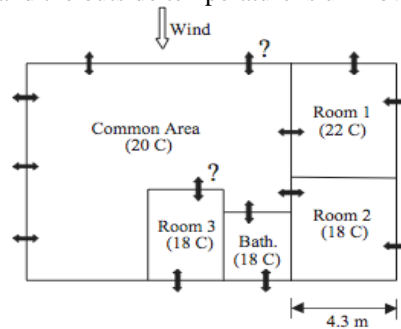

Fig. 1 Plan of the five-room building. The arrows represent windows or doors; question marks indicate unknown open or closed status. Windows in the bathroom and Room 2, and interior door for Room 1 are open. The wind blows at a steady 3 m/s.

3.2 Airflow and Pollutant Transport Simulation

We used the COMIS model [Feustel 1999] to predict indoor airflow and pollutant transport. COMIS predicts the steady-state airflows, and the dynamic transport of pollutants, by representing the building as a collection of well-mixed zones. Air flows between zones via cracks, doors, and windows (and also fans and ductwork, although we did not use these features here). Though we selected a multi-zone model for this application, the data interpretation algorithm may be used with any suitable airflow and pollutant transport model.

As part of the pre-event calculation, we generated five thousand airflow and pollutant simulations, each of them equally likely, using Latin Hypercube sampling [Iman and Conover 1980]. Means and variances for several sample sizes were tested to ensure that 5000 realizations adequately sampled the problem space.

3.3 Description of Synthetic Data

We generated synthetic data to represent measurements that might stream in from air monitoring sensors placed in the building. The synthetic data are based on an airflow and pollutant transport simulation that represents a possible pollutant release event. This simulation was excluded from the library of 5000 simulations.

We added measurement error to the synthetic data using a two-part error structure: (1) a normally distributed error associated with a standard deviation proportional to the true value, i.e., a fixed coefficient of variation, and (2) a

normally distributed random error that is independent of the magnitude of the measurement.

We generated high-, medium-, and low-quality synthetic data with progressively larger magnitudes of error components in the error structure. If adding the error generated a negative value for the pollutant concentration, we set the simulated measurement to zero. Details are described in Sohn et al. [2002], and not given here for brevity.

In addition to the three levels of data quality, we also evaluated two different data collection plans. The first, *concurrent sampling*, provides synthetic sensor data to the BMC algorithm from all five zones simultaneously, at five-minute intervals. The second, *sequential sampling*, provides measurements sequentially, from one zone at a time, at five-minute intervals. Sequential sampling might represent a situation where a single (expensive) sensor is multiplexed to several sampling tubes. The rooms were sampled in the order (1) CA zone at t=5 min.; (2) Room 1 at t=10 min.; (3) Room 2 at t=15 min.; (4) bathroom at t=20 min.; and (5) Room 3 at t=25 min. Starting at t=30 min., sampling returns to CA zone.

3.4 Data Interpretation

Figure 2 shows the estimation of the source location for the three qualities of data. Though it is tempting to interpret Figure 2 purely in terms of the success or failure of the interpretation approach, it is important to emphasize that the results merely illustrate the types of data interpretation and "what-if" analyses that may be conducted using Bayesian updating. Cases where the algorithm cannot identify the release scenario with high probability may be interpreted as due to insufficient information in the measurements or the BMC library.

With concurrent sampling of medium- or high-quality data, the interpretation correctly identified the source location at t=5 min., when five measurements were obtained. With low-quality data, the identification of the source location was slower, requiring more measurements, and thus time, to overcome the error in the data. Again, the medium- and high-quality data permit dramatic uncertainty reductions at t=5 min., in all cases converging to the correct answers.

Sequential sampling collects data five times slower than concurrent sampling. In consequence, the medium- and high-quality data did not locate the source until all of the rooms were sampled once (t=25 min.), though reasonably good estimates were generated as early on as t=10 min. The low-quality data, however, did not locate the source even after 30 min.
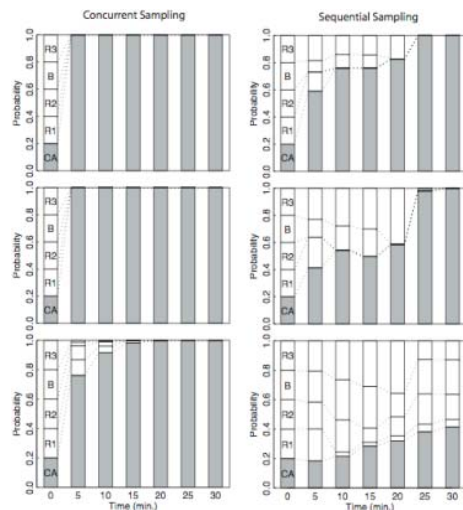
Fig. 2 Locating the source using (a) high-, (b) medium-, and (c) low-quality measurements. Concurrent sampling draws a measurement from each zone every five minutes. Sequential sampling draws a measurement from one room at a time every five minutes in the order: (1) common area at t=5 min., (2) Room 1 at t=10 min., (3) Room 2 at t=15 min., (4) bathroom at t=20 min., (5) Room 3 at t=25 min., and (6) common area at t=30 min. The equal probabilities at t=0 are before data interpretation.

4. Second Illustrative Application

Now we focus on trigger- or alarm-type sensors, rather than continuous-output devices. We base our case study on data from one of twelve tracer-gas experiments conducted at the Dugway Proving Ground, Utah [Sextro et al. 1999]. The paper from which this illustrative application is taken, [Sreedharan et al. 2006], examines how well sensors with different characteristics (threshold level, response time, and error rate), reconstruct the release event. The work demonstrates the importance of a systems perspective in selecting sensors with desirable sensor characteristics. However, for lack of space, only one illustrative example will be shown here.

4.1 Problem Description

We consider the following problem. A contaminant is released somewhere in a building, or near its indoor air intakes. A network of trigger or alarm-type sensors operates to identify the release. We seek to understand how sensor characteristics such as threshold level and response time affect the ability of the BMC interpretation algorithm to quickly detect and characterize the contaminant release, including the release location and mass.

For the purposes of this paper, we assume that each zone has one sensor, with a single threshold (meaning the output is either 0 for "below threshold," or 1 for "above threshold"). We characterize sensor performance by (i) threshold level; (ii) response time (also called integration time); and (iii) error rates. In this study, we compare several possible threshold levels; note that, to avoid false positives, a threshold may be above the detection limit of the sensor.

4.2 Building Characterization and Data Collection

The study focuses on one unit in a multi-unit building located. The unit consists of 660 m^3 of interior volume and approximately 280 m^2 of floor area on three levels. A mechanical air-handling unit (AHU) supplies air to the first and second floors. The AHU is a 100% recirculating unit (i.e., there is no deliberate outside air intake), and it returns air from the first floor.

A library of 5000 realizations was generated using a COMIS model of the building. The model was based on detailed experimental measurements [Sextro et al. 1999]. For the library, we sampled from statistical distributions of a set of key input parameters, as described earlier.

We generated hypothetical threshold sensor data by interpreting the tracer data from the experiment [Sextro et al. 1999] as if they were concentrations to which surface acoustic wave (SAW) sensors were exposed. SAW sensors are piezoelectric devices, often configured to provide alarms based on whether the incoming concentrations are above or below a predefined trigger or threshold level. False positive and false negative alarms may occur, according to the performance characteristics of each sensor, or the inability of the sensor to distinguish between the contaminant of interest and interfering chemicals that also may be present in the air. Three sensor attributes were varied: threshold level, response time, and error.

The threshold levels were chosen relative to the measured concentrations during the first 120 minutes of the release event. The lowest threshold would cause 98% of the data to trigger an alarm, while the highest threshold would trigger an alarm for only 1% of the data. However, for presentation purposes, we normalize both threshold levels and concentration data by the concentration that would be found in the release zone if the entire release amount instantaneously mixed throughout that zone. That is, thresholds and concentrations are reported in terms of the theoretical maximum peak concentration that could be measured in the system under the perfectly well-mixed assumption. With this normalization, the lowest threshold level was 0.02% of the maximum peak, and the highest was 16%.

Sensor response times ranged from 20 seconds to 180 seconds. In the simulations, concentrations are averaged over the response time, and then compared to the appropriate threshold level. Note that averaging over the response time corresponds to an assumption that the SAW desorption cycle is brief relative to its adsorption cycle. Our simulations ignore the duration of the desorption cycle (i.e., each sensor started integrating the next cycle of data as soon as it reported an alarm or no-alarm condition).
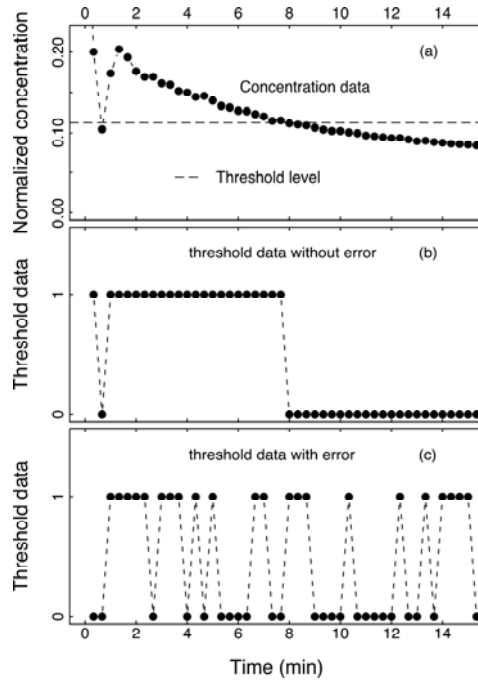
Fig. 3 Sample conversion of tracer gas concentration to threshold data: (a) concentration data; (b) threshold data without simulated error added; and (c) threshold data with simulated error added.

Simulations were run using data with and without synthetically added error. For simulations with added error, we generated sensor signals according to the following assumptions: (1) if the actual concentrations are within 25% of the sensor threshold level, the signal will be false 50% of the time; and (2) for concentrations outside of this range, the signal will be false either 10% or 30% of the time, depending on the assumed sensor error.

Figure 3 illustrates the conversion of measured concentration data to simulated threshold data. Figure 3(a) shows normalized time-averaged concentrations, with the threshold level indicated. Figure 3(b) shows the sensor signal that would result from a threshold sensor with no error and instantaneous response. Figure 3(c) shows the sensor signal, corrupted with false negatives and false positives. Because the false readings are generated stochastically, different realizations of the data in Figure 3(c) would have different output.

In this implementation of Bayes' rule, the likelihood function is based on the probability used to generate the false positives and negatives. For example, for data generated with a 30% error, the likelihood is 0.3 when the modeled concentration is more than 25% above the threshold level and the sensor has not signaled on; conversely, the likelihood is 0.7 if the sensor has signaled on. For the simulations using data without synthetically added error, we assume 5% error for all data. We did not assign 100% confidence to this data because of inherent uncertainty and variability. In practice the designer of the sensor system should

have reliable information on the sensor's actual rate of false positives and false negatives.

For the case-study examples, we systematically varied the sensor characteristics, as described above. In cases where error was specifically investigated, for each sensor attribute, we generated 50 sets of error-added threshold data, analogous to those displayed in Figure 3(c), for each sensor in the system. Each combination of threshold level, response time, and error produced a data stream with which to challenge the Bayes Monte Carlo algorithm. The algorithm was used to determine the release location and release magnitude; the time of release was assumed known.

4.3 Results for Triggered Sensors

For space considerations, this paper only shows the ability of the sensor system to estimate the release location. Interested readers are referred to [Sreedharan 2006] for more results and discussion.

The information content in threshold sensor data is significantly less than that in direct concentration measurements. Nevertheless, the sensor system can successfully reconstruct the source, at least in some circumstances. We demonstrate this with an example in which the concentration data have been converted to threshold data using a threshold level of 2.3%, a sensor response time of 20 seconds, and without additionally-added error. We judge the sensor system performance by its ability to reduce the uncertainty in the estimates of the release location, mass, and duration parameters, and by the time required to do so.

Figure 4 depicts the time required to identify the release location (Room 1.2a). At time zero, every zone is assumed to be equally likely as the release location. As sensor data arrive, the Bayes algorithm adjusts these probabilities, locating the release location with greater than 90% confidence within one minute. If rapid response hinges on locating a source very quickly, this example suggests that threshold sensors under this network configuration and data quality may be acceptable for real-time monitoring.

5. CONCLUSION

Real-time environmental monitoring systems have the potential to help protect building occupants in the event of high-risk pollutant releases. Here, we have demonstrated — albeit for a limited set of circumstances — that a network of continuous or single-level threshold sensors can be used to determine the location and magnitude of the release within a Bayes Monte Carlo framework. More importantly, the Bayesian approach naturally produces a systems-level view of the sensor network, which may lead to better tradeoffs between sensor characteristics such as response time and error than might be the case when considering sensors individually.
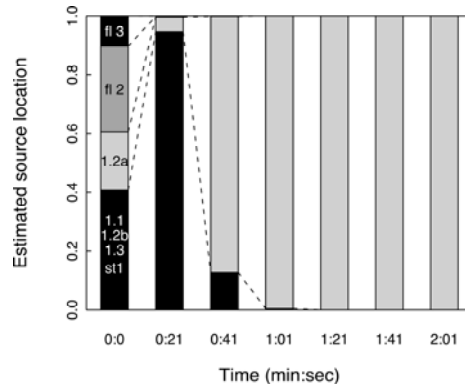
Fig. 4 Probability of source being in location indicated, as estimated with the Bayes Monte-Carlo algorithm using threshold data with response time of 20 s, threshold level of 2.3%, and without added error. The actual release location is Room 1.2a. Time is referenced to the instantaneous release event.

With more complex buildings, system characterization is technically more challenging, and also more expensive. Further research will be needed to test the feasibility in such buildings. For example, hybrid methods that augment prior knowledge with sensor system data that monitors building operations to learn about airflows and contaminant transport may improve overall system performance. Such advances would not only be beneficial for designing indoor monitoring systems, but may potentially be extended to help diagnose and interpret data from large-scale contaminant releases to the ambient atmosphere and to other environmental media. Such approaches also hold the promise of facilitating improvements in building performance with respect to energy use, thermal comfort, and indoor air quality.

References

Brand KP, Small MJ (1995) Updating Uncertainty in an Integrated Risk Assessment: Conceptual Framework and Methods; Risk Analysis; 15(6), 719-731.

Dakins ME, Toll JE, Small MJ, Brand KP (1996) Risk-Based Environmental Remediation: Bayesian Monte Carlo Analysis and the Expected Value of Sample Information; Risk Analysis. 1996; 16(1), 67-79.

Feustel HE (1999) COMIS — an international multizone air-flow and contaminant transport model. Energy and Buildings 30, 3-18.

Grewal MS, Andrews AP (2001) Kalman Filtering: Theory and Practice; John Wiley & Sons: New York, 2nd edition.

Iman RL, Conover WJ (1980) Small Sample Sensitivity Analysis Techniques for Computer Models, with an Application to Risk Assessment; Commun. Statist. - Theor. Meth; A9(17), 1749-1842.

Morgan MG, Henrion M (1990) Uncertainty, A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis; Cambridge University Press: New York.

Pinsky PF, Lorber MN (1998) A Model to Evaluate Past Exposure to 2,3,7,8,-TCDD; J. Exp. Anal. and Env. Epid. 1998; 8(2), 187-206.

Sextro RG, Daisey JM, Feustel HE, Dickerhoff DJ, Jump C (1999) Comparison of modeled and measured tracer gas concentrations in a multizone building. Proceedings of the 8th International Conference on Indoor Air Quality and Climate – Indoor Air 99, Vol 1, pp 696-701. Indoor Air 99, Edinburgh.

Sohn MD, Small MJ, Pantazidou M (2000) Reducing Uncertainty in Site Characterization Using Bayes Monte Carlo Methods; J. of Env. Engineering. 2000; 126(10), 893-92.

Sohn MD, Reynolds P, Singh N, Gadgil AJ. (2002) Rapidly Locating and Characterizing Pollutant Releases in Buildings. Journal of Air and Waste Management Association, 52: pp. 1422-1432.

Sreedharan P, Sohn MD, Gadgil AJ, Nazaroff WW (2006) Systems approach to evaluating sensor characteristics for real-time monitoring of high-risk indoor contaminant releases. Atmospheric Environment 40. pp. 3490-3502.

Wolfson LJ, Kadane JB, Small MJ (1996) Bayesian Environmental Policy Decisions: Two Case Studies; Ecological Applications. 1996; 6(4), 1056-1066.