Risk-Based Characterization for Vapour Intrusion at a Conceptual Brownfields Site: Part 1. Data Worth and Prediction Uncertainty

Xiaomin Wang¹, Andre J.A. Unger², Beth L. Parker³

¹²Department of Earth and Environmental Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada
³School of Engineering, University of Guelph, 50 Stone Road East Guelph, Ontario, Canada

"¹xm4wang@uwaterloo.ca; ²aunger@uwaterloo.ca; ³bparker@uoguelph.ca

Abstract-The focus of this paper is to present a methodology to assimilate soil core permeability and trichloroethylene (TCE) soil gas concentration data, and then to assess their worth in reducing prediction uncertainty with a numerical model. The specific problem involves a residential development impacted by indoor air exposure of TCE contamination originating from a groundwater plume. Three metrics are used to quantify the prediction uncertainty, namely: the ability to accurately predict the indoor air concentration within the houses at any point in time; the ability to reduce the standard deviation of predicted indoor air concentration within these houses; and, the ability to accurately forecast the probability of indoor air concentrations exceeding a regulatory limit. The data assimilation methodology involves generating multiple realizations of heterogeneous permeability fields conditioned upon a geostatistical analysis of the borehole data, combined with a discrete static Kalman filter to assimilate actual soil gas concentration data, to estimate soil gas and indoor air concentrations at those locations where the developer does not have any data but liability. The worth of using progressively more permeability and soil gas concentration data is quantified on the basis that it provides a statistically significant improvement in the three metrics used to measure prediction uncertainty.

Keywords: Brownfields; Vapour Intrusion; Prediction Uncertainty; Data Worth; Kalman Filter; Probability of Exceedance

I. INTRODUCTION

Brownfields are defined as “abandoned, idled, or under-used industrial and commercial facilities where expansion or redevelopment is complicated by real or perceived environmental contamination” by the U. S. Environmental Protection Agency (EPA). EPA estimates that there are more than 450,000 brownfields in the U. S. alone. The negative impacts of brownfields in the social, environmental and economic aspects initiate the concept of brownfields reclamation/redevelopment. Redevelopment of contaminated land presents a significant challenge in that people may become exposed to legacy contamination at the site. This contamination may continue to exist even after extensive remediation efforts given the difficult nature of characterizing the subsurface hydrogeological properties [1, 2], locating the sources of contamination [3, 4], and subsequent removal from the soil in compliance with regulatory criteria [1, 5-8]. The possibilities for people to suffer potential health impacts from exposure to legacy contamination could result in the developer’s receiving punitive damages of unlimited value. To promote the developer’s due diligence in remediating the site and to simultaneously limit the potential for future punitive damages, government agencies provide liability protection beyond a specified monitoring period [9-11].

Previous research studies [12, 13] discussed the hydrogeological process governing the fate and transport of TCE, a volatile chlorinated solvent, from groundwater to indoor air. As a first response, the developer would likely repurchase those houses directly overlying the contaminant plume given the likelihood that they will be adversely impacted, with the intent of minimizing their liability exposure. Thereafter, regulatory guidance regarding an exclusion zone [14] can be used to specify a critical distance beyond the lateral boundaries of the plume where it is no longer necessary to conduct pathway assessment and monitoring to determine whether the house is being affected by subsurface contamination. While this clearly specifies the developers’ due diligence in managing the housing development upon discovery of a groundwater plume, it does not limit their liability to the houses within the exclusion zone before the end of the liability period. The developer faces an additional difficulty in that they may not actually even have access to those houses in order to monitor the indoor air quality without directly purchasing them. With an aim to minimize their liability, the developer must be able to assess the probability that houses located within the exclusion zone may be adversely impacted given pathway assessment and monitoring data that they have collected. They must then reserve sufficient capital to be able to repurchase any adversely impacted house within the development provided that data become available to suggest that the indoor air quality exceeds a regulatory limit [15, 16].

Some studies provide a cost-risk-benefit framework that is directly amenable to address capital reserve issues posed by the developer [17-20]. Specifically, the framework presented by [13] combines both hydrogeological and economic information when pricing the reserve capital to cover the financial risk based on the probability that the indoor air concentrations of these houses may exceed a regulatory limit. Within the context of proposed brownfields development outlined above, the multiphase compositional numerical model CompFlow Bio [21-23] is used to simulate indoor air concentrations for houses within the exclusion zone, and subsequently the probability of indoor air concentrations exceeding a regulatory limit. Reference [24]
reviewed the body of work where mathematical models are used to investigate the exposure pathway of volatile organic compounds from the subsurface to indoor air. These models are then used to predict the indoor air concentrations as a line of evidence supporting risk management decisions when field data are limited. The difficulty in using numerical models follows from assumptions and limitations in constitutive relationships and parameters within their development. Furthermore, complications arise from bias and uncertainties in the simulated results due to applying a conceptual model to actual field sites [25-28].

Quantifying uncertainty as outlined in this brownfield problem has been addressed using both inverse modeling and optimization. References [29, 30] used inverse modeling by minimizing a square difference function between measured and simulated flow as well as transport variables. Their approach is used to decrease the uncertainty in the initial estimates of the parameters, followed by a data worth analysis to show the effect of additional measurements on the confidence of the model prediction. Reference [31] used optimization to design groundwater remediation strategies by employing the hydrogeological model to guide the decision-making process. The Monte Carlo approach [32] provides a stochastic method that simultaneously recognizes the uncertainty inherent in both the conceptual model and its parameters. When additional information on model parameters or field measurements becomes available, they are used to update the prior estimates of the statistics to posterior estimates in order to reduce the estimation uncertainty. The prior estimate is generally based on limited early data, expert judgment, or a combination of both. Bayes theorem is one of the methodologies used for this purpose. Within the Bayesian framework, [33] applied the Kalman filter algorithm to determine the least cost design of groundwater quality monitoring networks. This approach has also been used in other studies, e.g. [6, 34-39].

Following the work of [17, 22, 33], the motivation of this paper is to develop a methodology that utilizes kriging in combination with a discrete static Kalman filter to assimilate soil core permeability data and TCE soil gas concentration data. The objective of the data assimilation process is to assess its worth in terms of reducing the prediction uncertainty associated with forecasting the probability that indoor air concentrations will exceed a regulatory limit. Three metrics defining prediction uncertainty are introduced and quantified based on the mean, variance, and tail of the indoor air concentration distributions. The metrics are: first, the ability to correctly estimate the actual indoor concentration and unmonitored locations; second, the ability to reduce the standard deviation of the estimated indoor concentrations; and third, to accurately forecast the probability of indoor air concentrations exceeding a regulatory limit at unmonitored locations. As such, the question of data worth for reducing prediction uncertainty relates to changes in these metrics and hence statistical measures with various combinations of permeability and TCE soil gas concentration sampling strategies. This framework is particular useful to estimate potential indoor air concentrations at locations where the developer has liability but does not have direct access, under the constraint of the limited data availability.

II. PROBLEM DESCRIPTION

The highly stylized brownfields project commences with the proposition that a company owns a property that contains legacy dense non-aqueous phase liquid (DNAPL) contamination. They remediate the subsurface in compliance with governing regulatory requirements (Federal Contaminated Sites Action Plan, Canada; Superfund, or Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), USA). However, there is a possibility that some source zones of DNAPL contamination may remain below the water table. They now wish to construct multiple residential dwellings on the site. Followed [12], the central intent of this work revolves around the issue of assessing indoor air concentrations within a number of houses (six in total) depicted in Fig. 1 that directly overlie and are laterally offset from a groundwater plume emanating from a remnant TCE source zone. This TCE groundwater plume then results in soil gas concentrations of TCE in the vadose zone in the vicinity of the plume, causing indoor air impacts to the houses. To entice buyers, the developers guarantee that they will immediately repurchase any affected house at the original sale price appreciated by the US national home price index. Indoor air concentrations are monitored every $M_{monitor}$ days, and their guarantee lasts for the first $T$ days. Thereafter, the government indemnifies the developer [40, 41]. In return for the repurchase agreement and continuous monitoring, the residents agree not to blame any reasonably foreseeable health issues arising from potential long-term exposure of indoor air concentration on the developer given their due diligence. To maintain the development, the developer then remediates the subsurface to resolve the source of contamination, and demolishes/rebuilds/renovates any affected house to prevent further intrusion of soil gas. The house is then resold.

In the previous study, [12] focused on examining the fate and transport of TCE originating from a DNAPL source zone located below the water table, and their potential impact on multiple residential dwellings located down-gradient of the source zone using the three-dimensional three-phase multi-component compositional model CompFlow Bio. A series of sensitivity analysis were also conducted to understand the relative importance of factors such as: recharge, low permeability clay layer, extending the thickness of the vadose zone, and heterogeneity in the subsurface permeability structure of a stratigraphically continuous aquifer, on simulated indoor air concentrations. Reference [12] demonstrate that the lithological structure of the subsurface exerts a primary control on the fate and transport of TCE from groundwater to the indoor air, while the heterogeneous permeability structure of an individual lithological unit is of secondary but paramount importance. Therefore, in this work, the authors focus exclusively on heterogeneity as the primary contribution to prediction uncertainty. This is also consistent with [13] and [42] wherein uncertainty arising from the spatial variability of aquifer properties is a primary
component of the hydrogeological risk assessment process. Here, the authors assume the developer has already purchased houses 2 and 5 directly overlying the known TCE plume. As part of the developer’s due diligence in assessing their liability, they obtain a number of soil cores for the purpose of measuring the subsurface permeability structure and TCE soil gas concentrations. Despite the fact that the developer owns houses 2 and 5, it is impossible for them to drill vertically beneath the foundation without tearing them down (by way of preview, house 2 is not detrimentally impacted over the liability period and could be resold “as is” at a later date; however, house 5 is impacted above the regulatory limit). Furthermore, houses 1, 3, 4 and 6, which are all laterally offset from the plume, remain privately owned and inaccessible to the developer. The developer must reserve sufficient risk capital in order to purchase these houses should it become likely that any one of them is impacted.

![Diagram](image-url)

**Fig. 1** The three-dimensional problem geometry: a) top view, and b) vertical cross section

### III. PROPOSED METHODOLOGY

A key issue in this work is the assumption that although the developer has the liability for indoor air impacts to houses 1, 3, 4 and 6, the developer does not have access to these properties. This is often the case at real sites. To constrain the prediction uncertainty for indoor air impacts to the remaining houses, the developer then uses knowledge of the subsurface permeability structure obtained from the soil cores in combination with TCE soil gas concentration measurements within these same boreholes, as well as TCE soil gas concentration measurements from beneath the foundation slab of house 2.

#### A. Worth of Permeability Data

To evaluate the worth of permeability data to constrain the prediction uncertainty of indoor air concentrations arising from heterogeneity in the permeability structure of an aquifer, it begins by assigning the permeability realization denoted as “reality” in Fig. 2 as a representation of the point-to-point permeability structure of an actual site. This permeability realization was generated from the FGEN software [43] using the direct Fourier transform method, and is based on the geostatistics of the Borden aquifer without being conditioned by actual field data. In particular, the geometric mean permeability is $K = 2 \times 10^{-11}$ m$^2$, the variance of $Y = \ln K$ was assigned a value of $\sigma_Y = 1.0$ [−], and the correlation lengths of K in the principal, transverse and perpendicular bedding direction are $\lambda_1 = 5.0$ m, $\lambda_2 = 2.0$ m, and $\lambda_3 = 0.15$ m, respectively [2, 44, 45]. A Gaussian power spectral density function was used to distribute the geo-statistical parameters over the spatial domain. The variance $\sigma_Y$ was increased from that of the Borden aquifer to that of the more heterogeneous Cape Cod aquifer [46]. This change was motivated by the desire to pose a more difficult test (relative to the more homogeneous Borden aquifer) for a
realistically limited permeability sampling strategy to be able to reconstruct the structure of the aquifer over the entire conceptual model domain and effectively constrain prediction uncertainty.

![Logarithm of permeability realizations](image)

**Fig. 2** A visualization of the heterogeneous permeability realization chosen as reality, and a sample of a conditioned permeability realization generated using the 7K sampling strategy

Aquifer reconstruction followed by first sampling the “reality” aquifer in a series of patterns (i.e. 1K, 3K, 5K, 7K and 80K) is itemized on Table 1 as “K” with the actual borehole locations depicted in Figs. 1 and 2. Boreholes 1 to 7 used in the 1K, 3K, 5K and 7K sampling patterns are all located adjacent to a roadway separating the TCE source zone and the first row of houses 1, 2 and 3 which acts as a right-of-way and allows drilling without having to access private property. Sampling was conducted with a minimum borehole spacing of 5 m and a maximum spacing of 10 m between them. Each borehole is sampled for permeability along 0.1 m vertical increments across the entire thickness of the problem geometry. An obvious difficulty with this particular pattern of boreholes for the purpose of aquifer reconstruction is that they do not capture information over significant portions of the domain, and in particular, near all of the houses in the back row. Once again, the authors assume access limitations to houses 1, 3, 4 and 6 preclude mobilizing a drill rig onto these properties in order to obtain soil samples. Even though the developer owns houses 2 and 5, they choose not to drill within the development itself. An extreme and entirely unrealistic sampling strategy involving 80 boreholes (i.e. 80K) is used as a benchmark to illustrate the value of detailed and widespread permeability information across the entire conceptual model domain. Furthermore, 80K provides some perspective on the balance between two contrasting strategies, namely: being data rich at the expense of significant data collection expenditures with the intent of minimizing prediction uncertainty; and, alternatively conducting ones professional due diligence in characterizing the site while constrained by modest data collection expenditures, and accommodating a reasonable amount of prediction uncertainty.

Ordinary kriging is discussed in [47] as a means of conditioning a heterogeneous permeability relation using sample data within a data worth framework. Ordinary kriging is used to generate multiple statistically-equivalent heterogeneous permeability realizations to that of “reality” as shown in Fig. 2, where each realization is conditioned using the available borehole data from the 1K, 3K, 5K and 7K sampling patterns. Each realization is generated under the assumption of a Gaussian variogram. This follows from the work of [2] who used a statistical goodness-of-fit test to fit Gaussian, exponential and spherical variograms to Borden data [48-50]. While a single vertical borehole (1K) may be sufficient to characterize the subsurface lithology, it is probably entirely insufficient for the purposes of kriging in this context particular because it is impossible to estimate \(\lambda_1\) and \(\lambda_2\). Here, the assumption is that the same correlation length scale applies to different coordinate directions, namely isotropic correlation length. As such, 1K should be viewed as the “least” effort that one should use to investigate a site albeit the most likely one from a typical field investigation perspective, although it is not appropriate for the purpose of aquifer reconstruction as denoted here. In contrast, the authors postulate that 7K may be the “greatest” effort that one could reasonably use to investigate the site, although it lacks aquifer information in the transverse bedding direction and also assumes an isotropic correlation length. It is anticipated that the worth of the permeability data in terms of aquifer reconstruction will substantially diminish as one moves from the front row of the houses to the back row given that these houses are up to 40 m away from the row of boreholes along the right-of-way, and hence multiple values of \(\lambda_x\). The statistics of kriging variance \(\sigma_k^2(x_0)\), also the variance of the kriged permeability sample data \(\sigma_k^2\), for the 1K, 3K, 5K, 7K and 80K sample patterns are provided on Table 2. The mean value of the kriging variance decreases when additional samples are incorporated into the kriging system. 80K shows the smallest mean value of the kriging variance even though variability of the sample data is relatively high. Using the criterion of minimizing the average kriging variance, a more refined interpretation of the aquifer structure is obtained by both increased and widening the distribution of permeability data from boreholes.
To test whether the proposed sampling and kriging strategy have any merit in using the permeability data to reduce prediction uncertainty of the indoor air concentrations, \( n_{\text{realizations}} = 50 \) alternative permeability realizations for each of the 1K, 3K, 5K, 7K and 80K sample patterns are generated. CompFlow Bio is then used to simulate the fate and transport of TCE from groundwater to the indoor air of each of the six houses within the conceptual model domain in a Monte Carlo framework. It is expected that fifty alternative permeability realizations may not be a sufficient number of samples and more realizations may be needed to facilitate the analysis. However, it is not feasible to increase the number of realizations given limitations in the current computational resources. The runtime for one simulation was typically several days. Fig. 3 shows the resulting indoor air concentrations (denoted as \( c_{iA} \) and \( \tilde{c}_{iA} = \ln c_{iA} \) for each of the realizations, as well as that arising from reality (denoted as \( c_{iA\text{reality}} \)). Note that within Fig. 3, TCE soil gas concentration data were not used to reduce prediction uncertainty. Clearly, 1K does a poor job of capturing the variation of the indoor air concentrations with time in all of the six houses relative to reality. In contrast, both 7K and 80K capture the trend of the variation of the indoor air concentrations with time albeit with 7K exhibiting greater uncertainty than 80K. To this end, it appears that at least in a subjective sense the 7K scenario provides some value in using available permeability data to constrain prediction uncertainty. What remains is a more quantitative description of prediction uncertainty, especially with the inclusion of the TCE soil gas data.

### B. Worth of Soil Gas Concentration Data

The TCE soil gas sampling strategy involves collecting near-surface (at a depth of 0.1 m) samples within boreholes 2, 3, 4, 5 and 6, as well as beneath the foundation slab of houses 2 and 5. Various strategies are used to sample these locations and are itemized on Table 1 as 2C, 4C and 7C. For each sample location within a given sampling strategy, samples are regularly collected at each \( i \)th monitoring time interval \( (t_{i-1}, t_i] \) over the duration of the liability period. For all sampling strategies, borehole 4 and beneath the foundation slab of house 2 are always monitored given that they directly overlie the plume, are accessible, and would seemingly provide peak soil gas concentrations. The latter point is significant in that [12] established the idea that it is the mass flux of TCE crossing the foundation slab which largely controls the indoor air concentration. The mass flux of TCE is the product of the TCE soil gas concentration beneath the foundation slab times the volumetric flow rate of soil gas across the foundation slab. Consequently, the combined worth of TCE soil gas data to estimate peak concentrations with permeability data to estimate the volumetric flow of soil gas across the foundation slab are essential to constrain prediction uncertainty of indoor air concentrations within houses 1, 3, 4 and 6. Sampling strategy 2C denotes the minimal effort to monitor borehole 4 and house 2, and is combined with any of 1K, 3K, 5K, 7K and even 80K to evaluate the combined worth of both soil gas and permeability data. Sampling strategy 4C provides an incremental improvement over 2C by further sampling boreholes 3 and 5 which straddle the surface projection of the groundwater plume across the right-of-way. This strategy is motivated by [51] which concluded that sampling a plume along its lateral edges where the concentration gradient is large, serves to provide a great deal of information regarding plume location uncertainty. Finally, sampling strategy 7C further samples boreholes 2 and 6 as well as beneath the foundation slab of house 5. This strategy is motivated by [12]
which observed that indoor air impacts to houses decrease precipitously as a function of their lateral distance from the groundwater plume, generating narrow exclusion zones. Boreholes 2 and 6 are included to further resolve the lateral decline in soil gas concentrations in the transverse direction to the groundwater plume along the right-of-way. House 5 is included to provide TCE soil gas concentration information pertinent to the back row of houses. As with sampling strategy 2C, both 4C and 7C are combined with any of 1K, 3K, 5K, 7K and 80K to evaluate the combined worth of both soil gas and permeability data.

C. Kalman Filter and Its Application

Motivated by the work of [33], the discrete static Kalman filter (KF) is adopted as a tool to assimilate the spatial and temporal set of TCE soil gas data, arising from each sampling scenario identified on Table 1, to reduce prediction uncertainty of indoor air concentrations within houses 1, 3, 4 and 6. In general, the KF is used to estimate the state of a linear dynamic system as simulated by a stochastic model in association with noisy data and appropriate initial and boundary conditions. The KF describes how the system would respond after processing the noisy data in order to achieve an optimal estimation of its current state. As data become available, the KF can then deduce a minimum error estimate to improve future model prediction. A complete presentation of the discrete static KF can be found in [52]. Within this framework, the numerical model CompFlow Bio is used in combination with the heterogeneous permeability structure of the aquifer to represent the stochastic model. If all the information regarding the contaminant and subsurface properties and transport processes is available, the numerical model would predict the outcome exactly as seen in reality and uncertainty would not exist. However, perfect knowledge is impossible to obtain, so quantifying the probable situations and developing optimal estimates using the risk-based technique, KF, are needed. In this work, both groundwater and soil gas flow are steady-state processes. However, dissolution of TCE source zone with resulting transport of TCE from the groundwater, across the capillary fringe and into the vadose zone, and finally into the indoor air, is a transient process. Dissolution of the TCE source zone is the only process that imparts non-linearity to the transport of TCE and is isolated far from houses 1, 3, 4 and 6 where prediction uncertainty is aimed to be constrained. A critical assumption in the presentation of the KF is that the system is linear. Furthermore, a static KF rather than a dynamic KF is used because the system involves a forward-in-time prediction to account for the space-time correlations of model errors. Reference [39] stated that the traditional dynamic KF assumes that the model errors are not time-correlated. However, this assumption does not apply for contaminant transport model errors. The authors remind the reader that permeability data serve to condition the heterogeneous permeability structure of the aquifer. TCE soil gas concentrations constitute the noisy data, where the noise is imparted by; point-to-point variability in heterogeneous permeability structure of the aquifer as it impacts contaminant transport, potential field collection and laboratory analysis errors of the soil gas data.

1) Application of the Discrete Static Kalman Filter

An essential element in the application of the discrete static KF is that the noisy observed data conform to a Gaussian distribution. To confirm if this requirement is met, Fig. 4 presents the natural logarithm of simulated TCE soil gas
concentrations (denoted as \( c_{SG} \) and \( \hat{c}_{SG} = \ln c_{SG} \)) for 1K and 7K, and before assimilation of data arising from the 2C, 4C or 7C sampling scenarios. The notion of “before assimilation” for these concentrations is denoted as \( c_{SG}^- \). To test whether \( c_{SG}^- \) are Gaussian, it begins with the statistical hypothesis that for all Monte Carlo realizations, the TCE soil gas concentration \( c_{SG,i,j,l}^- \) at the \( t \)(th monitoring time interval \((t_{i-1}, t_i)\) where \( i = \{1, 2, ..., n_{times}\} \), the \( j \)th house, and with \( l = \{1, 2, ..., n_{realizations}\} \), are log-normally distributed. To test this statistical hypothesis, a one-sample Kolmogorov-Smirnov test (Massey, 1951; Miller, 1956) is used with the null hypothesis being that the vector of \( \hat{c}_{SG,i,j,l}^- \) follows a normal distribution. The alternative hypothesis is that the vector of \( \hat{c}_{SG,i,j,l}^- \) does not follow a normal distribution at the 5% significance level. The temporal transition where \( \hat{c}_{SG,i,j,l}^- \) becomes normally distributed is shown in Fig. 4 when the mean concentration \( \mu_{SG,i,j}^- \) is calculated as:

\[
\mu_{SG,i,j}^- = \frac{1}{n_{realizations}} \sum_{l=1}^{n_{realizations}} \hat{c}_{SG,i,j,l}^- \tag{1}
\]

changes from a dashed or a dash-dotted to a solid line. The dashed or dash-dotted line indicates the null hypothesis is not accepted, while the solid line indicates that it is accepted. In general, \( \hat{c}_{SG,i,j,l}^- \) are not normally distributed at early time because TCE soil gas concentrations beneath the houses for numerous realizations are lower than the accuracy of what can be reliably resolved with the CompFlow Bio model. Despite this predicament, \( \hat{c}_{SG}^- \) is still used within the KF over the entire liability period for reasons of continuity as well as necessity, as will be made clear shortly. The results show that 3K, 5K and 80K conform to the same behaviour. Finally, TCE soil gas concentrations \( \hat{c}_{SG}^- \) provided in Fig. 4 are post-processed to yield the indoor air concentrations \( c_{IA} \) depicted here in Fig. 3 based on the mass balance of TCE in the house, assuming no background concentration, rapid mixing of the indoor air, no indoor emission sources, and no indoor reactions [13].

\[
\begin{align*}
\hat{c}_{SG,i,j,l}^- &= \Phi c_{SG,i}^- + w_i \\
&= \Phi c_{SG,i}^- + \sum_{i=1}^{n_{realizations}} w_i
\end{align*}
\]
model with heterogeneous permeability fields largely dependent on a given spatial correlation, mean and variance, which could vary between alternative permeability sampling strategies. The vector of discrete estimates of the state variable here includes both spatial and temporal TCE soil gas concentrations, with the static discrete KF not differentiating between the space and time dimensions. Therefore, $c_{SG}^-$ is further simplified to $c_{SG}^-$. 

The TCE soil gas samples all arise from use of the “reality” aquifer shown in Fig. 2, with the resulting TCE soil gas concentrations beneath all six houses provided in Fig. 4 and denoted as $\hat{c}_{SG,\text{real}}$. These measurements are uniquely defined and known a priori everywhere within the computational domain in space and time, are assumed to be linearly related to the system, and are employed for the filtering process. A vector of measurement data $z$ that is corrupted by non-negligible measurement errors $v$ is extracted from $\hat{c}_{SG,\text{real}}$. These measurement errors are assumed to be uncorrelated in time and independent of the system estimation. Reference [53] suggested that the average error for different sampled soil gas data in shallow aquifer is $\pm 21\%$ which reflects both human/instrument performance and the effect of horizontal heterogeneities. Here, the assumption is that the acceptable relative error level for measurements varies within $\pm 25\%$ of the measured values and this error does not vary between different measurements. The measurement equation describing how the estimated $\hat{c}_{SG}^-$ are related to the actual system state $z$ is:

$$z = H \times \hat{c}_{SG}^- + v$$ (3)

where 

$z =$ the vector of $g$ noise-corrupted measurements, with dimension $g \times 1$, where in this case $g = n_{\text{sample locations}} \times n_{\text{times}}$ with $n_{\text{sample locations}} \in n_{\text{locations}}$, such that $n_{\text{sample locations}}$ includes only that subset of locations where concentration measurements are taken (i.e. where a ‘C’ is listed on Table 1);

$H =$ the measurement matrix, dimension $g \times m$;

$v =$ the vector of random measurement noise, $v \sim N(0,R)$, assumed to be uncorrelated with previous measurement errors, dimension $g \times 1$;

$R =$ the measurement noise covariance matrix, independent of the state variable, with dimension $g \times g$.

Elements of the measurement matrix $H$ are assigned a value of 1 if a sample is taken at location $j$ at the $i$th time interval; 0 otherwise. The diagonal elements of $R$ are assigned a value of $(25\%)^2$.

As measurement data $z$ become available, the concentration and its associated error covariance are updated by assimilating the data and its uncertainty. First, the Kalman gain matrix $K$ of dimension $m \times g$ which contains the model and measurements bias is calculated as:

$$K = P^- H^T (H P^- H^T + R)^{-1}.$$ (4)

Next, the posterior estimate of $\hat{c}_{SG}^+$ is updated using a linear function of the prior estimate and the measurement $z$ as:

$$\hat{c}_{SG}^+ = \hat{c}_{SG}^- + K (z - H \times \hat{c}_{SG}^-).$$ (5)

Finally, the corresponding error covariance $P^+$ is updated as:

$$P^+ = P^- - KH.$$ (6)

The optimal estimation for $\hat{c}_{SG}^+$ is obtained when the Kalman gain $K$ is maximized and the sum of the diagonal of the covariance matrix $P^+$ is reduced.

From Eq. (2), the first step in the application of discrete static KF is to propagate the state variables and error covariance of the system forward in time. The state matrix $\hat{c}_{SG}^-$ is built by the concentration variable $\hat{c}_{SG,ij}^-$ at location $j = \{1, 2, ..., n_{\text{locations}}\}$, the $i$th monitoring time interval $i = \{1, 2, ..., n_{\text{times}}\}$, and for a given permeability realization $l$. The dimension of the state matrix $\hat{c}_{SG}^-$ is $m \times 1$ and $m = n_{\text{locations}} \times n_{\text{times}}$. The variable $n_{\text{locations}}$ is the number of target locations to be investigated including measurement and non-measurement locations (6 houses and 5 boreholes, see Table 1) where $n_{\text{locations}} = 11$. The variable $n_{\text{times}}$ is the maximum number of monitoring time intervals at every 100 days from TCE injection to the end of liability period 5 years with $n_{\text{times}} = 18$. Note that $\hat{c}_{SG}^-$ is independently reconstructed for each permeability realization $l$. The vector of discrete estimates of the state variable is given as:

$$\hat{c}_{SG}^- = [\hat{c}_{SG,1,1}^-, ..., \hat{c}_{SG,n_{\text{times}},1}^-, \hat{c}_{SG,1,2}^-, ..., \hat{c}_{SG,n_{\text{times}},2}^-, ..., \hat{c}_{SG,n_{\text{locations}},n_{\text{times}}}^-]^T.$$ (7)

Each realization of $\hat{c}_{SG}^-$ for a given $l$ is an initial optimal estimate of the state vector for either the 1K, 3K, 5K, 7K or 80K scenarios.

The prior estimate of the error covariance $P^-$ before any soil gas samples that are taken using one of 2C, 4C or 7C is assembled using the covariance matrix elements $\text{cov}(c_{SG,ij}^-, c_{SG,j+\tau}^-)$ calculated by:
where \( \tau \) is the lag between two \( \hat{\varepsilon}_{SG,j} \) separated by both location \( j \) and monitoring time \( i \) in the vector \( \hat{\varepsilon}_{SG} \), and is of dimension \( m \times m \). Given that \( \hat{\varepsilon}_{SG} \) varies for each of the 1K, 3K, 5K, 7K and 80K scenarios, so does the covariance matrix.

The earlier hypothesis is reiterated that the progressive addition of permeability data to condition the heterogeneous structure of the aquifer should serve to increase the correlation between two values of \( \hat{\varepsilon}_{SG,j} \) separated by \( \tau \). Furthermore, the combined worth of permeability and concentration data should serve to reduce prediction uncertainty of the TCE soil gas concentrations beneath the foundation slab, and the flow of soil gas across the foundation slab into the indoor air of houses 1, 3, 4 and 6. The net result should be a reduction in the prediction uncertainty of indoor air concentrations within these same houses. Additional details on the implementation of the Kalman filter can be found in Appendix B of [54].

2) Interpretation of the Discrete Static Kalman Filter

Following the above application of the Kalman filter, a distribution of \( \hat{\varepsilon}_{SG}^+ \) arising from each permeability realizations \( l \) for a given 1K, 3K, 5K, 7K or 80K is inferred. This distribution must then be interpreted with regards to the notion of prediction uncertainty. Prediction uncertainty are defined using two metrics: first, the ability to estimate the actual indoor air concentration within houses 1, 3, 4 and 6 at any point in time at a level of confidence; second, the ability to use available soil gas concentration data to reduce the standard deviation of the estimated indoor air concentration within houses 1, 3, 4 and 6 at any point in time. These metrics are investigated by introducing the following moments of the posterior soil gas and indoor air concentration distributions:

\[
\begin{align*}
\mu_{SG,l,j}^+ &= \mathbb{E}\left[\hat{\varepsilon}_{SG,l,j}^+\right] = \frac{1}{n_{\text{realizations}}} \sum_{l=1}^{n_{\text{realizations}}} \hat{\varepsilon}_{SG,l,j}^+
\end{align*}
\]

\[
\begin{align*}
\mu_{IA,l,j}^+ &= \mathbb{E}\left[\hat{\varepsilon}_{IA,l,j}^+\right] = \frac{1}{n_{\text{realizations}}} \sum_{l=1}^{n_{\text{realizations}}} \hat{\varepsilon}_{IA,l,j}^+
\end{align*}
\]

\[
\begin{align*}
\sigma_{SG,l,j}^+ &= \sqrt{\mathbb{E}\left[\left(\hat{\varepsilon}_{SG,l,j}^+ - \mu_{SG,l,j}^+\right)^2\right]}
\end{align*}
\]

\[
\begin{align*}
\sigma_{IA,l,j}^+ &= \sqrt{\mathbb{E}\left[\left(\hat{\varepsilon}_{IA,l,j}^+ - \mu_{IA,l,j}^+\right)^2\right]}
\end{align*}
\]

While the indoor air concentrations are not directly estimated as part of the KF, values of the posterior \( \hat{\varepsilon}_{SG,l,j}^+ \) are post-processed as outlined by [12] and [13] to yield \( \hat{\varepsilon}_{IA,l,j}^+ \).

D. Probability of Indoor Air Concentrations Exceeding a Regulatory Limit

As discussed previously, the developer’s liability would occur when the indoor air concentration in one of these houses exceeds the regulatory limit, at which point the developer would need to purchase the house. To assess the developer’s liability, the third metric is defined associating with prediction uncertainty, namely; the ability to accurately forecast the probability of indoor air concentrations within houses 1, 3, 4 and 6 exceeding a regulatory limit. This third metric is directly related to the first two which relies on the first and second moments of \( \hat{\varepsilon}_{SG,l,j}^+ \) and \( \hat{\varepsilon}_{IA,l,j}^+ \) except that the “tail” of the distribution is now required to be evaluated.

The definition of probability of failure (also called the probability of exceedance) follows from [50] who used the criterion to delineate an exclusion zone. Failure occurs when the indoor air concentration of TCE at a single house exceeds the regulatory limit of 5 \( \mu \)g/m\(^3\) [15]. The probability of failure \( P_{lj} \) for the \( j \)th location is calculated as the sum of the number of Monte Carlo realizations for which the indoor air concentration of TCE \( c_{IA,l,j}^+ \) exceeds \( c_{\text{regu}} = 5 \mu \)g/m\(^3\), defined as \( n_{c_{IA,l,j}^+ > c_{\text{regu}}} \), divided by the total number of Monte Carlo realizations \( n_{\text{realizations}} \):

\[
P_{lj} = \frac{n_{c_{IA,l,j}^+ > c_{\text{regu}}}}{n_{\text{realizations}}}
\]

An alternative strategy to define \( P_{lj} \) is based on the work of [12] of “fitting” both a log-normal and a beta distribution to \( c_{IA,l,j}^+ \) in order to better capture the tail of the distribution with the limited number of Monte Carlo simulations. Both the log-normal and beta distributions make use of the first two moments of \( \hat{\varepsilon}_{IA,l,j}^+ \) given by \( \mu_{IA,l,j}^+ \) and \( \sigma_{IA,l,j}^+ \). The log-normal
distribution is characterized as being left-skewed with a heavy tail, and will therefore yield conservative estimates for the probability of exceedance relative to Eq. (10) for extreme (low probability) events. In contrast, the unimodal form of the beta distribution with a bounded domain does not exhibit a heavy tail, and will therefore yield less conservative estimates for the probability of exceedance relative to the log-normal distribution and possibly even Eq. (10).

IV. RESULTS

To show the performance of the discrete static KF, Figs. 5 and 6 depict the mean and standard deviation (see Eq. (9)) of the posterior soil gas and indoor air concentrations within all six houses over the liability period. These results are shown for four separate permeability and soil gas sampling combinations which would appear to have common use in practice, namely: 1K2C, 1K4C, 7K4C and 7K7C. They span the range of the least and greatest reasonable effort that one would consider when characterizing a site. The transition from 1K2C to 1K4C is meant to evaluate the worth of moderately increasing the amount of soil gas concentration data given limited permeability data. The transition from 1K4C to 7K4C is meant to evaluate the worth of significantly increasing the amount of permeability data given moderate amount of soil gas concentration data. Finally, the transition from 7K4C to 7K7C is meant to demonstrate the value of a detailed site investigation.

For each scenario, the solid line is the mean value while the error bars provide the standard deviation. The reality soil gas concentration $\hat{c}_{SG, reality}$ is also provided for reference.

![Diagram](image)

Fig. 5 The mean $\mu_{SG}$ (solid lines with symbols) and the standard deviation $\sigma_{SG}$ (error bars) of the natural logarithmic of soil gas TCE concentration after data assimilation $\hat{c}_{SG}$ using alternative sampling strategies for: a) house 1, b) house 2, c) house 3, d) house 4, e) house 5, and f) house 6. The red dashed line indicates reality $\hat{c}_{SG, reality}$.

To demonstrate the contribution of the KF, consider the soil gas concentration beneath the foundation slab of house 2 (see Fig. 5b) which is monitored for each of 1K2C, 1K4C, 7K4C and 7K7C. Despite the fact the prior $\hat{c}_{SG,l}$ (see Fig. 4) shows a great deal of variability for each permeability realization $l$, with this variability for 1K being unreasonably large relative to 7K and reality, the KF is able to reduce prediction uncertainty dramatically. However, the indoor air concentration within house 2 (see Fig. 6b) shows unreasonably large prediction uncertainty for both 1K2C and 1K4C, particularly at early time. In contrast, 7K4C and 7K7C show near identical prediction uncertainty demonstrating the value of characterizing the subsurface heterogeneous permeability structure in order to estimate the flow of soil gas into the house across the foundation slab and hence the mass flux of TCE into the indoor air within an acceptable range. This same pattern is repeated for house 5 (see Fig. 5e), except that the soil gas is only sampled for 7K7C. As a consequence, soil gas concentration prediction uncertainty for 7K4C increases slightly relative to 7K7C, while it visibly deteriorates for both 1K2C and 1K4C. The resulting pattern is further degraded for the indoor air concentrations shown in Fig. 6e.
The KF significantly underestimates \( \hat{\mu}_{IA,j} \) using alternative sampling strategies for: a) house 1, b) house 2, c) house 3, d) house 4, e) house 5, and f) house 6. The red dashed lines indicate reality \( \hat{\epsilon}_{IA,\text{reality}} \) while the dark grey dashed lines denote the regulatory limit.

Of particular interest is the prediction uncertainty within houses 1, 3, 4 and 6, which the developer does not possess but has liability. Soil gas concentrations within each of these four houses show a great deal of prediction uncertainty (see Figs. 5a, 5c, 5d and 5f), to the degree that it seems impossible to visibly note any trend that would differentiate between any of 1K2C, 1K4C, 7K4C and 7K7C. In contrast, the differences in indoor air concentrations between the sampling schemes become apparent (see Figs. 6a, 6c, 6d and 6f). For 1K2C and 1K4C, the KF significantly underestimates \( \mu_{IA,j}^+ \) relative to reality with very large values of \( \sigma_{IA,j}^+ \). This problem is exasperated for houses 1 and 4, which are both on one side of the groundwater plume relative to houses 3 and 6. While 7K7C appears to do a slightly better job at reducing prediction uncertainty relative to 7K4C, both 7K7C and 7K4C show vast improvements over 1K2C and 1K4C. The trend where a reduction in prediction uncertainty is favored for houses 3 and 6 over houses 1 and 4 still appears to be present, but is relatively subdued.

To initially quantify the prediction uncertainty, only the first metric will be focused. As such, the root-mean-square-error (RMSE\(_{ij}\)) is used, which is defined as:

\[
RMSE_{ij} = \sqrt{\frac{\sum_{t=1}^{n\text{times}} (\mu_{IA,j}^+ - \hat{\epsilon}_{IA,\text{reality},ij})^2}{n\text{times}}}
\]

where \( j \) is one of houses 1, 3, 4, 5 and 6, and for all sampling scenarios listed on Table 1. Results for all possible permeability and concentration sampling combinations are shown in Fig. 8a. With 7C, a clear trend develops for the progression of 1K, 3K, 5K, 7K and 80K where the RMSE\(_{ij}\) decreases, and reaches a minimum at 80K, for all houses. This same trend holds as the concentration sampling strategy is decreased from 7C to 4C. This trend is in support of the first metric. As the concentration sampling strategy is further decreased to 2C, this trend becomes erratic. The KF does not appear to be able minimize the RMSE\(_{ij}\) based on the combined worth of the permeability and concentration data when only two concentration sampling points are used.

A. Assessing the Significance in Improved Predictions for Indoor Air Concentrations at Unmonitored Houses

The monetary commitment required to collect increasing amounts of data could, in part, be motivated by the need to demonstrate a statistically significant improvement in the reduction of prediction uncertainty at the locations where the developer has liability but no access. In this section the first metric measuring prediction uncertainty will be focused. First, a statistical test is applied to infer whether there is a reduction in the absolute difference between \( \mu_{SG,ij}^+ \) and \( \hat{\epsilon}_{SG,\text{reality},ij} \) as well as between \( \hat{\mu}_{IA,j}^+ \) and \( \hat{\epsilon}_{IA,\text{reality},ij} \) for the pairs of incrementally increasing sampling data strategies represented by; 1K2C to 1K4C, 1K4C to 7K4C, and finally 7K4C to 7K7C. Within the context of the statistical inference test, each pair of sampling strategies represents a population of \( \hat{\epsilon}_{SG,ij}^+ \) and \( \hat{\epsilon}_{IA,j}^+ \) at each house \( j = 1, 3, 4 \) and 6, within time interval \( (t_{i-1}, t_i] \). These
populations are generated by the $l$ permeability realizations.

The test statistic used here is based on the normal distribution, which requires the use of the natural logarithm of the soil gas concentration and indoor air concentrations. For brevity, the presentation which follows is for indoor air concentrations but is equally applicable to soil gas concentrations. Let $\mu_{IA_{1j}} + \hat{\epsilon}_{IA_{reality_{lj}}}$ and $\mu_{IA_{2j}} + \hat{\epsilon}_{IA_{reality_{lj}}}$ represent the means of two populations “1” and “2”, respectively. Let $\Delta_0$ be the hypothetical difference in the absolute error, while $\hat{\epsilon}_{IA_{reality_{lj}}}$ and $\hat{\epsilon}_{IA_{reality_{lj}}}$ are the absolute errors obtained from two samples drawn randomly from each population. The statement for the hypothesis $H_0$ test is as follows:

Null hypothesis: $H_0 : \left\{ \begin{array}{l} \mu_{IA_{1j}} + \hat{\epsilon}_{IA_{reality_{lj}}} = \mu_{IA_{2j}} + \hat{\epsilon}_{IA_{reality_{lj}}} \\ \Delta_0 = 0 \end{array} \right.$ \hspace{1cm} (12)

Test statistic: $z_0 = \frac{\hat{\epsilon}_{IA_{reality_{lj}}} - \hat{\epsilon}_{IA_{reality_{lj}}}}{\sqrt{\frac{\sigma^2_{IA_{1j}} + \sigma^2_{IA_{2j}}}{n_{IA_{1j}} + n_{IA_{2j}} - 2}}}$ \hspace{1cm} (13)

Alternative hypotheses: Rejection criterion:

$H_1 : \left\{ \begin{array}{l} \mu_{IA_{1j}} + \hat{\epsilon}_{IA_{reality_{lj}}} > \mu_{IA_{2j}} + \hat{\epsilon}_{IA_{reality_{lj}}} \implies \Delta_0 > 0 \\ 0 < z_0 < \alpha \end{array} \right.$

$H_2 : \left\{ \begin{array}{l} \mu_{IA_{1j}} + \hat{\epsilon}_{IA_{reality_{lj}}} < \mu_{IA_{2j}} + \hat{\epsilon}_{IA_{reality_{lj}}} \implies \Delta_0 < 0 \\ 0 < z_0 < \alpha \end{array} \right.$

$H_3 : \left\{ \begin{array}{l} \mu_{IA_{1j}} + \hat{\epsilon}_{IA_{reality_{lj}}} \neq \mu_{IA_{2j}} + \hat{\epsilon}_{IA_{reality_{lj}}} \implies \Delta_0 \neq 0 \\ 0 < z_0 < \alpha \end{array} \right.$

The parameter $\alpha$ denotes the level of statistical significance.

The results of this statistical test are reported on Table 3 for the pairs of incrementally increasing sampling data strategies represented by: 1K2C to 1K4C, 1K4C to 7K4C, and finally 7K4C to 7K7C. A value of “T” denotes acceptance of hypothesis $H_2$, “F” denotes acceptance of hypothesis $H_3$, and “0” denotes acceptance of hypothesis $H_0$, all at a level of significance of $\alpha = 0.05$ or $5\%$. In other words, “T” denotes that the first prediction uncertainty metric is achieved via the worth of adding more data, “F” denotes the data with the opposite impact, and “0” denotes that no statement of significance can be made. A dash “-” denotes that the test could not be conducted because one or both of the populations do not conform to a normal distribution.

### Table 3 Results of the Statistical Test (See Eqs. (12-14)) to Determine Whether Pairs of Data Sampling Strategies 1K2C to 1K4C, 1K4C to 7K4C and 7K4C to 7K7C Improve Estimates for Soil Gas (SG) and Indoor Air (IA) Concentrations Relative to Reality: “T” Hypothesis is True, “F” Hypothesis is False; “-” Normal Distribution Does Not Apply

<table>
<thead>
<tr>
<th>Time (days)</th>
<th>m ∈ House No.</th>
<th>m ∈ House No.</th>
<th>m ∈ House No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>\mu_{IA_{1j}} - \hat{\epsilon}<em>{IA</em>{reality_{lj}}}</td>
<td>/</td>
<td>\mu_{IA_{2j}} - \hat{\epsilon}<em>{IA</em>{reality_{lj}}}</td>
</tr>
<tr>
<td>t_{ij}</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>0(100)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>100,200</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200,300</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>300,400</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>400,500</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>500,600</td>
<td>-</td>
<td>T/T</td>
<td>-</td>
</tr>
<tr>
<td>600,700</td>
<td>F/F</td>
<td>T/T</td>
<td>-</td>
</tr>
<tr>
<td>700,800</td>
<td>F/F</td>
<td>0/F</td>
<td>F/F</td>
</tr>
<tr>
<td>800,900</td>
<td>0/0</td>
<td>T/F</td>
<td>T/0</td>
</tr>
<tr>
<td>900,1000</td>
<td>T/0</td>
<td>T/F</td>
<td>T/0</td>
</tr>
<tr>
<td>1000,1100</td>
<td>T/0</td>
<td>T/F</td>
<td>T/0</td>
</tr>
<tr>
<td>1100,1200</td>
<td>T/0</td>
<td>T/F</td>
<td>T/0</td>
</tr>
<tr>
<td>1200,1300</td>
<td>T/0</td>
<td>T/F</td>
<td>T/0</td>
</tr>
<tr>
<td>1300,1400</td>
<td>T/0</td>
<td>T/F</td>
<td>T/0</td>
</tr>
<tr>
<td>1400,1500</td>
<td>T/0</td>
<td>T/F</td>
<td>T/0</td>
</tr>
<tr>
<td>1500,1600</td>
<td>T/0</td>
<td>T/F</td>
<td>T/0</td>
</tr>
<tr>
<td>1600,1700</td>
<td>T/0</td>
<td>T/F</td>
<td>T/0</td>
</tr>
<tr>
<td>1700,1800</td>
<td>T/0</td>
<td>T/F</td>
<td>T/0</td>
</tr>
</tbody>
</table>
For the pair of sampling strategies 1K2C to 1K4C, Table 3 indicates that there appears to be a consistent reduction in prediction uncertainty in the soil gas concentration beneath houses 3 and 4. However, there is no consistent reduction in prediction uncertainty for the indoor air concentrations for any of the houses. These results follow from the RMSE_{ij} analysis, and imply that only one permeability core data is insufficient for the KF to assimilate the available concentration data. For sampling strategies 1K4C to 7K4C, houses 3, 5 and 6 (and possibly 1) all show a significant reduction in the prediction uncertainty for soil gas concentrations. More importantly, all houses show a significant improvement for the indoor air concentrations. Finally, for sampling strategies 7K4C to 7K7C, there is a significant reduction in prediction uncertainty for soil gas concentrations beneath houses 1, 4 and 5. However, only house 5 shows a reduction in the prediction uncertainty of the indoor air concentrations. This occurs because 7C is the only strategy that samples the soil gas beneath house 5. In summary, there is a clear balance between the types and quantities of data that should be obtained as part of a site characterization effort in order to maximize their combined data worth. In the case of the problem at hand, 7K4C appears to provide the optimal balance.

B. Assessing the Significance in Uncertainty Reduction for Predicted Indoor Air Concentrations at Unmonitored Houses

The second metric measuring prediction uncertainty is addressed by applying a statistical test to infer whether there is a reduction in \( \sigma_{\text{SI}_{ij}}^2 \) and \( \sigma_{\text{AI}_{ij}}^2 \) for the pairs of incrementally increasing sampling data strategies represented by; 1K2C to 1K4C, 1K4C to 7K4C, and finally 7K4C to 7K7C.

The statistical test is based on the inference of variances of two normal populations “1” and “2”, with variances \( \sigma_{\text{AI}_{1ij}}^2 \) and \( \sigma_{\text{AI}_{2ij}}^2 \). For brevity, the presentation which follows is for indoor air concentrations but is equally applicable to soil gas concentrations. The development of this test hypothesis requires the introduction of the \( F \)-distribution, which is defined as follows:

\[
F = \frac{W/u}{Y/v}
\]

where \( W \) and \( Y \) are independent chi-square random variables with \( u \) and \( v \) degrees of freedom, respectively. The \( F \)-distribution has the probability density function:

\[
f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)} \left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1} x^{v/2} 0 < x < \infty
\]

where \( \Gamma \) is the gamma function. The statement for the hypothesis test is as follows:

Null hypothesis: \( H_0 : \sigma_{\text{AI}_{1ij}}^2 = \sigma_{\text{AI}_{2ij}}^2 \)  
Test statistic: \( f_0 = \frac{\sigma_{\text{AI}_{2ij}}^2}{\sigma_{\text{AI}_{1ij}}^2} \)

Alternative hypotheses:  
\( H_1 : \sigma_{\text{AI}_{1ij}}^2 \neq \sigma_{\text{AI}_{2ij}}^2 \)  
\( H_2 : \sigma_{\text{AI}_{1ij}}^2 > \sigma_{\text{AI}_{2ij}}^2 \)  
\( H_3 : \sigma_{\text{AI}_{1ij}}^2 < \sigma_{\text{AI}_{2ij}}^2 \)

Rejection criterion:  
\( f_0 > f_{1-\alpha,n_{\text{AI}_{1ij}}-1,n_{\text{AI}_{2ij}}-1} \) or \( f_0 < f_{1-\alpha,n_{\text{AI}_{1ij}}-1,n_{\text{AI}_{2ij}}-1} \)

\( f_1 - \alpha,n_{\text{AI}_{1ij}}-1,n_{\text{AI}_{2ij}}-1 = F^{-1}(\alpha)_{n_{\text{AI}_{1ij}}-1,n_{\text{AI}_{2ij}}-1} \)

The parameter \( \alpha \) denotes the level of statistical significance and \( F^{-1}(\alpha) \) is the quantile function of the \( F \)-distribution.

The results of this statistical test are reported on Table 4 for the pairs of incrementally increasing sampling data strategies represented by; 1K2C to 1K4C, 1K4C to 7K4C, and finally 7K4C to 7K7C. A value of “T” denotes acceptance of hypothesis \( H_2 \); “F” denotes acceptance of hypothesis \( H_4 \), and “0” denotes acceptance of hypothesis \( H_0 \), all at a level of significance of \( \alpha = 0.05 \) or 5%. In other words, “T” denotes that the second prediction uncertainty metric is achieved via the worth of adding more data, “F” denotes the data have the opposite impact, and “0” denotes that no statement of significance can be made. A dash “-” denotes that the test could not be conducted because one or both of the populations do not conform to a normal distribution.
Comparison of Table 4 to 3 indicates that the reduction in prediction uncertainty as represented by the second metric follows the same pattern as that of the first one, albeit with more “noise” (i.e. tests of “0” or “F”). The two patterns typically become aligned at late time near the end of the liability period. In summary, the results show that there is a clear balance between the types and quantities of data that should be obtained as part of a site characterization effort in order to maximize their combined data worth. Once again, 7K4C appears to provide the optimal balance.

C. Assessing the Significance in Improved Predictions for Probability of Exceedance at Unmonitored Houses

In Fig. 7, $P_{i,j}$ for $c_{iA_{i,j}}^+$ for sampling data strategies represented by: 1K2C, 1K4C, 7K4C and 7K7C are provided. For reference, the actual indoor air concentration $c_{iA_{i,j}}^{real}$ is provided in Fig. 3 as the red dashed line. Because house 2 is monitored for all sampling scenarios, there is uncertainty regarding its indoor air concentration. Houses 1, 3 and 6 which are all laterally offset from the groundwater plume, and for which the developer has liability, exhibit indoor air concentrations that in reality never exceed the regulatory limit. In contrast, house 5 which the developer already owns exceeds the regulatory limit only after 1100 days. Fig. 7 shows that 1K2C yields a probability of exceedance (using Eq. (10)) for houses 1, 3 and 6 which over-predicts reality, and correctly estimates $P_{i,j}$ for house 4. In addition, 1K2C underestimates $P_{i,j}$ for house 5. Surprisingly, the transition to 1K4C estimated the correct $P_{i,j}$ for all houses except for; one event in house 6 at 600 days, and house 5 for the entire monitoring period. The authors remind the reader that 1K4C performs poorly based on the first two performance metrics regarding prediction uncertainty. As the transition to 7K4C, $P_{i,j}$ is now correctly estimated for houses 1, 3 (except for one minor occurrence at 1100 days), 5 (after 1100 days) and 6. House 4 shows some minor values of probability of exceedance between 1100 and 1800 days. Finally, as the transition to 7K7C, the values of $P_{i,j}$ in house 4 are reduced as well as the values occurring before 1100 days in house 5. It is noted that the soil gas beneath house 5 is monitored for 7K7C, and yet the exact probability of exceedance outcome as occurs in reality cannot be obtained. This is due to 7K permeability sampling strategy incorrectly estimating the permeability below house 5 which leads to an incorrect estimate in the flux of TCE across the foundation slab and into the indoor air.
Fig. 7 The probability of exceedance $P_{ij}$ calculated using Eq. (10) for houses 1, 3, 4, 5 and 6 under alternative sampling strategies.

It is anticipated that through the process of collecting progressively more permeability and concentration data, the difference between the estimated probabilities of exceedance $P_{ij}$ and whether actual indoor air concentration exceeds the regulatory limit in reality $P_{\text{reality},ij}$ should tend to be zero. It is quantified as:

$$RMSE_{p,j} = \sqrt{\frac{\sum_{i=1}^{n_{\text{times}}} (P_{ij} - P_{\text{reality},ij})^2}{n_{\text{times}}}}$$  \hspace{1cm} (20)$$

where:

$$P_{\text{reality},ij} = \begin{cases} 1, & \text{if } c_{ij,\text{reality}} > c_{\text{regu}} \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (21)$$

and $RMSE_{p,j}$ denotes the root mean square error of the probability of failure for the $j$th house. Fig. 8b shows results for $RMSE_{p,j}$ with $P_{ij}$ calculated using Eq. (10). The results do not clearly show the anticipated pattern which is largely due to insufficient Monte Carlo realizations in order to adequately represent the “tail” of the distribution via Eq. (10). For instance, 1K seems to exhibit a lower $RMSE_{p,j}$ for various concentration sampling combinations relative to 3K, 5K and 7K. However, the results from Figs. 6 and 8a show that this is resulted from 1K erroneously underestimating indoor air concentrations. This in turn leads to a smaller tail on the probability distribution. In general, the anticipated pattern is apparent for 7C with the progression from 1K to 80K. Despite the “noise” in $RMSE_{p,j}$ where lines of different K samples cross, it does appear that for a
given sample line K there is a general decline in $RMSE_{p_j}$ with increasing values of C. This supports the idea that the combined worth of permeability and concentration data contributes to a reduction in prediction uncertainty as it applies to the third metric.

To show how the alternative method defining the probability of exceedance (fitting indoor air concentrations into a log-normal or a beta distribution) affects the improvement of prediction at unmonitored locations, Figs. 9 and 10 provide $P_{i,j}$ for $c_{iA_{l,j}}$ for sampling data strategies 1K2C, 1K4C, 7K4C and 7K7C using the log-normal and beta distributions, respectively.

For 1K2C, the $P_{i,j}$ pattern using both the log-normal and beta distributions is identical to that established using Eq. (10) (Fig. 7). Surprisingly, the log-normal distribution yields a lower $P_{i,j}$ at houses 3 and 6 relative to Eq. (10). The $P_{i,j}$ at house 5 is poorly predicted using all three methods. As the transition to 1K4C, both the log-normal and beta distributions now correctly estimate $P_{i,j}$ for all houses except for house 5. 1K4C performs poorly based on the first two metrics regarding to the prediction. In summary, little value in using either the log-normal or beta distributions to extrapolate the probability distribution for extreme events when $\mu_{iA_{l,j}}$ and $\sigma_{iA_{l,j}}$ are poorly constrained by insufficient data can be obtained. As the transition to 7K4C, the anticipated pattern for $P_{i,j}$ emerges. Using $P_{i,j}$ as calculated by Eq. (10) and shown in Fig. 7 as a benchmark, the log-normal distribution does provide slightly greater estimates of $P_{i,j}$ while the beta distribution provides slightly lower estimates for houses 1, 3, 4 and 6. For house 5, all three distributions appear to predict $P_{i,j}$ equally well. Finally, as the transition to 7K7C, the same pattern established by 7K4C is retained albeit with slightly lower values of $P_{i,j}$ relative to 7K4C. Of particular note is that the beta distribution for 7K7C yields $P_{\text{reality}_{i,j}}$ exactly for houses 1, 3 and 6, and exhibits only minor deviations between 1600 and 1800 days for house 4. As such, the beta distribution combined with sufficient data yields the most accurate prediction uncertainty estimate based on the third metric for those houses offset from the groundwater plume.

Figs. 8c and 8d show the $RMSE_{p_j}$ with $P_{i,j}$ calculated using the log-normal and the beta distributions. In general, the pattern of $RMSE_{p_j}$ for both the log-normal and beta distributions follows that of Eq. (10) and is subject to the same
interpretation. However, one additional feature becomes evident: once transitioning to a 7K sampling strategy, the $RMSE_{P_f}$ for the log-normal distribution appears greater than either that from Eq. (10) or the beta distribution, for any given C sampling strategy. Furthermore, 7K4C and 7K7C appear to perform equally well at positioning the $RMSE_{P_f}$ for the beta distribution as being greater than that of Eq. (10) but less than that of the log-normal distribution.

![Graph](image)

Fig. 9 The probability of exceedance $P_{ij}$ calculated using a log-normal distribution for houses 1, 3, 4, 5 and 6 under alternative sampling strategies
A data assimilation methodology was employed for optimal estimation of indoor air concentration. It first involved generating multiple Monte Carlo realizations of heterogeneous permeability fields conditioned upon measured values and a geo-statistical analysis of the borehole data. These realizations were then used with a numerical model to simulate distributions of soil gas and indoor air concentrations at locations throughout the simulation domain. These distributions were used within a discrete static Kalman filter to enable “actual” soil gas concentration data obtained as part of a sampling strategy to be used to estimate soil gas and indoor air concentrations at those locations where the developer does not have any data, but the liability. Actual soil gas concentration data were obtained by reserving a single heterogeneous permeability realization and denoting it as reality, and then simulating the fate and transport of TCE concentrations within it using the numerical model CompFlow Bio. By adopting the discrete static Kalman filter, an optimal estimate of the state vector of soil gas concentrations was produced. The state vector did not involve simultaneously estimating permeability. Instead, kriging was employed to generate multiple heterogeneous permeability realizations conditioned upon the available permeability data.

V. DISCUSSION AND CONCLUSIONS

The objective of this paper is to present a methodology to assimilate soil core permeability and TCE soil gas concentration data, and then to assess their worth in reducing the numerical model prediction uncertainty. The specific problem involves a residential development impacted by indoor air exposure of TCE contamination originating from a groundwater plume. The developer is assumed to own the houses above the plume, but not those adjacent to it. Furthermore, while the developer has the liability for those houses should their indoor air concentration exceed a regulatory limit, the developer is unable to access those houses and directly monitor their indoor air concentration. Hence, the developer must predict their indoor air concentrations based upon available data from neighbouring houses as well as soil gas concentrations within the aquifer.

A data assimilation methodology was employed for optimal estimation of indoor air concentration. It first involved generating multiple Monte Carlo realizations of heterogeneous permeability fields conditioned upon measured values and a geo-statistical analysis of the borehole data. These realizations were then used with a numerical model to simulate distributions of soil gas and indoor air concentrations at locations throughout the simulation domain. These distributions were used within a discrete static Kalman filter to enable “actual” soil gas concentration data obtained as part of a sampling strategy to be used to estimate soil gas and indoor air concentrations at those locations where the developer does not have any data, but the liability. Actual soil gas concentration data were obtained by reserving a single heterogeneous permeability realization and denoting it as reality, and then simulating the fate and transport of TCE concentrations within it using the numerical model CompFlow Bio. By adopting the discrete static Kalman filter, an optimal estimate of the state vector of soil gas concentrations was produced. The state vector did not involve simultaneously estimating permeability. Instead, kriging was employed to generate multiple heterogeneous permeability realizations conditioned upon the available permeability data.
The authors proposed three metrics to quantify the prediction uncertainty, namely: the ability to accurately predict the indoor air concentration within the houses at any point in time; the ability to reduce the standard deviation of predicted indoor air concentration within these houses; and the ability to accurately forecast the probability of indoor air concentrations exceeding a regulatory limit. Using the first and second metrics, the results imply that there is a great deal of value in collecting permeability data as part of a detailed site characterization (i.e. 7K) when attempting to use the KF to assimilate even a modest amount of soil gas concentration data (i.e. 4C) to improve the prediction accuracy and to reduce the prediction uncertainty of indoor air concentrations. In other words, once a detailed site investigation has been conducted, there is a point of diminishing returns where the further addition of soil gas concentration data (i.e. 4C to 7C) does little via the KF to significantly reduce the prediction uncertainty of indoor air concentrations. To assist with qualifying the first and second performance metrics, statistical inference tests to quantify statistically significant differences between sampling population sets were used. The fifty Monte Carlo permeability realizations were used to generate the sample populations. Although not shown, whether the fifty alternative permeability realizations were sufficient samples to estimate the first and second moments of the indoor air concentration distribution within all six houses at each tth monitoring time interval (t1−t, t2], were tested. While the first moment shows convergence, the second moment remains sensitive to additional realizations. The computational effort required to simulate fate and transport precluded further increasing the number of realizations.

The probability of exceedance as defined using Eq. (10) was used to estimate the third metric. In addition, both the log-normal and beta distributions were fit to the sample data in order to better resolve the low probability events characterized by the tail of the distribution given the limited number of Monte Carlo realizations. Following from the definition of the RMSE_P given by Eq. (20), one can conclude that for the lateral houses offset from the groundwater plume (i.e. houses 1, 3, 4 and 6) the log-normal distribution provides the most conservative estimates for prediction uncertainty as characterized by the probability of exceedance. In contrast, the beta distribution and Eq. (10) provide intermediate and least conservative estimates. The results show that extensive site characterization of the subsurface permeability structure is an essential prerequisite for estimating the probability of indoor air concentrations exceeding a regulatory limit, where the exceedance occurs as an extreme event.

In summary, the authors demonstrate that this approach is capable of providing the mean and standard deviation of estimated soil gas and indoor air concentrations. Furthermore, the worth of using progressively more permeability and soil gas concentration data is quantified on the basis that it provides a statistically significant improvement using the three metrics as measures for prediction uncertainty. In the companion paper, the authors will outline a methodology to price the worth of the data based on minimizing the project cost of the guarantee period of a brownfields redevelopment project. The project cost is the sum of all expected costs associated with repurchasing/renovating/reselling affected houses, the cost of collecting data, and a risk surcharge that is proportional to prediction uncertainty.

ACKNOWLEDGMENTS

Support for this work was provided by the University Consortium for Field-focused Groundwater Contamination Research. The authors greatly appreciate discussions with Larry Deschaine, Yingqi Zhang, George Pinder and David Landrault. The authors appreciate the advice of the anonymous reviewers who helped to improve this work.

REFERENCES

Journal of Civil Engineering and Science

Sept. 2014, Vol. 3 Iss. 3, PP. 150-170


