Integration of a Contaminant Source Land Use Regression Model in the Bayesian Maximum Entropy Spatiotemporal Geostatistical Estimation of Groundwater Tetrachloroethylene Across North Carolina

Kyle P. Messier

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirement for the degree of Master of Science in the Department of Environmental Science and Engineering

Chapel Hill

2010

Approved by:

Marc L. Serre, Advisor

Gregory Charakalis, Reader

Jacqueline MacDonald Gibson, Reader

Rebecca C. Fry, Reader

ABSTRACT

Kyle P. Messier

Integration of a Contaminant Source Land Use Regression Model in the Bayesian Maximum Entropy Spatiotemporal Geostatistical Estimation of Groundwater Tetrachloroethylene Across North Carolina

(Under the direction of Marc L. Serre)

The assessment of groundwater tetrachloroethylene (PCE or PERC) exposure across North Carolina is currently hindered due to limited statewide spatiotemporal contaminant maps. In this study we incorporate data from multiple sources to create estimation maps of groundwater PCE. A land use regression (LUR) mean trend model was developed as a function of exponentially decaying contribution from contaminant sources in North Carolina. This mean trend model was integrated in a Bayesian Maximum Entropy (BME) framework to produce informative space/time (S/T) maps. We compare our method with standard geostatistical methods (i.e. kriging and BME with constant mean trends) and find a 25 % reduction in cross-validation mean

square error. Our results suggest that dry cleaning and hazardous waste generator sites influence groundwater at distances of 1 km and 800 m respectively. This work introduces a novel integrated LUR and BME approach which produces accurate visual representations of PCE exposure across North Carolina.

Table of Contents

List of Tables

List of Figures

1. INTRODUCTION

Tetrachloroethylene (PCE or PERC) is a chlorinated solvent that is commonly used for dry cleaning of fabrics and for metal degreasing operations[1], and "likely carcinogenic to humans" according to the United States Environmental Protection Agency (USEPA) [2]. PCE is associated with both acute and chronic human exposures which can likely lead to health effects including nausea, headache, and cancer of the liver, lungs, and kidney[1]. In addition, PCE is one of the most frequently detected volatile organic compounds in groundwater in the United States [3-5]. The USEPA delegates private well standards to the states; North Carolina uses a groundwater quality standard for PCE of 0.7 ppb [6], designed to protect the health of private well owners. In North Carolina, at least 1,500 sites are estimated to be contaminated with PCE or similar solvents [7].

The current groundwater PCE management program in North Carolina is divided between the Department of Environment and Natural Resources (NCDENR) and the Department of Health and Human Services (NCDHHS). While this program is sufficient for post-hoc case by case management, it is limited for statewide exposure assessment and lacks predictive capabilities. One approach for modeling large-scale environmental exposure, which combines Space/Time Random Field (S/TRF) theory and Bayesian Maximum Entropy (BME), has proven successful in the statistical space/time estimation in surface water [8] and in air quality[9]. Another approach for modeling environmental exposure is land use regression, which has also proven successful in the statistical estimation of air quality contaminants [10].

A statewide groundwater PCE exposure assessment can help state agencies better protect public health; however, budget constraints, sparse data, and the extensive manpower required for well monitoring make statewide assessments difficult. In our study, we combine data from NCDENR, NCDHHS, and USGS to propose an integrated land use regression and BME approach, which leads to a cost-effective statewide PCE exposure assessment. To the authors' knowledge, an approach has not been implemented for PCE estimation which combines land use regression with S/TRF theory and BME.

Space/time random field theory provides a framework to model the variability and uncertainty of environmental parameters (e.g. groundwater pollutants) across space and time in terms of a probability distribution function (PDF) [11]. Space/time BME is a modeling technique that allows one to incorporate general knowledge (e.g. covariance) and site-specific knowledge about the spatial process of interest to produce maps that represent the distribution of the parameter at any unsampled point of interest, resulting in informative maps of water quality [11]. Furthermore, the BME framework allows for the general knowledge to be informed by a physically meaningful mean trend, such as a land use regression model.

This research proposes an approach within the space/time epistemic BME framework in conjunction with a land use regression model based on pollution sources. Specifically, instead of using constant global or local constant mean trend models, we define a mean trend model that is based on land use regression principles. We use groundwater quality data from multiple publicly available data sources encompassing the full range of site types: (i) monitoring wells near known contaminated sites, (ii) private wells, and (iii) ambient monitoring wells. We also incorporate contaminant source variables for the land use regression model as explanatory variables for PCE concentration. This approach was used to assess groundwater PCE concentration across the state

and predict potential undiscovered areas of contamination. The presented work includes (i) a land use regression mean trend that accounts for the effect of contaminant sources on groundwater PCE concentration; (ii) BME integration of the developed land use model and general and site specific knowledge bases about concentration residuals that yield informative space/time maps describing the distribution of groundwater PCE across North Carolina; and (iii) a cross-validation model comparison against geostatistical methods with constant mean trends. Finally, we conclude on the policy relevance of this work for groundwater PCE exposure.

2. MATERIALS & METHODS

2.1 Tetrachloroethylene Data Sources

Data on groundwater PCE were compiled from three sources, which are detailed as follows:

1. ***DSCA EDD Monitoring Wells***

North Carolina monitors PCE through the Dry Cleaning and Solvent Cleanup Act (DSCA) section of the N.C. Division of Waste Management, which was established to help fund cleanup of PCE contamination[12]. DSCA maintains contracts with private companies to construct monitoring wells, which in turn provide DSCA with an electronic data deliverable (EDD) that contains the locations of PCE concentrations in monitoring wells. There are approximately 207 DSCA sites distributed across the state, but EDD's are not available for all the sites yet. For this study, we have data from 48 DSCA monitoring sites, collected from 1999-2010, resulting in 641 monitoring wells with 709 space/time samples. It should be noted that the DSCA monitoring

sites are spatially clustered since all of the monitoring wells around a known polluted site are approximately within a square kilometer area.

2. DHHS Geocoded Private Wells

The North Carolina Department of Health and Human Services collects organic VOC data from North Carolina homeowners. Prior to 2007 the data collected were from homeowners who voluntarily had their well tested. Starting in 2007 all new wells built were required by law to be tested [13]. The data are analyzed at the Department of Public Health State Lab, where a paper report for each well is created and stored. There is no standard for providing GPS coordinates in the report; however, the well address is provided. Consequently, we digitized the paper reports by hand and then applied a geocoding scheme to obtain geographic coordinates. Using the address locator tool of ArcGIS ™, data were assigned coordinates in a multi-stage process using a North Carolina point reference file (courtesy of NCDHHS Spatial Analysis Group), followed by a North Carolina Department of Transportation line reference, then with a U.S. street address line reference file (Tele Atlas Dynamap Transportation, 2003). The locational error of geocoded addresses with a match score (A number between 0 and 100 that represents the overall accuracy of the address located datum.) of 70 and above have previously been shown to not be significantly different than those with a 100 match score using these reference files, therefore all geocoded addresses with a match score of 70 and above were included in the dataset[14]. The address geocoding resulted in 2,411 geocoded wells with 2,874 space/time samples from the years 2003-2010 that were previously unavailable.

3. USGS National Water Information Systems Wells

We downloaded all of the PCE well data available from the USGS NWIS website (http://nwis.waterdata.usgs.gov). We obtained 71 monitoring wells with 94 space/time samples from 2001-2010 distributed across the state.

The dataset post-processing is housed in an electronic database which contains the following fields: PCE value (ppb), longitude and latitude (North American Datum 83), data source (figure S1), site ID for EDD data, well ID (a unique identifier for every well; ID's given by an organization are maintained), sample date, and sample detection limit. Our blending of data sources resulted in 3,123 unique wells with 3,650 space/time samples.

2.2 Land Use Regression Model

2.2.1 Dependent Variable

The global mean trend of groundwater PCE was estimated by a land use regression model, where the dependent variable is the log-transformed PCE concentration obtained above. By taking the log-transformation we reduce the skewness from 21.34 to 2.62. Our PCE monitoring data contained below detect data; therefore a method to account for samples without detectable PCE was necessary. There are a variety of acceptable methods to handle left-censored below detect environmental data, including assigning the below detect a value of half the detection limit [8] or performing the analysis based on detection frequency [4]. In this study we model the probability distribution function (PDF) of log-PCE using a Gaussian distribution with a mean μ and variance σ^2 such that the cumulative distribution function (CDF) at the detection limit and the 95th percentile produce values equal to the percent of samples below detect and the 95th percentile of the sampled values, respectively. A full numerical description for the technique is described in the supplementary material. Once the full PDF of log-PCE is obtained, we assign below detect

data to the mean of the truncated normal (Gaussian) distribution, truncated at the detection limit (Figure 1).

2.2.2 Known and Potential Sources of PCE

PCE almost always occurs because of anthropogenic causes[1, 15], thus we constructed the independent variable based on the locations of sites that are known or potential sources of PCE. The location and associated information for land use variables were obtained from NC Division of Waste Management GIS personnel [15] and from NC Onemap [16], a public online database for GIS data. We incorporate the following land use variables into our contaminant source database: dry cleaners including DSCA and non-DSCA sites; Resource Conservation and Recovery Act (RCRA) hazardous waste generator sites; Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA or Superfund) sites; National Pollutant Discharge Elimination System (NPDES) sites; septage land application sites/ septage detention or treatment facility sites (Septage); brownfield sites; landfills (current and pre-regulatory); and manufacturing gas plants (MGP) sites.

2.2.3 Independent Variables Based on Contamination Sources

As mentioned above the occurrence of PCE in groundwater is mainly associated with anthropogenic sources. It is generally believed that major types of sources include dry cleaners, hazardous waste generators and Superfund sites, but other types of sources cannot be discounted [1]. For each type of pollution source *l*, (e.g. *l*=dry cleaners) we construct an explanatory variable calculated as the cumulative exponentially decaying contribution from each polluted site of that type , which can be expressed as

$$X_i^{(l)} = \sum_{j=1}^{n} \exp\left(-3 * \frac{D_{ij}}{a_l}\right), \quad (1)$$

where $X_i^{(l)}$ is the contamination contribution at well *i* from source *l*, D_{ij} is the distance between well *i* and polluted site *j*, *n* is the total number of polluted sites of type *l*, and a_l is the exponential decay range defining the pollution length-scale of that type of pollution source. The exponential operator in the model ensures concentration decreases quickly as the distance increases from the contaminant source. The cumulative aspect of the model accounts for the density of contaminant sources.

2.2.4 Contaminant Source Land Use Regression Model

The dependency of groundwater PCE log-concentration, Z, with different types of known sources can be expressed for sample *i* as

$$Z_i = \beta_0 + \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \cdots + \beta_m X_i^{(m)} + \epsilon_i \qquad (2)$$

where Z_i is the log-PCE concentration estimate for sample *i*, $X_i^{(1)}$ through $X_i^{(m)}$ are explanatory variables representing the cumulative exponentially decaying contribution from different types of contaminant sources, $\beta_0, \ldots, \beta_m$ are linear regression coefficients, and ϵ_i is an error term. This model allows investigation into the effects of various types of contaminant sources as well as the value for the decay range, a_l, associated with each type of source, which describes the distance corresponding to a 95 percent reduction in log-PCE. First, we investigate the effect of each decay range individually by constructing a series of univariate models for each pollution type *l*, and exploring how the univariate coefficient of determination r^2 changes as a function of each decay range a_l. Then we explore the interaction of decay ranges by examining how r^2 changes in the multivariate model (Eq. 2) as a function of various combinations of decay ranges. We ultimately choose the multivariate regression model with maximum r^2 obtained with physically meaningful (i.e. positively valued) and statistically significant regression coefficients. The decay ranges and

corresponding regression coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_m$ obtained for that model can then be used to construct the land use regression model $L_Z(\boldsymbol{s})$ of log-PCE concentration at any spatial location $\mathbf{s}=(s_1,s_2)$ as

$$L_Z(\boldsymbol{s}) = \hat{\beta}_0 + \hat{\beta}_1 X^{(1)}(\boldsymbol{s}) + \hat{\beta}_2 X^{(2)}(\boldsymbol{s}) + \cdots + \hat{\beta}_m X^{(\mathrm{m})}(\boldsymbol{s}), \quad (3)$$

where $X^{(1)}(\boldsymbol{s}), \ldots, X^{(\mathrm{m})}(\boldsymbol{s})$ are the cumulative exponentially decaying contribution from each type of pollution sources calculated for the spatial location $\boldsymbol{s}$. .

2.3 Bayesian Maximum Entropy Estimation Framework for Space/Time Mapping Analysis

In this study we use the BME method of modern spatiotemporal geostatistics [17,21] to estimate the concentration of groundwater PCE across space and time. *BMElib* [18,11] , a powerful MATLAB numerical toolbox of modern spatiotemporal geostatistics implementing the BME theory , was used to create space/time maps of PCE concentration across North Carolina. This framework has been successfully applied to groundwater [19,20] and environmental contaminants[8, 9, 21]. As shown in these studies, BME is a space/time geostatistical estimation framework grounded in epistemic principles that reduces to the space/time simple, ordinary, and universal kriging methods as its linear limiting case when considering a limited, Gaussian, knowledge base , while also allowing the flexibility to process a wide variety of additional knowledge bases (physical laws, empirical relationships, non-Gaussian distributions, hard and soft data, etc.) that are beyond the reach of the kriging methods of linear geostatistics. We only provide the fundamental BME equations for mapping PCE; the reader is referred to other works for more detailed derivations of these equations [17,18,22,11].

The theory of space/time random field (S/TRF) is used to model the variability and uncertainty associated with the distribution of PCE concentration across space and time. Our notation for variables will consist of denoting a single random variable Z in capital letter, it realization, z, in lower case; and vectors and matrices in bold faces, e.g. $\boldsymbol{Z} = [Z_1, \dots, Z_n]^T$ and $\boldsymbol{z} = [z_1, \dots, z_n]^T$. Let $Y(\boldsymbol{p})$ be the S/TRF describing the distribution of PCE concentration across space and time, and let $Z(\boldsymbol{p}) = \log Y(\boldsymbol{p})$ be its log-transform, where $\boldsymbol{p} = (\mathbf{s}, t)$, $\boldsymbol{s}$ is the space coordinate and t is time. The log-transformed residual S/TRF is defined as

$$X(\boldsymbol{p}) = Z(\boldsymbol{p}) - m_Z(\boldsymbol{s}) \tag{4}$$

where $m_Z(\boldsymbol{s})$ is a global geographical trend that can be modeled using various models. In this work, we first use a constant global geographical trend, and we then compare that approach with using $m_Z(\boldsymbol{s}) = L_Z(\boldsymbol{s})$, which allows to integrate the land use model in the geostatistical estimation analysis. Equation (4) then expresses that the S/TRF $X(\boldsymbol{p})$ models the space/time variability and uncertainty associated with the difference between the S/TRF $Z(\boldsymbol{p})$ and its global geographical trend model.

The knowledge available is organized in the general knowledge base ($\mathcal{G}$-KB) about the S/TRF $X(\boldsymbol{p})$ (e.g. describing its space/time variability, mean, covariance, etc.) and the site-specific knowledge base ($\mathcal{S}$-KB) corresponding to the hard and soft data available at a set of specific space/time points $\boldsymbol{p}_d$. The BME fundamental set of equations for modeling the S/TRF $X(\boldsymbol{p})$ is [22, 23,21]

$$\begin{cases} \int d\boldsymbol{x}\, (\boldsymbol{g}(\boldsymbol{x}) - E[\boldsymbol{g}]) e^{\mu^T \boldsymbol{g}(\boldsymbol{x})} = 0 \\ \int d\boldsymbol{x} f_S(\boldsymbol{x}) e^{\mu^T \boldsymbol{g}(\boldsymbol{x})} - A f_K(x_k) = 0 \end{cases} \tag{5}$$

where $\boldsymbol{x}$ is a vector of log-transform residual PCE concentrations at mapping points $\boldsymbol{p}$ consisting of the union of the data points $\boldsymbol{p}_d$ and the estimation point $\boldsymbol{p}_k$, $\boldsymbol{g}$ is a vector of functions selected

such that their expected values $E[\boldsymbol{g}]$ is known from the $\mathcal{G}$-KB, $f_S(\boldsymbol{x})$ is a PDF characterizing the knowledge and uncertainty associated with the $\mathcal{S}$-KB, A is a normalization constant, and f_K is the BME posterior probability density function describing residual PCE concentration at the estimation point $\boldsymbol{p}_k$, where the subscript $K = \mathcal{G} \cup \mathcal{S}$ means that f_K is based on the blending of the $\mathcal{G}$- and $\mathcal{S}$-KB.

The $\mathcal{G}$-KB for the S/TRF $X(\boldsymbol{p})$ describes its local space/time trends and dependencies. In this work, the general knowledge consists of the space/time mean trend function $m_x(\boldsymbol{p}) = E[X(\boldsymbol{p})]$, and the covariance function $C_X(\boldsymbol{p}, \boldsymbol{p}')$=$E[[X(\boldsymbol{p}) - m_x(\boldsymbol{p})][\mathrm{X}(\mathbf{p}') - m_x(\boldsymbol{p}')]]$ of the S/TRF $X(\boldsymbol{p})$. .

A key conceptual difference in this work and that of classical geostatistical estimation techniques is how we treat the below detect data to obtain $\mathcal{S}$-KB. In the classical kriging case, and to calculate $C_X(\boldsymbol{p}, \boldsymbol{p}')$, we harden the below detect to the truncated Gaussian mean as explained earlier. On the other hand in the BME approach we are able to rigorously account for the measurement uncertainty associated with any below detect by selecting a PDF $f_S(x_{\mathrm{soft}})$ that takes the full shape of the Gaussian distribution of PCE concentrations truncated above the detection limit (figure 1), which for sample i is given by

$$f_S(x_{soft}; \mu, \sigma, b_i) = \frac{1}{\sigma}\phi\left(\frac{x_{soft} - \mu}{\sigma}\right) / \Phi\left(\frac{b_i - \mu}{\sigma}\right) \qquad (6)$$

for $x_{\mathrm{soft}} < b_i$ and 0 otherwise, where ϕ is the standard normal PDF, Φ is its CDF, μ and σ are the mean and standard deviation of PCE estimated from left censored PCE data (see Fig. 1), $b_i = log(DL_i) - m_Z(\mathbf{s}_i)$, DL_i is the detection limit for sample i, and $m_Z(\mathbf{s}_i)$ is its global geographical trend value (Eq. 4). It follows that the site-specific knowledge consists of the hard data points, $\boldsymbol{X}_{hard}$, that is points measured above their detection limit, and soft data points, $\boldsymbol{X}_{soft}$, that is

points measured below their detection limit. The overall knowledge bases considered consist of $\mathcal{G} = \{m_X(\boldsymbol{p}), C_X(\boldsymbol{p}, \boldsymbol{p}')\}$, and $\mathcal{S} = \{f_S(.), \boldsymbol{X}_{hard}\}$. In this case the BME fundamental set of equations reduces to

$$f_K(x_k) = A^{-1} \int d\mathbf{x}\ f_S(\mathbf{x}) f_G(\mathbf{x}) \qquad (7)$$

where $f_G(\mathbf{x}) = e^{\mu^T \mathbf{g}(\mathbf{x})}$ is the Gaussian PDF for $\boldsymbol{X}$ obtained from the $\mathcal{G}$-KB, $\mathbf{x}$ is a realization of $\boldsymbol{X}$, $f_S(\mathbf{x})$is the truncated Gaussian PDF of $\boldsymbol{X}_{soft}$ and A is a normalization constant.

In this study we average measurements by the year they were sampled; thus we model the yearly average of PCE concentrations. General and site-specific knowledge were processed as described above by use of BMElib to obtain BME estimates of log-transformed residual S/TRF $X_y(\boldsymbol{p})$ across North Carolina for each year of the study period. The BME estimate for a given year is a function of data collected in that year, as well as years prior to and after that year. The estimation error associated with BME estimate $X_y(\boldsymbol{p})$ is fully characterized by the BME posterior PDF. The expected value and corresponding estimation error variance of the corresponding PCE concentration estimate at that estimation point is obtained by adding the global geographical trend $m_Z(\boldsymbol{s})$, and back log-transforming the BME posterior PDF for $X_y(\boldsymbol{p})$. This results in BME maps showing the space/time distribution of yearly PCE concentration across North Carolina.

2.4 Cross-Validation

Our approach has two distinct advantages over classical kriging techniques. First, we account for the full distribution of below detect by modeling it as truncated Gaussian soft data (eq. 6). Second, we use a land use model based on contaminant sources (eq. 3) to better inform the

estimation maps with a physically meaningful global geographical trend. Hence, we expect a gain of information in each step of our analysis.

In order to investigate the gain of information with each step of our approach, we calculate the mean square error (MSE) for some step (k) of the analysis as

$$MSE^{(k)} = \frac{1}{n}\sum_{j=1}^{n}\left(Z_j^{*(k)} - Z_j\right)^2 \quad (8)$$

where n is the number of data points, Z_j is the jth measured log-transformed yearly average PCE concentration, and $Z_j^{*(k)}$ is its corresponding estimate at stage (k). At each stage Z_j^* is estimated by removing Z_j from the data and re-estimating it using other data points. The MSE provides a measure of model estimation standard deviation. Using the cross-validation MSE we compare three estimation approaches consisting of (a) using a classical simple kriging technique where the global geographical trend is constant, i.e. $m_Z(\boldsymbol{s})=m$, and where below detect data are hardened to the truncated Gaussian mean; (b) using a simple BME technique where $m_Z(\boldsymbol{s})=m$ and with truncated Gaussian soft below detect data; and (c) using a LUR/BME approach the same as (b) but setting the global geographical trend to the land use model, i.e. $m_Z(\boldsymbol{s})= L_Z(\boldsymbol{s})$. We let MSE_{SK}, MSE_{BME}, and MSE_{LUR} be the mean square error for scenarios (a), (b), and (c), respectively. We define the percent change in mean square error *PCMSE* between two scenarios i and j as

$$PCMSE_{i/j} = \left(\frac{MSE_j - MSE_i}{MSE_i}\right) * 100 \quad (9)$$

Where i/j can be set to a/b or b/c. A negative PCMSE indicates a decrease in MSE, which corresponds to the percent improvement in estimation accuracy resulting from incorporating truncated Gaussian soft data and a contaminant source land use mean trend.

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

We find the mean and standard deviation for groundwater log-PCE to be -3.47*log-ppb* and 5.56*log-ppb* respectively. The minimum value was -7.4063*log-ppb* (exp(-7.4063)≈0.0006*ppb*), which was calculated as the truncated mean from a below detect observation with a detection limit of -0.6931*log-ppb* (≈0.5*ppb*). The maximum observed value was 10.6213*log-ppb* (≈ 41,000*ppb*). We expect the population mean of groundwater PCE to be low since it is not a ubiquitous contaminant, which we see with a mean of -3.47*log-ppb* (≈0.031*ppb*) well below the North Carolina groundwater standard. The large standard deviation of 5.56*log-ppb*)) is most likely due to the large range of detected values, from -1.5*log-ppb* (≈0.22*ppb*) to 10.6213*log-ppb* (≈41,000*ppb*).

3.2 Contaminant Source Land Use Regression Model

Contaminant source land use regression coefficients and statistics were calculated at regular intervals for the decay range in univariate and multivariate models (Eq. 3). We classify the explanatory variables according to their decay range r^2 curves (i.e. plot of r^2 versus the decay range) obtained for the univariate regression model (Figure 2). In the univariate case, the explanatory variables constructed from dry cleaners and RCRA sites explained the most variability in log-PCE concentration with r^2 values reaching a maximum of 0.20 and 0.17, respectively, for decay ranges of 1.25*km* and 0.67*km*, respectively (Table 1). We therefore classify the dry cleaners and RCRA sites into *Class 1* contaminant sites. *Class 1* contaminant sites are ones corresponding to high r^2, positive β_1 values and short decay ranges, all together

indicating they are actual local sources of groundwater PCE. We then classify within *Class 2* those contaminants sites corresponding to explanatory variables that explain between two and ten percent of the variability in log-PCE concentration, have positive β_1 values, and have decay ranges of 10-60 km (Figure 2). *Class 2* contaminant sites are not themselves direct, local sources of contamination, but represent surrogates for the presence of direct sources. We note that the Brownfield variable, a *Class 2* variable, has a small first peak at a short range indicating the possibility that it is a local source of PCE, but the peak in r^2 is not the absolute maximum and it has a lower value than our *Class 1* variables. Lastly, *Class 3* contaminant sites are ones that explain less than 2 % of the variability in log-PCE (Figure 2).

In the bivariate case, we did not see a significant increase (> 0.02) in r^2 for all possible combinations except when *Class 1* variables were combined. We found that when the dry cleaners and RCRA sites explanatory variables were combined there was a 0.02 increase in r^2 to 0.22 (Figure S2, Table S1). The resulting model has a high r^2, highly significant and positive coefficients, and accounts for the interaction between the two variables. When going from the univariate models to the bivariate model, the regression coefficients change from 3.83 to 3.07 and 1.89 to 0.64 for dry cleaners and RCRA respectively, while the corresponding decay ranges change from 1.25*km* to 0.99*km* and 0.67*km* to 0.80*km*, respectively. Multivariate models beyond two explanatory variables do not yield significantly higher r^2values, thus our land use regression modeling process stopped at the bivariate case. However, in other situations (i.e. different contaminant or different geographical location), multivariate models could provide additional information; therefore it is recommended that regression models should be calculated until there is no significant gain in percent of variance explained when adding variables.

Distance decay range curves (Figure 2, Figure S2) have been shown in previous studies to identify the range of influence for contaminant sources [10]. For groundwater PCE, logistic regression analysis has shown moderate associations with RCRA and CERCLA sites within 1 km [4]. This study is the first to quantify the distance of influence of PCE sources exhibit in North Carolina. Our findings suggest that the method outlined from equations 1-3, or similarly in Su et al. 2009, is a sound approach to identify ranges of influence for groundwater PCE. Based on our findings, we suggest that wells in North Carolina used for drinking water be set back farther than 1 km from a dry cleaner. This is a substantially larger distance than required by North Carolina code and generally farther than required by DSCA for known contaminated sites [24]. Our recommendation is substantially larger because (1) the reported dry cleaner locations may not always correspond to the exact location of the plume, (2) the zone of influence includes the main segment of the PCE plume and its 95 % removal distance at the edge of the plume, and (3) our maps are the average of the S/TRF realizations. Our results also highlight the cumulative effect of contaminant sources; hence density and distance of contaminant sources should be considered when establishing screening guidelines indicating which wells should be tested for PCE. For instance, in Figure S4 areas with only one or two RCRA sites nearby are estimated below the groundwater standard; however, dense clusters of RCRA sites lead to high estimated concentrations that can exceed the standard.

3.3 Space/Time Covariance Model

Exploratory data analysis confirmed that when setting the global geographical trend $m_Z(\boldsymbol{s})$ equal to the land use regression model $L_Z(\boldsymbol{s})$ then the residual field $X(\boldsymbol{p})$ can reasonably be modeled as being homogeneous/stationary because $L_Z(\boldsymbol{s})$ captures the main non-homonegeous trends in

PCE. As a result the covariance of $X(\boldsymbol{p})$ between points $\boldsymbol{p} = (\boldsymbol{s}, t)$ and $\boldsymbol{p}' = (\boldsymbol{s}', t')$ can be modeled as being only a function of the spatial lag $r = \left|\left|\boldsymbol{s} - \boldsymbol{s}'\right|\right|$ and the temporal lag $\tau = |t - t'|$. Using a numerical algorithm we developed to handle data unevenly distributed over space and time, we calculate experimental covariance values for $X(\boldsymbol{p})$ by finding pairs $(\boldsymbol{p}, \boldsymbol{p}')$ of measurement events that are separated by various values of r in distance and τ in time. We then used the experimental values to fit the nonseparable space/time covariance model

$$C_X(r, \tau) = c_1 \exp\left(-\frac{3r}{a_{r_1}}\right)\exp\left(-\frac{3\tau}{a_{\tau_1}}\right) + c_2 \exp\left(-\frac{3r}{a_{r_2}}\right)\exp\left(-\frac{3\tau}{a_{\tau_2}}\right) \quad (10)$$

where $c_1 = 7.49\ \sigma_X^2$, $a_{r_1} = 0.01$ Km, $a_{\tau_1} = 4$ years, $c_2 = 2.50\ \sigma_X^2$, $a_{r_2} = 3$ Km, and $a_{\tau_2} = 8$ years. The covariance model (eq 10, figure S3) provides useful information about the variability of detrended PCE in the groundwater of North Carolina. We see a very short spatial covariance range of only 0.01 Km in the first covariance structure, which describes the large variability of PCE within a short distance of dry cleaning and RCRA point sources. We also see a long spatial range in the second covariance structure, which describes the larger geographical extant of areas with non-detected PCE concentrations. We see long ranges in both temporal covariance structures because PCE can persist in the groundwater or soil for many months or years with little biodegradation [1]. Our experimental covariance calculations (figure S3) suggest that it can persist for years at detectable levels.

3.4 Space/Time Bayesian Maximum Entropy Maps

The general and site-specific knowledge was processed in BMElib to obtain the BME posterior PDF of PCE at any location and year of interest. The BME estimates can be used to construct maps describing the spatial distribution of groundwater PCE across North Carolina for a sampling year of interest. Figures 4a, 4b and 4c show maps of groundwater PCE concentrations

across North Carolina in 2009 estimated using methods a (kriging), b (BME) and c (LUR/BME) described earlier. Figure 4a, obtained using the Kriging approach, shows most areas well below the North Carolina groundwater standard, but it also shows small areas (i.e. in Guilford, Durham, Wake, etc. counties) above the standard. Figure 4b, obtained using BME, is similar to Figure 4a, but it provides a more detailed visualization of the groundwater PCE distribution for values below the detection limit. Finally figure 4C, obtained using the LUR/BME approach, is further improved with the incorporation of a meaningful global geographical mean trend based on the LUR model. This map estimates concentrations near or above the groundwater standard at places far from where any type of monitoring data exists. The map in Figure 4c can therefore be used to identify areas where contamination likely exists above the standard.

3.5 Cross-Validation

In this study, we present an integrated approach for modeling groundwater PCE at the statewide scale. We compare our integrated approach, which incorporates a contaminant source land use mean trend, with two geostatistical approaches that implement a constant global geographical trend. The cross-validation mean square errors for all three approaches are summarized in Table 2. We find a 23.75 percent decrease in MSE when using the BME approach (b) accounting for the full truncated Gaussian distribution for the below detect data, compared with the kriging approach (a), which hardens the below detect values (Both use the same constant global geographical trend.). This MSE reduction demonstrates the advantage of using BME over kriging, which is explained by the fact that BME provides a rigorous non-Gaussian statistical representation of the possible values a below detect datum can take.

When using the LUR/BME approach, which incorporates the contaminant source land use global geographical trend in the estimation framework, we observe an additional 25.46

percent decrease in MSE compared to the BME approach with a constant mean trend. This demonstrates the benefit of integrating a physically meaningful land use regression geographical trend into geostatistical estimation techniques.

3.6 Further Research

We incorporate a meaningful land use model based on anthropogenic sources of PCE; however local and regional hydrogeologic features, soil sorption, and hydraulic gradients also play a role in the occurrence of groundwater contamination [4]. Future research could incorporate these features as explanatory variables in the land use regression. Since there is no limitation to the types of data BME can incorporate, other soft data could be included in the analysis. Such data might include modeled PCE based on the degradation by-products of PCE (TCE, DCE isomers, Vinyl Chloride). More epidemiologic studies are needed to assess the health impacts of PCE [2]. BME methodology would allow one to account for uncertainties in data providing a more accurate assessment of exposure for use in such studies.

4. FIGURES

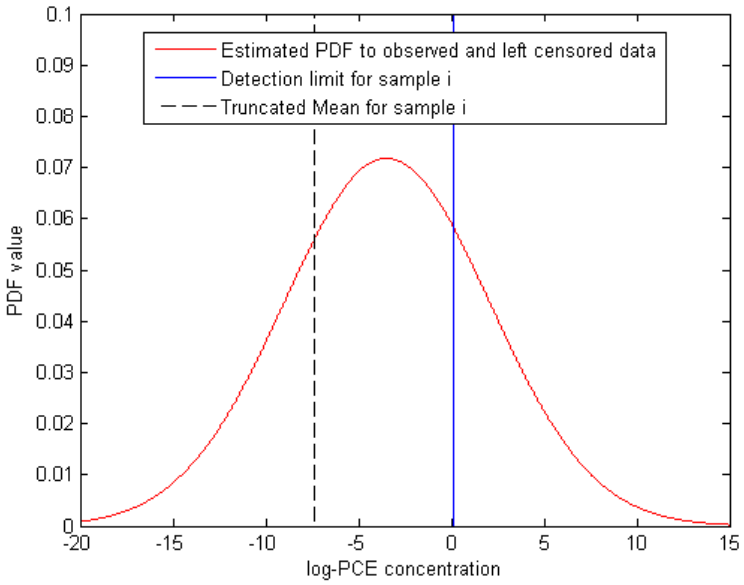

Figure 1. PDF of log-PCE with mean and variance estimated from observed and left censored data (see supplementary information), showing a sample detection limit and corresponding truncated Gaussian mean

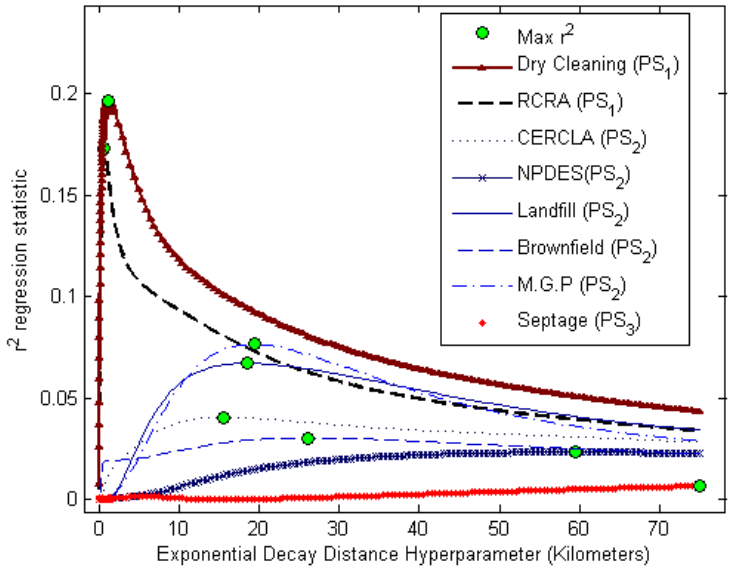

Figure 2. r^2 regression statistics as a function of the exponential decay range

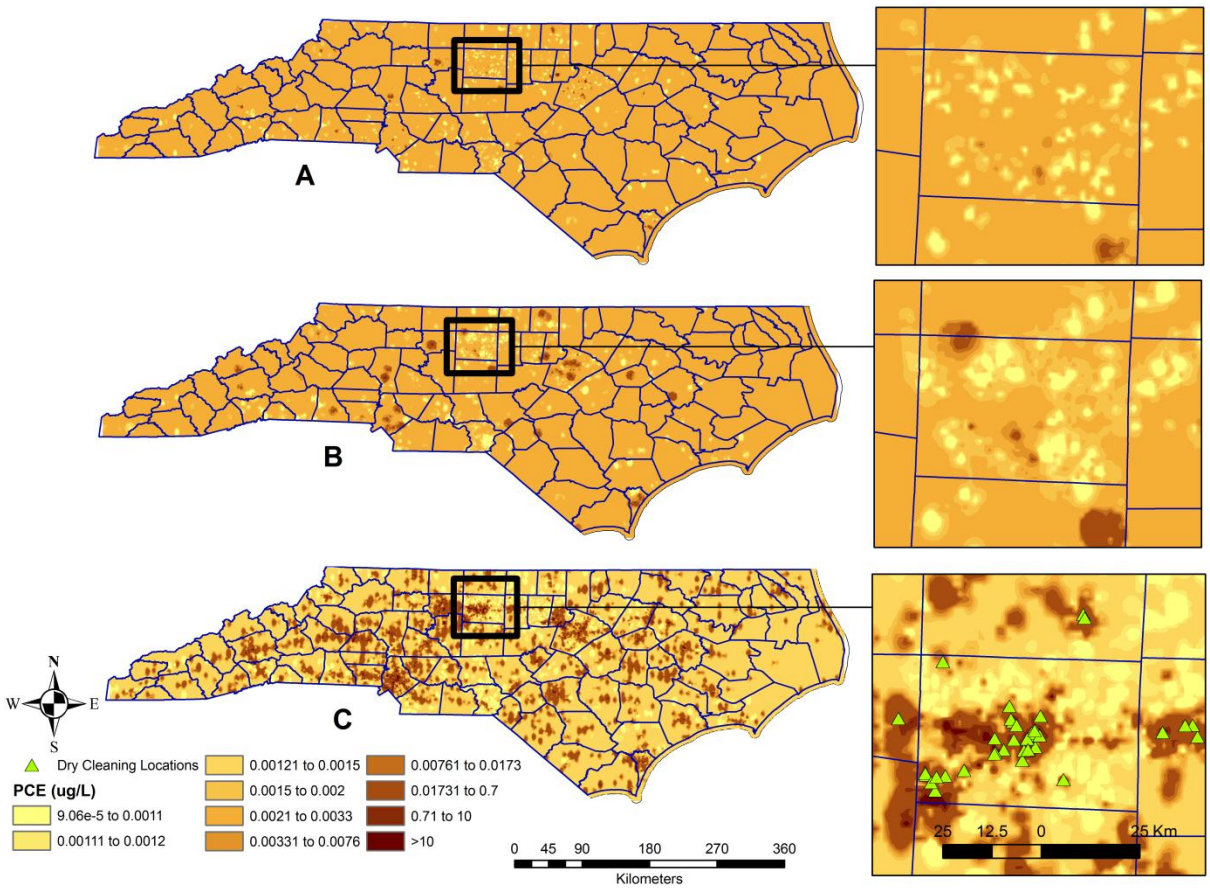

Figure 3. Groundwater PCE estimates using (A) Kriging with hardened below detects, (B) BME with below detects treated as a truncated Gaussian PDF, and (c) BME with below detects treated as a truncated Gaussian PDF and a land use regression mean trend based on dry cleaners and RCRA sites

5. TABLES

Table 1. Statistics for univariate land use regression models obtained for the decay range corresponding to the maximum r^2 value

	Exponential decay range in *Km*	r^2	P-value (F-Stat)	Beta 1 (95% CI)
Dry Cleaners	1.25	0.2	< 0.0001 (1147)	3.83 (3.61-4.05)
RCRA	0.67	0.17	< 0.0001 (982.2)	1.89 (1.78-2.01)
CERCLA	15.5	0.04	<0.0001 (197.9)	0.15 (0.13-0.17)
NPDES	59.5	0.02	<0.0001 (111.7)	0.03 (0.02-0.04)
Landfill	18.5	0.07	<0.0001 (337.7)	0.63 (0.56-0.69)
Brownfield	26.0	0.03	<0.0001 (146.9)	0.12 (0.10-0.014)
M.G.P.	19.5	0.08	<0.0001 (388.1)	2.99 (2.70-3.29)
Septage	75.0	0.007	<0.0001 (30.98)	0.09 (0.05-0.12)

Table 2. Cross-Validation Mean Square Error and Percent Change in Mean Square Error

	Kriging	BME	BME with LUR	PC12	PC23
MSE	22.98	17.52	13.06	-23.75	-25.46

6: Supporting Information

Pages: 6

Figures: 5

Tables: 1

This supporting information provides (a) a map of the data used for the analysis, (b) a representation of the land use regression model used, (c) a covariance model plot, (d) a map of the land use regression mean trend, (e) a summary statistics table of the land use model used, and (f) a detailed description of the method used to model the probability distribution function of log-PCE.

In figure S1 below we show all 3 of the data sources used in the study. Data sources came from the North Carolina Department of Environment and Natural Resources (NCDENR) division of waste management (DWM) Dry Cleaning Solvent and Cleanup Act (DSCA) branch.

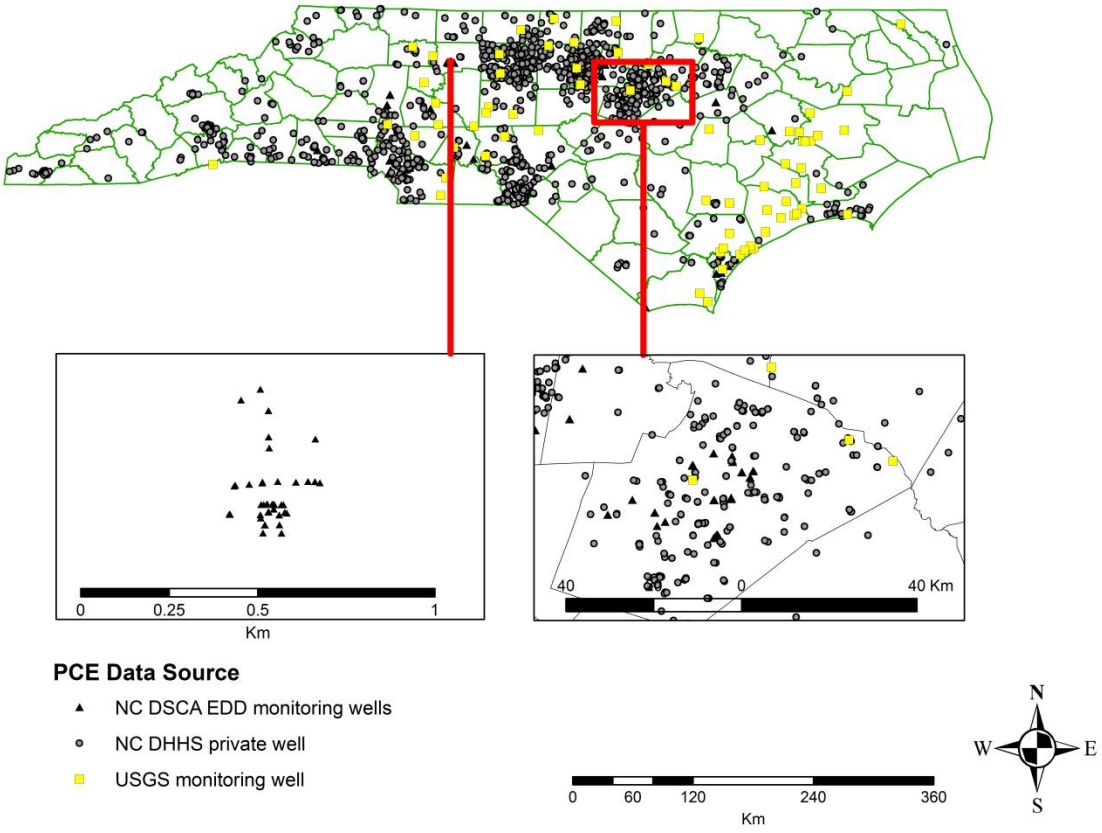

Figure S 1. Groundwater PCE data locations in North Carolina from three publicly available sources

Figure S2 below is a colormap reprenstation of r^2as a function of both the exponential decay range of RCRA sites and Dry Cleaning sites. The r^2 is represented by the varying colors and the axes represent the decay ranges. We select the regression model that corresponds to the decay ranges at the absolute maximum.

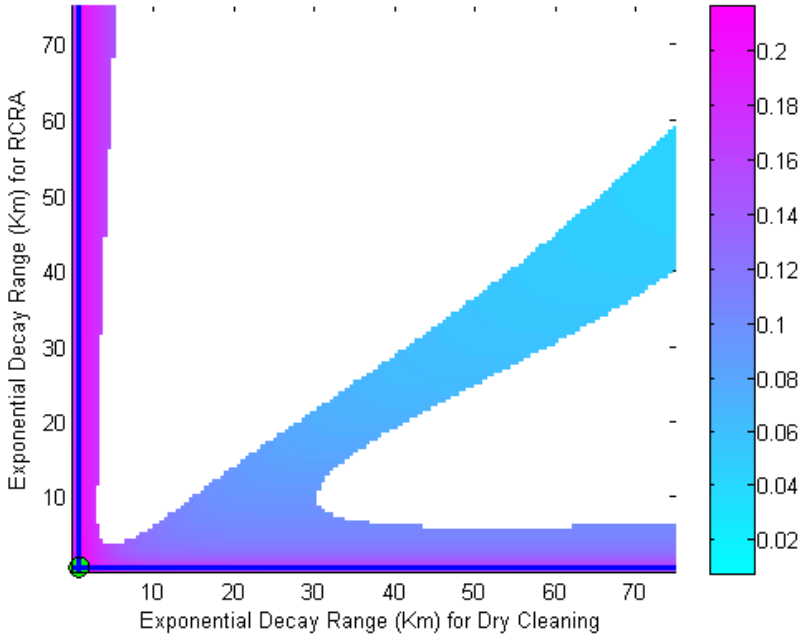

Figure S2. r^2 as a function of decay range for dry cleaners and RCRA sites. The color scale corresponds to the respective r-squared value.

Figure S3 below shows the covariance model used for the LUR/BME maps shown in this paper. The model has a short range component and a long range component, and it does not contain a nugget effect. We use expert judgment to fit a model to the experimental covariance data, although a least-squared approach will be implemented prior to submission to a journal.

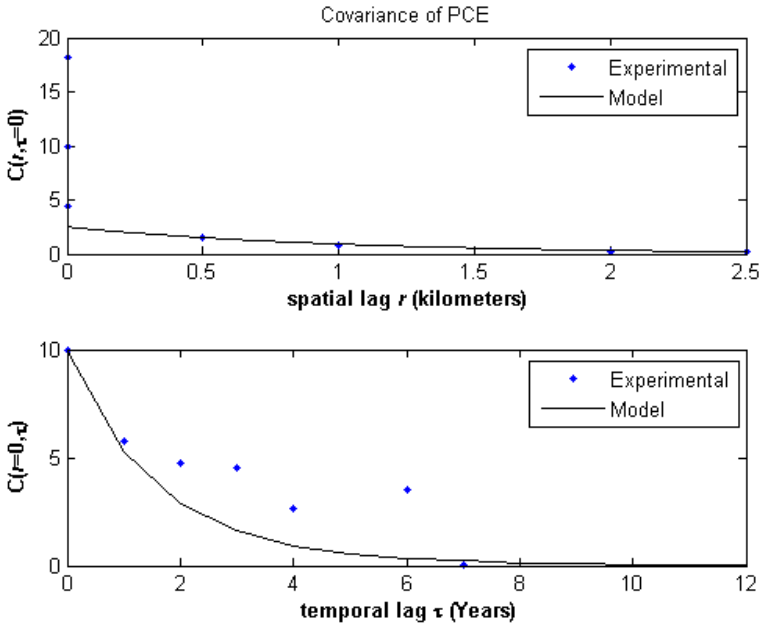

Figure S 3.Experimental and Modeled covariance for land use mean trend removed PCE

Figure S4 below maps the global land use regression mean trend used in the LUR/BME analysis.

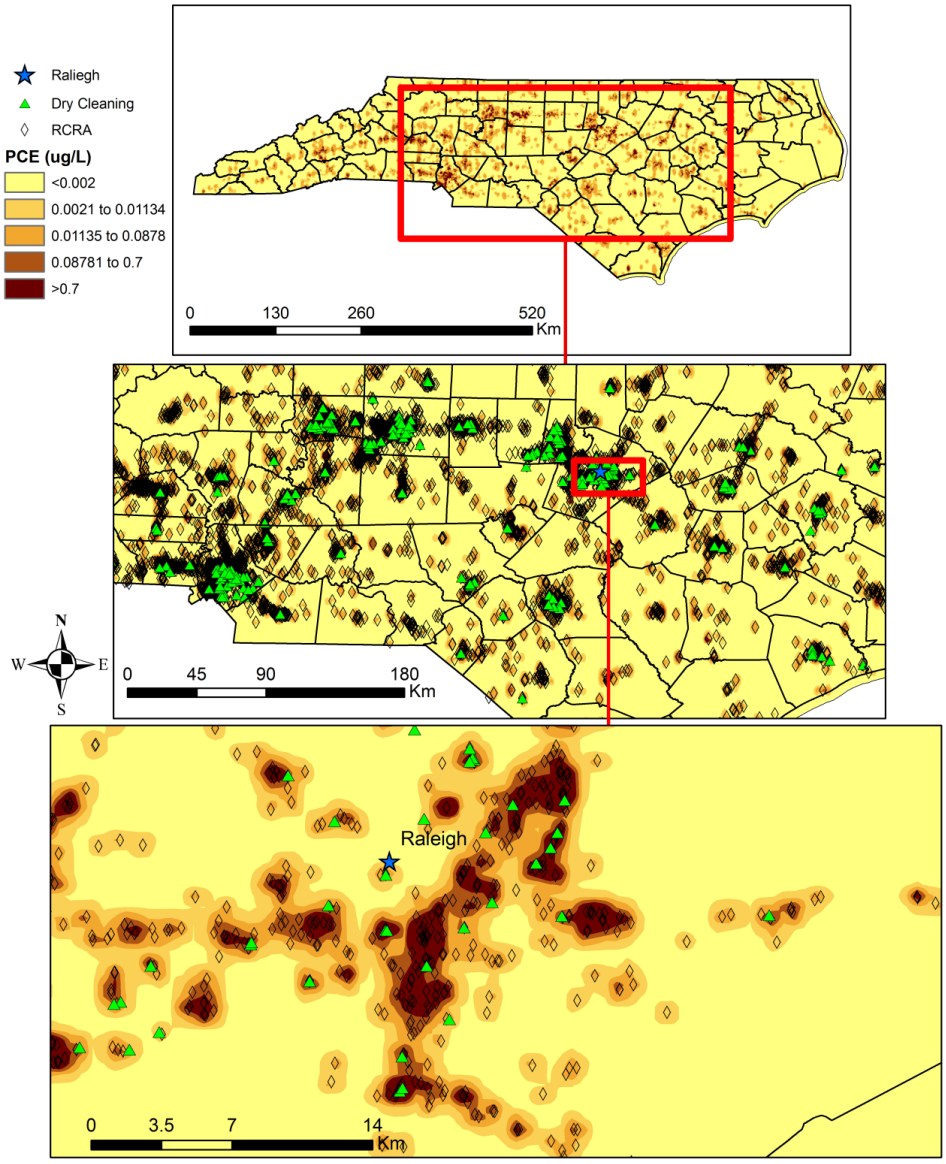

Figure S 4. Land Use Regression Mean Trend based on cumulative exponentially decaying contamination from Dry Cleaners and RCRA sites.

The following analysis provides a quantification of the policy implications of the study. We calculate a probability of the LUR/BME estimate being in exceedance of the North Carolina groundwater standard of 0.7 *ppb*. Figure S5 is a map displaying this probability across North Carolina.

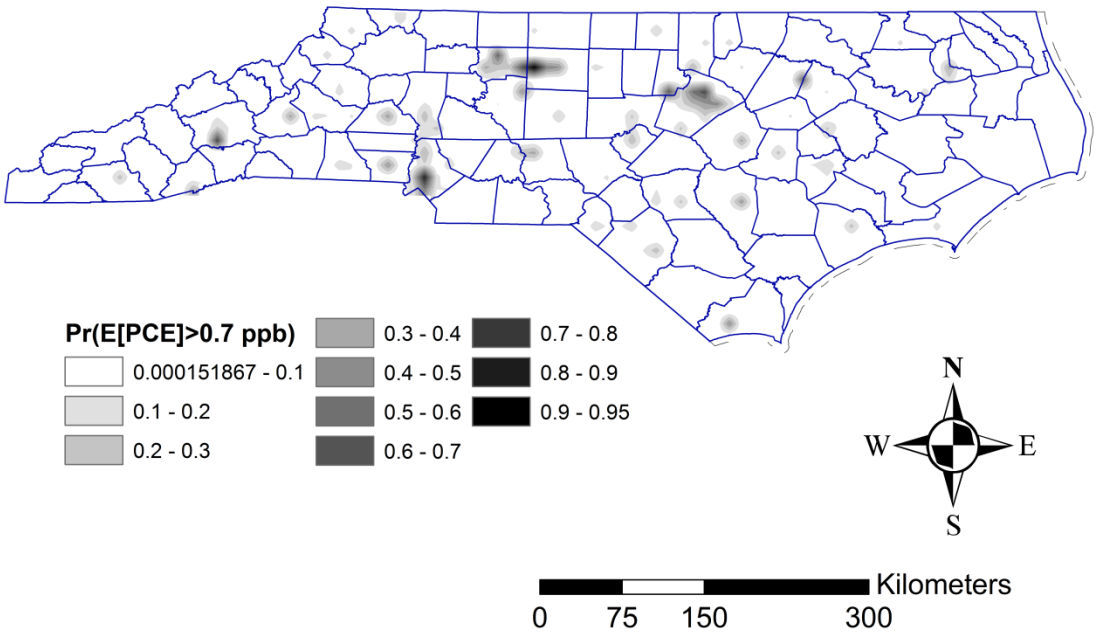

Figure S 5. The probability of the expected value of LUR/BME estimations will exceed the North Carolina groundwater standard of 0.7 ppb. It is calculated from the mean and variance of the LUR/BME posterior PDF.

Table S 1. Statistics for the bivariate regression model with Dry Cleaners and RCRA explanatory variables

r^2	Decay Range for Dry Cleaners	Decay range for RCRA	β_1 (95% CI). Slope for dry cleaners.	β_2 (95% CI). Slope for RCRA .	*p*-value for β_1	*p*-value for β_2
0.22	0.99 Km	0.83 Km	3.07 (2.72-3.42)	0.64 (0.53-0.74)	<0.0001	<0.0001

Estimating the PDF for PCE

We assume PCE to be a log-normal distributed environmental contaminant, so that the natural log of PCE concentration Z has a Gaussian PDF $f(z;\mu,\sigma^2)$ with mean μ and variance σ^2. Let n be the total number of PCE data, let p be the number of PCE data below the detection limit (DL), i.e. p is the number of left-censored data, and let $Z_{0.95}$ be the 95 percentile of PCE data (which in this work is above the DL). We seek μ and σ^2 such that

$$\begin{cases} \int_{-\infty}^{DL} dz f(z;\mu,\sigma^2) = p/n \\ \int_{-\infty}^{Z_{0.95}} dz f(z;\mu,\sigma^2) = 0.95 \end{cases}$$

We solve this problem numerically in the *MATLAB* computational platform by defining the following objective function

$$ObjFun = \left(\int_{-\infty}^{DL} dz f(z;\mu,\sigma^2) - p/n \right)^2 + \left(\int_{-\infty}^{Z_{0.95}} dz f(z;\mu,\sigma^2) - 0.95 \right)^2$$

and using the *MATLAB fmin* routine that finds the values for the μ and σ^2 pair which minimize that objective function.

7. REFERENCES

1. ATSDR, *Toxicological profile for tetrachloroethylene.* U.S. Department of Human Services, 1997.

2. NRC, *Review of the Environmental Protection Agency's Draft IRIS Assessment of Tetrachloroethylene*, in *Committee to Review EPA's Toxicological Assessment of Tetrachloroethylene*. 2010, The National Academics Press: Washington, D.C.

3. Moran, M.J., et al., *Occurrence and status of volatile organic compounds in groundwater from rural, untreated, self-supplied domestic wells in the United States,1986-99*, in *Water-Resources Investigations Report*. 2002, U.S. Geological Survey.

4. Moran, M.J.Z., John S.; Squillance, Paul J. , *Chlorinated Solvents in Groundwater of the United States.* Environmental Science & Technology, 2007. **41**: p. 74-81.

5. Squillace, P.J., et al., *Volatile Organic Compounds in Untreated Ambient Groundwater of the United States, 1995.* Environmental Science & Technology, 1999. **33**(23): p. 4176-4187.

6. NCAC, *Classifications and Water Quality Standards Applicable to the Groundwaters of North Carolina*, in *Title 15ANCAC 2L*. 2010: United States.

7. DSCA. *Dry-cleaning Solvent Cleanup Act (DSCA) Program*. 2010 [cited June 1, 2010]; The website home for NC DWM DSCA program]. Available from: http://portal.ncdenr.org/web/wm/dsca.

8. Akita, Y., G. Carter, and M.L. Serre, *Spatiotemporal Nonattainment Assessment of Surface Water Tetrachloroethylene in New Jersey.* Journal of Environmental Quality, 2007. **36**: p. 508-520.

9. Puangthongthub, S., et al., *Modeling the Space/Time Distribution of Particulate Matter in Thailand and Optimizing Its Monitoring Network.* Atmospheric Environment, 2007. **41**: p. 7788-7805.

10. Su, J.G., M. Jerrett, and B. Beckerman, *A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures.* Science of the Total Environment, 2009: p. 3890-3898.

11. Christakos, G., P. Bogaert, and M.L. Serre, *Temporal GIS: Advanced Functions for field-based Applications*, ed. Springer-Verlag. 2002, New York, NY.

12. GANC, *Dry Cleaning Solvent Cleanup Act of 1997*. 1997.
13. GANC, *House Bill 2873*, N.C.G. Assembly, Editor. 2006.

14. Sanders, A.P., et al., *Spatiotemporal Assessment of Arsenic Levels in North Carolina Domestic Well Waters*, in preparation for *Environmental Health Perspectives*. 2010.

15. Division of Waste Management, N.C., E.a.N. Resources, Editor. 2010: Raleigh.

16. NCOnemap, *Geographic Data Serving A Statewide Community*. 2010.

17. Christakos, G., *A Bayesian/Maximum-Entropy View To The Spatial Estimation Problem.* Mathematical Geosciences, 1990. 22(7): p. 763-776.

18. Serre, M.L. and G. Christakos, *Modern Geostatistics: Computational BME in the light of uncertain physical knowledge-- the equus beds study.* Stochastic Environmental Research Risk Assessment, 1999. **13**(1): p. 5-18.

19. LoBuglio, J. N., G. W. Characklis, and M. L Serre (2007) Cost-effective water quality assessment through the integration of monitoring data and modeling results, *Water Resour. Res.*, Vol. 43, No. W03435, pp. 1-16, doi:10.1029/2006WR005020.

20. Coulliette, A.D., E. Money[D], M.L. Serre**,** R.T. Noble (2009) Space/Time Analyses of Fecal Pollution and Rainfall in an Eastern North Carolina Estuary, *Environmental Science & Technology*, Vol. 43(10) pp. 3728-3735.

21. De Nazelle, A., S. Arunachalam, and M.L. Serre, *Bayesian Maximum Entropy Integration of Ozone Observations and Model Predictions: An Application for Attainment Demonstration in North Carolina.* Environmental Science & Technology, 2010. 44(15): p. 5707-5713.

22. Christakos, G., *Modern Spatiotemporal Geostatistics*. 2000: Oxford University Press.

23. Christakos, G., *Advanced Mapping of Environmental Data: Geostatistics, Machine Learning, and Bayesian Maximum Entropy*, ed. J.W. Sons. Vol. Chapter 6: Bayesian Maximum Entropy. 2008, New York, NY.

24. GANC, *Well Construction Standards.* 2009