SPATIAL STATISTICS AND REGRESSION ANALYSIS OF ENVIRONMENTAL EXPOSURE AND DISEASE: FROM AIR POLLUTION AND MICROBIAL GROUNDWATER CONTAMINATION ASSESSMENT TO DIARRHEA DISEASE MAPPING

Yasuyuki Akita

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Environmental Sciences and Engineering, School of Public Health.

Chapel Hill
2010

Approved by:

Marc L. Serre

William G. Vizuete

Michael E. Emch

Jiu-Chiuan Chen

Andrew D. Gronewold

ABSTRACT

Yasuyuki Akita: Spatial Statistics And Regression Analysis Of Environmental Exposure And Disease: From Air Pollution And Microbial Groundwater Contamination Assessment To Diarrhea Disease Mapping
(Under the direction of Marc L. Serre)

Recent technological advances in temporal geographic information systems (TGIS) include the Bayesian Maximum Entropy (BME) method, which accounts for the composite space/time variability and the wide variety of soft data characterizing many environmental and health processes. However, there are still several unaddressed implementation issues in the application of BME in environmental and health studies. In this work, the BME approach is applied to an air and a water environmental exposure assessment study where several unaddressed implementation issues are addressed.

First, a moving-window implementation of the BME method was numerically implemented and applied to the assessment of long-term exposure to ambient $PM_{2.5}$ across the contiguous U.S. Results for this work indicate that the moving-window BME method provides an efficient framework to account for the non-stationarity of the air pollutant variability and for the incompleteness of daily $PM_{2.5}$ measurements, which leads to estimates that are about 10 to 20% more accurate than those of classical approaches.

In a second study a two-stage estimation framework is implemented to estimate the concentration of E. coli across the tubewells in Bara Haldia, Bangladesh. The first stage of this framework consists in a latrine hydrological regression model, while the second BME stage of this estimation framework rigorously accounts for the uncertainty associated with the Most Probable Number (MPN) estimation of the density of microorganisms using data from multiple dilution series. The findings of this work indicate that latrines are a potential source of contamination of tubewells and thus have a significant impact on the spatial distribution of E. coli across tubewells.

Both applications show that the estimation framework based on the BME method successfully reduces estimation error compared with conventional geostatistical methods and provide highly informative maps.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of a number of people. Foremost, I would like to express my sincere gratitude to my advisor, Dr. Marc Serre, for his excellent guidance, patience, and encouragement throughout my Ph. D. work. His profound insight into science and enthusiasm in research had always motivated me.

I would also like to thank my Ph.D. committee members, Dr. William G. Vizuete, Dr. Michael E. Emch, Dr. Jiu-Chiuan Chen, and Dr. Andrew D. Gronewold. Without their remarkable assistance, advice, and criticism, this study would not have been successful.

I would acknowledge to all members of EID project. Without their support I would not be able to complete my analysis. A special thank goes to Veronica Escamilla for her GPS survey data set that I used extensively in my work.

It is a pleasure to express my gratitude wholeheartedly to all of my friends in the U.S. and in Japan for their continuous support and encouragement. I was extraordinarily fortunate to have great friends here in Chapel Hill. I also gratefully thank former and current lab mates at BMElab for their friendship and their willingness to share their thoughts with me during my Ph.D. work.

Finally, I would like to heartily thank all my family for all the support they provided during my Ph.D. period. I would like to express my deepest appreciation to

my wife, Naomi, for her endless support throughout the entire doctoral process. Without her support, I could not complete my Ph.D. work. I also thank my mother, Junko Akita, who sincerely raised me with her caring and gently love and my father, Yasuhiro Akita, who inspired me to pursue a doctoral degree. He passed away in 2004 after a long-time battle with cancer. But I believe that he is still living in my heart.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike information criterion
APE	Arithmetic mean of the Prediction Error
AQS	Air Quality System
ASE	Arithmetic mean of the Standard Error
ASPE	Arithmetic mean of the Standard Prediction Error
BME	Bayesian Maximum Entropy
EID	Ecology of Infectious Disease
FIB	Fecal Indicator Bacteria
GIS	Geographical Information Systems
HDSS	Health and Demographic Surveillance System
ICDDR,B	the International Centre for Diarrhoeal Disease Research, Bangladesh
MPN	Most Probable Number
MSE	Mean Square prediction Error
PDF	Probability Density Function
RMSE	Root Mean Square Error
RMSS	Root Mean Square Standardized
S/TRF	Space/Time Random Field
SRF	Spatial Random Field
WHO	World Health Organization

CHAPTER 1

Introduction

Spatial information is one of the key components of many environmental epidemiological researches. Generally both environmental exposures and associated disease outcomes depend heavily on the location where the study subjects reside. Thus an analysis ignoring spatial information tends to lead to erroneous results. Geographical Information Systems (GIS) play an important role in the analysis of spatial environmental and epidemiological data. They provide useful basic functions such as address geocoding and overlay operations to enhance the use of spatial information in the analysis (Vine, Degnan, and Hanchette 1997), as well as advanced data analysis functions such as surface creation and spatial statistics. In environmental epidemiological research, a GIS is primarily used in exposure assessment and disease mapping. Outputs are used as inputs for regression analysis to evaluate the strength of associations between exposure and disease. The general usage of GIS in the environmental epidemiology field is summarized in several studies (Elliott, and Wartenberg 2004; Jarup 2004; Nuckols, Ward, and Jarup 2004).

In exposure assessment, many studies have implicitly assumed a homogeneous distribution of the exposure field. For example, the average concentration of all the data within an area surrounding a study subject, or the

concentration of the nearest monitoring location to that study subject, are often used as measures of exposure. Recently the local scale variability in exposure has been taken into account in environmental epidemiologic studies. This local scale variability is often estimated using spatial interpolation methods. These methods are generally divided into two categories; deterministic and geostatistical methods. Inverse distance weighted interpolation and polynomial interpolation, which are both implemented in common GIS packages, are examples of deterministic methods. These methods interpolate the measurement values based on simple functions of the distance between an unmonitored point and its surrounding data points. One of the disadvantages of deterministic methods is, however, that they do not provide any measure of uncertainty associated with prediction.

Geostatistical methods provide estimates of the value at unmonitored locations, together with standard errors quantifying the associated estimation uncertainty. Kriging methods of linear geostatistics have, for instance, been widely used to estimate concentrations of environmental contaminants across space. Several types of kriging methods have been developed in order to take into account the underlying characteristics of the observational data. Unlike deterministic methods, kriging uses not only the distances but also the autocorrelation among spatial data in the estimation process. Thus, geostatistical methods generally outperform deterministic spatial interpolation methods. However, conventional kriging approaches have two major limitations. First, kriging methods rely only on exact measurements (referred to as hard data) and on data with normally distributed errors. However, in general, data with associated errors, which is referred to as soft

data, are generally not normally distributed. For example ambient air pollutant concentrations, which cannot take negative values, may be best represented using a normal distribution truncated below zero, which leads to a type of soft data that cannot be processed with linear kriging methods. Second, most of the kriging methods implemented in the common GIS packages do not fully take into account the temporal aspect of the data and only focus on the spatial distribution of the data. However many environmental monitoring data display composite spatial and temporal variability. Thus, accounting for the temporal dynamics of environmental and health data is indispensable in improving the accuracy of estimation.

The Bayesian Maximum Entropy (BME) method provides a rigorous mathematical framework that overcomes the limitations described above (Christakos 2000; Christakos, Bogaert, and Serre 2001; Christakos, and Li 1998). The BME method together with space/time random field (S/TRF) theory (Christakos 1992) takes into account the composite space/time variability and processes all available monitoring data distributed over space and time. Moreover, the BME method provides an efficient framework to rigorously assimilate any type of soft data into the estimation procedure. By using the BME framework, we can integrate data coming from multiple data sources with various levels and types of uncertainty. This approach has been used in several exposure assessment studies and was shown to successfully reduce estimation error (Akita, Carter, and Serre 2007; Puangthongthub et al. 2007). Because of these capabilities, the BME framework is an indispensable tool for environmental epidemiological research.

Even though the BME approach provides a powerful framework for space/time estimation, there are several unaddressed implementation issues in its application to environmental and health studies. In this study, the BME approach is applied to an air and a water environmental epidemiologic study where these unaddressed implementation issues are addressed.

In chapter 2, exposure to long-term ambient $PM_{2.5}$ concentration across the contiguous U.S. was modeled using the BME approach. A major issue for applying the geostatistical techniques to large geographic scale spatial process is that spatial dependency of the data is often assumed to be stationary over the study domain. In other words, spatial autocorrelation is assumed to remain the same across locations, and a single covariance model calculated from the whole data set is used for entire estimation domain. If the study area is sufficiently small, this assumption is generally appropriate. However, in a country-wide study, spatial dependency is expected to vary with locations, and the stationary assumption is inappropriate. To address this implementation issue, a moving-window BME approach is developed to take into account the non-stationarity of long-term $PM_{2.5}$ concentrations across the contiguous U.S.

From chapter 3 to 5, the BME approach was employed in an environmental epidemiologic study to investigate microbial contamination in groundwater and diarrheal disease occurrence in Bangladesh. These works were performed as part of an Ecology of Infectious Disease (EID) project, which tries to elucidate the complicated relationships between groundwater arsenic concentration, hydrogeological and environmental microbiological factors, and diarrheal diseases in

Bangladesh. In order to assess these relationships, collaborators were brought together with expertise in a wide range of topics including hydrogeology, microbiology, geography, and environmental sciences. The following institutions were involved in this project: Columbia University; University of Tennessee, Knoxville; the International Centre for Diarrheal Disease Research, Bangladesh (ICDDR, B); University of Dhaka; and University of North Carolina at Chapel Hill.

As a preliminary analysis for the subsequent chapters, we examine in chapter 3 the influence of rainfall on the spatial variability of fecal indicator bacteria (FIB) in tubewell water. The data used consisted in the E. coli concentrations measured from samples collected in Matlab, Bangladesh.

We then conduct in chapter 4 a detailed space/time mapping analysis of FIB concentrations in tubewell water over the same study site. When estimating the value at unmonitored location, a global mean trend is generally removed from the data before performing a geostatistical analysis. To estimate this global mean trend, polynomial functions and local smoothing methods are widely used. Both of these data driven approaches are solely based on the measurement values and their locations. However, environmental processes are sometimes governed by extraneous factors, resulting in trends that cannot be adequately captured by the data-driven approaches. In order to model such environmental processes, we need to build a global mean trend model that accounts for the effects of these extraneous factors.

Hence, in chapter 4, the concentration of E. coli, one of the commonly used FIB, is estimated across tubewells in Matlab, Bangladesh, using a latrine

hydrological regression model. In the latrine hydrological regression model, location of latrines, population density and short term rainfall are considered as extraneous factors governing the global trend of FIB across space and time. In addition, to further improve the quality of the estimation, a space/time knowledge synthesis framework based on the BME approach was developed and implemented. In this framework, soft data models for the measurement error due to the E. coli sampling procedure were integrated into the estimation of E. coli concentration.

In Bangladesh, diarrheal disease is still a severe problem which accounts for more than 50000 child deaths annually. In chapter 4, the relationship between microbial contamination of tubewell water and environmental factors was investigated. However, diarrheal disease events are not directly studied. In chapter 5, we, therefore, investigate how environmental factors affect the spatial distribution of diarrheal disease in Bangladesh. Based on previous studies we selected, arsenic concentration, depth of the tubewell, flood protection, socioeconomic status, temperature, population density and rainfall as possible risk factors for diarrheal disease, and the association between these factors and diarrheal disease was evaluated using regression analysis.

CHAPTER 2

Moving-window Bayesian maximum entropy space/time mapping of annual $PM_{2.5}$ ambient concentration across the U.S.

2.1. Background

Several epidemiologic studies have demonstrated that long-term exposure to fine-particulate matter ($PM_{2.5}$) is associated with increased morbidity and mortality (Beelen et al. 2008; Boldo et al. 2006; Eftim et al. 2008; Kunzli et al. 2005; Pope, Ezzati, and Dockery 2009). In most of these studies, long-term exposure was estimated by either the local average of $PM_{2.5}$ concentrations measured at monitoring stations near the study subject or the concentration observed at the nearest monitoring station. These exposure estimates implicitly assume a uniform distribution of concentration across the area surrounding the study subject, and the local exposure gradient between the resident location and its closest monitoring site(s) has not been taken into account. Recent studies have addressed this issue and accounted for the small scale spatial variability of $PM_{2.5}$ by applying a geostatistical interpolation method (Kunzli et al. 2005; Liao et al. 2006) or some spatial regression techniques (Brauer et al. 2003; Henderson et al. 2007). Geostatistical techniques, in particular, have been widely used in air pollution epidemiologic studies. For instance, a stronger association between long-term

exposure to $PM_{2.5}$ and chronic health effects relative to previous studies was found by estimating within-city exposure using a kriging geostatistical approach over the Los Angeles metropolitan area (Jerrett et al. 2005).

Although the local scale spatial variability can be successfully estimated by geostatistical techniques, there are several issues that arise when directly applying a geostatistical approach to long-term $PM_{2.5}$ exposure assessment at the national scale (i.e., over the entire U.S.). A major issue is that the spatial dependency of the long-term exposure to $PM_{2.5}$ is usually assumed to be stationary across the entire study area. In other words, the spatial autocorrelation of long-term $PM_{2.5}$ concentration is assumed constant across geographic locations, and a single spatial dependency model - such as variogram or covariance function - obtained from the whole data set is used for the entire estimation domain. In a national-scale study, however, spatial dependency is expected to vary with location and the stationarity assumption seems inappropriate. Thus, in order to perform national-scale exposure assessment, a framework that accounts for non-stationarity is needed.

Another issue pertaining to the assessment of long-term exposure to $PM_{2.5}$ is the completeness criterion used to reliably estimate long-term exposure. Long-term exposure to $PM_{2.5}$ is generally approximated by taking the average of $PM_{2.5}$ daily concentrations observed over some time period of exposure (e.g., yearly or monthly time periods), only if there are enough daily measurements within that time period to construct a reliable long-term exposure (Miller et al. 2007; Pope et al. 2002). For example, Pope et al. defined yearly average concentration based on $PM_{2.5}$ daily concentrations collected in 1999 and the first three quarters of 2000, only if a

monitoring site meets the following completeness criteria: At least 50% of the sixth-day samples are available for each quarter in either 1999 or 2000, and at least 45 total sampling days are available at that monitoring site. All average concentrations not satisfying that completeness criteria were then eliminated from the subsequent analysis due to the lack of the methodological framework to handle the uncertainty associated with yearly average concentrations.

In addition, long term exposure to $PM_{2.5}$ is often estimated from the average concentrations based on calendar years, so that the same long term exposure is assigned to a study subject regardless of the exact time of the health event within a given year. However, exposure misclassification can be reduced by accounting for the timing of disease occurrence. In other words, a time window of exposure based not only on its duration, but also on the exact beginning/ending times should be constructed to estimate the long-term exposure accurately.

Thus, the overall goal of this study is to conduct a national-scale assessment of long-term exposure to ambient $PM_{2.5}$ that takes into consideration all of the aforementioned issues. The yearly average concentration of ambient $PM_{2.5}$ over the contiguous 48 United States and District of Columbia was estimated using a moving-window implementation of a geostatistical estimation framework based on the Bayesian Maximum Entropy (BME) method. In this framework, the $PM_{2.5}$ yearly average concentration at all monitoring sites on an estimation date of interest were calculated as the average of $PM_{2.5}$ daily concentrations measured at that site over 365 days prior to the estimation date. In order to estimate the yearly average concentration at unmonitored points across the study region, the calculated yearly

average concentrations are then processed in the moving-window BME method, either as hard data (i.e., data with no error), or as soft data (i.e., data with associated measurement errors), based on the completeness criteria. The moving-window approach provides an efficient and easily implementable framework to account for the non-stationarity of a spatial random process (Haas 1990, 1995), while the BME method (Christakos 2000; Christakos, and Li 1998) rigorously processes the uncertainty of the $PM_{2.5}$ yearly average concentration due to the incompleteness of $PM_{2.5}$ daily concentrations within the year period of interest.

2.2. Materials and Methods

2.2.1. $PM_{2.5}$ Monitoring Data

$PM_{2.5}$ daily concentrations measured from 1999 to 2008 were obtained from the Air Quality System (AQS) maintained by the U.S. Environmental Protection Agency (U.S. EPA 2009). Since $PM_{2.5}$ daily concentrations reported to the AQS can be negative because of small measurement errors at low $PM_{2.5}$ daily concentrations, these negative values were replaced by zero. In addition the $PM_{2.5}$ daily concentrations which exceeded the federal maximum sample value (500$\mu g/L^3$) were regarded as outliers and removed from the data (U.S. EPA 2008). If multiple monitors were operating at the same monitoring location on the same day, the resulting co-located daily concentrations were treated as duplicated measurements and were averaged. The daily average concentrations measured from 2001 to 2003

were used in this study to perform the moving-window BME estimation of yearly $PM_{2.5}$ concentration at any space/time location in 2003.

2.2.2. The Moving-window Approach

The moving-window estimation approach described by Haas (1990, 1995) accounts for the non-stationarity of a spatial process over a large geographic domain by localizing the estimation procedure to regions small enough so that the spatial process may be assumed stationary within each small region. Our implementation of the moving approach in this work consists in calculating a covariogram at each estimation point of interest using only the data points within the region around that estimation point. Then the geostatistical analysis for that estimation point is conducted using the location-specific covariogram and the data around the estimation point. This region around the estimation point is referred to as the estimation “window” and moves with the estimation point. The size of the window has to be small enough to assure stationarity of the spatial process within the window, but also large enough so that it contains enough data points to model the covariogram. In this study, we used a window containing 100 monitoring sites, based on the minimum sample size expected to produce a reliable sample covariogram estimate (Olea 2006).

2.2.3. The Bayesian Maximum Entropy Method

The BME method introduced by Christakos (Christakos 1990; Christakos 2000) provides a mathematically rigorous framework that integrates a variety of available knowledge bases (e.g., spatial dependency model, empirical relationships,

scientific model) with data having varying levels of epistemic uncertainty. These data are categorized in hard data corresponding to exact measurements of the process, and soft data, which may have an uncertainty characterized by a probability density function (PDF) of any type (e.g., Gaussian, Uniform). A full description of the epistemic underpinnings and numerical implementation of the BME method can be found elsewhere (Christakos et al. 2001; Serre, and Christakos 1999). In brief the BME method can be viewed as a two-stage knowledge processing procedure: At the prior stage, maximum entropy theory is used to process the general knowledge base at hand and produce a prior PDF describing spatial process. Then at the posterior stage, an operational Bayesian conditionalization rule is used to update this prior PDF with respect to the site specific hard and soft data available, which produces a BME posterior PDF describing the value of the spatial process at any estimation point of interest.

Let $Z(\boldsymbol{s})$ be a spatial random field (SRF) representing the $PM_{2.5}$ yearly average concentration at some spatial location $\boldsymbol{s}$ (Christakos 1992). We will denote as Z_k the random variable representing the SRF at estimation point $\boldsymbol{s}_k$ (i.e., $Z_k = Z(\boldsymbol{s}_k)$), and similarly Z_h and Z_s are vectors of random variables representing the SRF at the hard data points $\{\boldsymbol{s}_h\}$ and the soft data points $\{\boldsymbol{s}_s\}$, respectively. By convention, lower case variables (e.g. z_h, z_s, or z_k) will denote realizations or deterministic values taken by their corresponding upper case random variables (e.g. Z_h, Z_s or Z_k)

In the case that the general knowledge base G about the SRF $Z(\boldsymbol{s})$ consists in its mean trend $m_z(\boldsymbol{s}) = E[Z(\boldsymbol{s})]$ and covariance function

$$c_Z(\boldsymbol{s}, \boldsymbol{s}') = E[(Z(\boldsymbol{s}) - m_Z(\boldsymbol{s}))(Z(\boldsymbol{s}') - m_Z(\boldsymbol{s}'))] \quad (2.1)$$

where $E[]$ is the expected value operator, then the BME fundamental equation reduces to

$$f_K(z_k) = A^{-1} \int d\boldsymbol{z}_s f_G(\boldsymbol{z}_h, \boldsymbol{z}_s, \boldsymbol{z}_k) f_S(\boldsymbol{z}_s) \quad (2.2)$$

where A is a normalization constant, the prior PDF f_G obtained from entropy maximization on $G = \{m_Z(.), c_Z(.)\}$ is multivariate normal with mean and covariance given by $m_Z(.)$ and $c_Z(.)$, respectively, the vector of deterministic values $\boldsymbol{z}_h$ corresponds to the hard data, and f_S is a PDF characterizing the epistemic uncertainty of the soft data. The BME posterior PDF is denoted with a subscript $K = G \cup S$ representing the knowledge blending (or union) of the general knowledge $G = \{m_Z(.), c_Z(.)\}$ and site specific knowledge $S = \{\boldsymbol{z}_h\ , f_S(.)\}$.

The expected value of the BME posterior PDF provides an estimate of yearly $PM_{2.5}$ concentration at the estimation point, and the corresponding BME posterior variance provides a useful characterization of the associated estimation uncertainty. The Strength of the BME method is that it considers epistemic uncertainty for the soft data represented by a PDF $f_S(.)$ of any type. Hence any combination of non-Gaussian distributions is automatically integrated in the estimation process. For example, if the soft data includes some points with Gaussian distributions while others have uniform distributions, then the BME posterior PDF is non-Gaussian, and the corresponding BME estimator is a non-linear combination of the hard and soft

data. Another advantage of the BME framework is that in the limiting case where only hard data are included in the estimation process and the SRF is stationary with a constant mean, then the BME estimator is simply the kriging estimator. This makes BME a consistent extension of the widely used kriging estimator when one needs to integrate non-Gaussian soft data, as is the case in this work.

2.2.4. The hard and soft $PM_{2.5}$ yearly average concentration data

In the context of an exposure assessment, we defined $PM_{2.5}$ yearly average concentration at some estimation time t_k as the average of $PM_{2.5}$ daily concentrations over the 365 days preceding time t_k. An exact yearly average concentration value is, therefore, given by the average of 365 daily average concentrations over one year preceding estimation time t_k. Since at most of the monitoring sites $PM_{2.5}$ daily concentrations were collected on a three-day cycle during the study period, an exact $PM_{2.5}$ yearly average concentration is rarely obtained. Thus, in most epidemiologic studies $PM_{2.5}$ yearly average concentrations satisfying some acceptable data completeness criterion are treated as hard data for the exact yearly average concentration. In this study, we used the completeness criterion that there must be more than 75% of intended measurements in each quarter of the year prior to t_k to ensure that the observations are evenly distributed throughout the yearly period. If the completeness criterion were satisfied, the hard data $z_{h,i}$ for the $PM_{2.5}$ yearly average concentration at monitoring site i and time t_k is simply defined as

$$z_{h,i} = \mu_{h,i} = \sum_{j=1}^{n_i} \frac{y_{i,j}}{n_i} \tag{2.3}$$

where $y_{i,j}$ is the j-th daily concentration measured at site i over the yearly period prior to t_k, and n_i is the number of $y_{i,j}$ daily values. These hard data are processed in identical fashion by the kriging method and BME method.

If the completeness criterion described above was not met, then the yearly average concentration was treated as a soft data if there were more than 10% of intended measurements in each quarter. Following the notation introduced above, let $Z_{s,i}$ be a random variable representing the yearly average concentration at site i, and let S be the site specific knowledge base provided by the incomplete set of daily measured values $y_{i,j}$. In the BME framework, the epistemic uncertainty associated with the incomplete daily concentrations is characterized by the PDF $f_S(z_{s,i})$. In this work, we assume that an adequate choice for f_S is a truncated normal distribution given by the following equation.

$$f_S(z_{s,i}) = \frac{\frac{1}{\sqrt{2\pi{\sigma_{s,i}}^2}} \exp\left(\frac{-(z_{s,i} - \mu_{s,i})^2}{2{\sigma_{s,i}}^2}\right)}{\Phi\left(\frac{b - \mu_{s,i}}{\sigma_{s,i}}\right) - \Phi\left(\frac{a - \mu_{s,i}}{\sigma_{s,i}}\right)} I_{[a,b]}(z_{s,i}) \tag{2.4}$$

where Φ is the standard normal cumulative probability distribution, $\mu_{s,i}$ is the average of the daily measure $y_{i,j}$ over the 365 days preceding time t_k, and $I_{[a,b]}(z_{s,i})$ is the indicator function

$$I_{[a,b]}(z_{s,i}) = \begin{cases} 1 & \text{if } a \leq y \leq b \\ 0 & \text{otherwise} \end{cases} \quad (2.5)$$

Since the yearly average concentration cannot be negative, the lower and upper bounds are $a = 0$ and $b = +\infty$, respectively. The epistemic uncertainty associated with this soft datum arises from the difference between the arithmetic average of all 365 daily concentrations, and the arithmetic average calculated from an incomplete sample of size n_i randomly selected out of a finite population of size 365. Therefore, a reasonable value for the standard deviation $\sigma_{s,i}$ of the truncated normal distribution $f_S(z_{s,i})$ is

$$\sigma_{s,i} = \sqrt{\frac{365 - \mathrm{n}_{s,i}}{365}} \times \sqrt{\frac{\sum_{j=0}^{\mathrm{n}_{s,i}} (y_{ij} - \mu_{s,i})^2}{\mathrm{n}_{s,i}}} \quad (2.6)$$

where the first term of this equation is a finite population correction factor that linearly decreases to 0 as n_i increases to the finite population size 365, and the second term quantifies the variability of measured daily concentrations within the yearly period.

Yearly average concentrations satisfying the completeness criterion were regarded as the exact yearly average concentration and treated as hard data. However, all yearly average concentrations, except for the one based on 365 daily $PM_{2.5}$ concentrations, have an associated uncertainty. This uncertainty can be evaluated by using eq. (2.6) to calculate the data standard deviation σ_h at any hard

data points. At each monitoring site i, the standard deviation of soft data $\sigma_{s,i}$ smaller than the maximum of the hard data standard deviations were replaced by the maximum of $\sigma_{h,i}$, in order to make sure that uncertainty associated with soft data is at least as large as the uncertainty associated with the hard data.

2.2.5. Estimation of spatial autocorrelation

The spatial autocorrelation of the SRF $Z(\boldsymbol{s})$ is characterized by means of its covariance function. The covariance function of a stationary SRF can be expressed in terms of the distance r between two location $\boldsymbol{s}$ and $\boldsymbol{s}'$, i.e. $c_Z(\boldsymbol{s}, \boldsymbol{s}') = c_Z(r = ||\boldsymbol{s} - \boldsymbol{s}'||)$. In this study, the method of moment estimator was employed to estimate the experimental covariogram at various spatial lags r (Cressie 1993; Curriero et al. 2002). The experimental covariogram was then used to fit a positive definite covariance model using an automated weighted least square procedure (Jian, Olea, and Yu 1996; Olea 2006). The following three parametric covariance models were tested: (1) exponential model, (2) Gaussian model, and (3) spherical model. In addition, the covariance model that best fit the experimental covariogram at each window among the aforementioned three covariance models was also selected based on the smallest Akaike information criterion (AIC).

2.2.6. Cross-validation analysis

In order to compare the model performance of the moving-window BME approach implemented in this study with other conventional methods, a cross-validation analysis was conducted that comparing the following three methods: (1)

the BME hard data only analysis assuming stationarity of $PM_{2.5}$ yearly average concentrations across the entire U.S., (2) the moving-window BME hard data only analysis, and (3) the moving-window BME soft data analysis. In method (1), a single covariogram was calculated for a given estimation time t_k using the all the data points throughout the U.S. On the contrary, in methods (2) and (3), the covariogram was calculated at each estimation point using only the data for the 100 monitoring sites closest to the estimation point. Only the hard data points was considered for the estimation in methods (1) and (2), which correspond to the conventional kriging approach, whereas in method (3) both hard and soft data were used for the estimation in a way that rigorously accounts for the uncertainty associated with yearly average concentrations failing the completeness criterion.

Leave-one-out cross-validations are performed using the yearly average concentrations that met the completeness criterion as the validation data set for 10 randomly selected days in 2003. Model performance was evaluated using the following cross-validation statistics: arithmetic mean of the prediction error (APE), arithmetic mean of the standardized prediction error (ASPE), arithmetic mean of the standard error (ASE), root mean square standardized (RMSS), and mean square prediction error (MSE). The prediction error is equal to the difference between the predicted and observed $PM_{2.5}$ yearly average concentration and the standardized prediction error is equal to prediction error divided by its estimated standard error The APE and ASPE are measures of bias of estimation and should be close to 0. RMSS which is defined as the standard deviation of standardized prediction error measures the accuracy of the estimated standard error and should be close to 1. For

an accurate model, the ASE and MSE should be as small as possible. In addition, the Pearson correlation coefficient and Spearman's rank correlation were also calculated to evaluate the linear correlation and rank order of the predicted and observed $PM_{2.5}$ yearly average concentrations.

2.2.7. Simulation

In this study only a small fraction of $PM_{2.5}$ yearly average concentrations did not meet the completeness criterion over the study period which leads to a small ratio of soft to hard data points. However epidemiologic studies in other countries, over other study periods, or for other air pollutants that have frequent missing daily concentration measurements may lead to much higher ratio of soft to hard data points. In order to explore the performance of the aforementioned estimation methods under this situation, four simulated $PM_{2.5}$ daily concentration data sets were constructed by randomly removing 5%, 10%, 15%, and 20 % of $PM_{2.5}$ daily concentrations from the original daily concentration data set. Using these realistic simulated data sets, the hard and soft data for $PM_{2.5}$ yearly concentrations were re-constructed, which resulted in a substantially larger fraction of soft to hard data points. Finally, these simulated yearly average concentrations were used to re-run the cross validation analysis to evaluate the model performance.

2.2.8. Space/Time Sensitivity Analysis

In a geostatistical estimation framework, the optimal selection of the estimation neighborhood consists in selecting data points that (1) are correlated with the estimation point, and (2) that are independent from one another. $PM_{2.5}$ yearly

average concentrations for a given site, however, are highly correlated from one day to the next because of the overlapping of all but one daily concentrations used to calculate the yearly average concentration. A particularity of our proposed approach including both hard and soft data points in the BME analysis is that the optimal estimation neighborhood therefore consists in selecting the (hard or soft) data point for each monitoring station corresponding to the estimation day of interest, which essentially correspond to a purely spatial analysis. This can be explained by the fact that once we have included all the hard and soft data points corresponding to an estimation day of interest, then adding data from preceding or following days will only result in information that is highly redundant with that which is already in the spatial only estimation neighborhood. As a result, even though our approach can be easily extended to a space/time context, we do not anticipate that this would result in a substantial decrease of estimation error over a purely spatial analysis. In order to investigate this point, we conducted a sensitivity analysis consisting in comparing the model performance of the space/time implementation of methods (1), (2) and (3), which we refer to methods (4), (5) and (6), respectively, i.e. method (4) is the BME space/time hard data only analysis assuming country wide stationarity, method (5) is the moving-window BME space/time hard data only analysis, and method (6) is the moving-window BME space/time soft data analysis. For this sensitivity analysis, the $PM_{2.5}$ yearly average concentration was modeled as a homogeneous/stationary space/time random field (S/TRF) (Christakos 1992). Then, the space/time dependency amongst $PM_{2.5}$ yearly average concentrations was modeled using a space/time separable covariance model. Furthermore the estimation neighborhood

for the space/time estimation methods included all the data points used in the spatial only estimation methods, as well as three additional $PM_{2.5}$ yearly average concentrations observed at days preceding or following the estimation day of interest for the three monitoring stations in the spatial only estimation neighborhood that are closest to the estimation point (in terms of a space/time metric) (Christakos et al. 2001). The cross-validation analysis was, then, conducted for the same 10 randomly selected days in 2003 to compare model performance. All analyses were conducted using the Matlab R2008a (MathWorks Inc.) and BMElib, suite of the Matlab libraries for the BME analysis (Christakos et al. 2001).

2.3. Result

2.3.1. $PM_{2.5}$ yearly average concentration

Of the 1515 $PM_{2.5}$ monitoring sites that operated from 1999 to 2008, 1239 had $PM_{2.5}$ daily concentrations during the 2001-2003 study period. Figure 2.1 (a) shows the entire 1515 $PM_{2.5}$ monitoring sites over the continental U.S. $PM_{2.5}$ yearly average concentrations calculated for an estimation date of December 31, 2003, which uses all the $PM_{2.5}$ daily measurements observed during 2003, are shown in Figure 2.1 (b). The yearly average concentrations that met the completeness criterion are shown in circles. They were treated as hard data in the BME analysis. In contrast, those shown in squares did not meet this criterion and were treated as soft data.

(a) (b)

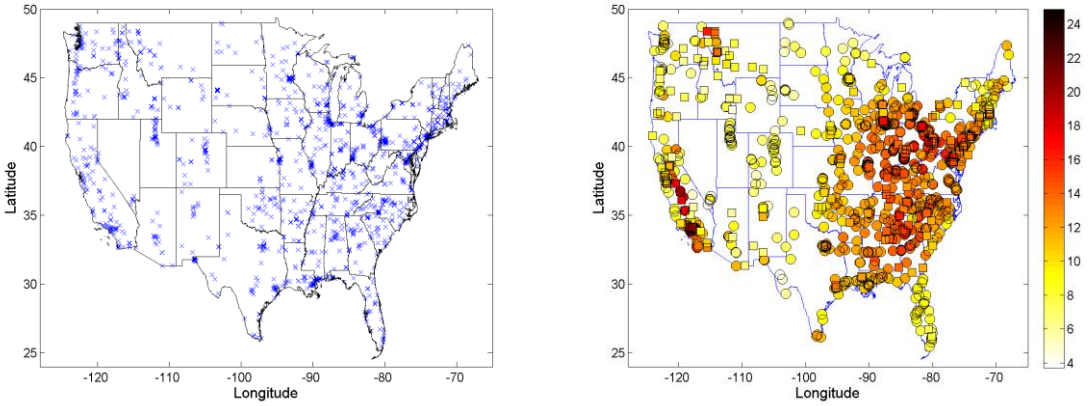

Figure 2.1: (a) $PM_{2.5}$ monitoring sites over the continental U.S. and (b) $PM_{2.5}$ yearly average concentration on December 31, 2003

Figure 2.2 displays time series of $PM_{2.5}$ daily and corresponding yearly average concentrations in 2003 at two monitoring sites: (a) 41-029-2129 and (b) 41-029-1001. $PM_{2.5}$ yearly average concentrations are shown in blue and green lines. The blue line shows yearly average concentrations that met the completeness criterion, whereas the green lines show yearly average concentrations which did not meet the completeness criterion, and their corresponding 95% confidence interval based on the soft data standard deviation $\sigma_{s,i}$ given by eq. (2.6). The $PM_{2.5}$ daily average concentrations are shown in red dotted line. At monitoring site 41-029-2129, all the yearly average concentrations calculated each day of 2003 met the completeness criterion, therefore a hard datum is shown for each of these days. On the other hand, at site 41-029-1001, most of the yearly average concentrations obtained during 2003 did not meet the completeness criterion and were treated as soft data because of the incompleteness of daily observations.

(a) (b)

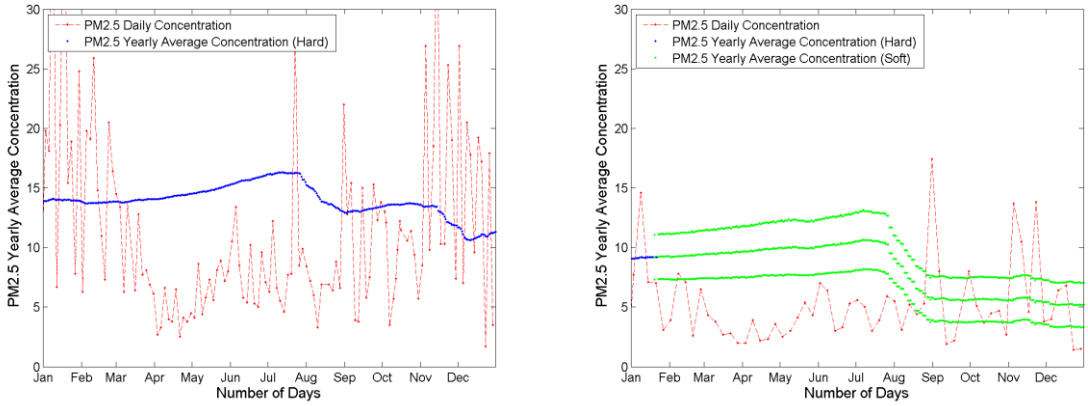

Figure 2.2: Time series of $PM_{2.5}$ daily and yearly average concentrations at monitoring site (a) 41-029-2129 and (b) 41-029-1001.

A histogram of all the $PM_{2.5}$ yearly average concentrations obtained in 2003 is shown in Figure 2.3. Although $PM_{2.5}$ yearly average concentrations were slightly positively skewed (coefficient of skewness: 0.357), their distribution is more symmetric than that of log-transformed yearly average concentrations (coefficient of skewness: -0.805).$PM_{2.5}$ yearly average concentrations were thus not log-transformed prior to the following estimation analysis.

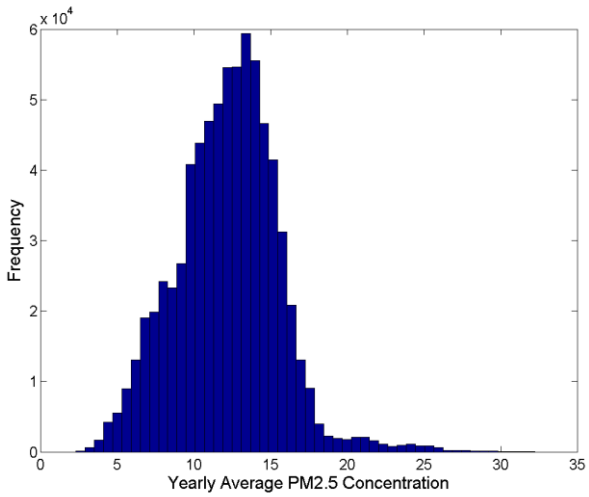

Figure 2.3: Histogram of all $PM_{2.5}$ yearly average concentrations obtained in 2003.

2.3.2. Cross-validation analysis

Table 2.1 shows the cross validation statistics obtained for method (1) – (3). The moving-window BME hard data only analysis (method (2)) reduced the MSE by 11% relative to the method under country wide stationarity assumption (method (1)). This indicates that using a moving-window approach to account for the non-stationarity of the process leads to 11% improvement in estimation performance over a method that assumes country wide stationarity. The moving-window BME soft data analysis (method (3)) further reduced the MSE by 18% relative to method (1), which indicates that there was a cumulative improvement in estimation performance when using the moving-window approach and accounting for soft data.

The APE and ASPE were generally close to 0, even though both values from method (1) were slightly closer to 0 relative to the moving-window approaches (method (2) and (3)). In contrast, the ASE from the moving-window approaches were

about 20% smaller than that from method (1). Similarly, the RMSS for the moving widow approaches were substantially closer to 1 than that for method (1). Likewise, the Pearson's correlation and Spearman's rank correlation were both higher for the moving-window approaches than for method (1). However, those from method (3) were the best among all three methods.

Table 2.1: Cross validation statistics obtained by the following three methods based on exponential covariance model: method (1) the BME hard data only analysis assuming stationarity across the U.S. (first column), method (2) moving-window BME hard data only analysis (second column), and method (3) moving-window BME soft data analysis

Method	(1)	(2)	(3)
MSE	2.459	2.186	1.998
APE	0.054	0.128	0.114
ASPE	0.012	0.052	0.044
ASE	1.939	1.570	1.540
RMSS	0.801	1.044	1.077
Pearson's Corr.	0.878	0.893	0.903
Spearman's Rank Corr.	0.886	0.894	0.902

The MSE based on other covariance models (exponential, Gaussian, spherical, and best fit covariance model) are listed in Table 2.2. In terms of MSE, the exponential covariance model outperformed the other covariance models. Moreover, the performance of three estimation methods (1) – (3) exhibits the same trend regardless the covariance model. Method (2) improves the estimation performance

over method (1) and method (3) further reduced the MSE. Other cross validation statistics are listed in Appendix A.

Table 2.2: Mean square error (MSE) of three estimation methods (1) – (3) based on four covariance functions; exponential (first row), Gaussian (second row), spherical (third row), and best fit model (forth row),

Method	(1)	(2)	(3)
Exponential Model	2.459	2.186	1.998
Gaussian Model	3.442	2.480	2.281
Spherical Model	3.050	2.231	2.066
Best Fit Model	3.050	2.302	2.100

2.3.3. Space/Time Sensitivity Analysis

Table 2.3 shows the MSE for the spatial only estimation methods (1) – (3) and for the corresponding three space/time estimation methods (4) – (6) based on exponential covariance model. The second and third columns list the change in percent (%) relative to method (1), and relative to the corresponding spatial only (SO) model, respectively. The space/time methods (4) and (5) reduced the MSE by approximately 4% relative to their corresponding spatial only methods. On the other hand, method (6) did not improve the estimation over method (3). This indicates that when using only hard data the space/time estimation framework leads to a performance improvement regardless of whether one is using nationwide or a local covariance model. This can be explained by the fact that on the estimation day of interest several monitoring stations do not meet the completeness criterion. As a

result the corresponding data points are not used in the spatial only analysis, while the space/time analysis is able to include data for these stations at some days following or preceding the estimation day when the completeness criterion is met. On the other hand, as explained earlier, we did not expect an improvement in estimation accuracy when implementing the BME method with soft data, which explains why the model performance does not improve from method (3) to (6).

Table 2.3: Mean square error (MSE) of spatial only estimation methods (1) – (3) and space/time estimation methods (4) – (6) based on the exponential covariance model. The change in MSE (in percent,%) relative to method (1) and to corresponding spatial only (SO) method are shown in column 3 and 4, respectively.

Method	MSE	Change (%) (Relative to (1))	Change (%) (Relative to SO)
(1)	2.459		
(2)	2.186	-11.1	NA
(3)	1.998	-18.75	NA
(4)	2.362	-3.94	-3.94 (Relative to (1))
(5)	2.088	-15.07	-4.47 (Relative to (2))
(6)	2.006	-18.43	0.4 (Relative to (3))

2.3.4. Simulation Study

In 2003 the fraction of soft data points for $PM_{2.5}$ yearly average concentrations was only about 18% of all the hard and soft data points. This fraction, however, increases as daily $PM_{2.5}$ observations were progressively removed from the original data set, reaching a fraction of 77% of soft data points when 20% of daily $PM_{2.5}$ concentrations were removed (Table 2.4). The MSEs from methods (1) – (3)

based on these simulated data sets are also shown in Table 2.4. The MSEs obtained from method (1) and (2) which relied only on hard data points increased as the ratio of soft data increased. In contrast, the MSE did not drastically change in method (3) which processes both the hard and soft data available. For example, when Simulated Data 4 was used, the MSE increased by about 90% in method (2), whereas the MSE from method (3) increased by only about 8%.

Table 2.4: Fraction of the $PM_{2.5}$ yearly concentration soft data (in %) and MSEs for methods (1) – (3) obtained for the true (i.e. uncensored) dataset (first row) and for the simulated datasets generated by randomly censoring 5% (second row), 10% (third row), 15% (fourth row), and 20% (fifth row) of the daily $PM_{2.5}$ observations

	Soft Data (%)	MSE (1)	MSE (2)	MSE (3)
True Data	17.9	2.459	2.186	1.998
Simulated Data 1 (5%)	24.7	2.543	2.267	2.009
Simulated Data 2 (10%)	37.5	2.834	2.543	2.024
Simulated Data 3 (15%)	57.2	3.177	2.901	2.105
Simulated Data 4 (20%)	76.7	4.216	4.149	2.156

2.3.5. Estimation Map

Figure 2.4 shows maps of the estimated $PM_{2.5}$ yearly average concentration in California on December 31st, 2003 obtained by (a) method (1) and (b) method (3) in California. The concentration map created by method (1) has a smoother distribution compared with that obtained by method (3). This result indicates that the moving-window BME method provides a description of spatial variation of $PM_{2.5}$ yearly average concentrations at a substantially finer resolution than that provided

by method (1). This result can be explained by the fact that ignoring the non stationarity and the soft data for $PM_{2.5}$ yearly average concentrations may lead to a loss of information that result in a loss of ability to describe detailed spatial gradients in long term exposure to $PM_{2.5}$. To visually inspect the accuracy of the estimated $PM_{2.5}$ yearly average concentrations, the yearly average concentration across the U.S. obtained by method (3) on December 31st, 2003 and U.S. EPA AirData annual summary of $PM_{2.5}$ concentration are shown in Figure **2.5** (U.S. EPA 2009). AirData annual summary concentrations (colored circles) and the estimated concentration (background color) show a good agreement.

(a) (b)

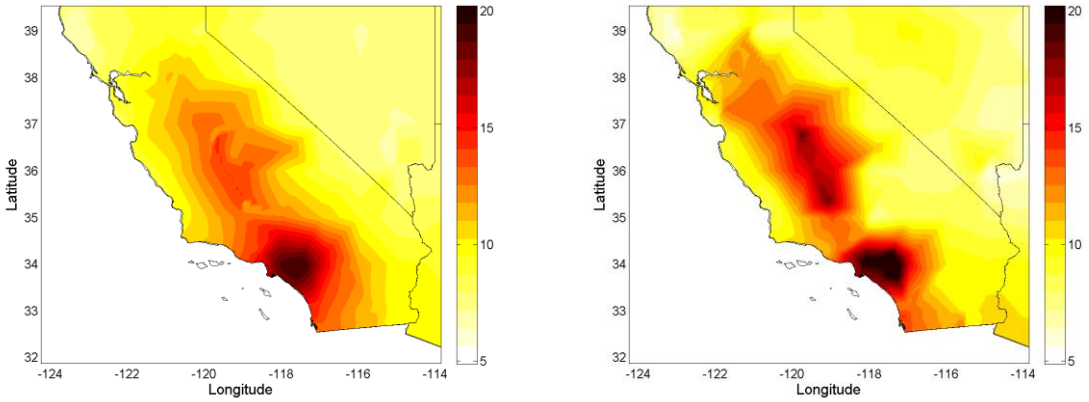

Figure 2.4: Map of the estimated $PM_{2.5}$ yearly average concentrations ($\mu g/m^3$) in California on December 31st, 2003 obtained by (a) method (1) and (b) method (3)

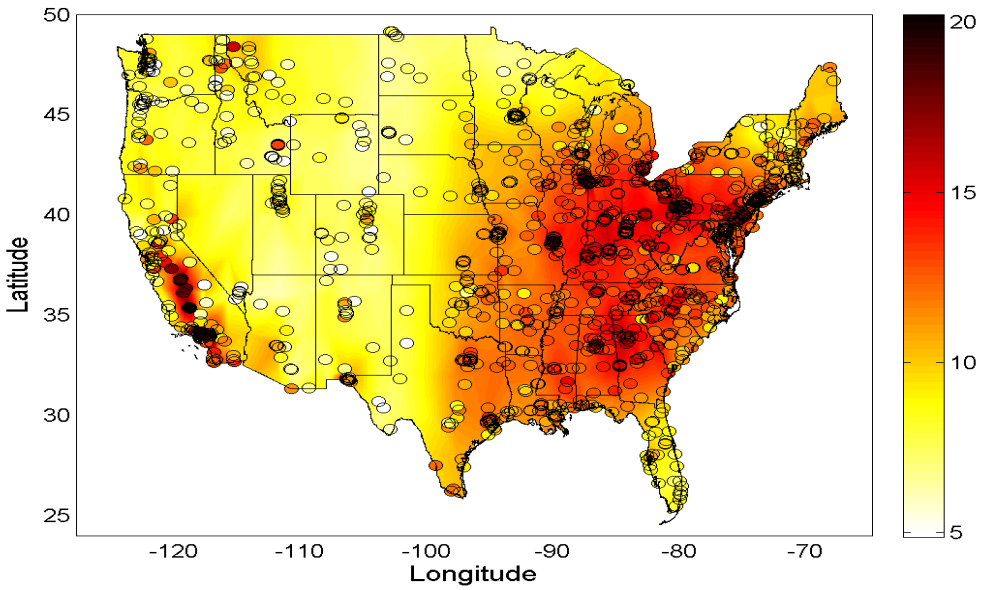

Figure 2.5: Map of the estimated $PM_{2.5}$ yearly average concentrations ($\mu g/m^3$) across the U.S. obtained using method (3) on December 31st, 2003. The U.S. EPA AirData annual summary of $PM_{2.5}$ concentration shown in colored circles.

2.4. Discussion

Classical linear geostatistical methods such as kriging are widely used to estimate individual-level exposure to air pollutants in many epidemiologic studies. Relative to conventional deterministic values such as the local average concentration or the concentration at the nearest monitoring station, or compared to deterministic interpolation techniques such as the inverse distance weighted average, geostatistical methods generally provide better estimates for individual-level exposure by taking into account the spatial dependency amongst the measured concentrations. However, there are several limitations when using classical linear

geostatistical methods to assess the long-term exposure to an air pollutant such as $PM_{2.5}$ over a large geographic region.

In the U.S., the concentration of $PM_{2.5}$ and its chemical composition show high spatial and temporal variability. Concentrations are generally higher in winter months on the west coast, whereas the level peaks in the summer on the east coast. Sulfate and other components of $PM_{2.5}$ also display clear spatial pattern (Bell et al. 2007). The spatial dependency of the $PM_{2.5}$ concentration is, therefore, expected to vary across space so that standard geostatistical methods relying on the stationarity assumption is inappropriate. Thus, a methodological framework that is capable of handling the non-stationarity of the spatial process is needed. Several studies addressed this issue in air pollution epidemiologic studies. Liao et al. (2006), for instance, conducted regional-scale kriging to estimate daily PM_{10} concentration over the U.S. by dividing the U.S. continent into five regions and calculating the semivariogram parameters in each region. However, based on cross-validation statistics they recommended using a national-scale kriging approach that assumes nationwide stationarity rather than a regional-scale kriging approach because of the insufficient model performance at the estimation points near the regional borders. That approach, however, is inadequate to account for the non-stationarity of the spatial process, since all estimation points in each region shared the same semivariogram parameters, which is equivalent to assuming within-region-stationarity. In the present study, we employ a moving-window approach to account for this issue. As shown in Table 2.1, the moving-window BME hard data only analysis (method (2)) led to 11% reduction of the MSE relative to the method

assuming country wide stationarity (method (1)). Other cross-validation statistics obtained from method (2) were also generally better than those obtained from method (1). The RMSS from methods (1) and (2) were 0.801 and 1.044, respectively. The RMSS from method (2) is pretty close to 1, which indicates that estimated standard error is a valid estimate. In contrast, the RMSS from method (1) is smaller than 1 suggesting an overestimation of the estimated standard error. In addition, the Spearman's rank correlation from method (2) is also higher than that from method (1), which suggests that the estimation results from method (2) better preserve the ranking of true concentrations. Thus, the moving-window approach accounting for the non-stationarity of the spatial process is superior to the method based on nationwide stationarity.

In addition to the moving-window approach, several different approaches have been introduced to deal the non-stationarity of the covariogram. Even though the moving-window approach is suitable for geostatistical estimation, it might not be appropriate for other applications of spatial statistics (Fuentes 2003). However, since the primary goal of this study is to better estimate the long-term exposure to $PM_{2.5}$ at unmonitored locations, we believe that the moving-window approach is a reasonable choice to handle the non-stationarity of the covariogram over a large spatial domain because it is a simple method to implement, which minimizes risk of implementation error in epidemiologic studies, and because it is shown in this study to significantly improve the estimation of long term exposure to $PM_{2.5}$, which, to our knowledge, has not been demonstrated to the same extend for other methods using non-stationary covariograms.

The reliability of the yearly average concentration is, generally, assessed by the completeness criterion based on the number of daily concentrations used to calculate the yearly average concentration. In many epidemiologic studies, the yearly average concentrations not satisfying the completeness criterion were simply eliminated from the analysis to avoid the possibility of obtaining misleading results. On the other hand, the BME method used in this study provides a flexible methodological framework that is capable of rigorously incorporating uncertain observations expressed as soft data characterized by any form of probability density function. In 2003, about 18% of the data did not meet the completeness criterion and were treated as soft data (Table 2.4). By accounting for these soft data method (3) reduced the MSE by 8% relative to method (2) which disregarded these soft data (Table 2.1). Other cross-validation statistics obtained from method (3) were also generally better than those obtained from method (2) in terms of estimation accuracy and ranking order preservation. Thus, overall, the moving-window BME soft data method which accounts for both the non-stationarity of the covariogram and for $PM_{2.5}$ yearly average concentrations not meeting the completeness criterion performs the best among the three methods investigated in this work, and is therefore the method recommended to minimize exposure misclassification in epidemiological studies investigating the effect of long term exposure to $PM_{2.5}$ on health.

To further investigate how the fraction of soft data points affects model performance, we also conducted the simulation study summarized in Table 2.4. We see from these results that the MSE obtained from method (3) is remarkably stable

even when up to 75% of the data do not meet the completeness criterion. By contrast, the MSE for methods (1) and (2) increase by a factor of almost 2 as compared to the MSE obtained for the true dataset. This means that our proposed approach (method 3) will continue to provide reliable assessment of long term exposure to an air pollutant even when the number of intended daily concentration decreases, while that is not at all the case in for the classical approach used in methods 1 and 2, which completely disregard the useful information provided by yearly averages that do not meet the completeness criterion.

We find from Table 2.3 that, as expected, a space/time estimation method slightly improves model performance relative to its corresponding spatial only method when only hard data are used (i.e. methods 4 and 5 improved upon methods 1 and 2, respectively), while the space/time estimation did not improve model performance when both hard and soft data are used (i.e. method 6 did not improve upon method 3). This result indicates that when using both hard and soft data (i.e. method 3), then the optimal estimation neighborhood consists in the (hard and soft) data for the estimation day of interest, which essentially means that the spatial estimation framework is optimal and little improvement is expected from implementing a full space/time estimation framework. This has the useful implication that in a practical epidemiological setting, the BME approach we are presenting in this work (method 3) will be easier to implement (and likewise less computationally intensive) than an approach (such as that used in methods 4 or 5) that would require the implementation of a full space/time estimation framework. In fact, of all the 6

spatial and space/time methods tested in this work, method 3 is the one with the smallest MSE, even though that method is a spatial only method.

This study introduces a window-based implementation of the BME method for long term exposure assessment to $PM_{2.5}$ that rigorously accounts of the uncertainty associated with incomplete daily $PM_{2.5}$ observations. This work provides methodological developments that complement those presented in recent studies using the BME method for air pollution estimation, and can be extended in the future to investigate the applicability of the framework presented here to assess long term exposure to variety of air pollutants. For example, in a study conducted in the Carolinas (states of North and South Carolina) to estimate the long-term exposure to ozone and PM_{10} (Yu et al. 2009) used the histogram of daily observations to construct the soft data. By contrast we use in this work a truncated Gaussian distribution with a standard deviation (eq. 2.6) that explicitly incorporates a finite population correction factor. Both the window-based implementation of the BME method and the finite population correction factor present alternative model specifications that can offer modelers with a flexible conceptual framework that can enhance future models used for the space/time estimation of long term exposure to air pollutants.

CHAPTER 3

Influence of rainfall on the spatial variability of fecal indicator bacteria across tubewells in a village of Matlab, Bangladesh

3.1. Background

3.1.1. Diarrheal disease in the developing countries

Despite great progress in improving water quality and sanitation in many parts of the world, diarrheal disease remains a severe problem among children. Even though the number of annual deaths from diarrheal disease has gradually decreased over the past two decades, morbidity and mortality of diarrheal disease remain high especially in the developing countries. The burden of diarrheal disease in developing countries was estimated to be more than 200 times higher than that in developed countries (Pruss et al. 2002). Currently diarrheal disease is the second leading cause of deaths among children under five years of age and it accounts for approximately 2.5 million deaths among children (Kosek, Bern, and Guerrant 2003). Diarrhea is a typical symptom of gastrointestinal infections which can be caused by various bacteria, viruses and parasites. Most of these pathogenic organisms spread through fecal-oral transmission in which water is the primary medium to transport microbial pathogens (Ashbolt 2004). Thus in order to prevent diarrheal disease an access to clean water is essential.

3.1.2. Drinking water source and microbial contamination in Bangladesh

In many developing countries, groundwater is a preferable drinking water source to surface water, since groundwater is normally less contaminated with microbial pathogens. In addition, the use of the groundwater is generally the only economically feasible option to obtain clean water (Pedley, and Howard 1997). In Bangladesh, the most densely populated country in the world located in South Asia, people gradually have switched their drinking water source from highly contaminated pond and river water to groundwater to prevent outbreaks of waterborne diseases. Since 1970's, with the assistance of the United Nations Children's Fund (UNICEFF) and many NGOs, millions of tubewells have been installed across the country to provide an access to safe drinking water (Smith, Lingas, and Rahman 2000). Currently more than 90% of households in rural Bangladesh are using tubewells as their primary drinking water source (NIPORT 2005). Nevertheless, the switching of drinking water source was found to be insufficient to eliminate the risk of diarrheal disease (Black et al. 1982; Black et al. 1981; Chen, Rahman, and Sarder 1980; Levine et al. 1976). Diarrheal disease remains a severe problem in the country and more than fifty thousand children still die annually from it (UNICEF 2009).

Several studies reported frequent occurrences of microbial contamination in the groundwater due to the combination of poorly designed disposal of human feces and insufficient protection of water source (Macler, and Merkle 2000; Melian et al. 1999; Pedley, and Howard 1997). A study conducted in a rural area of Bangladesh, for example, found that water samples collected from five tubewells contained several microorganisms such as zooplankton, viable bacteria, and fecal coliforms (Islam et al. 2001). A more recent study (Luby et al. 2008) also confirmed the low

levels of fecal contamination in tubewell water using over 200 samples collected in three flood-prone districts in Bangladesh. Similar studies performed recently in Bangladesh also confirmed a moderate levels of fecal contamination in tubewell water (Hoque et al. 2006; Luby, Islam, and Johnston 2006). These results indicate that tubewell water might not be a safe drinking water source in Bangladesh. Consumption of tubewell water might be one of the primary routes of exposure to microbial pathogens. Thus, in order to reduce the burden of diarrheal disease in Bangladesh, it is essential to better understand the mechanism of groundwater microbial contamination.

3.1.3. Groundwater and rainfall

Groundwater recharge due to rainfall is one of the key factors for controlling microbial contamination in shallow aquifer. Previous studies reported an association between microbial contamination in groundwater and precipitation (Barrell, and Rowland 1979; Wright 1986). The level of microbial contamination in shallow protected springs in Kampala, Uganda, for example, was significantly associated with rainfall, especially with short-time rainfall events (Howard et al. 2003). These results suggest that transport of microbial pathogens driven by rainfall might determine how widely microbial contamination stretches in shallow aquifers.
In the present study, we test the hypothesis that the spatial extent of microbial contamination in shallow aquifers is associated with rainfall. The study was conducted in Bara Haldia, one of the villages in the Matlab field research area of the International Centre for Diarrheal Disease Research, Bangladesh (ICDDR, B). We collected monthly tubewell water samples from May, 2008 to April, 2009 and we

analyzed these samples for Escherichia coli (E. coli). The range of the covariance function of log-transformed E. coli concentration was estimated each month of the study period as an indicator of the spatial extent of microbial contamination in the shallow aquifer. Then, correlation between the covariance range and rainfall was calculated to evaluate the association between the spatial extent of microbial contamination and rainfall.

3.2. Material and Method

3.2.1. Study Area

The study was conducted in Bara Haldia, one of the villages in the Matlab field research area of ICDDR, B. Matlab is a subdistrict of Bangladesh located in the south-central part of the country approximately 50 km southeast of Dhaka (the capital of Bangladesh). Approximately 220000 people live in Matlab field research area. There are more than 10000 baris which are clusters of households connected through a patrilineal line. Flood protection embankment was built in the 1980's along the Dhonagoda River running from north to south through Matlab. Approximately half of the study area within the embankment is protected from flooding. Figure 3.1 shows (a) the location of Matlab within Bangladesh and (b) a satellite image of the Matlab subdistrict and the location of the Bara Haldia study area. A GPS survey was conducted in the study area to identify the location of all tubewells, latrines, and households (Escamilla unpublished data). The location of 307 households, 244 latrines, and 186 tube wells were recorded. Figure 3.2 shows a satellite image of the

study area with the location of monitoring tubewells, latrines, and households that were GPS surveyed. A road across the study area divides the village in its northern and southern parts. The northern area is more densely populated than the southern area.

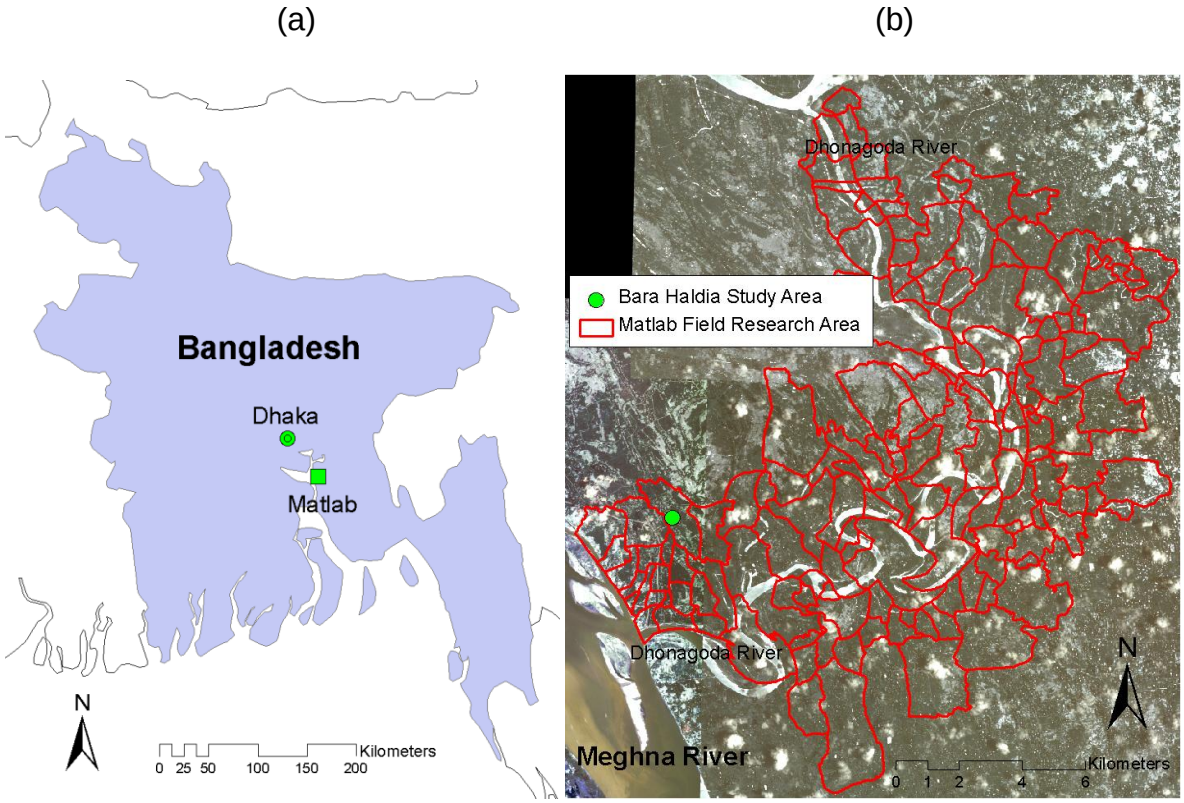

Figure 3.1: (a) The location of Matlab within Bangladesh and (b) A satellite image of the Matlab subdistrict and Bara Haldia study area

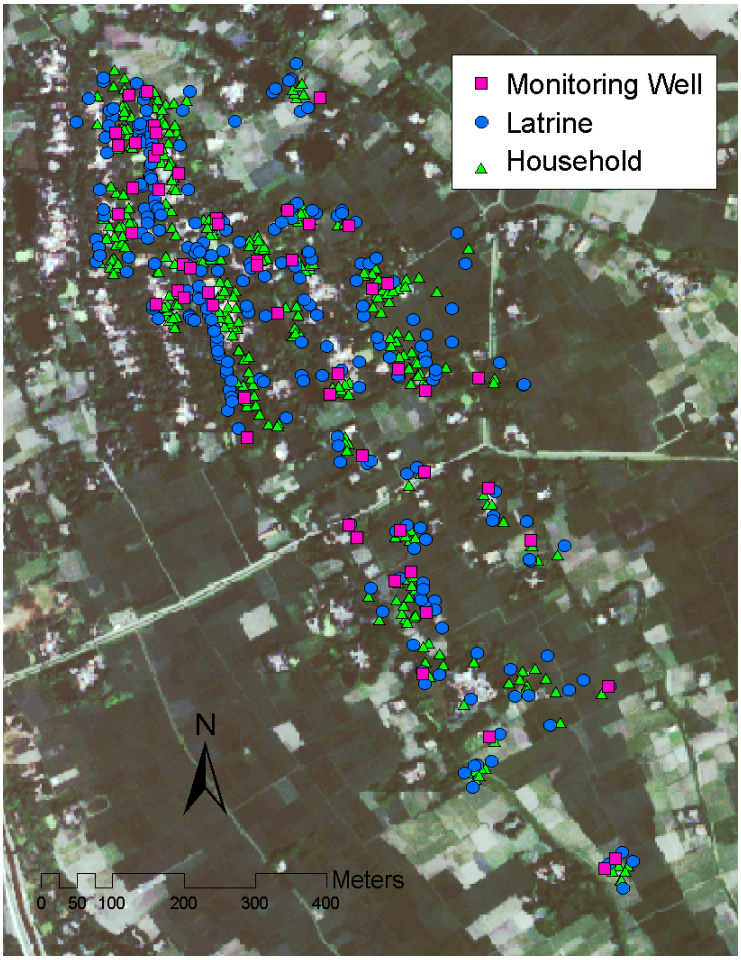

Figure 3.2: Satellite image of the Bara Haldia study area showing locations of monitoring tubewells, latrines, and households that were GPS surveyed

3.2.2. Rainfall Data

The tropical monsoon climate of Bangladesh is characterized by a monsoon season with heavy rainfalls from June to October and a dry pre- and post-monsoon season with virtually no rainfall. A HOBO weather station was installed in the Bara Haldia study area to collect precipitation and meteorological data (ONSET, Bourne, MA; http://www.onsetcomp.com/products/weather_stations). The weather station

started collecting rainfall data in June 2008. Hourly rainfall measurements were aggregated to obtain daily precipitations.

3.2.3. Tubewell Water Sample and Enumeration of Fecal Indicator Bacteria

We sampled a total of 55 shallow tubewells (depth less than 100 feet) for fecal indicator bacteria (FIB) in the Bara Haldia study area, which is a significant fraction of the total of 186 tubewells in use in that area, and provides a rich dataset to estimate the spatial extend of microbial contamination in the shallow aquifer.. Duplicate 100mL tubewell water samples were collected each month from May 2008 to April 2009. All samples were tested for two types of FIB: total coliform and E. coli. Total coliform and E. coli are widely used FIB to assess fecal contamination in the water. The presence of FIB indicates possible microbial contamination of the water. Both FIB were enumerated using a standard commercial kit (Colilert® reagent and the Quanti-Tray®/2000 manufactured by IDEXX Laboratories) based on the defined substrate technology (Rompre et al. 2002) The concentration of FIB based on a single sample is quantified as follows. (1) Colilert® reagent is mixed with the water sample. (2) The sample and reagent are poured into a Quanti-Tray®/2000 consisting of 49 large wells and 48 small wells. (3) The tray is sealed and incubated for 24 hours. (4) The number of wells positive for total coliform and for E. coli are counted. For total coliforms, the media turns yellow under ambient light; for E. coli the media also turns fluorescent blue under ultraviolet light. (5) The number of positive wells is converted into a concentration estimate and 95% confidence interval (CI) from the look-up table provided by IDEXX.

The concentration of FIB and the associated 95% CI were estimated using the most probable number (MPN) estimator (Cochran 1950; Hurley, and Roscoe 1983). A detailed description of the MPN method can be found in Appendix D. In brief, the MPN is a maximum likelihood point estimator given by the concentration μ (expressed in organisms/100mL) that maximizes the following likelihood function.

$$\mathrm{MPN} = \underset{\mu}{\operatorname{argmax}}\left(\prod_{i=1}^{r}\frac{n_i!}{s_i!\,(n_i - s_i)!}(1-\exp(-\mu V_i))^{s_i}(\exp(-\mu V_i))^{n_i - s_i}\right) \tag{3.1}$$

where r is the number of dilution sets, s_i is the number positive samples out of n_i total samples with volume V_i in the ith dilution set. This concentration μ is given by the root of the following equation (Appendix D).

$$\sum_{i=1}^{r}\frac{s_i V_i}{1-\exp(-\mu V_i)} = \sum_{i=1}^{r} n_i V_i \tag{3.2}$$

The MPN method assumes that the water sample is completely mixed so that the organisms in the liquid are randomly distributed. In other words, each dilution set has the same concentration. In this study, each single sample was assessed for this well-mixed assumption using the likelihood ratio test.

Since the probability distribution of the concentration μ is approximated by a log normal distribution, the standard error of log(MPN) is given by (Hurley, and Roscoe 1983)

$$SD_{\log(\mathrm{MPN})} = \left(\mathrm{MPN}^2 \sum_{i=1}^{r} \frac{{V_i}^2 n_i}{\exp(V_i \mathrm{MPN} - 1)} \right)^{-1/2} \tag{3.3}$$

Thus, the 95% confidence bound of MPN is obtained by the following.

$$95\%\ \text{Confidence Bound} = \exp(\log(\mathrm{MPN}) \pm 1.96 \times SD_{\log(\mathrm{MPN})}) \tag{3.4}$$

Since a Quanti-Tray®/2000 is a tray consisting of two dilution sets with 49 large wells of volume 1.86ml and 48 small wells of volume 0.186ml, the MPN is obtained by the root of the following equation.

$$\frac{s_1 V_1}{1 - \exp(-\mu V_1)} + \frac{s_2 V_2}{1 - \exp(-\mu V_2)} = n_1 V_1 + n_2 V_2 \tag{3.5}$$

where s_1 and s_2 are the number positive large and small wells, respectively. Similarly, $V_1 = 1.86$ and $V_2 = 0.186$ are the volume of large and small wells and $n_1 = 49$ and $n_2 = 48$ are the total number of large and small wells on the tray.

In order to estimate the concentration of duplicate samples, we assumed that both samples have the same concentration. With this well-mixed assumption for duplicate samples the MPN for duplicate sample is given by the root of the following equation (Appendix D).

$$\frac{(s_1 + s_3) V_1}{1 - \exp(-\mu V_1)} + \frac{(s_2 + s_4) V_2}{1 - \exp(-\mu V_2)} = 2n_1 V_1 + 2n_2 V_2 \tag{3.6}$$

where (s_1, s_2) and (s_3, s_4) are the number of positive large and small wells in first and second tray, respectively. The method can be applicable to any number of replicate samples (i.e., triplicate, etc.), as long as the well-mixed assumption holds (see equation in Appendix D). Each duplicate sample was also assessed for the well-mixed assumption by the likelihood ratio test (Haas 1999) using equations described in Appendix D.

3.2.4. Covariance function

In order to evaluate the spatial continuity of microbial contamination in the shallow aquifer, the covariance function of the log-transformed E. coli concentration was calculated for each month. The covariance function measures the spatial autocorrelation of the data as a function of the distance r between two locations $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$. In this study, following Curriero et al. (2002), the experimental covariance at various spatial lag r was calculated using the method-of-moments estimator that is given by the following equation.

$$\hat{C}\left(r = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|\right) = \frac{1}{|N(r)|}\sum_{i=1}^{N(\boldsymbol{r})}\left(Z(\boldsymbol{s}_i) - \bar{Z}\right)\left(Z(\boldsymbol{s}_j) - \bar{Z}\right) \tag{3.7}$$

where $\bar{Z}$ is an arithmetic mean of all values and $N(r)$ is the number of pairs of values a distance r apart (Cressie 1993). The experimental covariance was then used to fit the parameters of a positive definite covariance function. In this work, the powered exponential model was employed in order to account for the shape of the function

near the origin. The covariance range is a parameter of the covariance model corresponding to the separation distance at which the spatial correlation diminishes to a given fraction of the variance (Banerjee 2004). The powered exponential model is defined as

$$C(r) = C_1 \exp\left(-3\left|\frac{r}{a_r}\right|^b\right) \quad (3.8)$$

where r is a spatial lag, C_1 is the variance, also referred to as the sill parameter quantifying the variability of observations, a_r is spatial range at which the covariance decreases to 5% of the sill (or variance), and b $(0 < b \leq 2)$ is a power parameter which controls the shape of the covariance function. Both exponential and Gaussian covariance models are special cases of this model that correspond to $b = 1$ and $b = 2$, respectively. If the power parameter b is close to 2, the corresponding spatial distribution tends to be smooth. In contrast, the spatial field becomes irregular and patchy as the power parameter is close to 0. Shapes of powered exponential functions with different power parameter b are shown in Appendix B. All model parameters were estimated by an automated weighted least square procedure (Jian et al. 1996; Olea 2006).

3.2.5. Statistical analysis

The relationship between rainfall and the spatial continuity of microbial contamination in the shallow aquifer was evaluated using the Pearson's correlation coefficient between the covariance range of log-transformed E. coli concentration

and total rainfall observed over a given antecedent period. In order to evaluate the effect of the antecedent period, the correlation coefficient was computed multiple times by changing the antecedent period from 1 day to 21 days. All analyses were conducted using MATLAB R2008a (MathWorks Inc.).

3.3. Results

3.3.1. Quality Control Statistical Tests

During the study period, 526 duplicate 100ml water samples were collected monthly from the 55 monitoring tubewells in our study area, corresponding to a total of 526*2=1052 individual samples each analyzed in its own Quanti-Tray®/2000.

We performed quality control of our *laboratory* analysis procedures by statistically testing whether each Quanti-Tray®/2000 was well mixed according to the likelihood ratio test for individual samples (see Appendix D for a detailed description of this test). If both the large and small wells of the Quanti-Tray®/2000 for a given sample were all positive (i.e. $s_1 = n_1 = 49$ and $s_2 = n_2 = 48$) or all negative (i.e. $s_1 = 0$ and $s_2 = 0$), then the MPN cannot be uniquely determined and the likelihood ratio test is not applicable to that sample. Thus individual samples can be categorized into the following four groups (Table 3.1) based on the number of positive large and small wells and the result of the likelihood ratio test: (1) below detection limit (i.e. $(s_1, s_2) = (0,0)$), (2) above detection limit (i.e. $(s_1, s_2) = (49,48)$), (3) individual samples that passed the statistical test (i.e. for which the observed s_1 and s_2 can occur by chance under the assumption that organisms were well mixed

within the sample), and (4) individual samples that were rejected by the statistical test (i.e. for which the observed s_1 and s_2 cannot occur by chance under the well mixed assumption) at a significance level of $\alpha = 0.05$. We found that amongst the samples that could be tested (i.e. groups 3 and 4 above), the rate at which individual samples were rejected by the statistical test was 16/(373+16)=0.0411. This rejection rate is in good agreement with the rejection rate of $\alpha = 0.05$ that we expect to observe under the assumption that samples are well mixed. Thus this result supports that organisms were well-mixed within each Quanti-Tray®/2000 and that each organism exhibited growth when incubated in the culture medium, which provides quantitative evidence validating our laboratory procedures in the field.

Similarly we performed quality control of our *sampling* procedures by statistically testing whether the concentration of organisms were the same across duplicates for each of the 526 pairs of duplicate samples according to the likelihood ratio test for duplicate samples (see Appendix D for a detailed description of that test). As a result of this statistical test, duplicate samples were categorized into the following six groups (Table 3.2): (1) both duplicates are below detection limit (i.e. $(s_1, s_2) = (0,0)$ and $(s_3, s_4) = (0,0)$), (2) both duplicates are above detection limit (i.e. $(s_1, s_2) = (49,48)$ and $(s_3, s_4) = (49,48)$), (3) one duplicate is below detection limit (i.e. either (s_1, s_2) or (s_3, s_4) is equal to $(0,0)$, but not both), (4) one duplicate is above detection limit (i.e. either (s_1, s_2) or (s_3, s_4) is equal to $(49,48)$, but not both), (5) duplicate samples that passed the statistical test (i.e. for which the observed (s_1, s_2) and (s_3, s_4) can occur by chance under the assumption that duplicates have the same concentration of organisms), and (6) duplicate samples that were rejected

by the statistical test (i.e. for which the laboratory results cannot occur by chance under the assumption that duplicates have the same concentration of organisms) at a significance level of $\alpha = 0.05$. As can be seen from the table, amongst the samples that could be tested statistically (i.e. groups 5 and 6 above), the rate at which duplicate samples were rejected by this statistical test was 12/(135+12)=0.0816. This observed rejection rate is slightly greater but generally in good agreement with α=0.05, which means that the concentration of organisms were generally the same across duplicates. This result indicates that our sampling errors were successfully maintained to an acceptably small level compared to the analytical error of the Quanti-Tray®/2000, and thus provide quantitative evidence validating our sampling procedures in the field.

Table 3.1: Results of the likelihood ratio test classifying whether each of the 1052 individual samples were well mixed within sample at a significance level of $\alpha = 0.05$

Type	Number of individual samples
Below detection limit	654
Above detection limit	9
Organisms are well mixed within the sample at $\alpha = 0.05$	373
Organisms are not well mixed within the sample at $\alpha = 0.05$	16
Total Sample	1052

Table 3.2: Results of the likelihood ratio test classifying whether each of the 526 pairs of duplicate samples had the same concentration across duplicates at a significance level of $\alpha = 0.05$

Type	Number of duplicate samples
Both duplicates are below detection limit	280
Both duplicates are above detection limit	4
One duplicate is below detection limit	94
One duplicate is above detection limit	1
The concentration of organisms is the same across duplicates ($\alpha = 0.05$)	135
The concentration of organisms is different between duplicates ($\alpha = 0.05$)	12
Total Sample	526

3.3.2. Seasonal Variation of E. coli concentration

Observed MPN E. coli concentrations were grouped into five categories based on the classification scheme (Table 3.3) proposed by the World Health Organization (WHO 1997). During the monsoon season, the number of the clean tubewells (Category A) gradually decreased, whereas the number of high and very high risk tubewells increased. In post monsoon months, however, none of the tubewells were categorized as high or very high risk (Figure 3.3(a)). Figure 3.3 also shows the maps of the monitoring tubewells with WHO categories in monsoon month; (b) August, 2008, in end of the monsoon month; (c) November, 2008, and in post monsoon month; (d) March, 2009. Highly contaminated tubewells were located

in the densely populated northern part of the study area, whereas tubewells in southern part of the study area were generally clean during the study period.

Table 3.3: WHO classification scheme

E. coli MPN per 100mL	Category (Color Code)	Remarks
0	A (Blue)	
1-10	B (Green)	Low Risk
10-100	C (Yellow)	Intermediate Risk
100-1000	D (Orange)	High Risk
>1000	E (Red)	Very High Risk

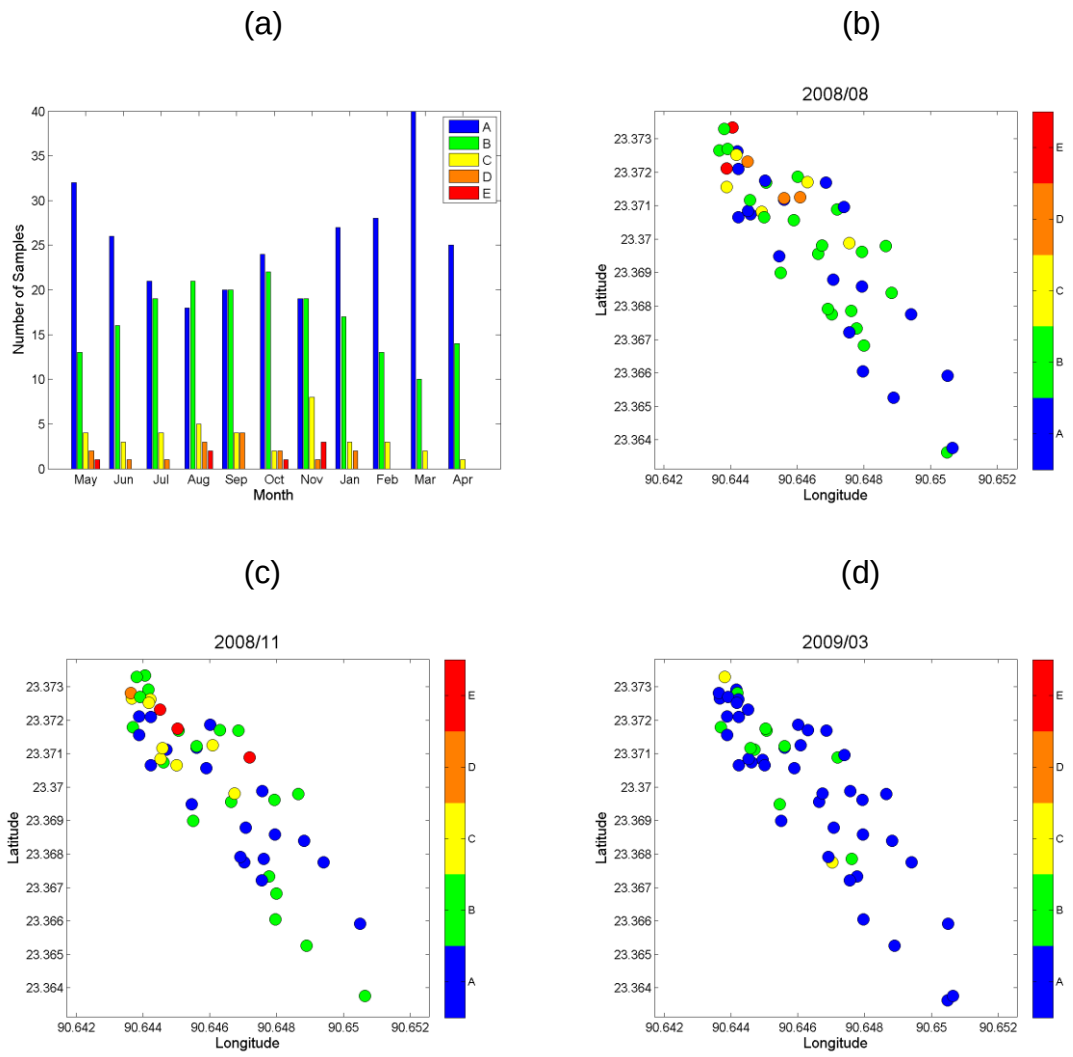

Figure 3.3: (a) temporal plot of WHO categories and spatial distribution of E. coli categories in (b) August, 2008, (c) November, 2008, and (d) March, 2009.

3.3.3. Covariance Model Parameters

Covariance model parameters estimated by an automated weighted least square procedure are listed in Table 3.4. All parameters showed a clear seasonal pattern. During the June to October monsoon season the covariance range is consistently covering a relatively long distance of about 100 m, and it reaches its

maximum length of about 170 m in July. In the post monsoon season, the covariance range quickly decreases to less than 50 m. The sill also shows a similar trend. Its value is generally higher in the monsoon season than it is in the post monsoon season, which indicates that the observed E. coli concentrations have higher variability during the rainy monsoon season. The power parameter is also exhibiting a clear seasonal pattern. Its value is generally small in the monsoon season except for September, whereas the value is close to its highest value two in the post-monsoon season.

Table 3.4: Covariance parameters during the study period

Month	Covariance Range (m)	Power Parameter	Sill
May	0.588	1.69	4.78
June	57.4	0.517	3.24
July	173	0.677	3.12
August	104	0.444	6.39
September	66.8	2	4.96
October	90.1	0.502	4.23
Nov/Dec	71.2	0.423	6.36
January	20.6	1.99	2.92
February	6.59	1.52	1.69
March	44.1	1.96	1.77
April	12.1	1.16	1.00

3.3.4. Statistical analysis

The covariance range of the log-transformed E. coli concentration is highly correlated with antecedent rainfalls. The Pearson's correlation coefficients and the associated p-values between the covariance range and the 1- to 21-day antecedent rainfalls are shown in Figure 3.4(a) and (b), respectively. The 13-day antecedent rainfall exhibited the highest correlation with the covariance range (r=0.885, p=0.00151). The temporal plot of the covariance range and the 13-day antecedent rainfall is shown in Figure 3.4(c).

The association between rainfall and the covariance range of the levels of the WHO classification scheme was also investigated (See Appendix C for more details). The covariance parameters showed similar trends, with long range autocorrelation in the monsoon season and shorter range in the post-monsoon season. The 9-day antecedent rainfall exhibited the strongest correlation with that covariance range (r = 0.9, p = 0.05).

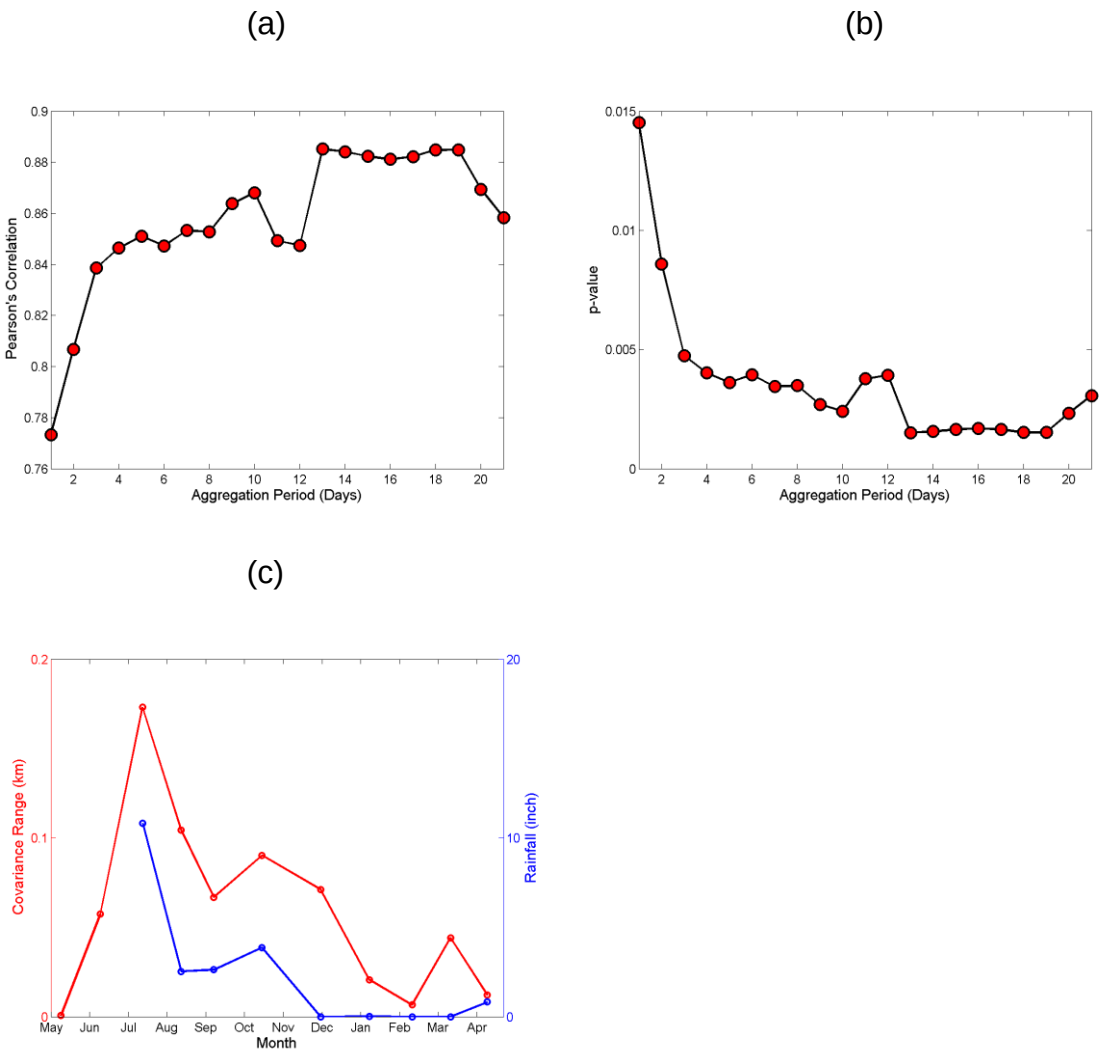

Figure 3.4: (a) Pearson's correlation coefficient and (b) associated p-value between the covariance range and the 1- to 21-day antecedent rainfalls (c) Temporal plot of covariance range and the 13-days antecedent rainfall.

3.4. Discussion

In the present study, we demonstrate that the spatial extent of microbial contamination in a shallow aquifer in Matlab, Bangladesh, is strongly associated with antecedent rainfall. The range of the covariance function of log-transformed E. coli

concentration was highly correlated with antecedent rainfalls, with the strongest association found for the 13-day antecedent rainfall.

Duplicate tubewell water samples were collected and analyzed using the Quanti-Tray®/2000 with Colilert® reagent, and their E. coli concentrations were calculated using the MPN method. The likelihood ratio test revealed that the rejection rate for individual and duplicate samples were about 4.11% and 8.16%, respectively, which was generally in good agreement with the rejection rate of α=5% expected under the assumption that samples were fully mixed and duplicates had the same concentrations. These results provide quantitative evidence indicating that our laboratory and sampling procedures were adequate in generating good quality data, and that the assumptions underlying our calculation of the MPN for duplicate samples are well founded for our dataset.

The equation we introduce to calculate the MPN for duplicate samples is a straightforward application of the MPN theory for multiple dilution series when the sampling error is small compared to analytical error. We showed in this work how to test the assumptions underlying our calculation of the MPN for duplicate samples, and we implemented numerically the MPN calculation into an easy-to-use spreadsheet that was made publicly available and is therefore available to new studies requiring MPN calculation for duplicate samples. The advantage of the framework we describe to calculate the MPN of duplicate samples when sampling errors are small is that it rigorously takes into account whether duplicate samples agree or disagree when calculating the CI, resulting in a better description of the uncertainty associated with the MPN than a naïve approach that would simply

average the MPN of each duplicate sample. A Bayesian hierarchical model was developed to account for large sampling errors, but this more complicated framework was not needed in this study and will therefore only be presented in future works where sampling errors are important relative to analytical errors.

The tubewell E. coli concentrations showed a clear seasonal pattern. The numbers of high and very high risk tubewells increased during the monsoon season. By contrast, none of the tubewells were categorized as high or very high in the post-monsoon season (Figure 3.3). This finding is broadly consistent with a study conducted in Uganda which found a statistically significant association between microbial contamination in shallow springs and short-time rainfall events (Howard et al. 2003). The correlation between rainfall and tubewell contamination can be explained through two different pathways. The first contamination pathway is simply through the tubewell itself. For example an inadequately sealed tubewell might allow contaminated overland runoff to penetrate in the tubewell and be downward transported in the annulus surrounding the pipe of the tubewell, which might further trigger the regrowth of bacteria adhering to the pipe. This contamination pathway is, however, tubewell specific and does not necessarily lead to contamination of the aquifer itself (i.e. the contamination is contained within isolated tubewells). The second contamination pathway may consist in the vertical subsurface microbial transport of fecal matter from contamination sources on the ground. This subsurface microbial transport might result from a downward hydraulic gradient induced by rainfall. This contamination would lead to a contamination of the aquifer itself (i.e. it is not contained within tubewells), which should result in widespread patterns of

contamination. A study conducted in Araihazar Upazila, Bangladesh, revealed that the age of groundwater, measured using the tritium-helium ($^3H/^3He$) groundwater dating technique, ranges from less than 1 year old to more than 30 years old (Stute et al. 2007). Thus microbial transport from contamination sources on the ground to the underlying shallow aquifer would appear to take at least a few months. This might suggest that widespread contamination in the subsurface originates from a small portion of microbial pathogens transported from highly contaminated sources on the ground by preferential flow path (Taylor et al. 2004). Hence the correlation between rainfall and tubewell contamination can be explained either by a contamination through broken tubewells seals, which would not lead to a widespread contamination of the shallow aquifer, or by direct subsurface downward transport from high strength point sources on the ground surface, which would lead to widespread contamination of the subsurface. To the best of our knowledge, no study has been able to distinguish between these two contamination pathways.

This work is the first study in which many tubewells were sampled for FIB over a relatively small area in Bangladesh, so as to generate a rich dataset allowing to precisely model the covariance of tubewell microbial concentrations. As a result we believe that this is the first study investigating the relationship between rainfall and the covariance range of E. coli concentrations, or put in other words, whether heavy rainfalls lead to widespread contamination of the aquifer. Indeed, a covariance range covering a long distance indicates that tubewell concentrations are autocorrelated, which provides evidence that the aquifer itself must be autocorrelated and therefore contaminated over large areas.

As shown in Table 3.4, the covariance parameters of log-transformed E. coli concentration exhibited a clear seasonal pattern. In the monsoon season, the covariance function was generally characterized by long covariance range, large sill, and small power parameter. Thus, log-transformed E. coli concentrations were highly variable and not smoothly distributed, while being spatially correlated over long distances. This result indicates that the overall spatial process can be described as the combination of two separate spatial processes. One is a white noise random process induced by contaminations that are independent from one monitoring well to the next. The other is a smooth spatial process characterized by the long covariance range. In post monsoon season, the covariance function demonstrated the opposite traits; i.e. short covariance range, small sill, and large power parameter. This result is reasonably explained by low level contamination over the study area during the post-monsoon season.

The covariance range was highly correlated with the antecedent rainfall and the correlation was strongest with 13-day antecedent rainfall (r=0.885). This suggests that the smooth spatial process with long covariance range in the monsoon season corresponds to a widespread contamination of the shallow aquifer that is due to rapid microbial transport along preferential paths as explained above. The strongest correlation is observed with the 13-day antecedent rainfall, which might correspond to the time over which microbial organisms are transported from high strength surface fecal contamination sources on the ground surface to the shallow aquifer along the preferential path. On the other hand, the white noise spatial random effect in the monsoon season might be explained by tubewell specific

contaminations at inadequately sealed tubewells or by bacterial regrowth in tubewell pipes.

Hence this study is the first of its kind finding that rainfall is one of the key components controlling the spatial extent of fecal contamination in the shallow aquifers of Bangladesh, and that the corresponding spatial process can be described as the combination of a white noise random process and smooth spatial process characterized by long covariance ranges in the monsoon season. The later might be associated with downward subsurface microbial transport along preferential flow paths from high strength fecal contamination sources on the ground. Highly contaminated point source of microbial pathogens in the study area includes pit latrines and surface water bodies, such as ponds, canals, and ditches. Thus it is critical that future works explicitly explore the effect from these point sources on tubewell fecal contamination, which might be critical in understanding how to reduce fecal exposure and possibly Diarrhea in Bangladesh.

CHAPTER 4

Space/Time Statistical Estimation of Fecal Indicator Bacteria across Drinking Wells in Bangladesh using Latrine Locations and Rainfall

4.1. Background

As described in the previous chapter, in order to reduce the burden of diarrheal disease in Bangladesh it is essential to elucidate the mechanisms leading to groundwater microbial contamination. In this chapter, to further investigate these mechanisms, we conduct a mapping analysis of the space/time distribution of tubewell E. coli contamination across tubewells in the study area described in the previous chapter. In that chapter, the spatial extent of microbial contamination in the shallow aquifer was found to be strongly associated with rainfall. Vertical flow due to rainfall is one of the driving forces for microbial transport from surface fecal contamination sources to the shallow aquifer. Thus in order to adequately describe the space/time distribution of microbial contamination in the groundwater, we will in this chapter incorporate data on variables describing rainfall as well as the source, fate, and transport of microbial contamination.

Latrines are a potential source of fecal contamination. Ali et al. (2002) found a statistically significant positive association between a variable based on latrine usage and cholera incidence rate. Emch et al. (2008) investigated the risk factors for

shigellosis and reported that neighborhoods near bazaars with many non-septic latrines were at the highest risk for S. dysenteriae. On the other hand, several studies found that the presence of latrines near tubewells was not a significant risk factor for microbial contamination in tubewell (Godfrey, Timo, and Smith 2006; Howard et al. 2003; Luby et al. 2008). These results might indicate that latrines potentially influence the microbial contamination of tubewell water, but a simple assessment such as the presence or absence of latrines around a tubewell of interest might be inadequate to assess the effect of latrines on that tubewell.

Although there are several factors potentially affecting the microbial fate and transport of E. coli, population is considered as a key factor for groundwater microbial contamination in this work. Several studies found a positive association between variables based on population and the occurrence of diarrheal disease (Ali et al. 2002; Emch 1999; Emch et al. 2008). For instance, Emch (1999) found that population density within a 0.5km radius is positively associated with hospitalized cholera incidence.

In order to investigate the space/time distribution of microbial contamination in tubewell water, we develop in this chapter a two-stage geostatistical estimation model. In the first stage, we construct a linear regression model predicting log-transformed E. coli concentration as a function of latrine density, population and rainfall. Then the regression model output is integrated into a space/time knowledge synthesis framework based on the Bayesian Maximum Entropy (BME) theory to estimate E. coli concentration at any unmonitored space/time location.

4.2. Material and Method

4.2.1. Study Area/ Tubewell Water Sample/Precipitation Data

The study was conducted in Bara Haldia using the E. coli concentration data and the precipitation data described in the previous chapter. See sections 3.2.1, 3.2.2, and 3.2.3 for a detailed description of study site, tubewell water samples, and precipitation data.

4.2.2. Latrine hydrological regression model

A latrine hydrological regression model was constructed in order to assess the relationship between the microbial contamination and the three risk factors presented above (latrine density, population, and rainfall). The log-transformed E. coli MPN value for duplicate samples was the dependent variable for that model (Appendix D). Explanatory variables were constructed for each risk factor using various hyperparameters. The effect of latrine around a monitoring well was quantified by the latrine variable (lv), which accounts for the microbial loading from each latrine as a function of the distance between the latrine and the tubewell. Microbial filtration in the subsurface can be modeled with a first order rate expression

$$C = C_0 \exp(-\lambda d) \tag{4.1}$$

where C is the concentration at distance d, C_0 is the initial concentration, λ is the filtration rate, and d is the distance along flow path (Logan et al. 1995). Thus microbial loading from a latrine to a tubewell can be approximated by an exponentially decaying function (Figure 4.1(c)). In this study, the latrine variable at tubewell i (lv_i) was defined as the sum of the exponentially decaying contribution from each of the surrounding tubewells and was expressed by

$$lv_i = \sum_{j=1}^{m} \exp\left(-\frac{d_{lv\_ij}}{3r_{lv}}\right) \tag{4.2}$$

where m is the total number of latrine, d_{lv_ij} is the distance from tubewell i to latrine j, and r_{lv} is a hyperparameter, which we call the latrine microbial range, that corresponds to the distance traveled from a latrine to achieve 95% removal of FIB. The contour map shown in Figure 4.1(d) displays the spatial distribution of the latrine variable lv calculated over the study area using $r_{lv} = 120$ m. Similarly the effect of population at tubewell i is quantified by the population variable (pv) defined as the total number of people living in households located within a distance r_{pv} from tubewell i, which we call the population radius. Figure 4.1(b) shows the contour map of the population variable, pv, calculated over the study area using $r_{pv} = 25$ m.

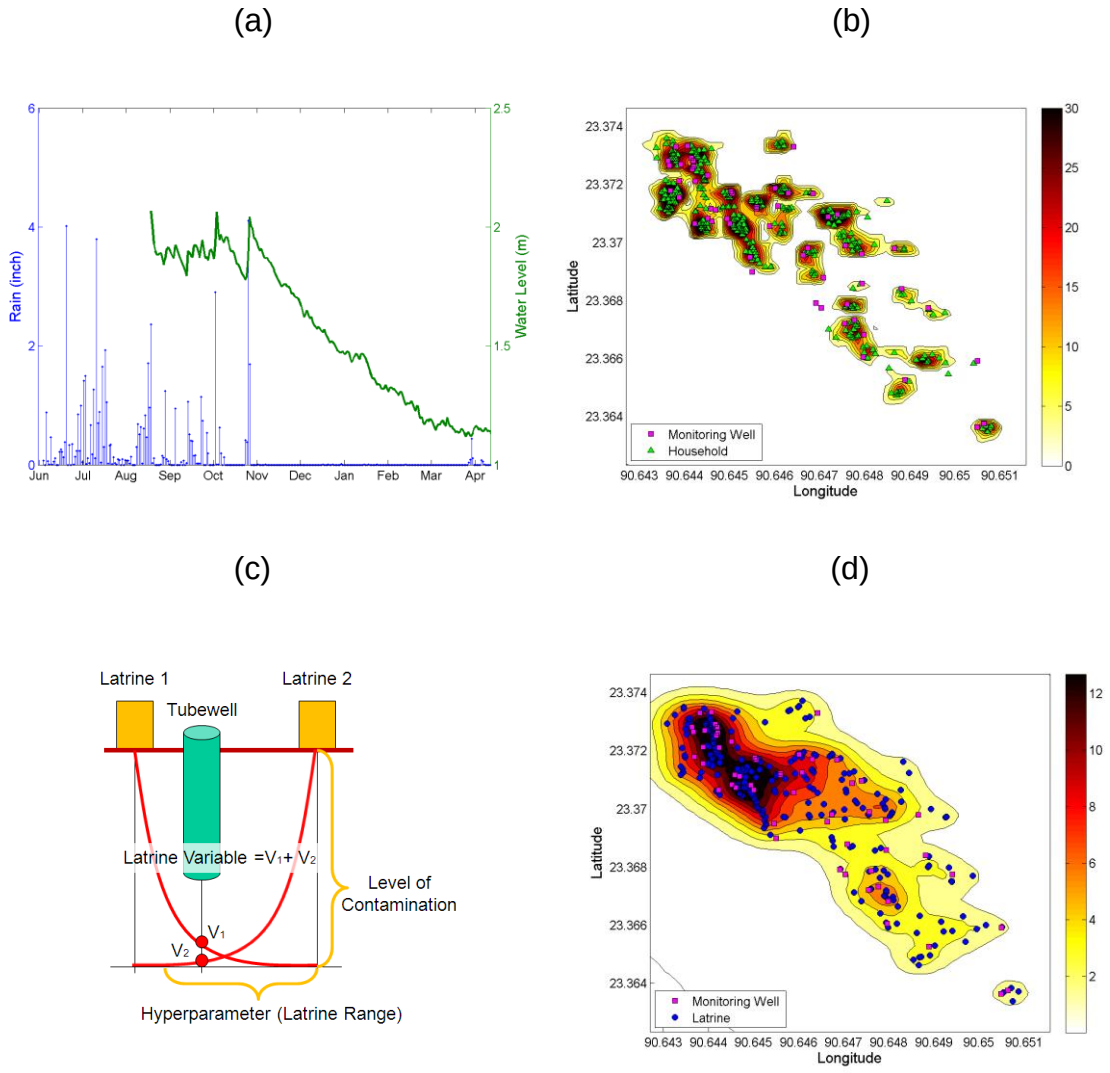

Figure 4.1: (a) Daily rainfall observed at the Bara Haldia weather station (b) Map of the population variable, pv, calculated over the study area using $r_{pv} = 25$ m (c) Latrine variable calculated at a tubewell as the sum of the exponentially decaying contribution from two latrines (d) Map of the latrine variable lv calculated over the study area using $r_{lv} = 120$ m.

The effect of rainfall on E. coli at some sampling date of interest is quantified by two rainfall variables denoted as $rv1$ and $rv2$. The rainfall variable 1, $rv1$, is

defined as a first antecedent rainfall (inches) characterized by two hyperparameters which are the rainfall duration dur_1 (days) and lag lag_1: (days) between the antecedent rainfall and the sampling date. Thus, $rv1$ (lag_1,dur_1) is defined as the sum of the daily rainfalls observed from $lag_1 + dur_1$ to lag_1 days prior to the sampling date. Similarly we constructed the second rainfall variable $rv2$ as the antecedent rainfall of duration dur_2that immediately precedes $rv1$. Thus, $rv2(dur_2)$ is the sum of daily rainfalls from day $lag_1 + dur_1 + dur_2$ to day $lag_1 + dur_1 + 1$ prior to the sampling event.

To see the effect of the second antecedent rainfall variable $rv2_i$, we constructed a first latrine hydrological regression model (LHM1) that did not include that variable and is expressed by the following equation

$$\log(\mathrm{MPN}_i) = \beta_0 + \beta_1 lv_i + \beta_2 pv_i + \beta_3 rv1_i + \varepsilon \qquad (4.3)$$

where MPN_i is the MPN of E. coli for tubewell i, lv_i , pv_i, and $rv1_i$ are the latrine variable, population variable, and first antecedent rainfall at tubewell i, respectively, β_0 through β_3 are regression coefficients, and ε is an error term. We then defined the second model (LHM2) as

$$\log(\mathrm{MPN}_i) = \beta_0 + \beta_1 lv_i + \beta_2 pv_i + \beta_3 rv1_i + \beta_4 rv2_i + \varepsilon \qquad (4.4)$$

where $rv2_i$ is the second antecedent rainfall variable for tubewell i. Linear regression theory was used to obtain the regression coefficients. Because of the

limited availability of rainfall data, only E. coli samples collected from July 2008 to April 2009 were used to obtain the regression coefficients.

4.2.3. Hyperparameter Selection

The hyperparameters were selected by finding the hyper parameter values that maximize the coefficient of determination (R^2) of the corresponding regression model, subject to the constraint that all the regression coefficients must be statistically significant and physically meaningful. Let us consider Figure 4.2(a) for illustration purposes. In this figure we show the R^2 obtained for the regression model LHM1 when the latrine microbial range r_{lv} varies from 0 to 600 m while the hyperparameters r_{pop} , $lag1$, and $dur1$ are kept at 25 m , 4 days and 1 day, respectively. Also shown on the figure in blue lines are the p-values associated with each of the three explanatory variables of LHM1. All three p-values are statistically significant (i.e. less than 0.05) in the region where r_{lv} ranges from 90 to 120 m. Within this region we seek the latrine microbial range r_{lv} that maximizes R^2, which leads to a selected value of 120 m for the latrine microbial range r_{lv}. While this graph provides an illustration of the procedure, in practice one can use a minimum search algorithm to find the combination of all four hyperparameter values for LHM1 which maximizes the R^2 subject to statistically significant and physically plausible regression coefficients.

4.2.4. Estimation at unmonitored location

Once the regression coefficients are estimated, the space/time trend for log transformed E. coli concentration m_Y at any space time location $\boldsymbol{p}$ based on LHM1 is obtained as

$$m_Y(\boldsymbol{p}) = \widehat{\beta_0} + \widehat{\beta_1} lv(\boldsymbol{p}) + \widehat{\beta_2} pv(\boldsymbol{p}) + \widehat{\beta_3} rv1(\boldsymbol{p}) \qquad (4.5)$$

where $\boldsymbol{p} = (\boldsymbol{s}, t)$ is the space/time location, $\boldsymbol{s} = (s_1, s_2)$ is the spatial coordinate, t is the time, and $\widehat{\beta_0}$ through $\widehat{\beta_3}$ are the estimates of β_0 through β_3 obtained from linear regression. A similar mean trend model can be constructed using LHM2 which includes the effect of the second antecedent rainfall variable.

In order to model the distribution of residual E. coli concentration over space and time, we developed a space/time knowledge synthesis framework based on the BME theory. In this framework, the theory of space/time random field (S/TRF) is employed to model the E. coli concentration (Christakos 1992). Let $Y(\boldsymbol{p})$ be the S/TRF modeling the distribution of log transformed E. coli concentration at space/time point $\boldsymbol{p}$, and $m_Y(\boldsymbol{p})$ be the latrine hydrological regression model defined above. Then, the S/TRF of the residual log transformed E. coli concentration $X(\boldsymbol{p})$ is defined as

$$X(\boldsymbol{p}) = Y(\boldsymbol{p}) - m_Y(\boldsymbol{p}) \qquad (4.6)$$

We then use the BME theory to estimate the residual concentration at any unmonitored space/time location. The BME framework to efficiently integrate general knowledge describing global characteristics of the S/TRF in terms of its mean trend and covariance, and site-specific knowledge consisting in the E. coli concentrations measured at specific space/time points. The details of the BME theory and its numerical implementation can be found elsewhere (Christakos 1990; Christakos et al. 2001).

The covariance of the S/TRF $X(\boldsymbol{p})$ is estimated using the residuals of log-transformed MPN at each monitoring well. Following a procedure described in Curriero et al. (2002), we calculated experimental covariance at various spatial lag r and temporal lag τ using the method-of-moments estimator (Cressie 1993). These experimental covariance values were then used to fit the parameters a positive definite covariance model. In this work, we used the powered exponential covariance model given by the following equation (Banerjee 2004).

$$C_x(r,\tau) = \begin{cases} {\sigma_0}^2 + {\sigma_X}^2 & \text{if } r = 0 \; and \; \tau = 0 \\ {\sigma_X}^2 \exp(-3\left|\frac{r}{a_r}\right|^{b_r}) \exp\left(-3\left|\frac{\tau}{a_\tau}\right|^{b_\tau}\right) & \text{otherwise} \end{cases} \quad (4.7)$$

All model parameters were estimated using an automated weighted least square procedure (Jian et al. 1996; Olea 2006).

As explained in the previous chapter, the MPN estimate for duplicate samples assumes that each individual sample was well mixed and that the duplicates had the same concentration. This assumption can be tested using the likelihood ratio test for

duplicate samples (Appendix D). In this work, only the samples that passed the likelihood ration test for duplicate samples (i.e. the 135 duplicate samples listed in group 5 of Table 3.2) were treated as hard data (i.e., data with no measurement error). All other duplicate samples were treated as soft data (i.e., data with measurement error). Since the sampling distribution of the duplicate sample MPN estimate is approximated by a log normal distribution (Hurley, and Roscoe 1983), the soft data of the residual log transformed E. coli concentration $X(\boldsymbol{p})$ were modeled as normal distributions with mean $\log(\mathrm{MPN}_i) - m_Y(\boldsymbol{p}_i)$ and standard deviation $\sigma_{\log(\mathrm{MPN})}$. However, when both duplicates are below detection limit (i.e. for the 280 duplicate samples listed in group 1 of Table 3.2), then if not possible to calculate the standard deviation $\sigma_{\log(\mathrm{MPN})}$, and consequently we used soft data such that we have

$$\exp(X(\boldsymbol{p}) + m_Y(\boldsymbol{p}))/ld \sim beta(3,3) \tag{4.8}$$

where $ld = 0.5$ organisms/100mL is a lower detection limit of the duplicate sample MPN estimator. Similarly when both duplicates are above detection limit (i.e. for the 4 duplicate samples listed as group 2 of Table 3.2) we used soft data corresponding to a uniform probability distribution between $\log(ud) - m_Y(\boldsymbol{p})$ and $\log(2ud) - m_Y(\boldsymbol{p})$, where ud=2455 organisms/100mL is the upper detection limit for the duplicate sample MPN estimator.

In order to evaluate the performance of our estimation framework, we conducted a cross validation analysis to compare the following three models

1. Latrine hydrological model only
2. Latrine hydrological model and the BME method with hard data (all samples are treated as hard data, thereby ignoring the uncertainty associated with duplicate samples that are below or above the detection limits)
3. Latrine hydrological model and the BME method with hard and soft data (where the soft data were constructed as described above)

Leave-one-out cross validations were performed using the hard data points as the validation data set. The model performance was evaluated by root mean square error (RMSE) statistics given by the following equation

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2} \qquad (4.9)$$

were Y are the hard data values and $\widehat{Y}_i$ are the cross validation estimates.

The regression analysis and geostatistical estimation were conducted using MATLAB R2008a (MathWorks Inc.).

4.3. Results

4.3.1. Hyperparameter Selection

The optimal hyperparameters values, corresponding R^2, and Akaike information criterion (AIC) are listed in Table 4.1. In terms of the goodness of fit (R^2) and parsimony (AIC) of the model, LHM2 performed better than LHM1, which indicates that the second rainfall variable should be included in the model. The corresponding regression coefficients and associated p-value are listed in Table 4.2.

Table 4.1: The optimal hyperparameter values, R^2, and AIC for LHM1 and LHM2

	LHM1	LHM2
Latrine Microbial Range (m)	120	110
Population Radius (m)	25	25
Time Lag (days)	4	4
Duration of Rainfall (days)	1	1
Duration of Rainfall (days)	NA	3
R^2	0.0641	0.0713
AIC	571.83	570.51

Table 4.2: Regression coefficients and associated p-value for LHM1 and LHM2

	LHM1		LHM2	
	Beta	p-value	Beta	p-value
Constant	-0.946	1.97E-05	-0.806	0.000386
Latrine Variable	0.0540	0.0159	0.0496	0.0488
Population Variable	0.0153	0.0499	0.0153	0.0492
Rainfall Variable1	1.03	0.00136	1.69	0.000264
Rainfall Variable2			-1.05	0.0480

Figure 4.2(a) shows the R^2 obtained for LHM1 as a function of the latrine microbial range, r_{lv}, while the other hyperparameters are fixed by their optimal values. The p-values associated with each of the three explanatory variables of LHM1 are also shown on the figure in blue lines. As explained earlier, the optimal value for the microbial range is 120m because that is the value at which the model reaches its greatest R^2 while being at the same time statistically significant (i.e. for which all three p-values are less than 0.05). Similarly, Figure 4.2(b) and (c) show the R^2 obtained for LHM1 as function of (b) the population radius, r_{pv}, and (c) the rainfall lag, lag_1, respectively. The contour map shown in Figure 4.2(d) displays the spatial distribution of E. coli concentration predicted by LHM1. Since both latrine and household were more densely located in the northern part of the study area, the estimated E. coli concentration was also higher in that part of the study area.

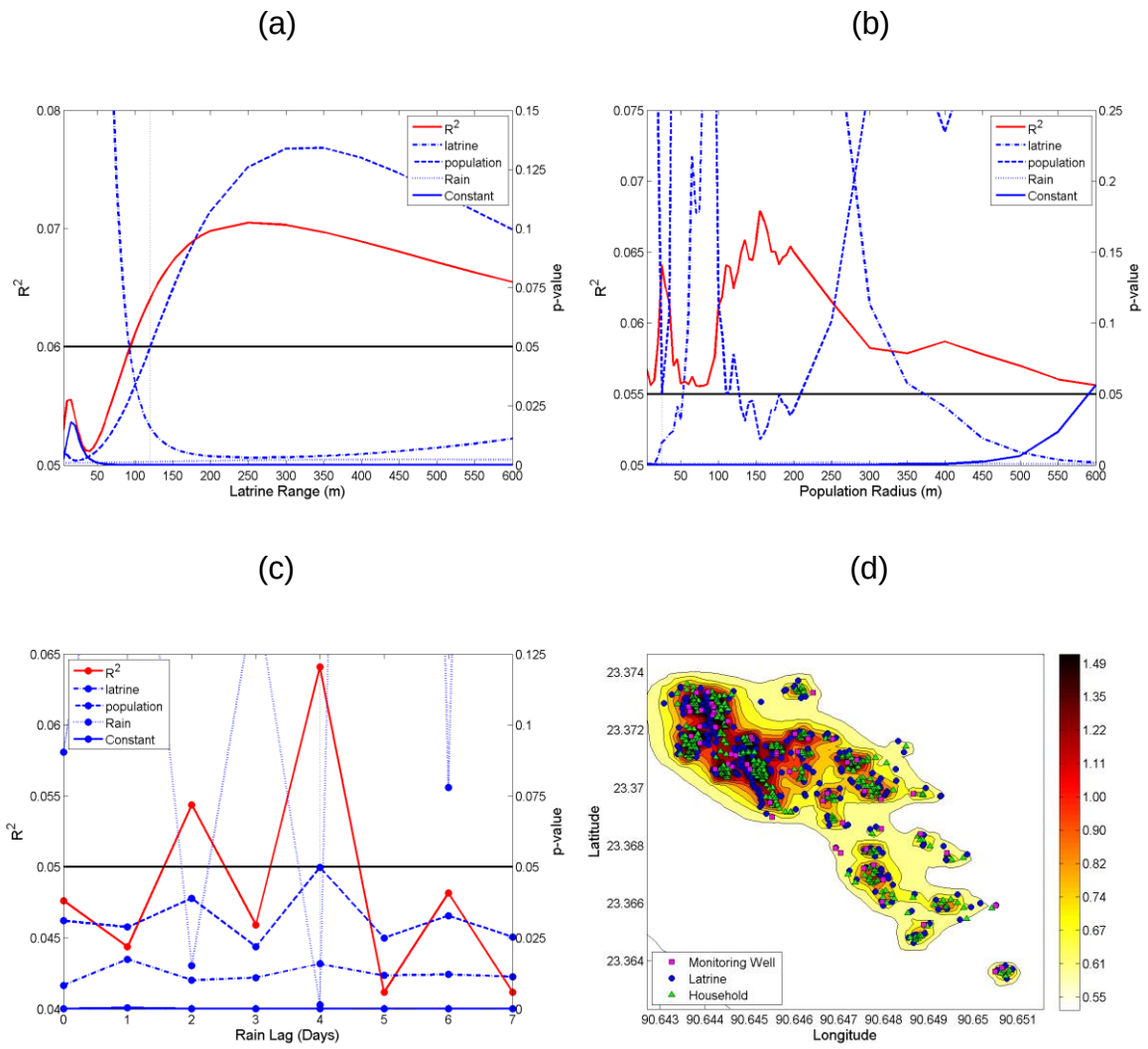

Figure 4.2: (a) Plot of the R^2 and p-values of the regression model LHM1 as a function of the latrine microbial range hyperparameter r_{lv} while fixing the other hyperparameters to their optimal values. The corresponding plots as a function of the population radius hyperparameter r_{pv} and the rainfall lag parameter lag_1 are shown in (b) and (c), respectively. (d) Map of the E. coli concentration the by LHM1

4.3.2. Covariance Function

The experimental covariance of the residual log transformed E. coli concentration $X(\boldsymbol{p})$ based on LHM1 is shown in red circle in Figure 4.3. Also shown

in solid green line is the fitted covariance model given by Eq. (4.7). The value of the covariance parameters obtained using LHM1 and LHM2 are listed in Table 4.3.

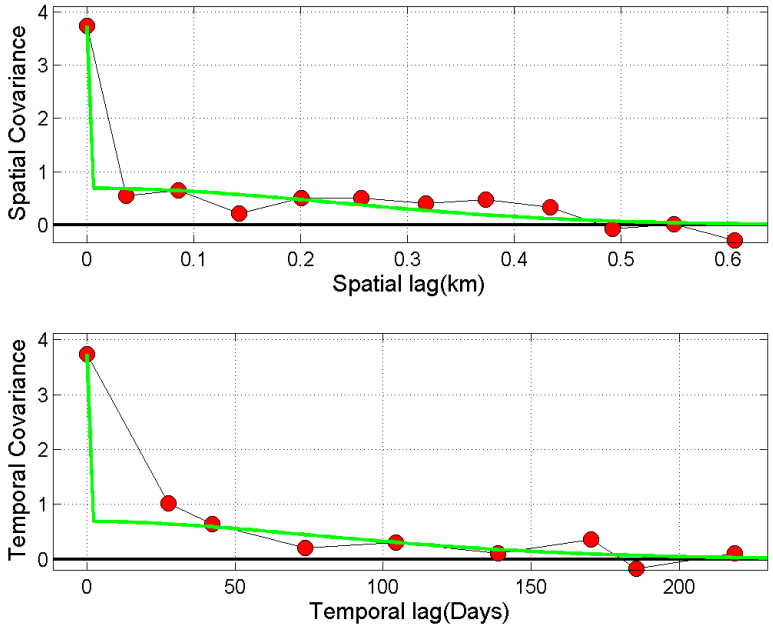

Figure 4.3: Space/time experimental covariance of the residual log transformed E. coli concentration $X(\boldsymbol{p})$ based on LHM1 (red circle) and fitted covariance function (green line).

Table 4.3: Covariance parameters for LHM1 and LHM2

	LHM1	LHM2
Nugget ($\sigma_0{}^2$)	3.0449	3.023
Sill ($\sigma_X{}^2$)	0.68871	0.68174
Spatial Range (a_r km)	0.56943	0.54035
Spatial Power (b_r)	2	2
Temporal Range (a_τ days)	213.0733	215.2395
Temporal Power (b_τ)	1.8219	1.8012

4.3.3. Cross validation

The results of the cross validation analysis of the three estimation methods described earlier are presented in Table 4.4 for LHM1 and for LHM2. The second column of that table shows the cross validation RMSE, which is a measure of the estimation error and the third column is the percent change in the RMSE relative to the latrine hydrological method. A negative percent change indicates a decrease in the RMSE, which corresponds to an improvement in performance estimation. Model 1-2 (LHM1 followed with BME estimation using hard data) reduced the RMSE by a notable 6% relative to model 1-1 (LHM1). Model 1-3 (LHM1 followed with BME estimation using hard and soft data) further improved model performance, as it reduced the RMSE by about 24% relative to model 1-1. Results were similar for the estimation methods based on LHM2. Interestingly, method 1-3 (LHM1 and BME estimation with hard/soft data) attained the smallest RMSE among all methods, even though LHM2 is a better regression model than LHM1 in terms of goodness of fit (R^2) and parsimony (AIC).

Table 4.4: The cross-validation RMSE of three estimation methods using a mean trend obtained from either LHM1 or LHM2

	Space/time estimation method	RMSE	Improvement (%)
1-1	Latrine Hydrological Model 1 (LHM1)	2.521	
1-2	LHM1 and BME estimation with hard data	2.368	-6.06
1-3	LHM1 and BME estimation with hard/soft data	1.906	-24.38
2-1	Latrine Hydrological Model 2 (LHM2)	2.511	
2-2	LHM2 and BME estimation with hard data	2.381	-5.17
2-3	LHM2 and BME estimation with hard/soft data	1.910	-23.93

4.3.4. Estimation at unmonitored location

The space/time knowledge synthesis framework based on the BME theory combine with the latrine hydrological model described in this work can be used to construct maps showing the spatial distribution of E. coli concentration across the study area for any particular day of interest or to produce plots showing how E. coli concentration changes over time at any particular tubewell location. Figure 4.4 shows map of E. coli concentration estimated using LHM1 and BME estimation with hard/soft data (model 1-3 in Table 4.4) on (a) November 30, 2008 and (b) March 10, 2009. A series of contour maps of E. coli concentration during the study period is also shown in Appendix E.

(a) (b)

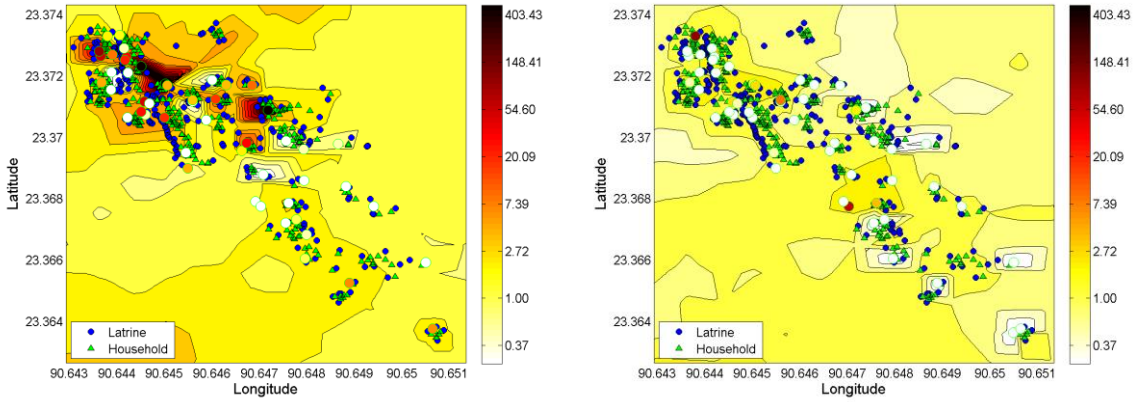

Figure 4.4: Map of E. coli concentration estimated by LHM1 and BME estimation with hard/soft data on (a) November 30, 2008 and (b) March 10, 2009

The plots of the E. coli concentration and associated 95% confidence interval obtained (using again estimation method 1-3) at tubewell 21783 and 21772 are

shown in Figure 4.5(a) and (b), respectively. The green dots on the plot shows the observed MPN values and the green solid lines display the corresponding CI used to construct the soft data. This plot is useful to visually identify the period in which the tubewell was highly contaminated with E. coli. Additionally, in order to quantify how often a tubewell was at risk of E. coli contamination during the study period, the estimated E. coli concentration was grouped into five risk categories (Table 4.5) based on the WHO classification scheme (WHO 1997). Using Figure 4.5 we find that tubewell 21783 was at the intermediate or higher risk (category C, D, or E) of E. coli contamination for approximately 30% of the study period, whereas tubewell 21772 had E. coli concentrations estimated to be in these high risk categories for only 2% of the study period.

Figure 4.6 shows a graph of the fraction of the WHO risk categories for each day of the study period. The plot was obtained by estimating the E. coli concentration for each day and each monitoring tubewells of the study domain. During the monsoon season, several monitoring wells were constantly at the intermediate or higher risk (category C, D, or E) of contamination, whereas all tubewells were categorized as low risk or safe (category A or B) during the post-monsoon season.

(a) (b)

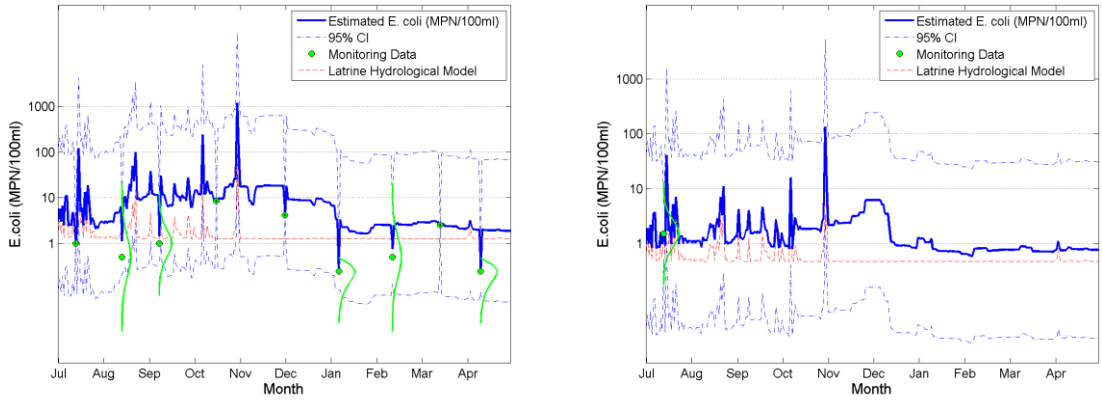

Figure 4.5: Plots of the E. coli concentration and associated 95% confidence interval predicted by LHM1 and BME estimation with hard/soft data at tubewell (a) 21783 and (b) 21772.

Table 4.5: The number of days during the study period in each WHO category

		21783		21772	
WHO Category	MPN/100ml	# of days	Rate (%)	# of days	Rate (%)
A (Safe)	0	3	0.987	147	48.4
B (Low Risk)	1-10	216	71.1	152	50
C (Intermediate Risk)	10-100	82	27.0	4	1.32
D (High Risk)	100-1000	2	0.658	1	0.329
E (Very High Risk)	>1000	1	0.329	0	0

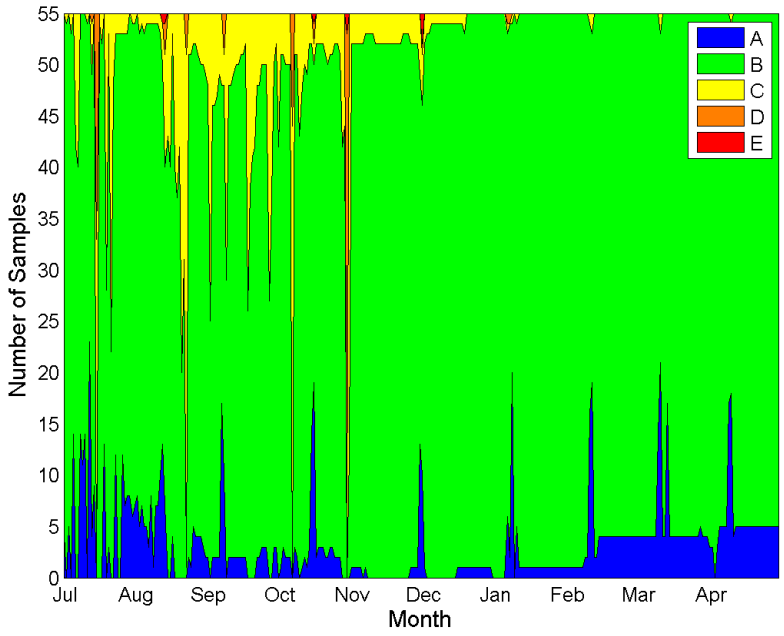

Figure 4.6: Graph showing the fraction of the monitoring tubewells in each WHO risk category

4.4. Discussion

In this work we developed a two-stage geostatistical estimation framework consisting of a latrine hydrological regression model that accounts for the effect of potential risk factors, and a space/time knowledge synthesis framework based on the BME method that integrates knowledge about the composite space/time variability and measurement errors associated with tubewell E. coli concentrations.

The latrine hydrological regression model investigates the effect of latrine density, population, and rainfall on E. coli contamination and reveals that all three are statistically significant risk factors (Table 4.2). The value of the hyperparameter

for each of these explanatory variables provides insight about the spatial and temporal scales of possible pathways of microbial contamination of the aquifer.

The value of latrine microbial range, r_{lv}, in LHM1 was 120 m which implies that a latrine located within that distance from a tubewell can still significantly affects the E. coli concentration in that tubewell. As shown in Figure 4.7, this fairly long travel distance might be explained by (a) indirect contamination from an intermediate pond, into which fecal matter is directly discharged from the latrine, (b) overland runoff, or (c) transportation by human. In addition, as shown in Figure 4.2(a), there is a small peak of R^2 around 15 m, even though the corresponding p-value is not statistically significant. This indicates the possibility of small scale direct underground transport from latrine to tubewell (Figure 4.7(d)), which should be investigated in future studies.

The optimal value of the population radius, r_{pv}, is 25 m (Table 4.1). This very short distance indicates that there is a very short pathway of contamination between people and tubewells. One possibility is that a higher usage of latrine may lead to failing latrines that directly contaminate the underlying aquifer through vertical subsurface transport. We found from our GPS survey that the average distance between households and their closest latrine is about 17.5 m on average in our study area, which added with the 15 m length scale identified for the direct underground transport from latrine to tubewell in Figure 4.7(d) is consistent with r_{pv}= 25 m. Hence our finding that the length scale of the microbial contamination from people to tubewells is 25 m provides additional evidence suggesting that latrines may be a potential source of contamination of the underlying shallow aquifer.

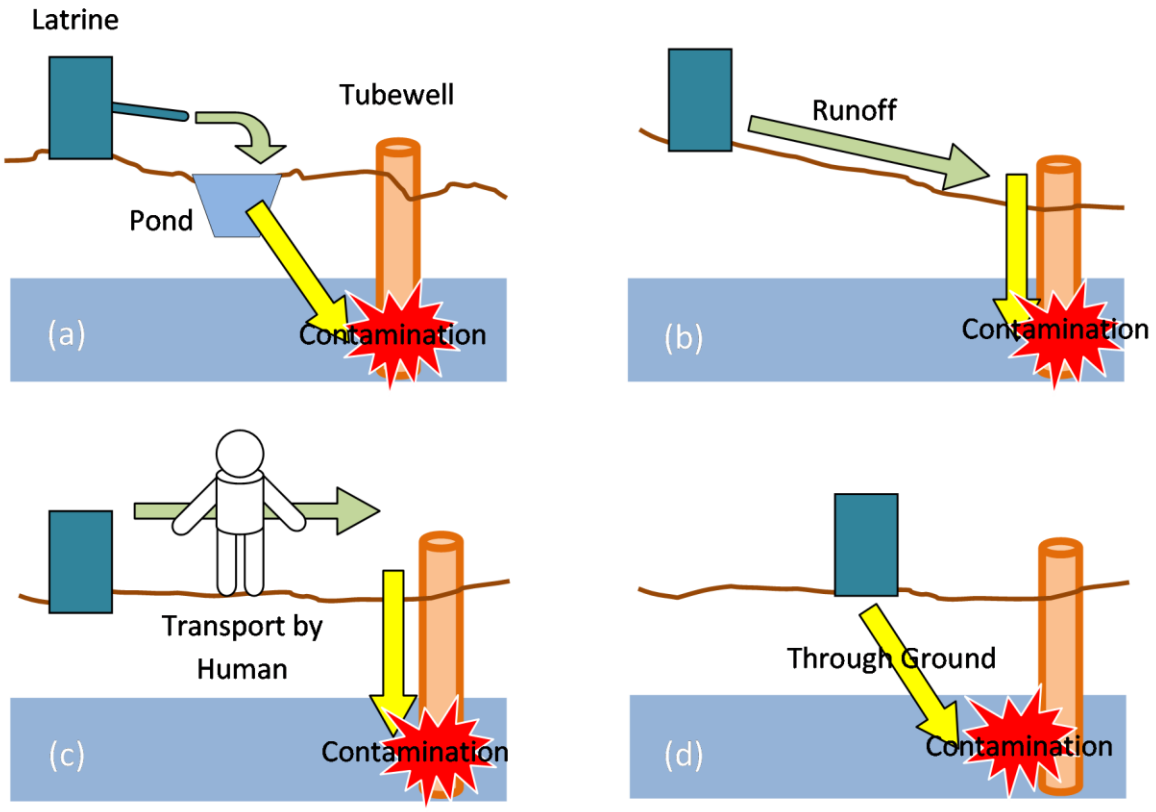

Figure 4.7: Effect of latrine on tubewell due to (a) indirect contamination from an intermediate pond, (b) overland runoff, (c) transportation by human, and (d) direct underground transport

The two hyperparameters, lag_1 and dur_1, for rain variable 1, rv_1, were 4 days and 1 day respectively. In other words, 1 day total rainfall observed 4 days prior to the sampling date leads to a statistically significant increase in the tubewell E. coli concentration. This result may imply that downward vertical gradient produced by a short rainfall may gradually carry pathogens from the ground to the aquifer. However, the hyperparameter dur_2 for the rain variable 2, rv_2, is 3 days for LHM2 and corresponding beta value is -1.05. Thus, if we have a rainfall in the preceding 3 days,

then the level of contamination is attenuated. Unlike the one day rainfall, this 3 consecutive day rainfall may indicate some washing out of microbes from the ground surface resulting in a protective effect as exhibited by the negative regression coefficient for the second rainfall variable. Overall these time scales observed for the effect of rainfall on tubewell contamination are quite short, suggesting that rainfall can rapidly lead to changes in microbial contamination of tubewells, and subsequently diarrhea for children coming into contact with the contaminated water either by drinking it from a tubewell, or playing in it if the water runoff from the rain is contaminated.

A space/time knowledge synthesis framework based on the BME theory was used to refine the LHM1 E. coli estimates by integrating knowledge about the composite space/time variability and measurement errors associated with observed tubewell E. coli concentrations. In particular the BME framework rigorously accounts for soft data modeling the uncertainty associated with the MPN estimation and with duplicate samples that were below or above detection limit. Our cross-validation analysis revealed that the estimation using soft data (methods 1-3 and 2-3 in Table 4.4) outperform the estimation methods relying on hard data (methods 1-2 and 2-2). The RMSE was 2.521, 2.368, and 1.906 for model1-1, 1-2, and 1-3, respectively (Table 4.4). Since the RMSE provides an assessment of the standard deviation of log FIB estimation errors (Coulliette et al. 2009), then there is a ratio of about (exp(2.521)-1)/(exp(2.368)-1)=1.18 in the estimation error for E. coli between methods 1-1 and 1-2, which means that accounting for the spatial autocorrelation among the E. coli concentrations can significantly improve the estimation accuracy

compared to using a regression based model alone. This ratio increases to about (exp(2.521)-1)/(exp(1.906)-1)=2 when comparing the estimation error for E. coli between methods 1-1 and 1-3, which indicates that the improvement in estimation accuracy is even more marked when BME is used to rigorously account for the measurement errors associated with MPN values. The maps (Figure 4.4) and temporal plots (Figure 4.5) created using the estimation method presented in this work provide a useful tool for exposure assessment in epidemiology studies of the effect of tubewell microbial contamination on Diarrhea as will be the case in future works for this study area. Additionally the maps created can help identify hot spots where risks of fecal contamination of tubewells may be high, which can be critical in designing effective public health intervention to minimize children exposure to fecally contaminated waters. Furthermore, we demonstrated that the estimation results can be used to summarize how frequently specific sites are at risk of contamination (Table 4.5) or how many monitoring tubewells are contaminated on a specific date (Figure 4.6), which can be useful for in an environmental management and policy context.

In this study, latrines were considered as the only source of fecal contamination. However, other surface water bodies, such as ponds and ditches, are also highly contaminated due to discharge from the latrines, and future works should also study their effect on tubewell contamination.

CHAPTER 5

Inter annual variability of community surveyed diarrheal disease among children from 2000 to 2002 in Matlab, Bangladesh

5.1. Background

As described previously, diarrheal disease remains a severe problem in most of the developing countries. Since most of the pathogenic organisms spread through fecal-oral transmission, an access to clean water is essential in order to prevent diarrheal disease. In rural Bangladesh contaminated tubewell water is considered as one of the primary routes of exposure to microbial pathogens. In previous chapters, the mechanism of microbial contamination in groundwater was investigated to help elucidate the mechanism of diarrheal disease outbreak. Rainfall, population density, and latrines were all found to be risk factors for the fecal contamination of shallow aquifers. However, the link between the occurrence of diarrheal disease and microbial contamination in groundwater has not been widely investigated. The overall goal of this study is, therefore, to advance our understanding of diarrheal disease etiology by analyzing the association between diarrheal disease events and microbial contamination in groundwater together with any possible extraneous factors that might influence disease occurrence and its interannual variability.

5.2. Material and Method

5.2.1. Study Area

The study was conducted in the Matlab field research area of the ICDDR, B. See previous chapter 3.2.1 for the description of Matlab field research area.

5.2.2. Demographic and health data

The health and demographic data were obtained from the Health and Demographic Surveillance System (HDSS) maintained by ICDDR, B. The HDSS is a longitudinal surveillance system that has demographic records of all Matlab residents since 1966. The data used in this study consist in a database of unspecified diarrhea incidence that occurred from 2000 to 2002 among children less than 5 years of age. ICDDR, B community health workers visited each household in Matlab monthly and recorded cases of diarrhea in the 24 hours prior to the community health worker's visit. Approximately 30000 cases were identified during the study period. Based on this data, the following binary variable was constructed for each month during the study period at each bari.

$$Y(\boldsymbol{p}) = \begin{cases} 1 & \text{if more than one child in the bari had diarrhea} \\ 0 & \text{otherwise} \end{cases} \qquad (5.1)$$

where $\boldsymbol{p} = (\boldsymbol{s}, t)$ is the space/time location, $\boldsymbol{s} = (s_1, s_2)$ is the spatial coordinate of bari, and t is the time expressed as a month during the study period.

5.2.3. Risk factors for diarrheal disease

Contaminated groundwater is one of our primary interests, even though there are many factors that could potentially govern the occurrence of diarrheal disease among children. Despite the many evidences of groundwater microbial contamination in Bangladesh, there is no microbial contamination data available for the community childhood diarrheal disease data we obtained. Thus, in order to investigate the effect of the groundwater microbial contamination on disease occurrence, we need a surrogate measure for microbial contamination. Local hydrogeological characteristics strongly affect the level of arsenic in the shallow aquifer in the Bengal Basin (Metral et al. 2008). Recent studies have also demonstrated that there is an inverse correlation between the level of dissolved arsenic and the rate of local groundwater recharge (Aziz et al. 2008). Hence, if the fate and transport of microbial pathogens is primarily driven by groundwater recharge, microbial contamination is expected to be inversely correlated with dissolved arsenic. Leber et al. (2007) found this inverse correlation using E. coli concentration data sampled in Arihazar, Bangladesh. Thus, low tubewell arsenic concentration was used as a surrogate measure for groundwater microbial contamination in this study. In addition, depth of the tubewell was also used as an indirect measure of microbial contamination, since microbial concentration decreases as function of depth in the subsurface because of microbial filtration.

Other key factors controlling the occurrence of diarrheal disease are rainfall and temperature. Several studies reported an association between diarrheal disease incidence and meteorological factors (Curriero et al. 2001; de Magny et al. 2007; Levy, Hubbard, and Eisenberg 2009; Pinfold, Horan, and Mara 1991). For example,

a study conducted in Dhaka, Bangladesh found that hospitalized non-cholera diarrheal disease was significantly associated with precipitation and temperature (Hashizume et al. 2007). In the previous chapter, we also demonstrated that the antecedent rainfall was significantly associated with tubewell E. coli concentrations in Bara Haldia, one of the villages in Matlab study area.

In addition, a study conducted in the same study area found that hospitalized diarrheal disease events were significantly associated with population density, socioeconomic status, and flood-control area (Ali et al. 2002; Emch 1999). Thus, all these factors were also considered as risk factors in this study.

5.2.4. Arsenic monitoring wells

We obtained arsenic concentration and well depth data observed at approximately 12,000 tubewells in over 6000 baris in 2002 and 2003. Since the arsenic concentration is expected remain constant over the time, it is reasonable to assume that the arsenic concentration data we obtained remained the same at each tubewell during our study period (Rahman et al. 2006). Depth and arsenic concentration were averaged in each bari. Map of arsenic concentration and well depth are shown in Figure 5.1(a) and (b).

5.2.5. Meteorological Data and Average Rainfall Variables

The daily rainfall data collected at several weather stations in Bangladesh were obtained from the Water Resources Planning Organization (WARPO). In addition, meteorological data were also downloaded from NOAA national data centers (NNDC) climate data online (http://www7.ncdc.noaa.gov/CDO/dataproduct).

Then the daily rainfall data obtained from these two sources were combined and used to calculate the average rainfall over aggregation periods (τ) ranging from 1 month to 12 months at each weather station. Since there was no weather station located in the study area, a geostatistical estimation framework based on the BME method was used to estimate rainfall over the study area based on the rainfall data available at surrounding weather stations. The details of the geostatistical method are described in Appendix F.

Temperature data were also downloaded from NOAA national data centers (NNDC) climate data online. Daily average temperatures were obtained by taking the average of all available observations over the country for each day of the study period. Then monthly average temperatures were calculated as a monthly average of the daily average temperatures.

Temporal plots of the estimated 2-month, 12-month, and 6-month rainfall averaged over the study area are shown in Figure 5.2 (a), (b) and (c), respectively. The temporal plot of temperature averaged over the country is also shown in Figure 5.2 (d).

5.2.6. Population Density Calculation

The population density variable at bari i ($pv_i = pv(\boldsymbol{p_i})$) was defined according to the following equation

$$pv_i = \frac{\sum_{j=1}^{l} n_j I_{ij}}{\sum_{j=1}^{l} a_j I_{ij}} \qquad (5.2)$$

$$I_{ij} = \begin{matrix} 1 & \text{if } d_{ij} \leq r_{\text{pv}} \\ 0 & \text{otherwise} \end{matrix} \quad (5.3)$$

where n_j is the number of people living in bari j, d_{ij} is the distance from the bari i to bari j, a_j is the area of bari j, and r_{pv} defines the radius around bari i. The population density variable was constructed using a hyperparameter r_{pv} value ranging from 5 m to 3000 m. Map of the population density with a short radius of $r_{pv} = 40$ m and a long radius of $r_{pv} = 2700$ m are shown in Figure 5.1 (e) and (f), respectively.

5.2.7. Socioeconomic status

A categorical socio-economic status (SES) score ranging from 1 to 5 was developed using factor analysis. The score was based on a composite of five binary variables representing ownership of household assets, two ordinal variables representing roof and wall material, and a categorical variable representing maternal education. An average of household SES scores was used as bari level SES score. Map of SES scores are shown in Figure 5.1 (d)

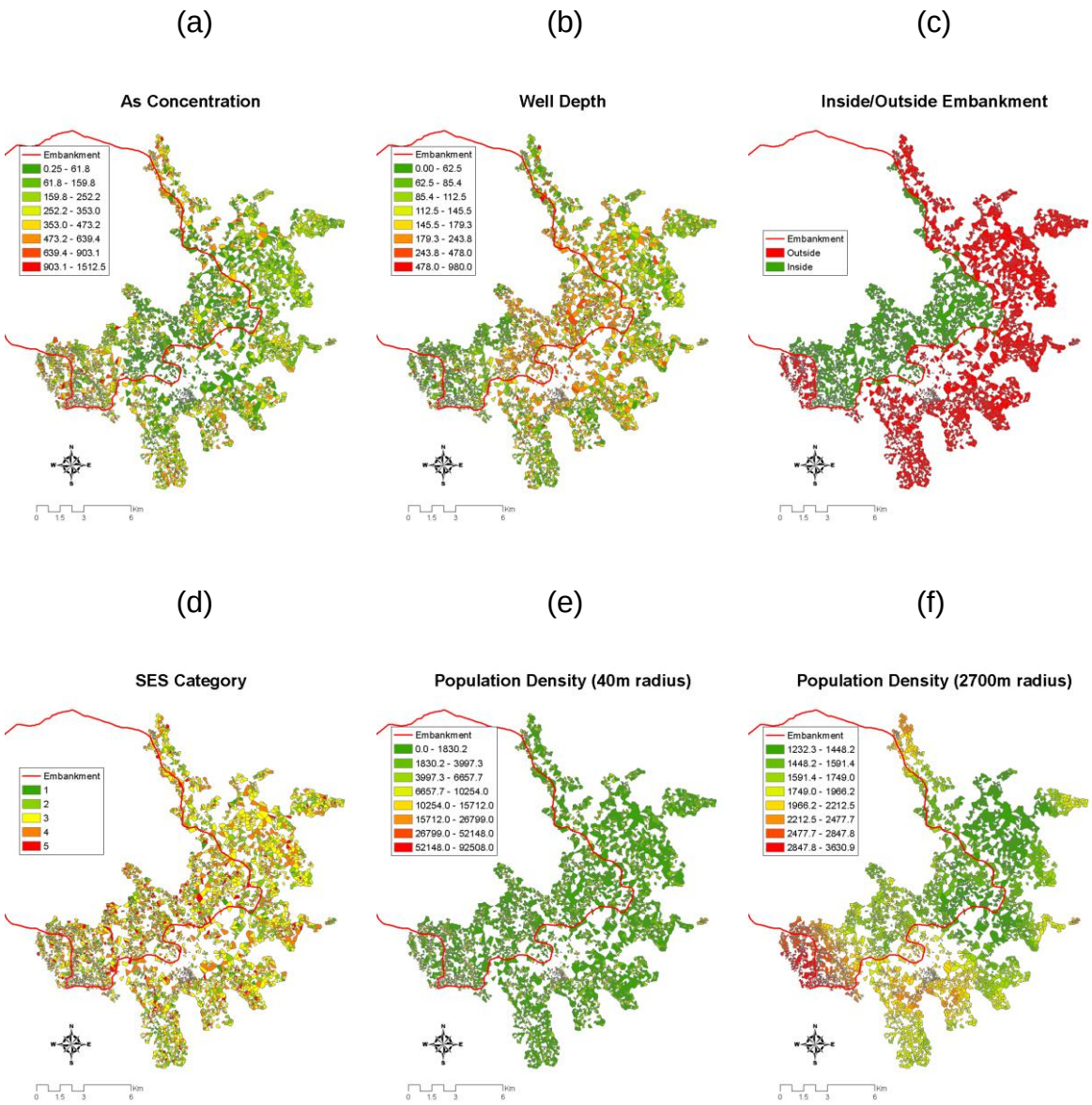

Figure 5.1: Map of explanatory variables: (a) arsenic concentration, (b) well depth, (c) inside/outside embankment, (d) SES score, (e) population density at $r_{pv} = 40$ m, and (f) population density at $r_{pv} = 2700$ m

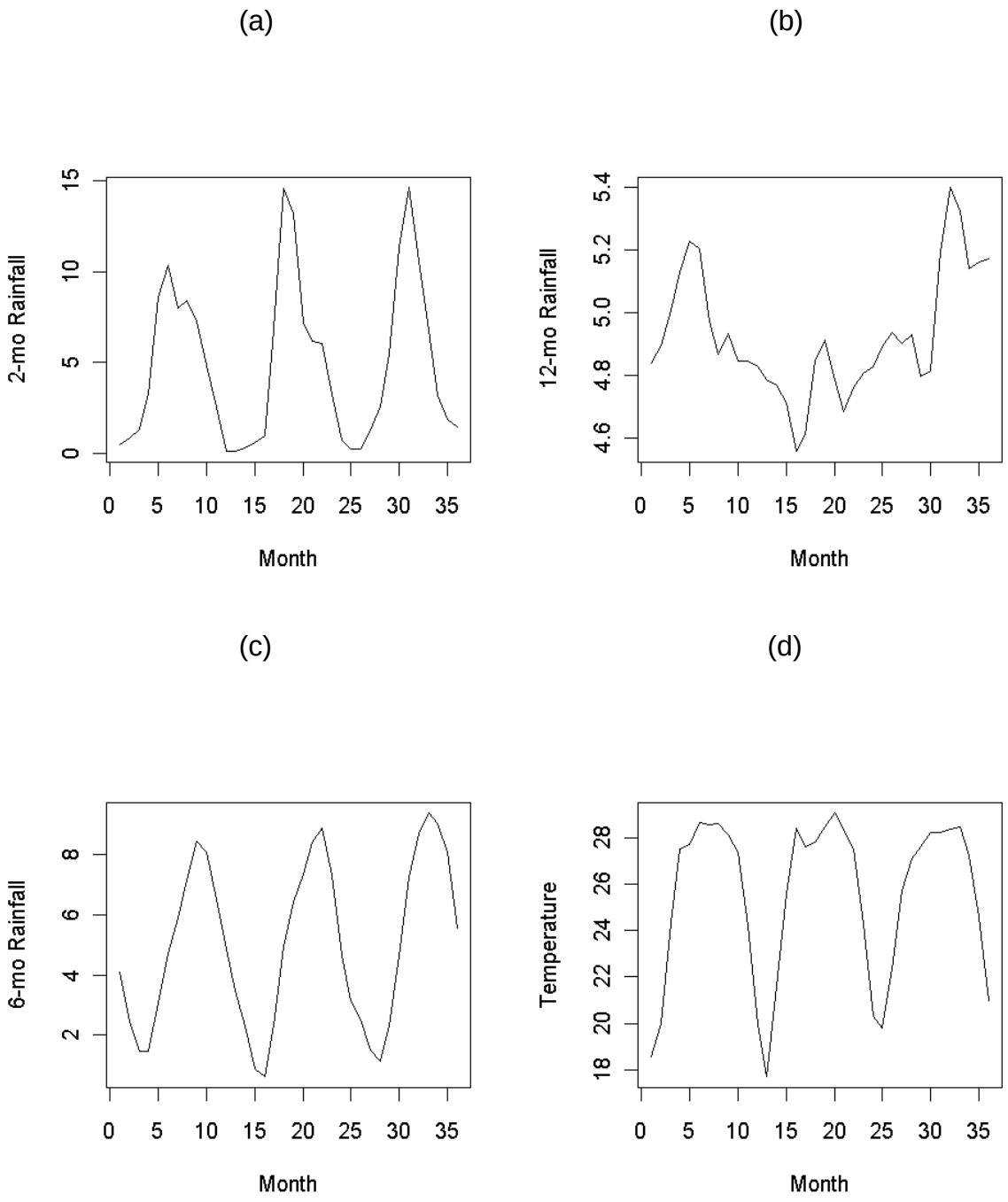

Figure 5.2: Temporal plot of (a) 2-month, (b) 12-month, and (c) 6-month average rainfall averaged over the study area, and (d) temperature averaged over the country.

5.2.8. Statistical Analysis

The relationship between childhood diarrheal disease events and the extraneous factors described above (i.e. arsenic concentration, well depth, rainfall, temperature, population density, socioeconomic status, and flood-control area) was analyzed using an univariate and a multivariate logistic regression model. The dependent variable of the model is the binary variable defined by Eq. (5.1). The data were also stratified by depth to investigate whether baris using shallow (depth<100ft) and deep (depth>100ft) tubewells have different disease etiology.

First, the relationship between childhood diarrhea and each explanatory variable was investigated using an univariate logistic regression model for each explanatory variable. Then the following multivariate logistic regression model was constructed.

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 \cdot \text{As Conc} + \beta_2 \cdot \text{Well Depth} + \beta_3 \cdot \text{Embank} + \beta_4 \cdot \text{SES} + \beta_5 \cdot \text{Rain1}(\tau_1) + \beta_6 \cdot pv1\left(r_{pv1}\right) \tag{5.4}$$

where P is the probability that the bari has at least one childhood diarrhea. In this model temperature was excluded due to the potential collinearity with rainfall variables. To explore the effect of the rainfall duration τ_1 and the population radius r_{pv1} the logistic regression Eq. (5.4) was performed multiple times so as to use all relevant combinations of these hyperparameters. The optimal values $\widehat{\tau_1}$ and $\widehat{r_{pv1}}$ for the hyperparameters τ_1 and r_{pv1} were selected based on the model with smallest

deviance subject to the constraint that all the corresponding regression coefficients must be statistically significant.

Once the best hyperparameter values $\widehat{\tau_1}$ and $\widehat{r_{pv1}}$ were selected, the effect of rainfall duration and the population radius was further investigated using the following multivariate regression model.

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 \cdot \text{As Conc} + \beta_2 \cdot \text{Well Depth} + \beta_3 \cdot \text{Embank} + \beta_4 \cdot \text{SES} + \beta_5 \cdot \text{Rain1}(\widehat{\tau_1}) + \beta_6 \cdot pv1\left(\widehat{r_{pv1}}\right) + \beta_7 \cdot \text{Rain2}(\tau_2) + \beta_8 \cdot pv2(r_{pv2}) \quad (5.5)$$

The values optimal values of the hyperparameters τ_2 and r_{pv2} were selected using the same procedure as that described above for τ_1 and r_{pv1}. All statistical analyses were conducted using R version 2.9.2.

5.3. Results

5.3.1. Univariate Logistic Regression Analysis

The linear regression coefficients and associated 95% confidence interval (CI) for the univariate logistic regression model corresponding to each explanatory variable is listed in Table 5.1. The corresponding results obtained when using baris with shallow or deep tubewells are also listed in Table 5.2 and Table 5.3, respectively.

Arsenic concentration has a protective effect (regression coefficient=-0.000260) and the odds ratio is 0.92 when the As concentration increases from the first As quantile (22.7 µg/L) to third As quantile (343.1 µg/L). The same effect was observed in baris using shallow tubewells (regression coefficient=-0.000157). In contrast, arsenic concentration was a risk factor in baris relying on deep tubewells (regression coefficient=0.000430).

Well depth, on the other hand, was a risk factor with an odds ratio of 1.006 for a 10 feet increase in depth. Well depth was also a risk factor in baris using shallow tubewells, whereas it had a protective effect in baris using deep tubewells.

For the following three explanatory variables, there was no difference between baris using shallow and deep tubewells. Baris located outside the embankment are more likely to have childhood diarrhea with an odds ratio of 1.67. Temperature is also a risk factor and the odds ratio is 1.31 for a 10°C increase in temperature. The socioeconomic status has a protective effect with an odds ratio of 0.68 if the SES score is increased one level.

The linear regression coefficient and associated 95% CI for the rainfall variable is shown as a function of rainfall duration in Figure 5.3. Rainfall over short aggregation periods (from 1-month to 5-month) is found to be a risk factor (positive regression coefficient). In contrast, rainfall over intermediate aggregation periods has a protective effect (negative regression coefficient), while rainfall over 12-month is a risk factor.

Figure 5.3 (b) shows how the linear regression coefficient and associated 95% CI for the population density variable changes as a function of population

radius. Similar to rainfall, the effect of population density is scale dependent: it is a risk factor when the population radius is short, whereas it is protective when the population radius is long.

Table 5.1: Regression coefficients and associated 95% confidence interval for the univariate logistic regression models

ALL	Regression Coef.	95% CI
As Concentration	-2.60E-04	(-3.36E-04, -1.83E-04)
Well Depth	6.15E-04	(4.33E-04, 7.97E-04)
Outside Embankment	5.12E-01	(4.79E-01, 5.44E-01)
Socioeconomic Status	-3.75E-01	(-3.93E-01, -3.58E-01)
Temperature	2.73E-02	(2.27E-02, 3.20E-02)

Table 5.2: Regression coefficients and associated 95% confidence intervals for shallow aquifer (depth < 100ft) based on the univariate logistic regression model

Shallow	Beta	95% CI
As Concentration	-1.57E-04	(-2.77E-04, -3.72E-05)
Well Depth	9.51E-03	(8.15E-03, 1.09E-02)
Outside Embankment	6.93E-01	(6.44E-01, 7.42E-01)
Socioeconomic Status	-3.75E-01	(-4.01E-01, -3.50E-01)
Temperature	2.61E-02	(1.93E-02, 3.30E-02)

Table 5.3: Regression coefficients and associated 95% confidence interval for deep aquifer (depth > 100ft) based on the univariate logistic regression model

Deep	Beta	95% CI
As Concentration	4.30E-04	(3.02E-04, 5.58E-04)
Well Depth	-1.62E-03	(-1.95E-03, -1.29E-03)
Outside Embankment	3.57E-01	(3.14E-01, 4.00E-01)
Socioeconomic Status	-4.22E-01	(-4.46E-01, -3.97E-01)
Temperature	2.84E-02	(2.21E-02, 3.47E-02)

(a) (b)

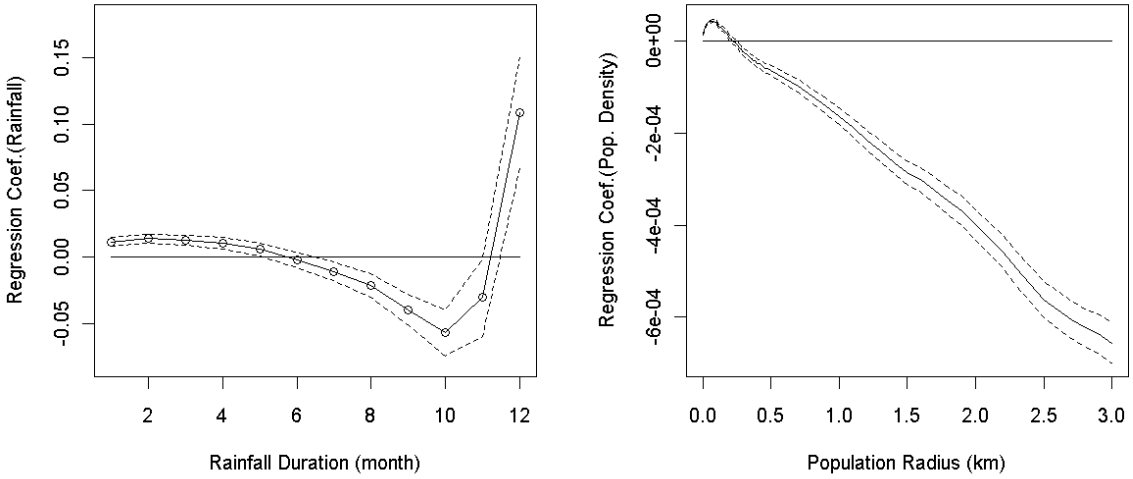

Figure 5.3: Regression coefficient and 95% confidence bound for (a) the rainfall variable as a function of rainfall duration and for (b) the population density variable as a function of population radius.

5.3.2. Multivariate Logistic Regression Analysis

$\tau_1 =$ 2-month and $r_{pv1} =$ 40 m were selected as the optimal hyperparameter values for the multivariate logistic regression model Eq. (5.4). Using these hyperparameter values, $\tau_2 =$ 12-month and r_{pv2}= 2700 m were selected as the optimal parameter values for the multivariate logistic regression model Eq. (5.5). The regression coefficients and associated 95% CI based on this model are listed in Table 5.4. In addition, the same tables for baris using shallow and deep tubewells are listed in Appendix G. In the multivariate model, arsenic concentration became a risk factor (regression coefficient=0.000172). Moreover in baris using shallow tubewells it is no longer a statistically significant variable. On the other hand, in baris using deep tubewells arsenic concentration is a statistically significant risk factor and has a stronger effect than that in univariate model.

Like in the univariate model, well depth is a risk factor in the multivariate model for baris using shallow tubewells, whereas it has a protective effect in baris using deep tubewells. For all the other explanatory variables, there are no differences between baris using shallow and deep tubewells. Baris located outside the embankment have a higher risk than those inside the embankment. The socioeconomic status has a protective effect and the odds ratio is 0.71 if the SES score is increased one level. Both the 2-month and 12-month rainfalls are risk factors and the odds ratio are 1.01 and 1.27 for a 1mm increase in the 2- and 12-month rainfall, respectively. The 40-meter-radius population density is a risk factor with an odds ratio of 1.006 for each 100 people/km^2 increase. In contrast, the 2700-

meter-radius population density has a protective effect with an odds ratio of 0.913 for each 100 people/km^2 increase.

Table 5.4: Regression coefficients and associated 95% confidence intervals based on the multivariate logistic regression model

	Beta	95% CI
As Concentration	1.72E-04	(7.88E-05, 2.65E-04)
Well Depth	3.65E-04	(1.27E-04, 6.01E-04)
Outside Embankment	5.38E-01	(5.04E-01, 5.72E-01)
SES	-3.33E-01	(-3.51E-01, -3.15E-01)
Rainfall 1 (2-mo)	1.07E-02	(7.13E-03, 1.43E-02)
Pop. Density 1	5.59E-05	(5.35E-05, 5.83E-05)
Rainfall 2 (12-mo)	2.38E-01	(1.94E-01, 2.81E-01)
Pop. Density 2	-9.05E-04	(-9.50E-04, -8.60E-04)

Thus far temperature was excluded from the logistic regression analysis due to the potential collinearity with rainfall variables. In order to evaluate the effect of temperature and other explanatory variables, temperature was added into the multivariate regression model Eq. (5.4) and (5.5). To evaluate the effect of collinearity among explanatory variables, the variance inflation factor (VIF) was calculated. The values of the hyperparameters τ and r_{pv} were selected using the same procedure as described earlier with the additional constraint that all the VIF must be smaller than five.

$\tau_1 =$ 6-month, $\tau_2 =$ 12-month, r_{pv1}= 40 m, and r_{pv2}=2700 m were selected as the optimal hyperparameter values. All selected hyperparameter values were the

same as the model without temperature, except for τ_1 due to the high correlation between 2-month daily rainfall and temperature (correlation coefficient=0.72). Regression coefficients and associated 95% CI based on this model are listed in Table 5.5. The same tables for baris using shallow and deep tubewells were also listed in Appendix H. The effect of each explanatory variable was generally the same as the model without temperature. Like in the univariate model, temperature is a risk factor with an odds ratio of 1.34 for a 10°C increase in temperature. Unlike the 2-month rainfall which was selected in the model without temperature, the 6-month average daily rainfall has a protective effect with an odds ratio of 0.978 for a 1mm increase in 6-month average daily rainfall.

Table 5.5: Regression coefficients and associated 95% confidence intervals based on the multivariate logistic regression model with temperature

	Beta	95% CI
As Concentration	1.74E-04	(8.06E-05, 2.67E-04)
Well Depth	3.60E-04	(1.23E-04, 5.98E-04)
Outside Embank	5.40E-01	(5.06E-01, 5.74E-01)
SES	-3.33E-01	(-3.51E-01, -3.15E-01)
Temperature	2.96E-02	(2.46E-02, 3.46E-02)
Rainfall 1	-2.22E-02	(-2.83E-02, -1.62E-02)
Pop. Density 1	5.60E-05	(5.36E-05, 5.84E-05)
Rainfall 2	2.83E-01	(2.39E-01, 3.27E-01)
Pop. Density 2	-9.08E-04	(-9.53E-04, -8.63E-05)

5.3.3. Estimated probability and disease rate map

The probability that a bari has at least one childhood diarrhea event is estimated using the following equation.

$$\hat{P} = \frac{1}{1 + \exp\left(-\left\{\widehat{\beta_0} + \widehat{\beta_1} \cdot \text{As Conc} + \cdots\right\}\right)} \qquad (5.6)$$

To visually verify the ability of the model to capture the interannual variability in diarrheal disease, the above estimated probability was calculated for each month of the study period and compared with the monthly rate of positive baris (i.e. baris with at least one reported case of childhood diarrhea). Plots of the observed rate of positive baris and of the average estimated probabilities based on (a) the multivariate logistic regression model Eq. (5.5) and on (b) the model Eq. (5.5) with temperature are shown in Figure 5.4.

(a) (b)

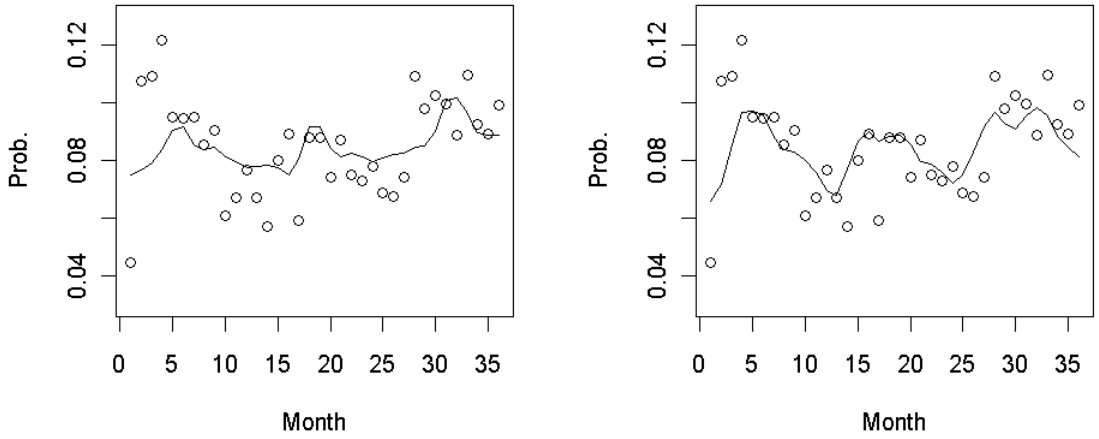

Figure 5.4: Plots of the observed rates of baris with childhood diarrhea (circles) and the corresponding probabilities (line) estimated using (a) the model without temperature and on (b) the model with temperature.

In this study, the binary health variable was constructed at each bari based on whether or not more than one child had a diarrhea event. Since about 80% of positive binary variables are based on only one childhood diarrhea event, the estimated probability is roughly equal to the probability of observing one case at the bari of interest. Thus the rate of childhood diarrhea can be approximated by the following equation.

$$\text{Rate of childhood diarrhea} = \frac{\hat{p}}{(\text{Bari Population}) \cdot (\text{Proportion of children})} \tag{5.7}$$

In rural Bangladesh, about 13% of the total population are children under five years of age. The estimated rate of childhood diarrhea averaged over the study period is shown in Figure 5.5. The average rate was obtained by calculating the average of estimated probability during the study period at each bari and then substituting $\hat{P}$ in Eq. (5.7) with the average estimated probability.

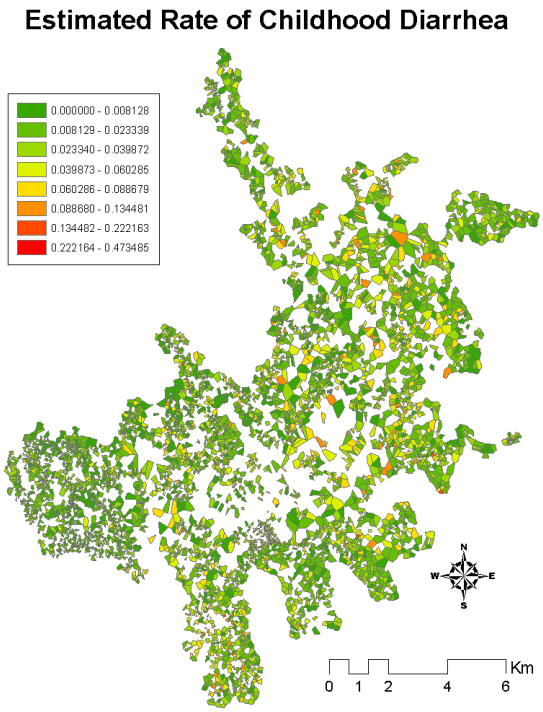

Figure 5.5: Estimated rate of childhood diarrhea averaged over the study period

5.4. Discussion

We find in this study that there is a significant association of non-specific childhood diarrhea with arsenic concentration, well depth, flood protection area, socioeconomic status, temperature, antecedent rainfall, and population density.

The effect of arsenic concentration is dependent on the well depth. In baris using shallow tubewells, arsenic concentration has a protective effect in the univariate model (Table 5.2). Because of the inverse correlation between the level of dissolved arsenic and microbial contamination through groundwater recharge at shallow depths, low arsenic in shallow tubewells is a proxy for microbial contamination, which is a potential risk factor for childhood diarrhea, thereby resulting in low arsenic being a risk factor for diarrhea at low depths for the univariate model. However, arsenic concentration is not a statistically significant predictor of childhood diarrhea in the multivariate logistic regression model (Appendix G and H). This can be explained by the fact that the inverse correlation between arsenic and microbial contamination is weak, and is easily compromised when controlling for other hydro geological variables. Indeed rainfall and population were found to be risk factors for microbial contamination in the previous chapter. Since rainfall and population density are included in the multivariate model, they may become a more significant proxy for microbial contamination than arsenic, and as a result low arsenic is “pushed out” of the model. Consequently arsenic is not a significant risk factor in the multivariate model.

In the deep aquifer arsenic is a risk factor for childhood diarrhea for both the univariate and multivariate models. This can be explained by the fact that low arsenic is not a proxy for microbial contamination in the deep aquifer, as hydraulic

recharge only plays a role in the inverse arsenic-microbial contamination relationship at shallow depths. Since there is no hypothesized causal relationship between arsenic and diarrhea, then arsenic in the deep aquifer must be a proxy for some other risk factors for diarrhea that is not controlled for in this study.

Well depth has a protective effect in baris using deep tubewells (Table 5.3). This might indicate that downward microbial filtration dominates microbial transport at high depths, resulting in lower microbial contamination at very high depth, and consequently lower diarrheal disease rates. On the other hand, in baris using shallow tubewells, well depth is found to be a risk factor for childhood diarrhea. Thus processes other than microbial filtration might come at play in the fate and transport of micro-organisms in shallow tubewells, such as bacterial regrowth or lateral groundwater transport originating from surface water bodies such as ponds or streams, which somehow might lead to ground water near the surface being less fecally contaminated than at intermediate depths. In any case we lack knowledge to explain exactly why well depth is a risk factor for diarrheal disease in shallow tubewells, and this therefore needs to be investigated in future works.

Baris located outside the embankment were found to be more likely to have childhood diarrhea. A study conducted in three flood-prone areas in Bangladesh found that tubewell E. coli contamination was significantly associated with a history of inundation (Luby et al. 2008). Thus, children living outside the embankment might be at higher risk of exposure to E. coli which results in higher rate of childhood diarrhea. However, the effect of the flood protection might vary geographically and it might not the same for different microbial pathogens. For instance, previous studies

have found that individuals living inside the embankment are at higher risk of cholera incidence (Ali et al. 2002; Emch 1999). In Bangladesh several microbial pathogens were confirmed as a cause of childhood diarrhea. Those include Vibrio cholerae, rotavirus, enterotoxigenic E. coli, and Shigella spp. (Albert et al. 1999). Thus future studies are needed to evaluate the effect of flood protection area on each microbial pathogen causing diarrheal disease.

Baris with lower socioeconomic status are found to be at a higher risk of childhood diarrhea, which is consistent with previous studies. The association between diarrheal disease and socioeconomic status was confirmed in many studies (Ackers et al. 1998). Higher socioeconomic status provides better ability to access clean water and good sanitation facilities or to maintain good hygiene practices, which are critical in reducing the risk of diarrheal disease. Thus children living in poor baris are more vulnerable to diarrheal disease.

High temperature increases the risk of childhood diarrhea, which is consistent with previous studies conducted in Dhaka, Bangladesh (Hashizume et al. 2007). This result indicates that drinking water and food might be more contaminated during hot month due to higher bacterial growth caused by high temperature. The effect of high temperature on diarrheal disease was also confirmed worldwide (Checkley et al. 2000; Lama et al. 2004; Singh et al. 2001). In bangladesh, higher temperature may also enhance potential bacterial regrowth of pathogens in tubewells.

As shown in Figure 5.3 (b), short-term (from 1-month to 5-month) rainfall are found to be a risk factor, whereas rainfall over intermediate time periods (from 7-month to 11-month) are protective. This might indicate that groundwater recharge

from short-term high rainfall may gradually carry pathogens from the ground to the aquifer, but continued rainfall might attenuate the groundwater contamination. However this attenuation effect might not be an important factor, since the multivariate model Eq. (5.5) reveals that the 2-month and 12-month rainfall are both positively associated with and have the largest effect on childhood diarrhea.

Population density also shows a strong scale dependence, albeit across space instead of time as was the case for rainfall. The population density over both a short and long radius significantly affects childhood diarrhea, since the 40-m and 2700m population densities were found to have the largest effect on childhood diarrhea in multivariate model Eq. (5.5). The population density over the short 40-m radius is a significant risk factor, whereas the population density over a large 2700-m radius has a significant protective effect. We have found in the previous chapter that the 25-m population variable has a statistically significant positive association with E. coli concentration, which was explained by the fact that people walk a short distance to use latrines, and a high usage of latrine may need to failing latrines and subsequent fecal contamination of the underlying aquifer. This therefore explains why the 40-m population density variable was found in this study to be a risk factor for diarrheal disease, as increased population density over this short radius may lead to increased latrine usage and subsequent increased fecal contamination of nearby tubewells, which is a potential risk factor for diarrhea. On the other hand the population density over the large 2700 m radius may be a proxy for entirely different effects, such as protective socio economic variables not fully captured by our coarse SES groupings, or as an indicator of the availability and maintenance of

infrastructures (e.g. sanitary, transport, deep wells, etc.) that have a protective effect on childhood diarrhea.

CHAPTER 6

Conclusion Remarks

Recent technological advance in geographic information systems have provided a data analysis framework to handle the large amount of spatial information usually associated with environmental epidemiological research. Many advanced functions are currently used for space/time data analysis in both exposure assessment and disease mapping. In exposure assessment, spatial interpolation techniques are widely used to account for the local scale variability in exposure. Geostatistical methods that account for the spatial autocorrelation amongst data points provide estimates of the value at unmonitored location together with standard errors that quantify the associated estimation uncertainty. Geostatistical methods generally outperform classical deterministic interpolation methods that do not provide any measure of uncertainty associated with prediction. Among geostatistical methods, the BME approach provides a modern mathematical framework that accounts for the composite space/time variability of the data and assimilates soft data with any type of distribution. Thus by applying the BME method, environmental monitoring data coming from multiple data sources distributed over space and time with various levels and types of uncertainty can be integrated into estimation process. Previous exposure assessment studies demonstrated that the BME method successfully reduced estimation error relative to conventional geostatistical

approaches, such as the classical kriging method derived as a linear limiting case of the more powerful BME method. Thus the BME method is an indispensable tool for environmental epidemiological research. However, there are several unaddressed implementation issues in its application to environmental and health studies. In this work, the BME approach was applied to an air and a water environmental epidemiologic study to take into account some of these unaddressed implementation issues.

Long-term exposure to ambient $PM_{2.5}$ concentration across the contiguous U.S. was modeled in chapter 2. A moving-window BME method was developed and numerically implemented to take into account the non-stationarity of long-term $PM_{2.5}$ concentrations and the uncertainty associated with long-term average concentrations calculated from incomplete daily measurements. A cross-validation analysis revealed that the moving-window BME method successfully improved mapping accuracy, resulting in a 18% reduction in MSE relative to a conventional kriging approach that assumed stationarity across the U.S. and relied only on the hard data available. Furthermore, a simulation study was conducted to investigate how the estimation performance changes as the ratio of soft to hard data increases. This simulation revealed that the improvement in estimation performance is even greater as more soft data become available.

In chapter 3 and 4, the space/time distribution of tubewell E. coli concentration was estimated using a geostatistical estimation framework. Chapter 3 provides the framework to calculate the MPN estimate of E. coli concentration from duplicate samples, and validates this framework for the E. coli data used in this

study. Chapter 3 also demonstrates that the spatial extent of microbial contamination in the shallow aquifer is strongly associated with antecedent rainfall. The covariance range of log-transformed E. coli concentration was found to be highly correlated with 13-day antecedent rainfall. The most likely explanation for this association is that one of the pathways for the contamination of tubewells includes downward subsurface microbial transport from high strength point source pollution on the ground to the underlying aquifer. Latrines constitute an important point source pollution in the study area, and they are therefore suspected to have an important effect on the spatial distribution of pathogens in the ground water.

In chapter 4 we develop and implement a two-stage geostatistical estimation framework consisting of a latrine hydrological regression model that accounts for the effect of potential risk factors including the density of latrines, and a space/time knowledge synthesis framework based on the BME method that integrates knowledge about the composite space/time variability and measurement errors associated with MPN E. coli concentrations. The latrine hydrological regression model investigates the effect of latrine density, population, and rainfall on E. coli contamination and reveals that all three are statistically significant risk factors. The value of the hyperparameter for each of these explanatory variables provides insight about the spatial and temporal scales of possible pathways of microbial contamination of the aquifer. The value of latrine microbial range is about 120 m, which implies that a latrine located within that distance from a tubewell can still significantly affects the E. coli concentration in that tubewell. We hypothesize that a latrine might have short length-scale effect on microbial contamination due to direct

microbial transport from latrine to tubewell, and a long length-scale effect which might be explained by indirect contamination through surface water body, overland runoff, or transportation by human.

In chapter 3 and 4, we confirm the frequent occurrence of microbial contamination of tubewells, especially in monsoon season. Downward vertical gradient produced by rainfall is hypothesized to be one of the driving forces of microbial transport from surface fecal contamination sources to the underlying shallow aquifer. In addition the level of contamination might be influenced by population density and location of the sources of contamination. Based on these findings, the potential links between diarrheal disease and microbial tubewell contamination as well as other potential risk factors are investigated in chapter 5. The health data used in this analysis consisted in non-specific community childhood diarrhea data collected in Matlab from 2000 to 2002. Arsenic concentration was used as a surrogate measure for microbial contamination in tubewells. Logistic regression analysis revealed that arsenic concentration, well depth, flood protection area, socioeconomic status, temperature, antecedent rainfall, and population density were all statistically significant predictor of non-specific childhood diarrhea. Overall in this work, the geostatistical framework based on the BME method was applied to a long-term $PM_{2.5}$ exposure assessment over the U.S. and the space/time modeling of microbial contamination in a shallow aquifer in Matlab, Bangladesh. The window-based BME method implemented for the $PM_{2.5}$ study was also used to estimate rainfall for each baris included in the study of the inter-annual variability of diarrheal disease. As described above, the implementations of the BME method that

were developed in this work were found to perform better than a conventional geostatistical approach that do not account for data uncertainty and relies on exact measurement (hard data). As a result the implementations of the BME method developed in this work provide highly informative maps to estimate the exposure to an air and water pollutants.

Appendix A: Cross-validation statistics

Following three tables shows cross validation statistics obtained from (1) the BME hard only analysis assuming nationwide stationarity (first column), (2) moving-window BME hard only analysis (second column), and (3) moving-window BME soft data analysis (third column) based on three different covariance model: Best fit covariance model (Table A.1), Gaussian covariance model (Table A.2), and spherical covariance model (Table A.3).

Table A.1: Cross validation statistics obtained by the following three methods based on best fit covariance model: method (1) the BME hard data only analysis assuming stationarity across the U.S. (first column), method (2) moving-window BME hard data only analysis (second column), and method (3) moving-window BME soft data analysis

Method	(1)	(2)	(3)
MSE	3.050	2.302	2.100
APE	-0.001	0.103	0.090
ASPE	-0.005	0.056	0.05
ASE	2.300	1.429	1.392
RMSS	0.756	1.739	1.743
Pearson's Corr.	0.847	0.887	0.897
Spearman's Rank Corr	0.863	0.887	0.895

Table A.2: Cross validation statistics obtained by the following three methods based on Gaussian covariance model: method (1) the BME hard data only analysis assuming stationarity across the U.S. (first column), method (2) moving-window BME hard data only analysis (second column), and method (3) moving-window BME soft data analysis

Method	(1)	(2)	(3)
MSE	3.442	2.480	2.281
APE	-0.036	0.072	0.059
ASPE	-0.017	0.039	0.035
ASE	2.412	1.456	1.421
RMSS	0.768	1.831	1.839
Pearson's Corr.	0.824	0.878	0.888
Spearman's Rank Corr	0.847	0.879	0.886

Table A.3: Cross validation statistics obtained by the following three methods based on Spherical covariance model: method (1) the BME hard data only analysis assuming stationarity across the U.S. (first column), method (2) moving-window BME hard data only analysis (second column), and method (3) moving-window BME soft data analysis

Method	(1)	(2)	(3)
MSE	3.05	2.231	2.066
APE	-0.001	0.097	0.083
ASPE	-0.005	0.037	0.033
ASE	2.3	1.544	1.514
RMSS	0.756	1.075	1.072
Pearson's Corr.	0.847	0.89	0.899
Spearman's Rank Corr	0.863	0.891	0.898

Appendix B: Shape of Powered Exponential Covariance Model

Figure B.1 shows shape of the powered exponential model defined as eq. (3.8) with sill $C_1 = 1.0$, spatial range $a_r = 5.0$, and several different power parameter b.

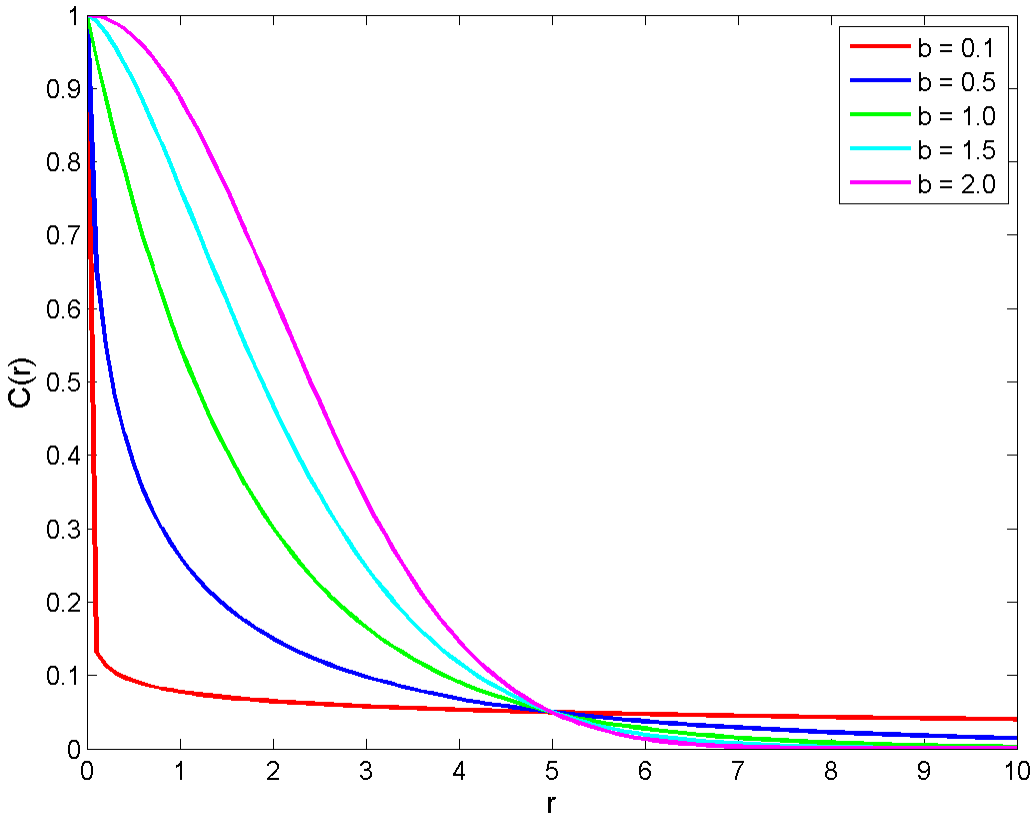

Figure B.1: Shape of the powered exponential model with sill $C_1 = 1.0$, spatial range $a_r = 5.0$, and several different power parameter b.

Appendix C: Rainfall and covariance range based on the levels of WHO classification scheme

The association between rainfall and the covariance range of the levels of WHO classification scheme was investigated. One is assigned to category A and five to category E. Covariance parameters (Table C.1) show the same trend as covariance range obtained for log-transformed concentration, with a long range in the monsoon season and shorter range in the post-monsoon season. As shown in Figure C.**1**, 9-day antecedent rainfall showed the strongest correlation ($r = 0.9$, $p = 0.05$).

Table C.1: Covariance parameters during the study period

Month	Covariance Range (m)	Shape Parameter	Sill
May	0.636	0.421	0.856
June	124	0.657	0.509
July	235	0.863	0.533
August	93.1	0.343	1.08
September	61.6	2	0.806
October	67.6	0.507	0.757
Nov/Dec	60.7	0.400	1.16
January	11.3	1.08	0.609
February	44.5	1.99	0.382
March	30.9	1.24	0.274
April	7.65	1.99	0.29

(a) (b)

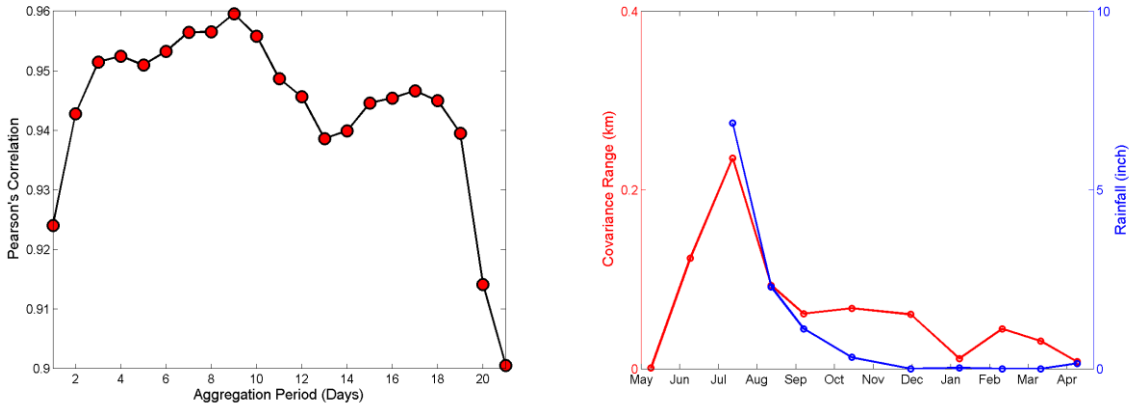

Figure C.1: (a) Pearson's correlation coefficient and (b) Temporal plot of covariance range and 9-days antecedent rainfall

Appendix D: Details of MPN Calculation

D.1. Basic assumptions of MPN method

The most probable number (MPN) is one of the most commonly used procedures to enumerate the microbial organisms in a water sample. The MPN is a maximum likelihood point estimator that maximizes the probability of observing the result of dilution series analysis. There are two basic assumptions for the MPN calculation. One is that the water sample is completely mixed, so that the organisms in the sample are randomly distributed (well-mixed assumption). The other is that each organism exhibits growth when incubated in the culture medium.

D.2. MPN for a single dilution set

The probability that a sample of volume V will contain N organisms is given by the Poisson distribution described by the following equation.

$$P(x = N) = \frac{(\mu V)^N}{N!}\exp(-\mu V) \quad \text{(D.1)}$$

where μ is the density of the organisms in the water that was sampled. Therefore the probability of not observing any organism, P_s, is given by

$$P_s = P(x = 0) = \exp(-\mu V) \quad \text{(D.2)}$$

and the probability observing one or more organisms, P_f,is given by

$$P_f = 1 - P(x = 0) = 1 - P_s = 1 - \exp(-\mu V) \quad \text{(D.3)}$$

In practice a water sample is subdivided into a set of n sub-samples of equal volume V each. We refer to this as a single dilution set because all sub-samples have the same volume. Under the well-mixed assumption, the organisms are randomly distributed across the sub-samples so that the density μ of organism is constant across sub-samples, and as a result the probability of observing s fertile sub-samples is given by the binomial distribution.

$$P(s|\mu) = \frac{n!}{s!\,(n-s)!} {P_f}^s \left(1 - P_f\right)^{n-s} \quad \text{(D.4)}$$

Substituting $P_f = 1 - \exp(-\mu V)$ into equation above, we obtain the following likelihood function.

$$P(s|\mu) = \frac{n!}{s!\,(n-s)!} (1 - \exp(-\mu V))^s (\exp(-\mu V))^{n-s} = L(\mu|s) \quad \text{(D.5)}$$

The most probable number (MPN) is defined as the maximum likelihood point estimator of the density μ that maximizes the likelihood function.

$$\text{MPN} = \underset{\mu}{\text{argmax}} \left(\frac{n!}{s!\,(n-s)!} (1 - \exp(-\mu V))^s (\exp(-\mu V))^{n-s} \right) \quad \text{(D.6)}$$

In other words, the MPN is the density of the organisms that maximizes the probability of obtaining the observed result, s fertile sub-samples out of n total sub-samples making up the single dilution set. The MPN is obtained by finding the root of the derivative of the log likelihood function.

$$\begin{aligned} l(\mu) &= \log\big(L(\mu|s)\big) = \log\left(\frac{n!}{s!\,(n-s)!}(1-\exp(-\mu V))^{s}(\exp(-\mu V))^{n-s}\right) \\ &= \log\left(\frac{n!}{s!\,(n-s)!}\right) + s\log(1-\exp(-\mu V)) + (n-s)(-\mu V) \end{aligned} \tag{D.7}$$

$$\frac{\partial l}{\partial \mu} = s\frac{V\exp(-\mu V)}{1-\exp(-\mu V)} - (n-s)V \tag{D.8}$$

Thus, the MPN is given by the following equation

$$\frac{sV}{1-\exp(-\mu V)} = nV \tag{D.9}$$

$$\therefore \mathrm{MPN} = -\frac{1}{V}\log\frac{n-s}{n} \tag{D.10}$$

D.3. MPN for multiple dilution sets

If all the sub-samples of a single dilution set are fertile or sterile, the MPN cannot be uniquely determined. Therefore, multiple dilution sets are usually used to estimate the density. Assume that the water sample is divided into r dilution sets, where the *i*-th dilution set is itself subdivided into n_i sub-samples of volume V_i. Under

the well mixed assumption, the joint probability of observing s_i $(i = 1 \cdots r)$ fertile sub-samples out of n_i sub-samples of volume V_i is given by the following likelihood function.

$$L(\mu|s_1, s_2, \cdots s_r) = \prod_{i=1}^{r} \frac{n_i!}{s_i!\,(n_i - s_i)!} (1 - \exp(-\mu V_i))^{s_i} (\exp(-\mu V_i))^{n_i - s_i} \qquad \text{(D.11)}$$

Therefore, the MPN is given by the density μ that maximizes the following likelihood function.

$$\text{MPN} = \underset{\mu}{\text{argmax}} \left(\prod_{i=1}^{r} \frac{n_i!}{s_i!\,(n_i - s_i)!} (1 - \exp(-\mu V_i))^{s_i} (\exp(-\mu V_i))^{n_i - s_i} \right)$$

The MPN is given by the root of the following equation.

$$\sum_{i=1}^{r} \frac{s_i V_i}{1 - \exp(-\mu V_i)} = \sum_{i=1}^{r} n_i V_i \qquad \text{(D.12)}$$

Since the probability distribution of the density μ is approximated by the log normal distribution, the standard error of the MPN is given by (Hurley, and Roscoe 1983)

$$SD_{\log(\text{MPN})} = \left(\text{MPN}^2 \sum_{i=1}^{r} \frac{{V_i}^2 n_i}{\exp(V_i \text{MPN} - 1)} \right)^{-1/2} \qquad \text{(D.13)}$$

Thus, the 95% confidence bound of the MPN is obtained by

$$\text{95\% Confidence Bound} = \exp(\log(\text{MPN}) \pm 1.96 \times SD_{\log(\text{MPN})}) \quad \text{(D.14)}$$

D.4. MPN for IDEXX Quanti-Tray®/2000

The IDEXX Quanti-Tray®/2000 with Colilert® reagent is a semi-automated method based on specific enzymatic activity to enumerate the density of coliforms, E. coli, and enterococci using the MPN theory (Rompre et al. 2002). A Quanti-Tray®/2000 is a tray consisting of two dilution sets. The first dilution set consists in 49 large wells of volume 1.86ml, and the second dilution set consists of 48 small wells of volume 0.186ml. When a water sample is collected in the field, that water sample is poured into a tray consisting of these two dilution sets. Thus, using eq. D.12, the MPN for the density of organisms in that water sample is given by the root of the following equation.

$$\frac{s_1 V_1}{1 - \exp(-\mu V_1)} + \frac{s_2 V_2}{1 - \exp(-\mu V_2)} = n_1 V_1 + n_2 V_2 \quad \text{(D.15)}$$

where s_1 and s_2 are the numbers of positive large and small wells, $V_1 = 1.86$ml and $V_2 = 0.186$ml are the volume of large and small wells, and $n_1 = 49$ and $n_2 = 48$ are the total number of large and small wells, respectively

Oftentimes, duplicate samples of the same water are collected in the field. Each sample is poured in its own Quanti-Tray®/2000. The MPN theory can be

applicable to enumerating the density of organisms in these duplicate trays if the well-mixed assumption holds, i.e. if the density of organisms is the same in each tray. Under that assumption, the MPN for the density is given by eq. D.12 as the root of the following equation

$$\frac{s_1 V_1}{1-\exp(-\mu V_1)} + \frac{s_2 V_2}{1-\exp(-\mu V_2)} + \frac{s_3 V_1}{1-\exp(-\mu V_1)} + \frac{s_4 V_2}{1-\exp(-\mu V_2)} = n_1 V_1 + n_2 V_2 + n_1 V_1 + n_2 V_2 \quad \text{(D.16)}$$

where (s_1, s_2) and (s_3, s_4) are the number of positive large and small wells in the first and second tray, respectively. This equation can be re-arranged as

$$\frac{(s_1 + s_3) V_1}{1-\exp(-\mu V_1)} + \frac{(s_2 + s_4) V_2}{1-\exp(-\mu V_2)} = 2n_1 V_1 + 2n_2 V_2 \quad \text{(D.17)}$$

which simply states that the MPN for duplicate trays is obtained by assuming that the sum of the number of positive wells are observed on double-sized trays.

The extension to the case of *k* replicate samples of the same water analyzed with the Quanti-Tray®/2000 (i.e. *k*=2 for duplicate samples, *k*=3 for triplicate samples, etc.) is straightforward and leads to the following equation to calculate the MPN

$$\frac{\left(\sum_{j=1}^{k} s_{1,j}\right) V_1}{1-\exp(-\mu V_1)} + \frac{\left(\sum_{j=1}^{k} s_{2,j}\right) V_2}{1-\exp(-\mu V_2)} = k n_1 V_1 + k n_2 V_2 \quad \text{(D.18)}$$

where $s_{1,j}$ and $s_{2,j}$ are the numbers of positive large and small wells of the *j*-th replicate sample.

D.5. Likelihood Ratio Test

We can test the well-mixed assumption using the likelihood ratio test. For a single water sample poured in a multiple dilution set, the null hypothesis is that all dilution sets have a common density μ and the alternative hypothesis is that each dilution set has a distinct density μ_i. As described above, the null likelihood function can be written as

$$L^0 = \prod_{i=1}^{r} \frac{n_i!}{s_i!\,(n_i - s_i)!} (1 - \exp(-\mu V_i))^{s_i} (\exp(-\mu V_i))^{n_i - s_i} \qquad \text{(D.19)}$$

Similarly, the alternative likelihood function can be given by

$$L^A = \prod_{i=1}^{r} \frac{n_i!}{s_i!\,(n_i - s_i)!} (1 - \exp(-\mu_i V_i))^{s_i} (\exp(-\mu_i V_i))^{n_i - s_i} \qquad \text{(D.20)}$$

Thus, the likelihood ratio is defined as the following

$$\lambda = \frac{L^0}{L^A} \qquad \text{(D.21)}$$

For large samples, $-2\log\lambda$ has approximately a chi-squared distribution with $r-1$ degrees of freedom.

$$-2\log\lambda = 2\sum_{i=1}^{r}\left[V_i(\mu-\mu_i)(n_i-s_i) - s_i\log\left(\frac{1-\exp(-\mu V_i)}{1-\exp(-\mu_i V_i)}\right)\right] \sim \chi^2(r-1) \qquad \text{(D.22)}$$

Therefore, the null hypothesis is rejected if $-2\log\lambda$ exceeds the $1-\alpha$ percentile of the chi-squared distribution, where α is the significance level of the statistical test. Usually the significance level is set to a small number, e.g. α=0.05, and the test is performed at that significance level for each sample analyzed in the laboratory. We then compare the rate at which samples are rejected (i.e., for which the test rejects the null hypothesis) at the significance level α. If the rejection rate is comparable with the significance level (e.g. if about 5% of the samples are rejected by the test when α=0.05), then the test is performing as expected under the null hypothesis. On the other hand, if the rejection rate is much larger than α, then the null hypothesis may not hold (e.g. samples may not have been well mixed), and the laboratory procedures need to be inspected.

For replicate field samples collected from the same water into different sampling containers (e.g. duplicate or triplicate samples), the null hypothesis is that all samples have a common density μ, and the alternative hypothesis is that each sample has a distinct density μ_j ,*j*=1…*k*, where *k* is the number of replicate samples. The difference in density μ_i reflects the possibility of sampling error, i.e. that there may be a different density of organisms in each replicate sample because of

problems with the sampling procedure, or due to contamination of the sample between sample collection and sample analysis. Similar to the single sample test, the test statistic for replicate samples is given by the following equation.

$$-2\log\lambda = 2\sum_{j=1}^{k}\sum_{i=1}^{r}\left[V_{ij}\left(\mu-\mu_j\right)\left(n_{ij}-s_{ij}\right) - s_{ij}\log\left(\frac{1-\exp(-\mu V_{ij})}{1-\exp(-\mu_i V_{ij})}\right)\right] \sim \chi^2(k-1) \tag{D.23}$$

where k is the number of replicate samples, while *r* is the number of dilution series for a given sample. For large samples, this test statistic has approximately a chi-squared distribution with $k-1$ degrees of freedom. Therefore, as explained above, the null hypothesis is rejected if $-2\log\lambda$ for this statistic exceeds the $1-\alpha$ percentile of the chi-squared distribution.

Appendix E: Space/Time Estimation Map of E. coli Concentration

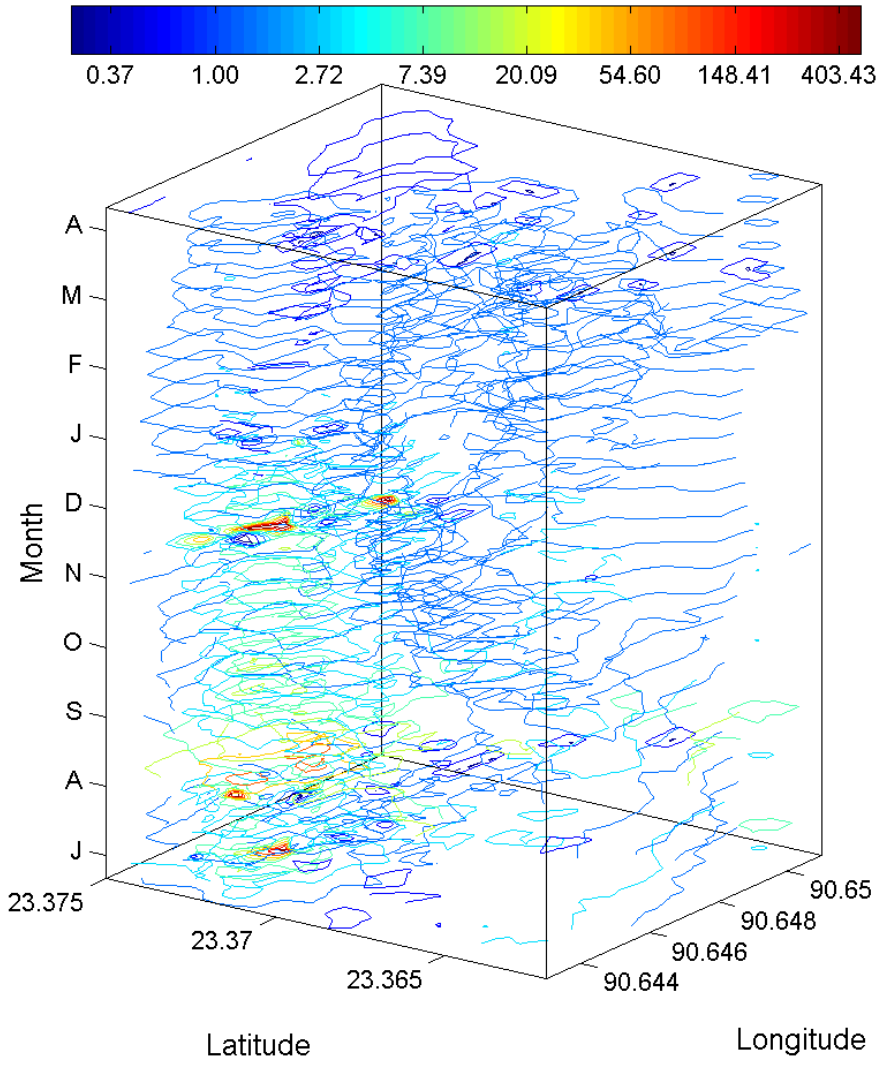

Figure E.1: A Series of contour maps of E. coli concentration during the study period estimated by LHM1 and BME estimation with hard/soft data

Appendix F: Rainfall Variable

F.1. Background

The space/time geostatistical estimation framework based on the Bayesian Maximum Entropy (BME) method was implemented to estimate the 1- to 12-month rainfall across the study area. First the mean trend of the average daily rainfall was estimated using the method introduced by Kyriakidis et al. (2004) and then removed from the average daily rainfall data. To estimate the 1- to 12-month rainfall at unmonitored study sites, the residual of the 1- to 12-month rainfall data were further processed using the BME method with a soft data model accounting for the number of observations used to aggregate daily rainfall values in the 1- to 12-month rainfall data.

F.2. Data Acquisition and cleaning

Daily rainfall data were obtained from the Water Resources Planning Organization (WARPO). Additional daily rainfall data were obtained from NOAA national data centers (NNDC) climate data online (http://www7.ncdc.noaa.gov/CDO/dataproduct) and combined with the WARPO data. Outliers and unreliable daily rainfall records were eliminated from the data based on the following criteria.

1. Daily rainfall records with values greater than 1835mm (world records of daily precipitation)
2. Daily rainfall records with values smaller than 0mm

3. Daily rainfall records obtained at a weather station where all daily rainfall records were 0mm
4. More than 365 days consecutive 0mm daily rainfall records

F.3. Mean trend model and Residual field

In order to model the distribution of the average daily rainfall over space and time, we employed a space/time geostatistical estimation framework based on the BME theory. In this framework, the theory of space/time random field (S/TRF) is employed to model the daily rainfall (Christakos 1992). Let $Y(\boldsymbol{p})$ be the S/TRF modeling the distribution of the average daily rainfall at space/time point $\boldsymbol{p}$. This variable was constructed at the last day of each month at each weather station and described as the following equation.

$$Y(\boldsymbol{p}_i) = \frac{1}{ma}\sum_{j=1}^{ma} b_{t-j+1}$$

where $\boldsymbol{p} = (\boldsymbol{s}, t)$ is the space/time location, $\boldsymbol{s} = (s_1, s_2)$ is the spatial coordinate, t is time, m is the number of the days in 1 month (=30 days), $a = 1 \cdots 12$ is the aggregation period in months, and b_k is the daily rainfall on kth day. A space/time deterministic mean trend of the average daily rainfall $m_Y(\boldsymbol{p})$ is modeled by the following linear regression model.

$$m_Y(\boldsymbol{p}) = \beta_0 + \beta_1 \bar{Y}$$

where $\bar{Y}$ is the spatial average of all daily rainfalls. Thus, the S/TRF of the residual average daily rainfall $X(\boldsymbol{p})$ is described as

$$X(\boldsymbol{p}) = Y(\boldsymbol{p}) - m_Y(\boldsymbol{p})$$

F.4. The BME Estimation

The space/time empirical covariance was calculated using the method-of-moments estimator (Cressie 1993), then used to fit the parameters of a powered exponential model (Banerjee 2004). All model parameters were estimated by an automated weighted least square procedure (Jian, Olea, and Yu 1996; Olea 2006). The aggregated rainfall data were considered as unreliable, and treated as a soft data with Gaussian distribution, if there was a month with less than five daily measurements during the aggregation period.

A 21-by-21 estimation grid of points was constructed over the study area for each month during the study period and the residual daily average rainfall was estimated using the BME method. The deterministic mean trend at each estimation grid point, $\widehat{m_Y}$, was obtained using the spatially interpolated the regression coefficients, $\widehat{\beta_0}$ and $\widehat{\beta_1}$.

$$\widehat{m_Y} = \widehat{\beta_0} + \widehat{\beta_1}\bar{Y}$$

Thus the daily average rainfall at each estimation grid point is given by

$$\hat{Y} = \hat{X} + \widehat{m_Y} = \hat{X} + \widehat{\beta_0} + \widehat{\beta_1}\bar{Y}$$

where $\hat{X}$ is the BME mean estimate of residual daily average rainfall. Finally, estimated the aggregated rainfall at estimation grid points was spatially interpolated at each bari location. The geostatistical estimation was conducted using MATLAB R2008a (MathWorks Inc.).

Appendix G: Multivariate Logistic Regression Model-1

Regression coefficients and associated 95% confidence interval (CI) based on multivariate logistic regression model given by Eq. (5.5)

Table G.1: Regression coefficients and associated 95% confidence interval based on the multivariate logistic regression model for baris using shallow tubewells

	Beta	95% CI
As Concentration		
Well Depth	3.61E-03	(2.23E-03, 4.99E-03)
Outside Embankment	6.04E-01	(5.52E-01, 6.56E-01)
SES	-2.88E-01	(-3.15E-01, -2.61E-01)
Rainfall 1	1.02E-02	(4.86E-03, 1.55E-02)
Pop. Density 1	5.94E-05	(5.54E-05, 6.33E-05)
Rainfall 2	2.01E-01	(1.38E-01, 2.64E-01)
Pop. Density 2	-8.97E-04	(-9.56E-04, -8.38E-04)

Table G.2: Regression coefficients and associated 95% confidence intervals based on the multivariate logistic regression model for baris using deep tubewells

	Beta	95% CI
As Concentration	5.32E-04	(3.85E-04, 6.80E-04)
Well Depth	-5.05E-04	(-8.52E-04, -1.58E-04)
Outside Embankment	4.62E-01	(4.14E-01, 5.10E-01)
SES	-3.78E-01	(-4.04E-01, -3.53E-01)
Rainfall 1	1.13E-02	(6.45E-03, 1.62E-02)
Pop. Density 1	5.14E-05	(4.83E-05, 5.45E-05)
Rainfall 2	2.69E-01	(2.06E-01, 3.32E-01)
Pop. Density 2	-7.76E-04	(-8.53E-04, -7.00E-04)

Appendix H: Multivariate Logistic Regression Model-2

Regression coefficient and associated 95% confidence interval (CI) based on the multivariate logistic regression model given by Eq. (5.5) with temperature as additional explanatory variable.

Table H.1: Regression coefficients and associated 95% confidence intervals based on multivariate logistic regression model for baris using shallow tubewells

	Beta	95% CI
As Concentration		
Well Depth	3.61E-03	(2.23E-03, 4.99E-03)
Outside Embank	6.05E-01	(5.53E-01, 6.57E-01)
SES	-2.88E-01	(-3.15E-01, -2.61E-01)
Temperature	2.79E-02	(2.05E-02, 3.52E-02)
Rainfall 1	-1.84E-02	(-2.74E-02, -9.31E-03)
Pop. Density 1	5.95E-05	(5.55E-05, 6.34E-05)
Rainfall 2	2.39E-01	(1.75E-01, 3.03E-01)
Pop. Density 2	-8.97E-04	(-9.56E-04, -8.38E-04)

Table H.2: Regression coefficients and associated 95% confidence intervals based on multivariate logistic regression model for baris using deep tubewells

	Beta	95% CI
As Concentration	5.39E-04	(3.92E-04, 6.87E-04)
Well Depth	-5.01E-04	(-8.48E-04, -1.54E-04)
Outside Embank	4.66E-01	(4.18E-01, 5.13E-01)
SES	-3.79E-01	(-4.04E-01, -3.53E-01)
Temperature	3.12E-02	(2.45E-02, 3.79E-02)
Rainfall 1	-2.50E-02	(-3.32E-02, -1.67E-02)
Pop. Density 1	5.15E-05	(4.84E-05, 5.46E-05)
Rainfall 2	3.20E-01	(2.57E-01, 3.84E-01)
Pop. Density 2	-7.84E-04	(-8.61E-04, -7.07E-04)

Reference

Ackers, M. L., R. E. Quick, C. J. Drasbek, L. Hutwagner, and R. V. Tauxe. 1998. "Are there national risk factors for epidemic cholera? The correlation between socioeconomic and demographic indices and cholera incidence in Latin America." *International Journal of Epidemiology* 27(2): 330-34.

Akita, Y., G. Carter, and M. L. Serre. 2007. "Spatiotemporal nonattainment assessment of surface water tetrachloroethylene in New Jersey." *Journal of Environmental Quality* 36(2): 508-20.

Albert, M. J., A. S. G. Faruque, S. M. Faruque, R. B. Sack, and D. Mahalanabis. 1999. "Case-control study of enteropathogens associated with childhood diarrhea in Dhaka, Bangladesh." *Journal of Clinical Microbiology* 37(11): 3458-64.

Ali, M., M. Emch, J. P. Donnay, M. Yunus, and R. B. Sack. 2002. "Identifying environmental risk factors for endemic cholera: a raster GIS approach." *Health Place* 8(3): 201-10.

Ashbolt, N. J. 2004. "Microbial contamination of drinking water and disease outcomes in developing regions." *Toxicology* 198(1-3): 229-38.

Aziz, Z., A. van Geen, M. Stute, R. Versteeg, A. Horneman, Y. Zheng, S. Goodbred, M. Steckler, B. Weinman, I. Gavrieli, M. A. Hoque, M. Shamsudduha, and K. M. Ahmed. 2008. "Impact of local recharge on arsenic concentrations in shallow aquifers inferred from the electromagnetic conductivity of soils in Araihazar, Bangladesh." *Water Resources Research* 44(7).

Banerjee, S. 2004. *Hierarchical modeling and analysis for spatial data*. Boca Raton: Chapman & Hall.

Barrell, R. A. E. and M. G. M. Rowland. 1979. "The relationship between rainfall and well water pollution in a West African (Gambian) village." *Journal of Hygiene* 83(1): 143-50.

Beelen, R., G. Hoek, P. A. van den Brandt, R. A. Goldbohm, P. Fischer, L. J. Schouten, B. Armstrong, and B. Brunekreef. 2008. "Long-term exposure to traffic-related air pollution and lung cancer risk." *Epidemiology* 19(5): 702-10.

Bell, M. L., F. Dominici, K. Ebisu, S. L. Zeger, and J. M. Samet. 2007. "Spatial and temporal variation in PM2.5 chemical composition in the United States for health effects studies." *Environmental Health Perspectives* 115(7): 989-95.

Black, R. E., K. H. Brown, S. Becker, A. R. Alim, and I. Huq. 1982. "Longitudinal studies of infectious diseases and physical growth of children in rural Bangladesh. II.

Incidence of diarrhea and association with known pathogens." *Am J Epidemiol* 115(3): 315-24.

Black, R. E., M. H. Merson, I. Huq, A. R. Alim, and M. Yunus. 1981. "Incidence and severity of rotavirus and Escherichia coli diarrhoea in rural Bangladesh. Implications for vaccine development." *Lancet* 1(8212): 141-3.

Boldo, E., S. Medina, A. LeTertre, F. Hurley, H. G. Mucke, F. Ballester, I. Aguilera, D. Eilstein, and G. Apheis. 2006. "Apheis: Health impact assessment of long-term exposure to PM2.5 in 23 European cities." *European Journal of Epidemiology* 21(6): 449-58.

Brauer, M., G. Hoek, P. van Vliet, K. Meliefste, P. Fischer, U. Gehring, J. Heinrich, J. Cyrys, T. Bellander, M. Lewne, and B. Brunekreef. 2003. "Estimating long-term average particulate air pollution concentrations: Application of traffic indicators and geographic information systems." *Epidemiology* 14(2): 228-39.

Checkley, W., L. D. Epstein, R. H. Gilman, D. Figueroa, R. I. Cama, J. A. Patz, and R. E. Black. 2000. "Effects of El Nino and ambient temperature on hospital admissions for diarrhoeal diseases in Peruvian children." *Lancet* 355(9202): 442-50.

Chen, L. C., M. Rahman, and A. M. Sarder. 1980. "Epidemiology and causes of death among children in a rural area of Bangladesh." *Int J Epidemiol* 9(1): 25-33.

Christakos, G. 1990. "A Bayesian Maximum-entropy View To The Spatial Estimation Problem." *Mathematical Geology* 22(7): 763-77.

Christakos, G. 1992. *Random field models in earth sciences*. San Diego: Academic Press.

Christakos, G. 2000. *Modern spatiotemporal geostatistics*. Oxford ;New York: Oxford University Press.

Christakos, G., P. Bogaert, and M. L. Serre. 2001. *Temporal GIS : advanced functions for field-based applications*. Berlin ;New York: Springer.

Christakos, G. and X. Y. Li. 1998. "Bayesian maximum entropy analysis and mapping: A farewell to kriging estimators?" *Mathematical Geology* 30(4): 435-62.

Cochran, W. G. 1950. "Estimation of bacterial densities by means of the "most probable number"." *Biometrics* 6(2): 105-16.

Coulliette, A. D., E. S. Money, M. L. Serre, and R. T. Noble. 2009. "Space/Time Analysis of Fecal Pollution and Rainfall in an Eastern North Carolina Estuary." *Environmental Science & Technology* 43(10): 3728-35.

Cressie, N. A. C. 1993. *Statistics for spatial data*. New York: J. Wiley.

Curriero, F., J. Patz, J. Rose, and S. Lele. 2001. “Analysis of the association between extreme precipitation and waterborne disease outbreaks in the United States, 1948-1994.” *Epidemiology* 12(4): S37-S37.

Curriero, F. C., M. E. Hohn, A. M. Liebhold, and S. R. Lele. 2002. “A statistical evaluation of non-ergodic variogram estimators.” *Environmental and Ecological Statistics* 9(1): 89-110.

de Magny, G. C., J. F. Guegan, M. Petit, and B. Cazelles. 2007. “Regional-scale climate-variability synchrony of cholera epidemics in West Africa.” *Bmc Infectious Diseases* 7.

Eftim, S. E., J. M. Samet, H. Janes, A. McDermott, and F. Dominici. 2008. “Fine particulate matter and mortality - A comparison of the six cities and American Cancer Society cohorts with a medicare cohort.” *Epidemiology* 19(2): 209-16.

Elliott, P. and D. Wartenberg. 2004. “Spatial epidemiology: Current approaches and future challenges.” *Environmental Health Perspectives* 112(9): 998-1006.

Emch, M. 1999. “Diarrheal disease risk in Matlab, Bangladesh.” *Soc Sci Med* 49(4): 519-30.

Emch, M., M. Ali, and M. Yunus. 2008. “Risk areas and neighborhood-level risk factors for Shigella dysenteriae 1 and Shigella flexneri.” *Health & Place* 14(1): 96-105.

Fuentes, M. 2003. “Statistical assessment of geographic areas of compliance with air quality standards.” *Journal of Geophysical Research-Atmospheres* 108(D24).

Godfrey, S., F. Timo, and M. Smith. 2006. “Microbiological risk assessment and management of shallow groundwater sources in Lichinga, Mozambique.” *Water and Environment Journal* 20(3): 194-202.

Haas, C. N. 1999. *Quantitative microbial risk assessment*. New York: Wiley.

Haas, T. C. 1990. “Lognormal And Moving Window Methods Of Estimating Acid Deposition.” *Journal of the American Statistical Association* 85(412): 950-63.

Haas, T. C. 1995. “Local prediction of a spatio-temporal process with an application to wet sulfate deposition.” *Journal of the American Statistical Association* 90(432): 1189-99.

Hashizume, M., B. Armstrong, S. Hajat, Y. Wagatsuma, A. S. G. Faruque, T. Hayashi, and D. A. Sack. 2007. “Association between climate variability and hospital

visits for non-cholera diarrhoea in Bangladesh: effects and vulnerable groups." *International Journal of Epidemiology* 36(5): 1030-37.

Henderson, S. B., B. Beckerman, M. Jerrett, and M. Brauer. 2007. "Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter." *Environmental Science & Technology* 41(7): 2422-28.

Hoque, B. A., K. Hallman, J. Levy, H. Bouis, N. Ali, F. Khan, S. Khanam, M. Kabir, S. Hossain, and M. S. Alam. 2006. "Rural drinking water at supply and household levels: Quality and management." *International Journal of Hygiene and Environmental Health* 209(5): 451-60.

Howard, G., S. Pedley, M. Barrett, M. Nalubega, and K. Johal. 2003. "Risk factors contributing to microbiological contamination of shallow groundwater in Kampala, Uganda." *Water Research* 37(14): 3421-29.

Hurley, M. A. and M. E. Roscoe. 1983. "Automated statistical analysis of microbial enumeration by dilution series." *Journal of Applied Bacteriology* 55(1): 159-64.

Islam, M. S., A. Siddika, M. N. H. Khan, M. M. Goldar, M. A. Sadique, A. Kabir, A. Huq, and R. R. Colwell. 2001. "Microbiological analysis of tube-well water in a rural area of Bangladesh." *Applied and Environmental Microbiology* 67(7): 3328-30.

Jarup, L. 2004. "Health and environment information systems for exposure and disease mapping, and risk assessment." *Environmental Health Perspectives* 112(9): 995-97.

Jerrett, M., R. T. Burnett, R. J. Ma, C. A. Pope, D. Krewski, K. B. Newbold, G. Thurston, Y. L. Shi, N. Finkelstein, E. E. Calle, and M. J. Thun. 2005. "Spatial analysis of air pollution and mortality in Los Angeles." *Epidemiology* 16(6): 727-36.

Jian, X. D., R. A. Olea, and Y. S. Yu. 1996. "Semivariogram modeling by weighted least squares." *Computers & Geosciences* 22(4): 387-97.

Kosek, M., C. Bern, and R. L. Guerrant. 2003. "The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000." *Bulletin of the World Health Organization* 81(3): 197-204.

Kunzli, N., M. Jerrett, W. J. Mack, B. Beckerman, L. LaBree, F. Gilliland, D. Thomas, J. Peters, and H. N. Hodis. 2005. "Ambient air pollution and atherosclerosis in Los Angeles." *Environmental Health Perspectives* 113(2): 201-06.

Kyriakidis, P. C., N. L. Miller, and J. Kim. 2004. "A spatial time series framework for simulating daily precipitation at regional scales." *Journal of Hydrology* 297(1-4): 236-55.

Lama, J. R., C. R. Seas, R. Leon-Barua, E. Gotuzzo, and R. B. Sack. 2004. "Environmental temperature, cholera, and acute diarrhoea in adults in Lima, Peru." *Journal of Health Population and Nutrition* 22(4): 399-403.

Leber, J. 2007. "The Extent and Variability of Microbial Contamination of Shallow Tube Wells in Two Villages in Bangladesh." New York: Columbia University.

Levine, R. J., M. R. Khan, S. Dsouza, and D. R. Nalin. 1976. "Failure of sanitary wells to protect against cholera and other diarrheas in Bangladesh." *Lancet* 2(7976): 86-89.

Levy, K., A. E. Hubbard, and J. N. S. Eisenberg. 2009. "Seasonality of rotavirus disease in the tropics: a systematic review and meta-analysis." *International Journal of Epidemiology* 38(6): 1487-96.

Liao, D., D. J. Peuquet, Y. Duan, E. A. Whitsel, J. Dou, R. L. Smith, H. M. Lin, J. C. Chen, and G. Heiss. 2006. "GIS approaches for the estimation of residential-level ambient PM concentrations." *Environ Health Perspect* 114(9): 1374-80.

Logan, B. E., D. G. Jewett, R. G. Arnold, E. J. Bouwer, and C. R. Omelia. 1995. "CLARIFICATION OF CLEAN-BED FILTRATION MODELS." *Journal of Environmental Engineering-Asce* 121(12): 869-73.

Luby, S., M. S. Islam, and R. Johnston. 2006. "Chlorine spot treatment of flooded tube wells, an efficacy trial." *Journal of Applied Microbiology* 100(5): 1154-58.

Luby, S. P., S. K. Gupta, M. A. Sheikh, R. B. Johnston, P. K. Ram, and M. S. Islam. 2008. "Tubewell water quality and predictors of contamination in three flood-prone areas in Bangladesh." *J Appl Microbiol* 105(4): 1002-8.

Macler, B. A. and J. C. Merkle. 2000. "Current knowledge on groundwater microbial pathogens and their control." *Hydrogeology Journal* 8(1): 29-40.

Melian, R., N. Myrlian, A. Gouriev, C. Moraru, and F. Radstake. 1999. "Groundwater quality and rural drinking-water supplies in the Republic of Moldova." *Hydrogeology Journal* 7(2): 188-96.

Metral, J., L. Charlet, S. Bureau, S. B. Mallik, S. Chakraborty, K. M. Ahmed, M. W. Rahman, Z. Q. Cheng, and A. van Geen. 2008. "Comparison of dissolved and particulate arsenic distributions in shallow aquifers of Chakdaha, India, and Araihazar, Bangladesh." *Geochemical Transactions* 8.

Miller, K. A., D. S. Siscovick, L. Sheppard, K. Shepherd, J. H. Sullivan, G. L. Anderson, and J. D. Kaufman. 2007. "Long-term exposure to air pollution and

incidence of cardiovascular events in women." *New England Journal of Medicine* 356(5): 447-58.

NIPORT. 2005. *Bangladesh demographic and health survey 2004*. Dhaka; Calverton, Maryland: National Institute of Population Research and Training (NIPORT) : Mitra and Associates; ORC Macro.

Nuckols, J. R., M. H. Ward, and L. Jarup. 2004. "Using geographic information systems for exposure assessment in environmental epidemiology studies." *Environmental Health Perspectives* 112(9): 1007-15.

Olea, R. A. 2006. "A six-step practical approach to semivariogram modeling." *Stochastic Environmental Research and Risk Assessment* 20(5): 307-18.

Pedley, S. and G. Howard. 1997. "The public health implications of microbiological contamination of groundwater." *Quarterly Journal of Engineering Geology* 30: 179-88.

Pinfold, J. V., N. J. Horan, and D. D. Mara. 1991. "SEASONAL EFFECTS ON THE REPORTED INCIDENCE OF ACUTE DIARRHEAL DISEASE IN NORTHEAST THAILAND." *International Journal of Epidemiology* 20(3): 777-86.

Pope, C. A., R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, and G. D. Thurston. 2002. "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution." *Jama-Journal of the American Medical Association* 287(9): 1132-41.

Pope, C. A., M. Ezzati, and D. W. Dockery. 2009. "Fine-Particulate Air Pollution and Life Expectancy in the United States." *New England Journal of Medicine* 360(4): 376-86.

Pruss, A., D. Kay, L. Fewtrell, and J. Bartram. 2002. "Estimating the burden of disease from water, sanitation, and hygiene at a global level." *Environ Health Perspect* 110(5): 537-42.

Puangthongthub, S., S. Wangwongwatana, R. M. Kamens, and M. L. Serre. 2007. "Modeling the space/time distribution of particulate matter in Thailand and optimizing its monitoring network." *Atmospheric Environment* 41(36): 7788-805.

Rahman, M., M. Vahter, N. Sohel, M. Yunus, M. A. Wahed, P. K. Streatfield, E. C. Ekstrom, and L. A. Persson. 2006. "Arsenic exposure and age- and sex-specific risk for skin lesions: A population-based case-referent study in Bangladesh." *Environmental Health Perspectives* 114(12): 1847-52.

Rompre, A., P. Servais, J. Baudart, M. R. de-Roubin, and P. Laurent. 2002. “Detection and enumeration of coliforms in drinking water: current methods and emerging approaches.” *J Microbiol Methods* 49(1): 31-54.

Serre, M. L. and G. Christakos. 1999. “Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge - the Equus Beds study.” *Stochastic Environmental Research and Risk Assessment* 13(1-2): 1-26.

Singh, R. B. K., S. Hales, N. de Wet, R. Raj, M. Hearnden, and P. Weinstein. 2001. “The influence of climate variation and change on diarrheal disease in the Pacific Islands.” *Environmental Health Perspectives* 109(2): 155-59.

Smith, A. H., E. O. Lingas, and M. Rahman. 2000. “Contamination of drinking-water by arsenic in Bangladesh: a public health emergency.” *Bull World Health Organ* 78(9): 1093-103.

Stute, M., Y. Zheng, P. Schlosser, A. Horneman, R. K. Dhar, S. Datta, M. A. Hoque, A. A. Seddique, M. Shamsudduha, K. M. Ahmed, and A. van Geen. 2007. “Hydrological control of As concentrations in Bangladesh groundwater.” *Water Resources Research* 43(9).

Taylor, R., A. Cronin, S. Pedley, J. Barker, and T. Atkinson. 2004. “The implications of groundwater velocity variations on microbial transport and wellhead protection - review of field evidence.” *Fems Microbiology Ecology* 49(1): 17-26.

UNICEF. 2009. *Diarrhoea : why children are still dying and what can be done*. Geneva: World Health Organization.

U.S. EPA. 2008. AQS Data Coding Manual v2.33. Research Triangle Park, NC: U.S. Environmental Protection Agency.

U.S. EPA. 2009. Air Quality System. Research Triangle Park, NC: U.S. Environmental Protection Agency. Available: http://www.epa.gov/ttn/airs/airsaqs/ [accessed 6 January 2009]

Vine, M. F., D. Degnan, and C. Hanchette. 1997. “Geographic information systems: their use in environmental epidemiologic research.” *Environ Health Perspect* 105(6): 598-605.

WHO. 1997. *Guidelines for drinking-water quality Volume 3: Surveillance and Control of Community Supplies*. Geneva: World Health Organization.

Wright, R. C. 1986. “The seasonality of bacterial quality of water in a tropical developing country (Sierra Leone).” *Journal of Hygiene* 96(1): 75-82.

Yu, H. L., J. C. Chen, G. Christakos, and M. Jerrett. 2009. “BME Estimation of Residential Exposure to Ambient PM10 and Ozone at Multiple Time Scales.” *Environmental Health Perspectives* 117(4): 537-44.