Improving Characterization of Vapor Intrusion Sites with A Deep Learning-based Data Assimilation Method

Jun Man[1], Yijun Yao[2], Jiangjiang Zhang[3], Junliang Jin[4], and Jianyun Zhang[4]

[1]Institute of Soil Science, Chinese Academy of Sciences
[2]Zhejiang University
[3]Hohai University
[4]Nanjing Hydraulic Research Institute

November 18, 2022

Abstract

Knowledge of soil properties is essential for risk assessment of vapor intrusion (VI). Data assimilation (DA) provides a valuable means to characterize contaminated sites by fusing the information contained in the measurement data (such as concentrations of volatile organic chemicals). Nevertheless, the application of DA in risk assessment of VI is quite limited. Moreover, soil heterogeneity is often overlooked in VI-related research. To fill these knowledge gaps, we apply a state-of-the-art DA method based on deep learning (DL), that is, $\mathrm{ES}_{(\mathrm{DL})}$, to better characterize the contaminated sites in VI risk assessment. The effectiveness of $\mathrm{ES}_{(\mathrm{DL})}$ is well demonstrated by three representative scenarios with increasing soil heterogeneity. The results clearly show that ignoring soil heterogeneity will significantly undermine one's ability to make reasonable decisions in VI risk assessment. As a preliminary attempt of applying an advanced DA method in VI research, this work provides implications for the potential of using DL and DA in complex problems that couple hydrological and environmental processes.

Improving Characterization of Vapor Intrusion Sites with A Deep Learning-based Data Assimilation Method

Jun Man[1,2], Yijun Yao[1,2], Jiangjiang Zhang[1,3], Junliang Jin[3], and Jianyun Zhang[3]

[1]Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, 210008, China,
[2]University of Chinese Academy of Sciences, Beijing, 100049, China,
[3]Yangtze Institute for Conservation and Development, Hohai University, Nanjing, 210008, China

Key Points:

- In risk assessment of vapor intrusion (VI), heterogeneity of soil properties is often overlooked
- We propose to use a deep learning-based data assimilation method to characterize complex VI sites
- Incorporation of site heterogeneity characterization significantly improves risk assessment of VI, even when an imperfect prior is employed

Corresponding author: J. Zhang, `zhangjiangjiang@hhu.edu.cn`

Abstract

Knowledge of soil properties is essential for risk assessment of vapor intrusion (VI). Data assimilation (DA) provides a valuable means to characterize contaminated sites by fusing the information contained in the measurement data (such as concentrations of volatile organic chemicals). Nevertheless, the application of DA in risk assessment of VI is quite limited. Moreover, soil heterogeneity is often overlooked in VI-related research. To fill these knowledge gaps, we apply a state-of-the-art DA method based on deep learning (DL), that is, $ES_{(DL)}$, to better characterize the contaminated sites in VI risk assessment. The effectiveness of $ES_{(DL)}$ is well demonstrated by three representative scenarios with increasing soil heterogeneity. The results clearly show that ignoring soil heterogeneity will significantly undermine one's ability to make reasonable decisions in VI risk assessment. As a preliminary attempt of applying an advanced DA method in VI research, this work provides implications for the potential of using DL and DA in complex problems that couple hydrological and environmental processes.

1 Introduction

Vapor intrusion (VI) is the exposure pathway that volatile organic chemicals (VOCs) migrate from the subsurface contaminated site (e.g., groundwater) into the building basement or foundation through soils (T. E. McHugh et al., 2012; Shirazi et al., 2019). When presented in the indoor air with certain levels, VOCs can pose risks to human health. In the past decades, VI has been identified in many different sites, and it has been drawing an increasing attention in the investigation methods, model development, and regulations, etc (Abreu & Johnson, 2005; DeVaull, 2007; Johnston et al., 2014; Ma et al., 2018, 2020; T. McHugh et al., 2017; Ström et al., 2019; Yao et al., 2013).

To assess the risk of VI, various models, ranging from analytical to numerical, from one-dimensional to three-dimensional, have been developed to simulate the process of VOC migration from the source to the receptor (Bozkurt et al., 2009; Pennell et al., 2009; Yao et al., 2011, 2012). In the simulation of VOC transport, there exist multiple sources of uncertainties, including those from environment, contaminant, and household properties (Johnston et al., 2014). Comprehensive analyses have shown that the movement of VOC in the vadose zone determines the distribution of vapor-phase contaminant in the soil profile, and has a considerable influence on the air quality (Abreu & Johnson, 2005; Yao et al., 2012, 2014). Thus, determining the soil properties is an indispensable step in risk assessment of VI. As demonstrated by previous works, soil hydraulic parameters like the porosity are important controlling factors of VI (Durner, 1994; Soto & Kiang, 2016), and soil heterogeneity plays an important role in many processes happening in the vadose zone, including the migration and mitigation of VOCs (Gao et al., 2019; Mousavi Nezhad et al., 2013; Nezhad et al., 2011; Reddy & Adams, 2001; Soto & Kiang, 2016; Verginelli et al., 2019). In such situations, accurately predicting VOCs movement from the subsurface site to the indoor space may be challenging. This necessitates characterization of heterogeneous soil properties in the risk assessment of VI sites.

When measurement data (e.g., VOC concentrations in the indoor air or soil profile) are available, one can utilize them to reduce uncertainties in the simulation of VI, and thus to better assess the health risk of this process (e.g., reduce false-negative or -positive identification of VI sites). One promising approach to fuse the measurement information into the VI model is data assimilation (DA; Carrassi et al., 2018). Nevertheless, rigorous quantification and reduction of uncertainties with DA is rather limited in VI-related research. One important work conducted by Johnston et al. (2014) used Markov chain Monte Carlo (MCMC, a well-known DA method) to update the VI model parameters (including soil hydraulic properties, air exchange rate, indoor-outdoor pressure difference, and building parameters, etc) from indoor VOC concentrations, to assist better decision making. In that work, homogeneous soil properties were assumed. In many cases, it is necessary to consider the soil heterogeneity (Bozkurt et al., 2009; Verginelli et al., 2019; Wang et al., 2020; Yao et al., 2017). Nevertheless, updating heterogeneous soil properties poses a significant challenge to DA methods. That might account for the reason why there is so few research (if there is any) that applies DA in heterogeneous field for VI risk assessment.

In geo- and environmental sciences, the most popular DA methods include MCMC, particle filter (PF), ensemble Kalman filter (EnKF) and its variants (Beven, 2010; Carrassi et al., 2018; Evensen, 2009; Klimova, 2018; Smith, 2013). However, MCMC and PF are not suitable for high-dimensional problems (Snyder et al., 2008; Zhang, Vrugt, et al., 2020), and EnKF, as well as its variants, are constrained by the Gaussian assumption (Evensen, 2009; Stordal et al., 2011; Zhang, Zheng, et al., 2020). To adequately characterize the site properties that govern the migration process of VOCs, one usually needs to handle a large number of unknown variables (which are intractable for MCMC and PF), and the model parameters, simulation outputs, and measurement errors involved may not be Gaussian-distributed (then EnKF and its variants are not applicable). In this situation, a more capable DA method is required. In the past decade, machine learning techniques, especially deep learning (DL), have been extensively used to solve tough problems in different research fields, including environmental risk assessment and protection (Aquilina et al., 2018; Mayr et al., 2016; Reichstein et al., 2019; Weichenthal et al., 2019). The success of DL comes from its power in extracting complex features automatically and simulate nonlinear relationships effectively from training data (Goodfellow et al., 2016; LeCun et al., 2015). Inspired by the advances in DL theories and applications, a new DA method, namely $\mathrm{ES}_{(\mathrm{DL})}$, is developed in our recent work (Zhang, Zheng, et al., 2020). $\mathrm{ES}_{(\mathrm{DL})}$ is efficient in solving high-dimensional DA problem that is free from the Gaussian assumption, and is thus utilized in the present work to quantify and reduce the uncertainty originated from the heterogeneous soil properties for VI risk assessment. As will be demonstrated in latter part of this paper, incorporating soil heterogeneity characterization can greatly improve one's knowledge about the transport process of VOCs, which is vital for decision making in environmental management.

The rest of this paper is organized as follows. In Section 2, we describe the concerned processes (i.e., transport of a representative VOC) and the corresponding governing equations. To improve our understanding of the underlying system from the measurement data, a state-of-the-art DA method, that is, $\mathrm{ES}_{(\mathrm{DL})}$, is formulated in the subsequent Section 3. In Section 4, we are concerned with benchmarking analysis of the benefit of considering soil heterogeneity when characterizing the contaminated site. Finally, we conclude and discuss this work in the Section 5.

2 Problem Description

2.1 Overview of The Study Area

In this study, we consider the migration of trichloroethylene (TCE) from the groundwater into a building through unsaturated soil. This building covers a wide area of 24 m×24 m with a 10 m×10 m square foundation 2 m below the ground. A 0.15 m×0.005 m crack at the foundation floor is the only way for TCE to enter the indoor air. Figure 1 depicts the sketch of the study domain, which is discretized into 933 nodes (the red dots signify the monitoring locations for TCE concentration in the soil profile) in the numerical model. Initially, the pressure head is −2 m on the ground surface and linearly increased to 0 m at the bottom (i.e., the groundwater level), representing a hydro-static condition. It means that the transport of liquid-phase TCE is not involved in the vadose zone. A volatile-type boundary condition is imposed on the top boundary and the crack, while the two lateral sides are impervious boundaries (Abreu & Johnson, 2005). At the bottom of the domain, the first-type (Dirichlet) boundary condition is imposed for solute transport. In all scenarios, TCE concentration (in the gas phase) at the pollution source is set as 1 mol/m^3. The simulation lasts for 500 days.

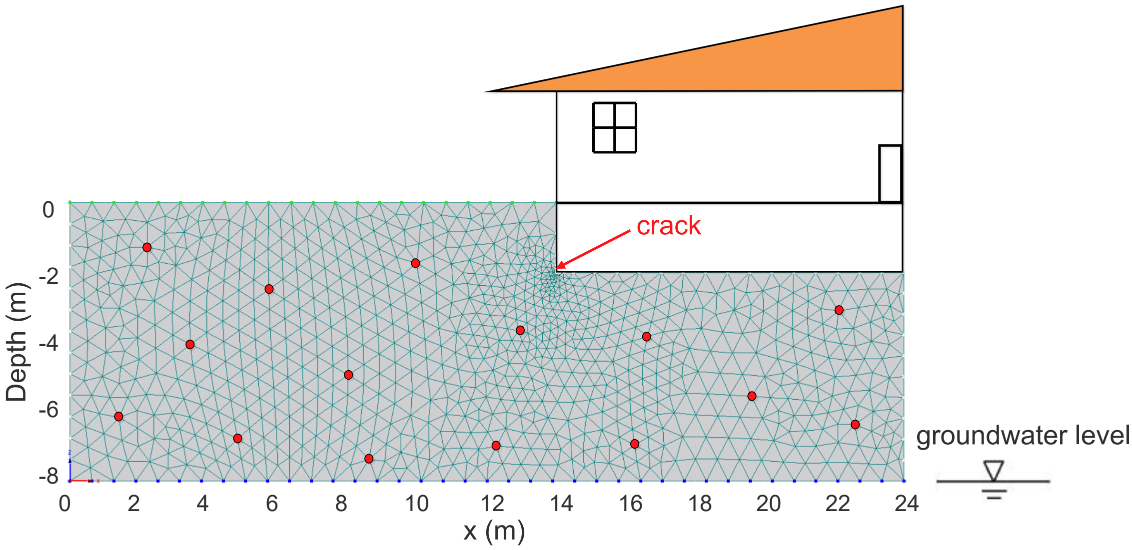

Figure 1. Sketch of the study domain. Concentrations of TCE are obtained at measurement locations denoted by the red dots.

2.2 Governing Equation

Here, we focus only on the diffusion and adsorption of TCE in the soil profile, which can be described by the following equation (Abreu & Johnson, 2005; Yao et al., 2013):

$$\left(\theta_{\mathrm{g}}+\frac{\theta_{\mathrm{w}}}{H}+\frac{K_{\mathrm{oc}}f_{\mathrm{oc}}\rho_{\mathrm{b}}}{H}\right)\frac{\partial C_{\mathrm{g}}}{\partial t}=\vec{\nabla}\cdot\left(D_{\mathrm{eff}}\vec{\nabla}C_{\mathrm{g}}\right), \tag{1}$$

where θ_{g} $[\mathrm{L}^3_{\mathrm{gas}}/\mathrm{L}^3_{\mathrm{soil}}]$ and θ_{w} $[\mathrm{L}^3_{\mathrm{water}}/\mathrm{L}^3_{\mathrm{soil}}]$ are the gas- and moisture-filled porosity, respectively; H is the Henry's law constant [-]; K_{oc} is the adsorption coefficient of TCE to soil organic carbon $[(\mathrm{M}/\mathrm{M}_{\mathrm{oc}})/(\mathrm{M}/\mathrm{L}^3_{\mathrm{water}})]$; f_{oc} is the mass fraction of soil organic carbon $[\mathrm{M}_{\mathrm{oc}}/\mathrm{M}_{\mathrm{soil}}]$; ρ_{b} is the soil bulk density $[\mathrm{M}_{\mathrm{soil}}/\mathrm{L}^3_{\mathrm{soil}}]$; C_{g} is the vapor concentration in soil (gas phase) $[\mathrm{M}/\mathrm{L}^3_{\mathrm{gas}}]$; t is the time [T]; $\vec{\nabla}$ is the del operator; and D_{eff} is the effective diffusion coefficient of TCE in soil $[\mathrm{L}^2/\mathrm{T}]$, which can be calculated as,

$$D_{\mathrm{eff}}=D_{\mathrm{g}}\frac{\theta_{\mathrm{g}}^{10/3}}{\theta_{\mathrm{T}}^2}+\frac{D_{\mathrm{w}}}{H}\frac{\theta_{\mathrm{w}}^{10/3}}{\theta_{\mathrm{T}}^2}, \tag{2}$$

where D_{g} and D_{w} are the diffusion coefficients of TCE in air and water $[\mathrm{L}^2/\mathrm{T}]$, respectively; $\theta_{\mathrm{T}}=\theta_{\mathrm{g}}+\theta_{\mathrm{w}}$ is the total soil porosity $[\mathrm{L}^3_{\mathrm{pores}}/\mathrm{L}^3_{\mathrm{soil}}]$; The relationship between θ_{w} and pressure head h is described by the van Genuchten model (Van Genuchten, 1980):

$$\frac{\theta_{\mathrm{w}}-\theta_{\mathrm{r}}}{\theta_{\mathrm{T}}-\theta_{\mathrm{r}}}=\begin{cases}\frac{1}{(1+|\alpha h|^n)^{1-1/n}}, & h<0\\ 1, & h\geq 0\end{cases}, \tag{3}$$

where θ_{r} is the residual moisture content [-], α [1/L] and n [-] are shape parameters related to the soil pore-size distribution.

Assuming the building as a well-mixed continuously stirred tank, the indoor TCE concentration, C_{in} $[\mathrm{M}/\mathrm{L}^3]$, at time t can be calculated by the following equation:

$$V_{\mathrm{b}}\frac{dC_{\mathrm{in}}}{dt}=M_{\mathrm{ck}}-C_{\mathrm{in}}A_{\mathrm{e}}V_{\mathrm{b}}, \tag{4}$$

$$C_{\mathrm{in},0} = \frac{M_{\mathrm{ck}}}{Q_{\mathrm{ck}} + V_{\mathrm{b}} A_{\mathrm{e}}} \approx \frac{M_{\mathrm{ck}}}{V_{\mathrm{b}} A_{\mathrm{e}}}, \tag{5}$$

where $C_{\mathrm{in},0}$ is the indoor TCE concentration at the initial time [M/L^3], chosen as the steady-state indoor TCE concentration; V_{b} is the building volume [L^3]; A_{e} is the indoor air exchange rate [1/T]; Q_{ck} is the soil gas flow rate to the enclosed space [L^3/T]; and M_{ck} is the contaminant entry rate through the crack [M/T], which can be estimated as:

$$M_{\mathrm{ck}} = \int_{\Omega} J_{\mathrm{ck}} d\Omega, \tag{6}$$

where J_{ck} is the mass flux of TCE through the crack [M/L^2/T]; and Ω is the cross-section area of the crack [L^2].

In the heterogeneous condition, analytical expression of equation (1) is usually not available. Here we use the finite element method to numerically solve the governing equation of TCE transport.

3 The Deep Learning-based Data Assimilation Method: $\mathbf{ES_{(DL)}}$

For the sake of simplicity, here we use the following compact form to represent the migration process of TCE , that is,

$$\widetilde{\mathbf{y}} = f(\mathbf{m}) + \epsilon, \tag{7}$$

where $f(\cdot)$ is the numerical model built with the finite element method; $\mathbf{m}$ is the vector for the model parameters, which include, but not limited to, spatially-distributed soil hydraulic parameters and variables that determine the transport of TCE vapor in the porous medium; and $\widetilde{\mathbf{y}}$ denotes the concentration measurements of TCE in the soil profile, which contain an error term ϵ.

To facilitate a better understanding of VI, it is essential to assimilate the measurements, $\widetilde{\mathbf{y}}$, to reduce the uncertainty of the model parameters, $\mathbf{m}$. Here, a state-of-the-art data assimilation method proposed in our recent work (Zhang, Zheng, et al., 2020) is adopted. This method, termed as $\mathrm{ES_{(DL)}}$, utilizes deep learning to handle non-linearity and non-Gaussianity encountered in many complex problems. Thus, it is a more general and flexible alternative of the widely-used EnKF (as well as its variants). In $\mathrm{ES_{(DL)}}$, we use the prior ensemble to represent our initial knowledge about the model parameters, that is, $\mathbf{M}^{(0)} = \left\{\mathbf{m}_1^{(0)}, ..., \mathbf{m}_{N_{\mathrm{e}}}^{(0)}\right\}$, where N_{e} is the ensemble size, and $\mathbf{m}_i^{(0)} \sim p(\mathbf{m}),\ i = 1, ..., N_{\mathrm{e}}$, and $p(\mathbf{m})$ is the prior distribution of model parameters. Then the corresponding model outputs are calculated as, $\mathbf{Y}^{(0)} = \left\{f(\mathbf{m}_1^{(0)}), ..., f(\mathbf{m}_{N_{\mathrm{e}}}^{(0)})\right\}$.

Essentially, data assimilation works by correcting the prior ensemble, $\mathbf{M}^{(0)}$, with the update vectors, $\Delta\mathbf{M}^{(0)} = \left\{\Delta\mathbf{m}_1^{(0)}, ..., \Delta\mathbf{m}_{N_{\mathrm{e}}}^{(0)}\right\}$, from the innovation vectors, $\Delta\mathbf{Y}^{(0)} = \left\{\widetilde{\mathbf{y}} - f(\mathbf{m}_1^{(0)}) + \epsilon_1, ..., \widetilde{\mathbf{y}} - f(\mathbf{m}_{N_{\mathrm{e}}}^{(0)}) + \epsilon_{N_{\mathrm{e}}}\right\}$, where $\epsilon_1, ..., \epsilon_{N_{\mathrm{e}}}$ are random realizations of measurement errors. Our new knowledge about the model parameters is represented by the updated ensemble, that is, $\mathbf{M}^{(1)} = \mathbf{M}^{(0)} + \Delta\mathbf{M}^{(0)}$. In $\mathrm{ES_{(DL)}}$, training data of innovation and update vectors are generated from $\mathbf{M}^{(0)}$ and $\mathbf{Y}^{(0)}$ as, $\mathbf{D}_{\mathrm{in}}^{(0)} = \left\{f(\mathbf{m}_i^{(0)}) - f(\mathbf{m}_j^{(0)}) + \epsilon_{ij} | i = 1, ..., N_{\mathrm{e}} - 1,\ i < j \leq N_{\mathrm{e}}\right\}$ and $\mathbf{D}_{\mathrm{out}}^{(0)} = \left\{\mathbf{m}_i^{(0)} - \mathbf{m}_j^{(0)} | i = 1, ..., N_{\mathrm{e}} - 1,\ i < j \leq N_{\mathrm{e}}\right\}$. From the training data $\mathbf{D}^{(0)} = \left\{\mathbf{D}_{\mathrm{in}}^{(0)}\ \mathbf{D}_{\mathrm{out}}^{(0)}\right\}$, a nonlinear mapping, $\mathcal{G}_{(\mathrm{DL})}[\cdot]$, from $\Delta\mathbf{Y}^{(0)}$ to $\Delta\mathbf{M}^{(0)}$ can be obtained with an adequate deep learning model, that is,

$$\Delta\mathbf{M}^{(0)} = \mathcal{G}_{(\mathrm{DL})}\left[\Delta\mathbf{Y}^{(0)}\right]. \tag{8}$$

Then each sample in $\mathbf{M}^{(0)}$ can be updated as,

$$\mathbf{m}_i^{(1)} = \mathbf{m}_i^{(0)} + \mathcal{G}_{(\mathrm{DL})}\left[\tilde{\mathbf{y}} - f(\mathbf{m}_i^{(0)}) + \epsilon_i\right], \tag{9}$$

where $\mathbf{M}^{(1)} = \{\mathbf{m}_1^{(1)}, ..., \mathbf{m}_{N_e}^{(1)}\}$ is the updated ensemble that contains the information assimilated from the measurement data. For highly nonlinear problems, using the measurements multiple times can be more effective, as doing local steps towards the measurements can make use of more linear fits to the data. Details about the $\mathrm{ES}_{(\mathrm{DL})}$ method can be found in (Zhang, Zheng, et al., 2020). As shown in Figure 2, we use the ResNet architecture (He et al., 2016) in $\mathrm{ES}_{(\mathrm{DL})}$ for the effective characterization of heterogeneous VI sites. Readers who are interested in the theories of deep learning and data assimilation are suggested to refer to (Goodfellow et al., 2016; LeCun et al., 2015) and (Carrassi et al., 2018; Evensen, 2009; Law et al., 2015), respectively.

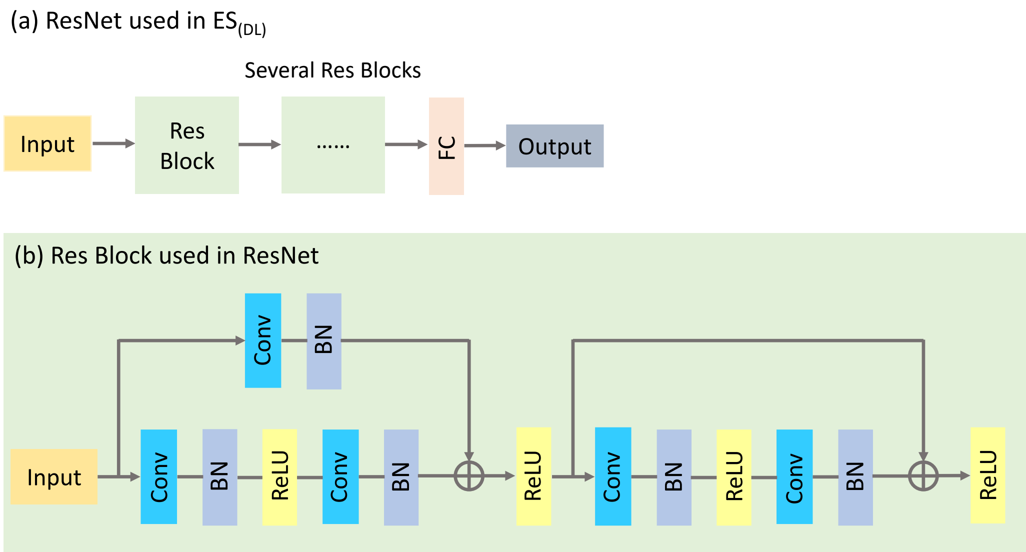

Figure 2. (a) The deep neural network (ResNet) used in $\mathrm{ES}_{(\mathrm{DL})}$ for data assimilation; (b) The residual (Res) block consisting of three layers, that is, convolutional layer (Conv), batch normalization layer (BN), and ReLU activation layer (ReLU). Here, FC denotes a fully-connected layer.

4 Illustrative Examples

Soil homogenization is common practice in risk assessment of VI (Friscia, 2014). Nevertheless, overlooking soil heterogeneity will certainly introduce some bias to the analysis results. To demonstrate whether ignoring soil heterogeneity will undermine the reliability of VI risk assessment, three representative VI scenarios are set up below.

4.1 Scenario 1: Layered Soil with The Accurate Prior

In the first scenario, we consider the migration of TCE in the soil consisting of three layers: sand, loamy sand and sandy loam (from top to bottom). The thicknesses of the three layers are 2 m, 3 m and 3 m, respectively. TCE concentrations at the monitoring locations (red dots in Figure 1) are measured at the 5th, 15th and 25th days. Three sets of θ_r, α and n corresponding to the three soils are unknown and to be inferred from these concentration measurements, while other parameters in the VI model are identified from experiments and/or literature, whose values are listed in Table 1.

Table 1. Available parameter values for the vapor intrusion model

Parameter	Symbol	Unit	Value
Foundation length	L	m	10
Foundation width	-	m	10
Foundation depth	d_{f}	m	2
Crack depth	d_{ck}	m	0.15
Crack width	w_{ck}	m	0.005
Space volume	V_{b}	m^3	174
Air exchange rate	$\mathrm{A_e}$	1/h	0.5
Henry's law constant	H	-	0.42
Gas diffusion coefficient	D_{g}	m^2/d	0.68
Liquid diffusion coefficient	D_{w}	m^2/d	7.86×10^{-5}
Longitudinal dispersivity	α_{L}	m	0.001
Transverse dispersivity	α_{T}	m	0.001
Concentration of pollution source (gas-phase)	C_{source}	$\mathrm{mol/m}^3$	1
Adsorption coefficient of organic carbon	K_{oc}	m^3/kg	0.126
Total soil porosity	θ_{T}	$\mathrm{m}^3/\mathrm{m}^3$	0.43

Carsel and Parrish (1988) proposed two Gaussian transformation approaches (LN: log-normal, and LR: log-ratio) to characterize the prior distributions of θ_{r}, α and n, that is,

$$\mathrm{LN}: Y = \ln(X), \tag{10}$$

$$\mathrm{LR}: Y = \ln[(X-u)/(v-X)], \tag{11}$$

where X is the parameter before transformation within the range of $[u, v]$; Y is the transformed parameter that is Gaussian distributed. The statistical parameters of sand, loamy sand and sandy loam used for distribution approximation are provided in Table 2. The factored covariance matrix ($\mathbf{L}^{\mathrm{T}}$) as shown in Table 3 is obtained by the Cholesky decomposition:

$$\mathbf{C}_{\mathrm{Y}} = \mathbf{L}\mathbf{L}^{\mathrm{T}}, \tag{12}$$

where $\mathbf{C}_{\mathrm{Y}}$ is the prescribed covariance matrix among the transformed parameters; $\mathbf{L}$ is a lower triangular matrix; and the superscript "T" denotes the transpose operator. Then random realizations of θ_{r}, α and n for the three soils can be generated as follows:

$$\mathbf{Y} = \mathbf{u} + \mathbf{L}\boldsymbol{\xi}, \tag{13}$$

where $\mathbf{u}$ is the mean vector for the transformed parameters and $\boldsymbol{\xi}$ is a standard Gaussian random vector.

Table 2. Statistical parameters used for distribution approximation (θ_r: m^3/m^3; α : $\times 10^2$/m; and "NO" means no transformation is needed for α in loamy sand)

Soil texture	Hydraulic parameter	Transformation type	Parameter ranges		Mean	Standard deviation
			u	v		
Sand	θ_r	LN	0	0.10	−3.120	0.224
	α	LR	0	0.25	0.378	0.439
	n	LN	1.50	4.00	0.978	0.100
Loamy sand	θ_r	LR	0	0.11	0.075	0.567
	α	NO	0	0.25	0.124	0.043
	n	LR	1.35	5.00	−1.110	0.307
Sandy loam	θ_r	LR	0	0.11	0.384	0.700
	α	LR	0	0.25	−0.937	0.764
	n	LN	1.35	3.00	0.634	0.082

Table 3. Correlation among the transformed hydraulic parameters represented by the factored covariance matrix

		θ_r	α	n
Sand	θ_r	0.182	0.258	−0.047
	α		0.143	−0.011
	n			0.017
Loamy sand	θ_r	0.522	0.017	−0.194
	α		0.014	0.019
	n			0.108
Sandy loam	θ_r	0.538	0.017	−0.194
	α		0.014	0.019
	n			0.108

Here, measurements of TCE concentration at the monitoring sites are obtained through running the numerical model with the true parameter values (red vertical lines in Figure 3) and perturbing the simulation results with errors that fit $\epsilon \sim \mathcal{N}(0, 0.01^2)$. Then we apply the $\mathrm{ES}_{(\mathrm{DL})}$ method to infer the unknown parameters in light of these measurement data. The network architecture used in $\mathrm{ES}_{(\mathrm{DL})}$ is presented in Figure 2, where only one residual block is employed. Prior and posterior distributions of the parameters of interest are presented in Figure 3. It can be seen that for most parameters, the uncertainty ranges are significantly reduced through assimilating the measurement data, and the true parameter values generally locate near the centers of these posterior curves.

In VI-related researches, to reduce the complexity of the problem, soil at the contaminated site was often assumed to be homogeneous (Friscia, 2014). Here, we test whether this simplification still hold in this layered-soil scenario. Three more cases are further tested, that is, assuming the soil as homogeneous and using the prior beliefs of sand, loamy sand and sandy loam respectively in data assimilation. In Figure 4, 95% confidence intervals of predicted indoor TCE concentrations with or without considering the layered soil heterogeneity are calculated and plotted against the simulation time. When the soil is assumed as a single layer of sand or loamy sand, indoor TCE concentrations will be over-estimated, which may lead to over-repair in practice; if the soil profile is treated as a single layer of sandy loam in the prior assumption, indoor TCE concentrations will be under-estimated, which may risk the human health; when the layered heterogeneity of soil is considered correctly (represented by the purple lines), the indoor TCE concentrations can be predicted accurately, indicated by a narrower and more accurate confidence interval. To provide quantitative comparisons, we calculate the root-mean-square errors (RMSEs) between the true TCE concentrations and the the predicted concentrations from different prior beliefs. The RMSE values are 4.07×10^{-6}, 1.15×10^{-6}, 1.26×10^{-6}, and 3.76×10^{-8} for the sand, loamy sand, sandy loam and layered assumptions, respectively. The result shows that the error can be reduced by two orders of magnitude if the layered-heterogeneity condition of soil is rigorously considered.

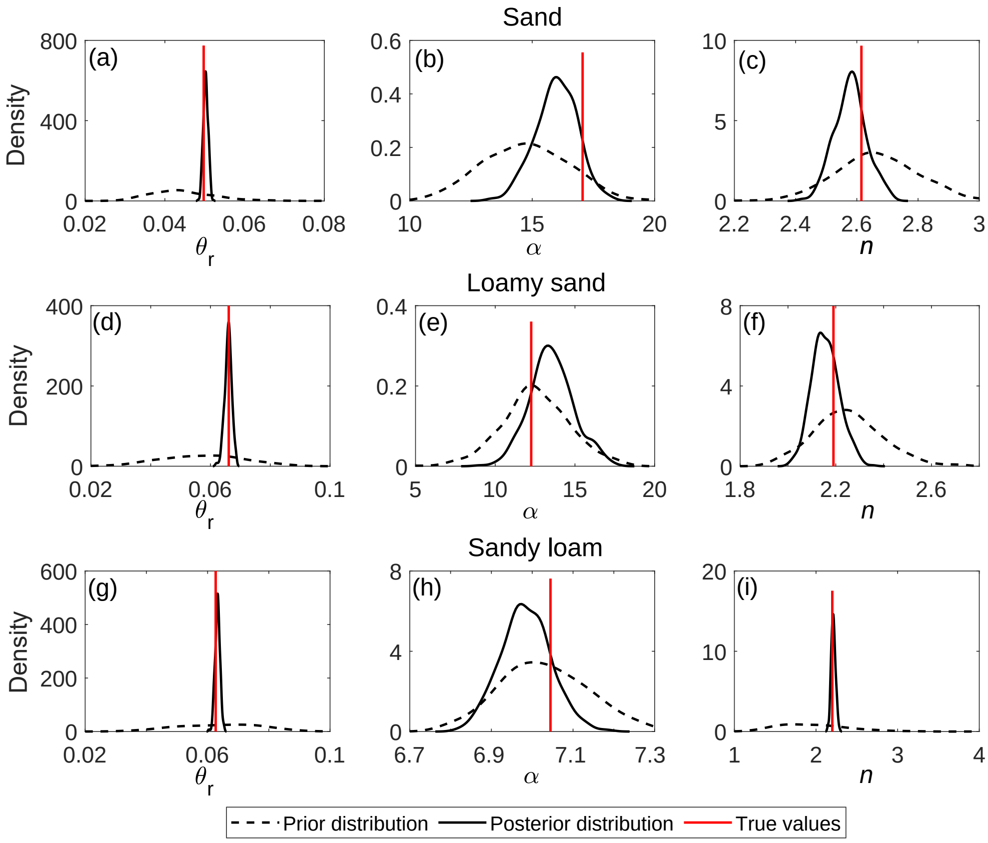

Figure 3. Prior and posterior distributions of the parameters of interest. Here the true parameter values are represented by the red vertical lines.

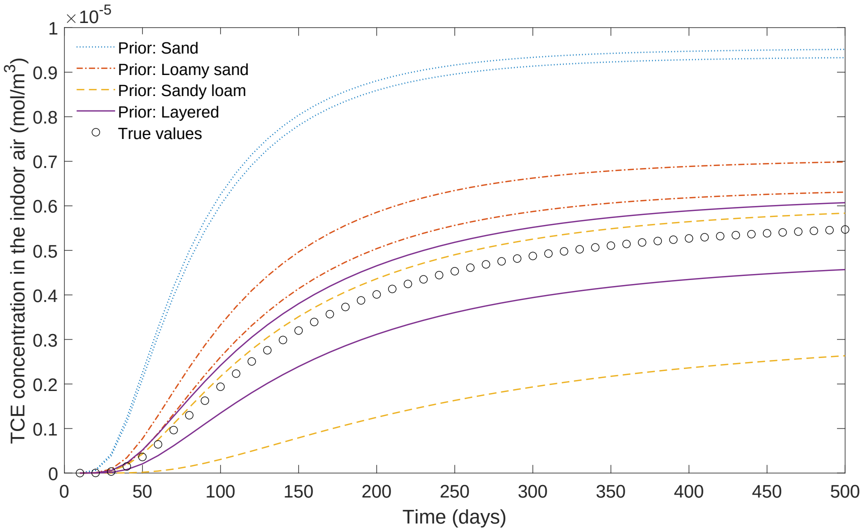

Figure 4. 95% confidence intervals of predicted TCE concentrations in the indoor air with different prior beliefs of the soil structure.

4.2 Scenario 2: Spatially Heterogeneous Field with The Accurate Prior

In addition to the layered structure considered in the prior section, below we set up a more complex scenario where the soil is spatially heterogeneous. Here, we assume that the uncertainty stems from the random field of α, whose prior is log-Gaussian (Li et al., 2009). That is to say, there are 931 (the number of model grids) unknown model parameters to be estimated. In the two-dimensional field of $Y = \ln(\alpha)$, the mean value is $\mu_Y = 2.21$, and the covariance between two arbitrary locations, (x_1, z_1) and (x_2, z_2), can be characterized by

$$C_Y\left[(x_1, z_1), (x_2, z_2)\right] = \sigma_Y^2 \exp\left[-\sqrt{\left(\frac{x_1 - x_2}{\lambda_x}\right)^2 + \left(\frac{z_1 - z_2}{\lambda_z}\right)^2}\right], \tag{14}$$

where $\sigma_Y^2 = 0.15$ is the variance of the Y field; $\lambda_x = 4$ m and $\lambda_z = 2$ m are the correlation lengths in the horizontal and vertical direction, respectively. With the given statistics, realizations of Y field can be generated using the Cholesky decomposition.

TCE concentrations at the monitoring locations (red dots in Figure 1) are measured every three days from the 3rd day to the 24th day. Thus, a total of 120 concentration measurements are obtained. Considering the complexity of the current scenario, we use two residual blocks in ResNet (Figure 2) for the $\mathrm{ES}_{(\mathrm{DL})}$ method. By assimilating the available measurements, we can obtain the mean estimate of Y field (Figure 5b) that can correctly capture the high and low regions of the true Y field (Figure 5a), yet their spatial extent is underestimated. This is caused by the sparsity of the measurement locations. Furthermore, we compare the 95% confidence intervals of predicted TCE concentrations in the indoor air calculated from the prior and posterior parameter ensembles respectively in Figure 6, which clearly indicate that uncertainty in the prediction can be greatly reduced. Based on the above results, we are confident to claim that the proposed method can effectively estimate the heterogeneous parameter field in light of the measurement data. For better estimation results, a larger and diverse data set is warranted.

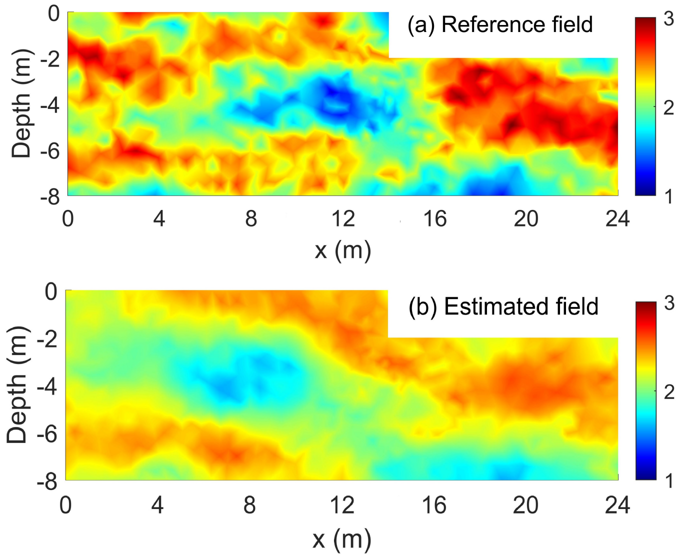

Figure 5. (a) Reference $\ln(\alpha)$ field, (b) Estimated mean $\ln(\alpha)$ field from data assimilation using the $ES_{(DL)}$ method.

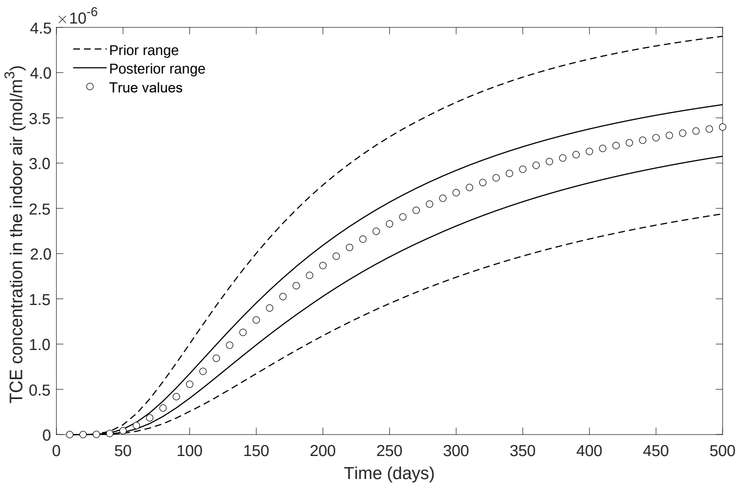

Figure 6. 95% confidence interval of predicted TCE concentrations in the indoor air for spatially heterogeneous field with the accurate prior.

4.3 Scenario 3: Spatially Heterogeneous Field with Imperfect Priors

In the above sections, we consider two representative scenarios of soil heterogeneity, that is, layered and spatially-heterogeneous, and accurate prior knowledge is available in advance. However, in many situations, accurate prior information is often difficult to obtain. Below we consider such a scenario where the exact prior information (spatially heterogeneous) is not known a priori, and imperfect assumptions (layered or homogeneous) are implied in data assimilation to reduce parameter uncertainties of θ_r, α and n.

The mean values of the three parameters are $\mu_{\theta_r} = 0.078\ \mathrm{m}^3/\mathrm{m}^3$, $\mu_\alpha = 3.6/\mathrm{m}$ and $\mu_n = 2.56$, and their coefficients of variation are $V_{\theta_r} = 0.2$, $V_\alpha = 0.4$ and $V_n = 0.4$, respectively. As $\theta_r > 0$, $\alpha > 0$ and $n > 1$, it is assumed that $\ln(\theta_r)$, $\ln(\alpha)$ and $\ln(n-1)$ are Gaussian distributed, whose statistical moments can be calculated as (Man et al., 2018):

$$\mu_Y = 2\ln(\mu_X) - 0.5\ln\left[\mu_X^2\left(1+V_X^2\right)\right], \tag{15}$$

$$\sigma_Y^2 = \ln\left(1+V_X^2\right), \tag{16}$$

where X and Y represent variables before and after the log-normal transformation, respectively. Using the Cholesky decomposition, random realizations of the three parameter fields can be obtained. Here the correlation lengths in the horizontal and vertical direction are $\lambda_x = 4$ m and $\lambda_y = 2$ m, respectively.

In risk analysis of VI, four imperfect prior assumptions are tested, that is, the homogeneous conditions using the prior of sand, loamy sand, and sandy loam, respectively, and the layered soil condition with three soil types. Based on the four priors, we implement data assimilation and make predictions. In Figure 7, 95% confidence intervals of the predicted indoor TCE concentrations are calculated and plotted against the simulation time. When the soil is assumed as sand or loamy sand, the indoor TCE concentrations will be overestimated; if the soil is assumed to be sandy loam, the uncertainty of predictions will be too high to inform decision making; although the layered assumption is still imperfect, much better predictions can be made, which is crucial for VI risk assessment. As for quantitative comparisons, the RMSEs between the true and predicted indoor TCE concentrations are calculated as 3.52×10^{-6}, 4.95×10^{-6}, 1.56×10^{-6}, and 1.46×10^{-7} for the above four prior beliefs, respectively. This indicates that the error can be reduced by at least one order of magnitude by applying the layered-soil assumption. The above results again demonstrate that characterizing soil heterogeneity is essential in VI-related research, even when accurate prior information is missing.

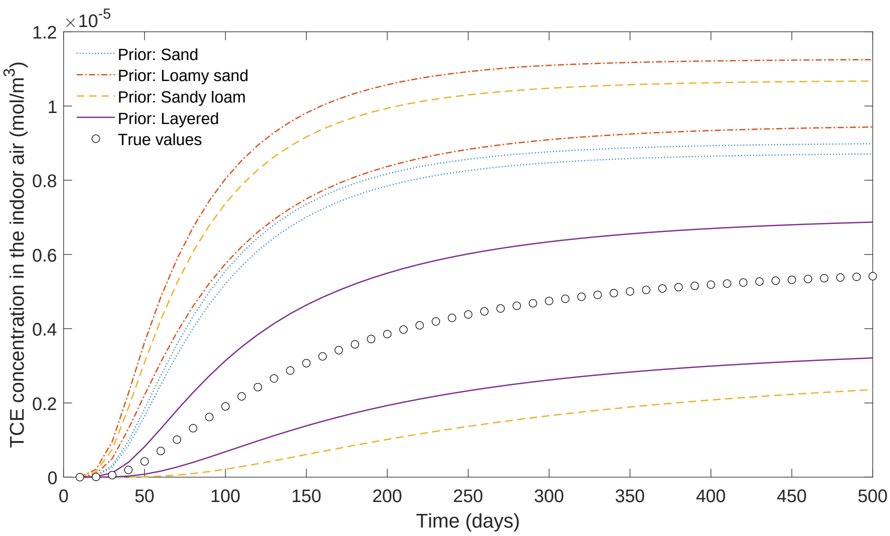

Figure 7. 95% confidence intervals of predicted TCE concentrations in the indoor air with four imperfect prior assumptions.

5 Conclusions

Heterogeneity is an inherent attribute of soil. In risk analysis of vapor intrusion (VI), assuming soil homogeneity is common practice, which can simplify the problem, but may undermine the effectiveness of decision making. Moreover, uncertainties are ubiquitous, yet the utilization of data assimilation to reduce uncertainties is lacking in VI-related research. To fill these gaps, we propose a deep learning-based data assimilation method, that is, $ES_{(DL)}$, and apply it to improve site characterization of VI in heterogeneous fields.

In this study, three representative scenarios, that is, layered soil with the accurate prior, spatially heterogeneous soil with the accurate prior, and spatially heterogeneous soil with imperfect priors, are tested. What we want to demonstrate through these case studies is that: if the soil heterogeneity is not reasonably treated, one's ability to understand the VI process and to make reasonable predictions will be compromised. It is true that considering soil heterogeneity will make the risk assessment of VI more challenging. Thus, It is strongly recommended to apply an adequate data assimilation method in VI-related research to fuse the information contained in the measurement data. This work is a preliminary attempt of using the deep learning-based data assimilation method to improve site characterization for VI risk assessment in heterogeneous soils, which provide important implications for both researchers and practitioners concerning risk assessment of VI at contaminated sites.

Here, only the uncertainties originated from measurement errors and model parameters are considered. In practice, another important source of uncertainty, that is, structural inadequacy of the VI model, should also be treated explicitly. To this end, different approaches can be adopted (Claeskens & Hjort, 2008; Evensen, 2019; Gupta et al., 2012; Xu & Valocchi, 2015). For example, in data assimilation, multiple competing VI models can be used simultaneously, and the degree of confidence of each model can be evaluated; one can also treat model structural errors as nuisance variables whose values are updated together with the model parameters; moreover, a data-driven model can be built for the model error, and this error model can be integrated into the system model to avoid over-confident predictions. These issues are very interesting and will be tested in future works.

Acknowledgments

In this work, Jun Man is supported by the Natural Science Foundation of Jiangsu Province (grant BK20201105) and the Innovation and Entrepreneurship Program of Jiangsu Province; Jiangjiang Zhang is supported by the National Natural Science Foundation of China (grant 41807006); Junliang Jin and Jianyun Zhang are supported by the National Natural Science Foundation of China (grant 51779144) and the National Key R&D Program of China (Grant 2017YFC1502706). The authors would also like thank Prof. Lingzao Zeng from Zhejiang University for providing insightful suggestions. Computer codes and data used are available at `https://www.researchgate.net/publication/348448583_DLDA_VI`.

References

Abreu, L. D., & Johnson, P. C. (2005). Effect of vapor source-building separation and building construction on soil vapor intrusion as studied with a three-dimensional numerical model. *Environmental Science & Technology*, *39*(12), 4550–4561. doi: 10.1021/es049781k

Aquilina, N. J., Delgado-Saborit, J. M., Bugelli, S., Ginies, J. P., & Harrison, R. M. (2018). Comparison of machine learning approaches with a general linear model to predict personal exposure to Benzene. *Environmental Science & Technology*, *52*(19), 11215–11222. doi: 10.1021/acs.est.8b03328

Beven, K. (2010). *Environmental modelling: An uncertain future?* Routledge, London; New York: CRC Press.

Bozkurt, O., Pennell, K. G., & Suuberg, E. M. (2009). Simulation of the vapor intrusion process for nonhomogeneous soils using a three-dimensional numerical model. *Groundwater Monitoring & Remediation*, *29*(1), 92–104. doi: 10.1111/j.1745-6592.2008.01218.x

Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, *9*(5), e535. doi: 10.1002/wcc.535

Carsel, R. F., & Parrish, R. S. (1988). Developing joint probability-distributions of soil-water retention characteristics. *Water Resources Research*, *24*(5), 755–769. doi: 10.1029/WR024I005P00755

Claeskens, G., & Hjort, N. L. (2008). Model selection and model averaging. *Cambridge Books*. doi: 10.1017/CBO9780511790485

DeVaull, G. E. (2007). Indoor vapor intrusion with oxygen-limited biodegradation for a subsurface gasoline source. *Environmental Science & Technology*, *41*(9), 3241–3248. doi: 10.1021/es060672a

Durner, W. (1994). Hydraulic conductivity estimation for soils with heterogeneous pore structure. *Water Resources Research*, *30*(2), 211–223. doi: 10.1029/93WR02676

Evensen, G. (2009). *Data assimilation: the ensemble Kalman filter.* Berlin, Germany.

Evensen, G. (2019). Accounting for model errors in iterative ensemble smoothers. *Computational Geosciences*, *23*(4), 761–775. doi: 10.1007/s10596-019-9819-z

Friscia, J. M. (2014). *Vapor intrusion modeling: limitations, improvements, and value of information analyses* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Gao, H., Zhang, J., Liu, C., Man, J., Chen, C., Wu, L., & Zeng, L. (2019). Efficient Bayesian inverse modeling of water infiltration in layered soils. *Vadose Zone Journal*, *18*(1), 1–13. doi: 10.2136/vzj2019.03.0029

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep Learning* (Vol. 1). Cambridge, MA: MIT press Cambridge.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*(8), W08301. doi: 10.1029/2011WR011044

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). Las Vegas, NV: IEEE. doi: 10.1109/CVPR.2016.90

Johnston, J. E., Sun, Q., & Gibson, J. M. (2014). Updating exposure models of indoor air pollution due to vapor intrusion: Bayesian calibration of the Johnson-Ettinger model. *Environmental Science & Technology*, *48*(4), 2130–2138. doi: 10.1021/es4048413

Klimova, E. (2018). Application of the ensemble Kalman filter to environmental data assimilation. In *Iop conference series: Earth and environmental science* (Vol. 211, pp. 1755–1307). Tomsk, Russian Federation.

Law, K., Stuart, A., & Zygalakis, K. (2015). *Data Assimilation: A Mathematical Introduction* (Vol. 62). Springer.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi: 10.1038/nature14539

Li, W., Lu, Z., & Zhang, D. (2009). Stochastic analysis of unsaturated flow with probabilistic collocation method. *Water Resources Research*, *45*(8), W08425. doi: 10.1029/2008WR007530

Ma, J., Jiang, L., & Lahvis, M. A. (2018). Vapor intrusion management in china: lessons learned from the united states. *Environmental Science & Technology*, *52*(6), 3338-3339. doi: 10.1021/acs.est.8b00907

Ma, J., McHugh, T., Beckley, L., Lahvis, M., DeVaull, G., & Jiang, L. (2020). Vapor intrusion investigations and decision-making: A critical review. *Environmental science & technology*, *54*(12), 7050–7069. doi: 10.1021/acs.est.0c00225

Man, J., Zhang, J., Wu, L., & Zeng, L. (2018). ANOVA-based multi-fidelity probabilistic collocation method for uncertainty quantification. *Advances in Water Resources*, *122*, 176–186. doi: 10.1016/j.advwatres.2018.10.012

Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, *3*, 80. doi: 10.3389/fenvs.2015.00080

McHugh, T., Loll, P., & Eklund, B. (2017). Recent advances in vapor intrusion site investigations. *Journal of Environmental Management*, *204*, 783–792. doi: 10.1016/j.jenvman.2017.02.015

McHugh, T. E., Beckley, L., Bailey, D., Gorder, K., Dettenmaier, E., Rivera-Duarte, I., . . . MacGregor, I. C. (2012). Evaluation of vapor intrusion using controlled building pressure. *Environmental Science & Technology*, *46*(9), 4792–4799. doi: 10.1021/es204483g

Mousavi Nezhad, M., Javadi, A., Al-Tabbaa, A., & Abbasi, F. (2013). Numerical study of soil heterogeneity effects on contaminant transport in unsaturated soil (model development and validation). *International Journal for Numerical and Analytical Methods in Geomechanics*, *37*(3), 278–298. doi: 10.1002/nag.1100

Nezhad, M. M., Javadi, A., & Rezania, M. (2011). Modeling of contaminant transport in soils considering the effects of micro-and macro-heterogeneity. *Journal of Hydrology*, *404*(3-4), 332–338. doi: 10.1016/j.jhydrol.2011.05.004

Pennell, K. G., Bozkurt, O., & Suuberg, E. M. (2009). Development and application of a three-dimensional finite element vapor intrusion model. *Journal of the Air & Waste Management Association*, *59*(4), 447–460. doi: 10.3155/1047-3289.59.4.447

Reddy, K. R., & Adams, J. A. (2001). Effects of soil heterogeneity on airflow patterns and hydrocarbon removal during in situ air sparging. *Journal of Geotechnical and Geoenvironmental Engineering*, *127*(3), 234–247. doi: 10.1061/(ASCE)1090-0241(2001)127:3(234)

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, *566*(7743), 195–204. doi: 10.1038/s41586-019-0912-1

Shirazi, E., Ojha, S., & Pennell, K. G. (2019). Building science approaches for vapor intrusion studies. *Reviews on Environmental Health*, *34*(3), 245–250. doi: 10.1515/reveh-2019-0015

Smith, R. C. (2013). *Uncertainty quantification: Theory, implementation, and applications* (Vol. 12). Philadelphia, PA: SIAM.

Snyder, C., Bengtsson, T., Bickel, P., & Anderson, J. (2008). Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, *136*(12), 4629–4640. doi: 10.1175/2008MWR2529.1

Soto, M. A., & Kiang, C. H. (2016). Vapor intrusion in soils with multimodal pore-size distribution. In *E3s web of conferences* (Vol. 9, p. 07002).

Stordal, A. S., Karlsen, H. A., Nævdal, G., Skaug, H. J., & Vallès, B. (2011). Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter. *Computational Geosciences*, *15*(2), 293–305. doi: 10.1007/s10596-010-9207-1

Ström, J. G., Guo, Y., Yao, Y., & Suuberg, E. M. (2019). Factors affecting temporal variations in vapor intrusion-induced indoor air contaminant concentrations. *Building and Environment*, *161*, 106196. doi: 10.1016/j.buildenv.2019.106196

Van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, *44*(5), 892–898. doi: 10.2136/sssaj1980.03615995004400050002x

Verginelli, I., Yao, Y., & Suuberg, E. M. (2019). Risk assessment tool for chlorinated vapor intrusion based on a two-dimensional analytical model involving vertical heterogeneity. *Environmental Engineering Science*, *36*(8), 969–980. doi: 10.1089/ees.2018.0468

Wang, G., Xiao, Y., Zuo, J., Wang, Y., Man, J., Tang, W., . . . Yao, Y. (2020). Physically simulating the effect of lateral vapor source-building separation on soil vapor intrusion: Influences of surface pavements and soil heterogeneity. *Journal of Contaminant Hydrology*, *235*, 103712. doi: 10.1016/j.jconhyd.2020.103712

Weichenthal, S., Hatzopoulou, M., & Brauer, M. (2019). A picture tells a thousand. . . exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. *Environment International*, *122*, 3–10. doi:

10.1016/j.envint.2018.11.042

Xu, T., & Valocchi, A. J. (2015). A Bayesian approach to improved calibration and prediction of groundwater models with structural error. *Water Resources Research*, *51*(11), 9290–9311. doi: 10.1002/2015WR017912

Yao, Y., Pennell, K. G., & Suuberg, E. M. (2012). Estimation of contaminant subslab concentration in vapor intrusion. *Journal of Hazardous Materials*, *231*, 10–17. doi: 10.1016/j.jhazmat.2012.06.016

Yao, Y., Shen, R., Pennell, K. G., & Suuberg, E. M. (2011). Comparison of the Johnson-Ettinger vapor intrusion screening model predictions with full three-dimensional model results. *Environmental Science & Technology*, *45*(6), 2227–2235. doi: 10.1021/es102602s

Yao, Y., Shen, R., Pennell, K. G., & Suuberg, E. M. (2013). A review of vapor intrusion models. *Environmental Science & Technology*, *47*(6), 2457–2470. doi: 10.1021/es302714g

Yao, Y., Verginelli, I., & Suuberg, E. M. (2017). A two-dimensional analytical model of vapor intrusion involving vertical heterogeneity. *Water Resources Research*, *53*(5), 4499–4513. doi: 10.1002/2016WR020317

Yao, Y., Yang, F., Suuberg, E. M., Provoost, J., & Liu, W. (2014). Estimation of contaminant subslab concentration in petroleum vapor intrusion. *Journal of Hazardous Materials*, *279*, 336–347. doi: 10.1016/j.jhazmat.2014.05.065

Zhang, J., Vrugt, J. A., Shi, X., Lin, G., Wu, L., & Zeng, L. (2020). Improving simulation efficiency of MCMC for inverse modeling of hydrologic systems with a Kalman-inspired proposal distribution. *Water Resources Research*, *56*(3), e2019WR025474. doi: 10.1029/2019WR025474

Zhang, J., Zheng, Q. A., Wu, L., & Zeng, L. (2020). Using deep learning to improve ensemble smoother: Applications to subsurface characterization. *Water Resources Research*, *56*(12), e2020WR027399. doi: 10.1029/2020WR027399