GEOSTATISTICAL ESTIMATION OF WATER QUALITY USING RIVER AND FLOW COVARIANCE MODELS

Prahlad Jat

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.

Chapel Hill
2016

Approved by:

Marc L. Serre

Gregory W. Characklis

Jacqueline MacDonald Gibson

William Vizuete

Lawrence E. Band

ABSTRACT

Prahlad Jat: Geostatistical Estimation of Water Quality using River and Flow Covariance Models
(Under the direction of Marc L. Serre)

Assessing water quality along rivers is vital for watershed management and to protect the public health. Monitoring water quality at every river mile is logistically impractical and prohibitively expensive. Geostatistical estimation offers a cost effective alternative that can be rapidly implemented to statistically model spatially dependent water quality parameters using the available monitoring data. Geostatistical modeling requires a covariance model to describe the variability and autocorrelation of the water quality along rivers. Three main classes of covariance models, namely the Euclidean, river, and flow-weighted covariance models, are commonly used in geostatistical water quality estimation.

In the first study we use a river covariance model to successfully characterize the space/time variability of chloride, an emerging contaminant, along rivers in Maryland. This method leads to a 24% reduction in mean square estimation error compared to the Euclidean method. In the next two studies we use the flow-weighted covariance for the estimation of fecal coliform (FC), and Dissolved Organic Carbon (DOC), respectively. Surprisingly, very few geostatistical water quality studies have successfully implemented the flow-weighted covariance model and improved estimation accuracy. To address this critical gap, we introduce the first implementation of a flow weighted covariance model that uses gradual flow, and we then use this model in a novel hybrid Euclidean/Gradual-flow covariance model to estimate FC in the Haw and Deep rivers in North Carolina, and DOC in three sub-basins in Maryland. Our novel

hybrid Euclidean/Gradual-flow covariance model captures variability coming from both terrestrial sources and hydrological transport, and it leads to a 12% and 15% reduction in mean square error for FC and DOC, respectively, compared to the traditional Euclidean covariance. This novel hybrid covariance model is widely applicable to any other study area and to other water quality parameters.

To my mentor and family. I couldn't have done this without you.

Thank you for all of your support along the way.

ACKNOWLEDGEMENTS

First and foremost I would like to express my deepest gratitude to my advisor, Marc Serre, for his guidance, inspiration and support during my tenure as a PhD student. I feel extremely privileged to work with Marc and I have learned a great deal from him, above and beyond his contributions to my research. He has always been willing to delve into the details of the work and talk through issues, sometimes long past a reasonable meeting length. I am very thankful to all of my doctoral committee members: Profs Gregory W. Characklis, Jacqueline MacDonald Gibson, William Vizuete, and Lawrence E. Band. Each of them offered great insight and suggestions that have enhanced this body of work.

Last, but not least, I would like to thank my loving family and friends for their encouragement and support throughout this process. Finally, I would like to acknowledge my colleagues in BMElab and in the Department of Environmental Sciences and Engineering for the Environment for sharing the ups and downs and providing constant motivations during the long journey of my PhD study

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 : INTRODUCTION

1. Literature review of geostatistical estimation of river water quality

Surface water quality is an essential component of the natural environment. Characterizing the surface water quality is often a daunting task, but it is an important one in verifying whether the observed water quality is suitable for its intended purpose and to meet the requirements of Section 305(b) of the Clean Water Act (1972). Monitoring water quality helps to determine trends and patterns in the water affected by the release of contaminants or due to other natural and anthropogenic activities. However, high monitoring costs limit the implementation of exhaustive water quality monitoring programs and therefore probability-based water quality surveys are typically needed to do the water quality assessment needed to meet the Clean Water Act requirements (Peterson et al., 2006). EPA's National Water Quality Inventory Report 2004 (EPA, 2009) stated that about 44% of streams, 64% of lakes, and 30% of estuaries assessed were not clean enough to meet the intended purposes in spite of the progress in cleaning up the nation's water.

Geostatistical modeling provides a convenient way to model spatially dependent observations. Typically, a geostatistical analysis assumes that nearby measurements are more strongly related than measurements observed far apart, as is the concept of the First Law of Geography (Tobler, 1970). Implementation of geostatistical models for analyzing spatially correlated data is well documented in the literature (Goovaerts, 1997; Heuvelink et al., 2010). In the recent past, there have been several studies that were successfully attempted to characterize water quality using geostatistical approaches. Many of these studies used traditional linear

kriging techniques or other interpolation and regression based methods with an Euclidean distance (Rasmussen et al., 2005; Tortorelli and Pickup, 2006; Cressie et al., 2005; Peterson and Urquhart, 2006).

Water quality is often dynamic and changes rapidly in space and time. Such space-time variability of water quality cannot be captured using purely spatial models. In the case of many water quality parameters, temporal variability plays a key role in understanding the overall impact on a basin-wide system. Hence, recent developments in geostatistics have moved beyond the purely spatial approach to include temporal variability as well (Stein 1986, Christakos 1992, Bogaert 1996, Kyriakidis and Journel 1999, Fuentes 2004, Kolovos et al. 2004). Space/time geostatistical models extend the concept of autocorrelation between nearby sites from the spatial dimension into the spatial and temporal dimensions, and they produce more accurate estimates at unmonitored space/time locations.

The Bayesian Maximum Entropy (BME) framework (Christakos 1990, 2000; Serre et al. 1998, Serre and Christakos, 1999) is a method of modern spatiotemporal geostatistics. The BME method has been successfully applied to a variety of environmental issues, including air quality (Christakos and Serre, 2000; Christakos et al. 2004; Wilson and Serre 2007), and disease mapping (Law et al. 2004, 2006). There have also been several interesting studies that involve the BME estimation of water quality (Serre et al. 2004, LoBuglio et al., 2007; Akita et al. 2007, Couillette et al., 2008). These studies have demonstrated that more accurate water quality maps can be produced using space/time BME than using a purely spatial analysis. For instance, Akita et al., (2007) use spatiotemporal methods to assess tetrachloroethene (PCE) in the rivers of New Jersey, and reported a 56% improvement in estimation accuracy when compared to a purely

spatial approach. This substantial improvement could most likely be due to the irregularity of the spatial and temporal sampling of PCE data.

Traditionally geostatistical water quality studies have used Euclidean distances to describe spatial autocorrelation (Rasmussen et al., 2005; Tortorelli and Pickup, 2006; Cressie et al., 2005; Peterson and Urquhart, 2006). Many studies have raised questions about the use of an Euclidean distance metric in the estimation of water quality along stream networks as it may fail to account for stream network topology (Money et al., 2009a). There have been several recent studies which proposed to use the river distance, i.e. the shortest distance along the river between sites, as an alternative to the Euclidean distance when studying the spatial autocorrelation amongst stream monitoring sites. This distance is called the "hydrologic distance" (Peterson et al., 2007), "stream distance" (Ver Hoef et al., 2006), "river distance" (Cressie et al., 2006; Money, 2009a) in the recent literature. Money et al., (2009b) reported that the use of the river distance in modeling the autocorrelation in fecal contamination along the Raritan River in New Jersey resulted in lower estimation errors compared to using the Euclidean distance. However, simply replacing the Euclidean distance with the hydrologic distance may violate geostatistical modelling assumptions and may yield an invalid (i.e., non-positive-definite) model of spatial statistical dependence. Ver Hoef et al., (2006) showed that the hydrologic distance with the spherical covariance model resulted in negative eigenvalues of the covariance matrix and hence the variances can be negative too. First Ver Hoef et al., (2006) and later Money et al., (2009a) showed that using the exponential covariance model with the river distance is a valid and permissible model.

In some mapping studies the river distance alone may not fully incorporate the dependency of the covariance function with the stream network topology and flow connectivity.

In these situations the river covariance model based solely on river distance may not adequately depict the unique spatial autocorrelation of water quality along stream networks (Peterson et al., 2013). Ver Hoef et al. (2006) showed that models using river distance and flow connectivity may be more appropriate than models that only use river distances. In the recent past, there have been successful attempts to develop valid spatial covariance models that incorporate both river distance and flow connectivity (Ver Hoef et al., 2010). A covariance function that use both river distance and flow connectivity was first introduced and derived by Ver Hoef et al. (2006), and further investigated by Fouquet and Bernard-Michel (2006), Bernard-Michel and de Fouquet (2006), Cressie et al. (2006), Money at al. (2010), and Ver Hoef and Peterson (2010). Their obvious advantage is that they incorporate flow connectivity in the model of spatial autocorrelation.

2. Classes of covariance models used to study water quality

2.1. Autocorrelation in water quality

The underlying spatial and temporal autocorrelation of a space/time random field describing a water quality parameter is determined by the characteristics and shape of its covariance model. Homogeneous and stationary covariance models describe the dependency between water quality measured at two space/time points as a function of the spatial distance and time difference between these measurement points. The spatial separation distance between observation sites is usually referred to as the spatial "lag". The spatial component of the covariance function provides a tool to quantify how autocorrelation decreases as a function of spatial lag. Since the factors driving spatial dependencies in water quality are often difficult or impossible to measure directly, the covariance model provides a practical tool to quantify these

dependencies along river networks. The covariance value at a zero lag (i.e. for a separation distance of zero) is called the covariance 'sill', and it is equal to the variance of the space/time random field. The covariance generally decreases as the lag increases, and the lag at which the covariance drops to 5% of the sill value is called the covariance spatial range. Hence the covariance spatial range indicates how quickly autocorrelation decays with separation distance. Practically, observed values are considered to be weakly correlated at separation distances exceeding the covariance range.

The way we calculate separation distance affects covariance modeling and there are a variety of distance measures to consider when dealing with water quality parameters distributed along the river network. These various distance metrics give rise to various permissible covariance models that can be used to study water quality along river networks. However, three main classes of permissible covariance models, namely the Euclidean covariance, river covariance, and flow-weighted covariance models, are the most commonly used covariance models in water quality studies.

Many known and unknown processes operating simultaneously within the river network and in its surrounding terrestrial landscape drive the autocorrelation in the water quality parameters. The choice of the proper covariance model can be influenced by these processes for specific mapping situations. Here we use terrestrial landscape sources and longitudinal hydrological transport as two examples of the many processes that can drive the autocorrelation in water quality along rivers. Figure1 is a cartoon illustrating how these two processes could give rise to autocorrelation described by (a) an Euclidean, (b) a river and (c) a flow covariance model, as well as (d) a hybrid Euclidean-flow covariance model. It should be emphasized that since

there can be many other processes driving the spatial autocorrelation in water quality, then figure 1 is just one of many examples that can give rise to each of these covariance models.

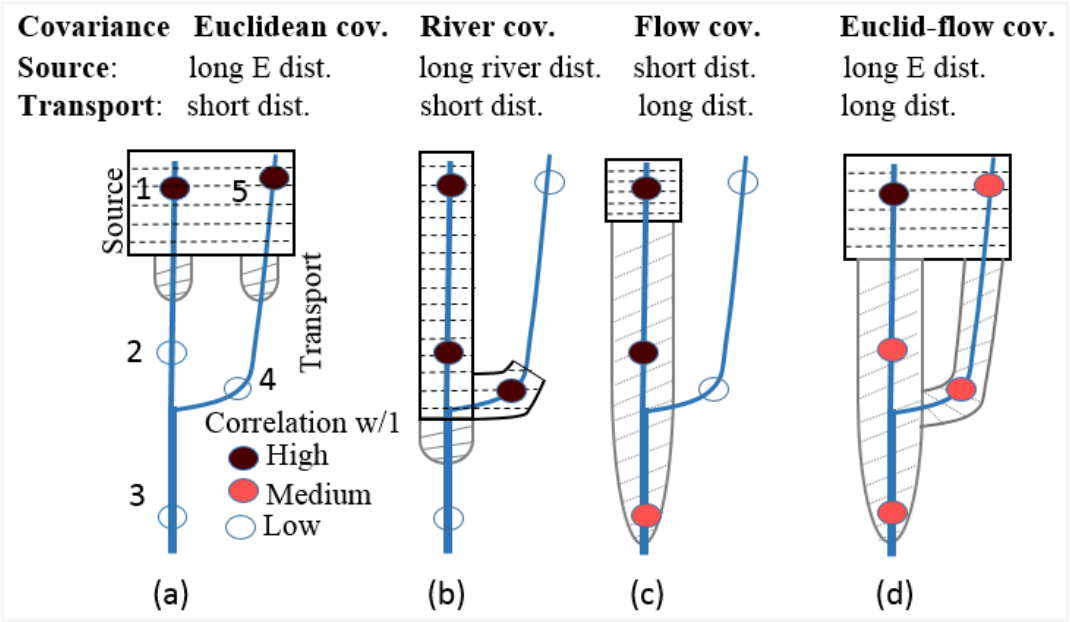


Figure 1.1: Schematic representation of how the autocorrelation in terrestrial contamination sources and the longitudinal transport distance along rivers can lead to various water quality covariance models: (a) contamination source autocorrelated across long Euclidean distances coupled with transport over short distances can lead to an Euclidean covariance model, (b) source autocorrelated along long river distances and transport over short distances can lead to a river covariance model, (c) point sources autocorrelated over short distances and transport along long river distances can lead to a flow covariance model and (d) contamination source autocorrelated across long Euclidean distances coupled with transport along long distances can lead to nested Euclidean-flow covariance model. For each covariance model the correlation between 4 sites (labeled 2 to 5) and site 1 is shown using color darkness.

In this figure, the strength of the autocorrelation in water quality between a reference monitoring site at site 1 and other sites located at sites 2-5 is shown using color darkness. Sites shown with the highest color darkness have the highest correlation with site 1, and sites with the lowest darkness have the lower correlation with site 1. Details about each covariance model are given next.

2.2. Euclidean covariance model

The Euclidean distance is used as the measure of the separation distance between sites in Euclidean covariance models. This traditional metric measuring the straight line distance between sites is commonly used in most geostatistical frameworks. This class of covariance models adequately describes spatial autocorrelation when water quality parameters are largely driven by terrestrial processes over long Euclidean distances, and when transport has little impact on the spatial distribution of water quality, as shown in figure 1 (a).

Figure 1(a) illustrates that since the Euclidean distance separating site 1 and site 5 is short then observations at these two sites are highly autocorrelated, even though they are separated by a long distance along the river. Conversely since sites 2, 3 and 4 are at long Euclidean distances from site 1, they are therefore weakly correlated with site 1. This could be a case when the pollutant sources are distributed over long Euclidean distances across the land and the longitudinal transport of the untransformed pollutant is very short. In other words, the Euclidean class of covariance models can better express autocorrelation in water quality parameters when the pollution source is distributed over long distances across land regardless of the river hydrography, and the transport of that water quality parameter is occurring over short distance because of lack of travel distance or because the pollutant is not persistent in the water. Hence site 5 is highly correlated with site 1 because they share the same pollution source, while sites 2, 3 and 4 are weakly correlated with site 1 because these sites do not share a pollution source with site 1 nor is the contaminant transported from site 1 to these sites in untransformed form.

Non-point pollution sources such as atmospheric deposition and large agricultural fields are good examples of pollutions sources autocorrelated over the long Euclidean distances. Likewise, there are many examples of pollutants experiencing short longitudinal transport

distances along rivers. For instance, regardless of the travel distance, pollutants can be transformed quickly in the river waters by several natural, physical, and biological processes, such as uptake by biotic or aquatic species, degradation, oxidation and reduction, settling etc., and as a result these pollutants are transformed before they can be transported over long distances. For instance under some environmental conditions the ammonia concentration in river waters may have an Euclidean covariance. This can for example happen when the source of ammonia are large agricultural fields stretching across parallel river branches, and when the river waters provide an environment for quick biological uptake or quick transformation/oxidation into other nitrogen forms.

2.3. River covariance model

Euclidean covariance models form a widely used class of permissible covariance models. However, many studies raise questions about the use of the Euclidean distance in the estimation of water quality along stream networks as it may fail to incorporate stream network topology.

The second class of permissible covariance models considered here are river covariance models. River covariance models quantify the autocorrelation between two points based on the river distance, i.e. the distance along the river reaches connecting the two points. This class of covariance models better describes autocorrelation when the river network topology needs to be taken into account when quantifying autocorrelation.

As shown in figure 1(b), sites 2 and 4 are at short river distances from site 1 and therefore they are highly correlated with site 1. Site 5 is at a long river distance from site 1, and therefore it is weakly correlated with site 1 even though it is at a short Euclidean distance from it. Site 3 is also at a long river distance from site 1 and therefore it is also weakly correlated with site 1. This

covariance model only accounts for the river distance between sites, not for flow connectivity. For example site 4 is on a parallel branch and therefore not flow connected with site 1. However it is highly correlated with site 1 because there is a short river distance when traveling downstream from site 1 along the main river reach and then traveling upstream along the side branch all the way to site 4. Hence the river covariance model describes autocorrelation governed by river distances regardless of flow connectivity.

Some processes can lead to a river covariance model. As illustrated in figure 1(b) this can happen when the pollution source is autocorrelated along long river distances and there is little longitudinal transport of the untransformed pollutant once it reaches the river waters. This can happen when the pollution source is distributed along elongated agricultural fields or roads that happen to follow the river topography, or if there are source attenuation processes such as green buffers that follow the river topography downstream and upstream of connected river reaches. For example, chloride from deicing salt applied along roads laid parallel to streams with strong buffer capacity can lead to a spatial distribution of chloride stream concertation that can be adequately quantified using river covariance models. This example is investigated in objective 1 of this dissertation.

2.4. Flow-weighted covariance model

The third class of covariance models for water quality parameters are flow-weighted covariance models, herein referred to as simply the flow covariance model. Flow-weighted covariance models account for both river topology and flow connectivity by incorporating river distance and flow in the covariance model. This kind of covariance models are useful to describe autocorrelation for persistent pollutants that travel over long downstream distances along rivers,

and therefore for which dilution of the pollutant along a river is an important driver for the autocorrelation exhibited by that pollutant.

Using the flow covariance model, the covariance value between two points is not only a function of the river distance separating these points, but also a factor that is equal to zero if the points are not flow connected, or that is equal to the ratio of upstream flow to the downstream flow when the points are flow connected. This factor is referred to as the flow ratio. It quantifies the proportion of the downstream flow that is coming from the upstream point, which essentially accounts for the dilution from side branches.

The combined effect of the river distance and flow ratio between points can be seen in figure 1(c) depicting correlation between site 1 and sites 2-5 using a flow covariance model. The correlation between the site 1 and site 2 is high because they are at a short river distance and because the flow ratio is high, since there is little dilution between site 1 and 2, or put in other words most of the flow in 2 is coming from 1. However the correlation drops as the downstream point moves down past the side branch. For instance the correlation between site 1 and 3 drops from high to medium, because of the dilution from the side flow which causes the flow ratio between 1 and 3 to drop appreciatively. Finally sites 4 and 5 are on a side branch that is not flow connected with site 1; therefore they have a zero correlation with site 1 since none of the flow at sites 4 and 5 is coming from site 1.

Figure 1(c) depicts an example of source and transport processes that can lead to a flow covariance model. In this example the contamination is coming from a point source, meaning that the contamination is localized and therefore autocorrelated over a very short distance. On the other hand in this example the transport occurs over a very long distance. This generally happens when there is sufficient flow to generate long travel distances, and the pollutant is persistent, i.e.

it is not removed from the stream water. In that case site 2 is highly correlated with site 1 because of transport from site 1 to 2. Site 3 has a medium correlation with site 1 because of dilution from the side branch, which brings in water with uncorrelated pollution concentration. Finally sites 4 and 5 are not correlated with site 1 because they are not flow connected.

This class of permissible covariance models is relatively new and has only recently been used to describe the autocorrelation in water quality along rivers. Flow covariance models are a suitable choice to describe autocorrelation in water quality when autocorrelation is driven by longitudinal hydrologic transport of persistent pollutants along rivers.

2.5. Mixture of Euclidean and flow covariance models

The three covariance models described above are suitable in many specific mapping situations. Using source and transport as illustrative processes driving the autocorrelation of water quality, the Euclidean and river covariance models are suitable when the autocorrelation in water quality is driven by autocorrelation in the pollution source but not its transport, and the flow covariance model is suitable when autocorrelation is driven by transport but not source. However there are other mapping situations that can combine traits from two or more covariance models. In this work we will specifically explore the use of a nested Euclidean and flow covariance model. Mathematically a nested Euclidean and flow covariance model is simply written as the linear combination of an Euclidean covariance model and a flow covariance model. The linear weight of each model describes the proportion of variability in water quality described by that model. For illustration purposes Figure 1(d) depicts the variability corresponding to a nested Euclidean and flow covariance model with an equal weight for the Euclidean and flow covariance models. In that case the correlation of sites 2-4 with site 1 is the

average of the correlation from the Euclidean model (figure 1a) and flow model (figure 1c). In that case sites 2 to 5 are all having a medium correlation with site 1, because they either share the same source (site 5) or are within transport distance (sites 2 to 4).

The advantage of using a nested Euclidean and flow covariance model is that it widens the range of mapping situations that can be adequately modeled. Using source and transport as an illustrative example, the Euclidean-flow covariance model adequately describes variability for water pollutants for which the contamination occurs across long Euclidean distances, such as large agricultural fields or other terrestrial features, and for which transport also occurs over somewhat long distances downstream of the contamination source. To our knowledge Euclidean-flow covariance models have not been used in the past and therefore this work will be the first to introduce this model.

An example may be the spatial distribution of fecal coliform along some river networks. Fecal coliforms are an indicator of fecal contamination. Its source includes grazing and agricultural fields that can extend across long Euclidean distances, and its transport may occur over intermediate to long distances downstream of the sources when fecal coliforms are present in the suspended solid transported at high flow during storm events. In this case both source and transport may occur over intermediate to long distances and therefore the Euclidean-flow covariance model may be more suitable then a purely Euclidean or purely flow covariance model. This case is explored in objective 2 of this dissertation.

Another example may be the spatial distribution of dissolved organic carbon (DOC) along rivers. DOC comes from terrestrial sources that may be autocorrelated over long Euclidean distances, and DOC may be transported over intermediate distances, and as a result the

variability of DOC may adequately be described using an Euclidean-flow covariance model. This case is explored in objective 3 of this dissertation.

3. Some Knowledge gaps in previous water quality studies

Several studies have successfully attempted to characterize water quality along rivers using geostatistical approaches as these approaches provide a convenient way to model spatially dependent water quality observations. Quantifying autocorrelation is a defining feature of geostatiscal modeling and the selection of the most appropriate covariance model is of a great significance. However, when modeling spatial dependence in river networks, there are many mapping situations for which there are significant knowledge gaps in knowing what covariance model should be used.

Several geostatistical water quality studies have used traditional Euclidean covariance models (Rasmussen et al., 2005; Tortorelli and Pickup, 2006; Cressie et al., 2005; Peterson and Urquhart, 2006). Euclidean covariance models fail to account for river connectivity and topology. More recently, flow covariance models have also been used in water quality parameters. Many past studies (Ver Hoef *et al.,* 2006, Cressie *et al.,* 2006, Peterson and Urquhart, 2006) explored and compared the Euclidean and flow covariance models to better quantify autocorrelation in water quality along river networks. Cressie *et al.* (2006) and Peterson and Urquhart (2006) found that the Euclidean covariance model performed better than the flow covariance model. However, these studies did not report results using river covariance models (i.e. covariance models based only on river distances but not flow ratio), unlike several other studies which successfully used river distances in other river networks (Gardner et al., 2003, Ganio et al., 2009, Yang and Jin, 2010, Money et al., 2011, Chen et al., 2012, and Cressie et al.,

2013). There may be situations where processes distributed along river networks (e.g. runoff from roads -a known pollution source, vegetation buffers -a known attenuation process, etc.) are important drivers of the water quality autocorrelation along rivers. Therefore, an important remaining question is whether the river covariance model better describes autocorrelation in water quality along river networks than the Euclidean and the flow covariance models. This knowledge gap will be addressed in the first objective of this research by implementing the river covariance model and comparing it to the Euclidean and flow covariance models when modeling the distribution of Chloride along rivers in Maryland. Another remaining a matter of investigation is whether water quality estimation maps obtained using a river covariance model lead to an assessment of impairment that is significantly different than that obtained using an Euclidean covariance model. This knowledge gap will also be addressed in the first objective of this research.

Many known and unknown processes such as degradation, biogeochemical processes, and hydrological interactions in river networks are very complex. Our understanding of these processes over the terrestrial landscape and in stream networks is still limited (McGuire et al., 2014). Euclidean and river covariance models may be better suited to describe autocorrelation in water quality arising from the spatial distribution of the contamination source across the terrestrial landscape, whereas flow covariance models may be better suited to describe autocorrelation driven by hydrological transport processes. Using a purely Euclidean, purely river or purely flow covariance model may fail to fully describe autocorrelation driven simultaneously by both terrestrial source and hydrologic transport processes. To the best of our knowledge, there are no studies to date that have used a mixture of the Euclidean, river and flow covariance models to better describe the autocorrelation in water quality. This is an important

knowledge gap to be addressed in order to improve water quality estimation along river networks using geostatistical approaches. This knowledge gap will be addressed using a mixture of the Euclidean and flow covariance models to study the space/time distribution of fecal coliforms along rivers in North Carolina in objective 2, and to study the space/time distribution of DOC along rivers in Maryland in objective 3.

4. Research objectives

4.1. Research objective 1: Bayesian Maximum Entropy Space/time Estimation of Surface Water Chloride in Maryland Using River Distances

Headwater streams and rivers are important sources of water for downstream ecosystem and human population. These streams comprise the vast majority of the streams and river miles. River network based geostatistical modeling approaches can be used to assess the space/time variations in headwater streams and rivers. Indeed, each water quality study needs to consider all classes of permissible covariance models (Euclidean, river, and flow-weighted), but not always the case. There are many known and unknown natural processes driving the autocorrelation in water quality parameters. The choice of a covariance model to explain the autocorrelation in water quality can be influenced by these processes for specific mapping situations.

Widespread contamination of surface water chloride and its effect on the ecosystem health are emerging environmental concern. The rate of urban development, changes in road salt application practices, and changing climate conditions may drive a variety of spatial and temporal patterns in chloride concentrations (Corsi et al., 2015). Accurate estimation of chloride is crucial to understand these patterns, to improve our understanding of the extent and nature of chloride contamination, and to design effective measures to control the chloride pollution.

Peterson and Urquhart (2006) found that the spatial autocorrelation of dissolved organic carbon (DOC) in Maryland is better described using a covariance based on Euclidean distances rather than using a flow-weighted river distance covariance. However, their work did not report results for a autocorrelation using only river distances unlike several other studies which successfully used river distances in other river networks (Gardner et al., 2003, Ganio et al., 2009, Yang and Jin, 2010, Money et al., 2011, Chen et al., 2012, and Cressie et al., 2013). Hence, an important remaining question is whether the river distance works better than the Euclidean for the geostatistical estimation of chloride concentration along rivers in Maryland. We hypothesize that processes that are distributed along river networks (e.g. highways -a known source of chloride, vegetation buffers -a known attenuation process, etc.), are important drivers of the distribution of chloride along rivers and this autocorrelation can be best described using river distance. The first objective of this dissertation is therefore to introduce a framework for the *BME space/time estimation of surface water chloride using river distances in three subbasins located in Maryland*, and to compare this method with alternate methods using Euclidean distances.

4.2. Research objective 2: Introducing a novel geostatistical approach combining Euclidean and flow-weighted covariance models to estimate fecal coliform along the Deep and Haw Rivers in North Carolina

The complexity of the spatial and temporal patterns in water quality along river networks has not been fully investigated. As described earlier, several natural processes may act simultaneously and with different intensities, resulting in spatial autocorrelation that is best described using a mixture of covariance models. Using a purely Euclidean, purely river or purely

flow covariance model may limit the ability to fully describe the variability of many water quality parameters.

Unlike most conventional water quality parameters, fecal coliform bacteria are living organisms. Fecal coliform bacteria can enter rivers through discharge of fecal material in surface run-off, combined sewer overflows, and point source discharges. They do not simply mix with the water and float downstream, instead they multiply quickly when conditions are favorable for growth, or die in large numbers when conditions are unfavorable. Because bacterial concentrations are dependent on specific conditions for growth, and these conditions change quickly, spatial and temporal patterns of fecal coliform bacteria can be very erratic and hence are not easy to model using mechanistic approaches. Geostatistical approaches on the other hand provide an ideal framework to statistically model the space/time variability of fecal coliform and obtain estimates and associated prediction confidence intervals at any unsampled points along the river network.

Modeling the space/time variability of fecal coliform requires choosing a covariance model that captures well its spatial variability along rivers. The spatial variability of fecal coliform is driven by two important factors. One is the source of fecal matter consisting in large parts of grazing fields or agricultural fields where manure is spread. These fields may extend over large Euclidean distances exceeding local watersheds, and in some instances covering areas that extend across watersheds. The covariance model that may best describe this terrestrial process is the Euclidean covariance model. The other important factor driving the spatial distribution of fecal coliform in the water is hydrological transport along rivers. This occurs when fecal coliforms are present in the suspended solid, which can be transported over long distances during storm events when flows are high and the water has a high turbidity. The

covariance model that can best describe this process is the flow covariance model. Because both processes (terrestrial source and hydrological transport) may act simultaneously, we hypothesize that using a mixture of Euclidean and flow covariance models will better characterize the true underlying autocorrelation in fecal coliform concentrations than using a purely Euclidean or purely flow covariance model. To the best of our knowledge, no previous study has used a mixture of Euclidean and flow covariance model to describe the space/time variability of a water quality parameter. Therefore, the objective 2 of this dissertation is the introduction of a novel geostatistical approach combining the Euclidean and flow-weighted covariance models to estimate fecal coliform along the Haw and Deep Rivers in North Carolina.

4.3. Research objective 3: Bayesian Maximum Entropy Space/time Estimation of Surface Water Chloride in Maryland Using River Distances

Dissolved organic carbon (DOC) is an organic matter that can pass through a filter (0.7 and 0.22 um). DOC is an important constituent of water quality due to the fact that it plays a central role in the dynamics of stream and river ecosystems, affecting processes such as metabolism, acidity and nutrient uptake. It forms complexes with trace metals and alters bioavailability and longitudinal transport of compounds that are toxic to aquatic organisms.

Headwater streams are important sources of water for downstream ecosystems. The spatial variability of the concentration of DOC in headwater streams is strongly influenced by the production of organic matter across the terrestrial environment. The Euclidean covariance model is a suitable choice to describe variability of DOC in the water that results from its production across the terrestrial landscape. For example Peterson and Urquhart (2006) found that the spatial autocorrelation in DOC concentrations in Maryland is better described using a covariance based

on Euclidean distances rather than using a flow-weighted covariance. On the other hand, DOC can be transported over long distance (kilometers) along rivers via stream flow (Worrall, Burt & Adamson, 2006; Kaplan et al., 2008) and hence the flow-weighted covariance seems to be an equally appropriate model to describe the autocorrelation of DOC. Hence the finding by Peterson and Urquhart (2006) that the spatial autocorrelation in DOC is better described using a purely Euclidean covariance than using a purely flow-weighted covariance may not tell the full story. It is possible that terrestrial sources and hydrological transport act simultaneously on the variability of DOC. In that case the finding by Peterson and Urquhart (2006) could be explained if the terrestrial sources of DOC is the dominant driver of the autocorrelation of DOC along rivers, or if there are few flow connected monitoring sites, because in that case using a purely flow covariance limits the ability to estimate DOC along river reaches that are not flow connected to any monitoring site. Therefore, the objective 3 of this dissertation is to perform a space/time estimation of DOC along rivers in Maryland using a combination of the Euclidean and flow-weighted covariance models.

CHAPTER 2 (PAPER 1): BAYESIAN MAXIMUM ENTROPY SPACE/TIME ESTIMATION OF SURFACE WATER CHLORIDE IN MARYLAND USING RIVER DISTANCES[1]

1. Introduction

Chloride contamination of rivers and its effect on the ecosystem health is a great environmental concern. During the winter snow, roads and sidewalks are treated with deicing salts. As the snow melts, more than 50 percent of the chloride in the deicing salt is transported to surface waters, leading to widespread effects on water chemistry (Church and Granato, 1996). Road salt application practices and a variety of other processes lead to complex spatial and temporal patterns in chloride concentrations (Corsi et al., 2015).

Geostatistical methods provide potential for water quality assessment. Several studies have characterized surface water quality using spatial linear kriging methods (Peterson and Urquhart, 2006 and Money et al., 2010). However, spatial kriging studies do not account for space/time autocorrelation and non-Gaussian 'soft' data (interval and censored data etc.). To address this issue, the Bayesian Maximum Entropy (BME)(George Christakos, 1990 and Christakos and Li, 1998) method is used here to estimate chloride concentration across space/time along a river network in Maryland.

1 This chapter previously appeared as an arcticle in the Journal Environmnetal Pollution. The original citation is as follows : ''Jat, P., M.L. Serre. 2016. Bayesian Maximum Entropy space/time estimation of surface water chloride in Maryland using river distances. *Environ. Pollut*. doi:http://dx.doi.org/10.1016/j.envpol.2016.09.020

BME is a nonlinear estimation method that rigorously accounts for space/time variability and non-Gaussian soft data, and leads to kriging as its linear limiting case (Christakos, 1990, Christakos and Li, 1998 and Christakos and Serre, 2000).

Peterson and Urquhart (2006) found that in Maryland the spatial autocorrelation of dissolved organic carbon (DOC) is better described using a covariance based on Euclidean distances rather than using a Weighted Asymmetric Hydrologic Distance (WAHD) covariance model, which is calculated based on the river distance (distance measured along the river network) and the proportion of flow shared between points (Peterson and Urquhart (2006), Money et al., 2009). Therefore, when considering other water quality parameters in Maryland, we expect that the Euclidean distance will better describe the spatial autocorrelation. However, their work did not report results for a autocorrelation using covariance based only on river distances (and not proportion of flow shared between points), unlike several other studies which successfully used river distances in other river networks (Gardner et al., 2003, Ganio et al., 2009, Yang and Jin, 2010, Money et al., 2011, Chen et al., 2012, and Cressie et al., 2013). Hence, an important remaining question is whether the river distance works better than the Euclidean for the geostatistical estimation of chloride along rivers in Maryland.

The objectives of this study are therefore to introduce a framework for the BME space/time estimation of surface water chloride using river distances in three subbasins located in Maryland, to compare this method with alternate methods using Euclidean distances, to do a sensitivity analysis of methods used to deal with censored data, and to perform a space/time statistical estimation of chloride concentration along all river miles in our study area using the BME method based on river distances.

2. Materials and Methods

2.1. Chloride and Hydrography Data

A total of 390 space/time chloride concentration values were obtained from the Maryland Biological Stream Survey (MBSS) dataset from 2005 to 2014 in stream waters located in the Gunpowder-Patapsco, Severn, and Patuxent subbasins (figure 1). The concentration values ranged from 1.5 mg/l to 3251.2 mg/l, with mean 93.69 mg/l and standard deviation 230.44 mg/l. Details on field sampling design, sampling methodology, and lab analysis procedures can be found elsewhere (Taylor-rogers, 1997).

The river network in our study area is described based on flow lines (figure 1) obtained from the USGS National Hydrography Data ("USGS Hydrography data," 2015). The impervious surfaces are described based on the National Land Cover Database published by the Multi-Resolution Land Characteristics Consortium for the conterminous United States. Details about the NHD flowlines and impervious surface data are provided in the Supplementary Information (SI).

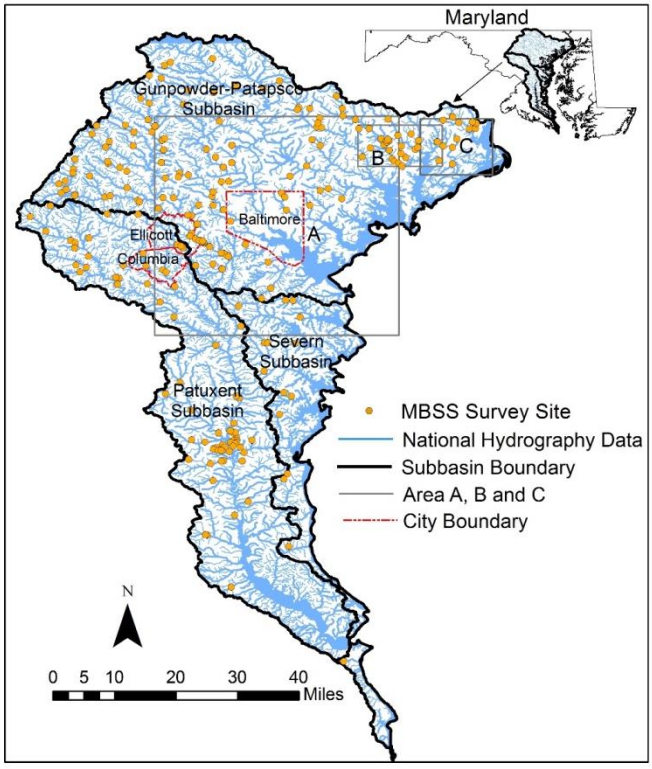


Figure 2.1: The Maryland Biological Stream Survey (MBSS) sites in the Gunpowder-Patapsco, Patuxent, and Severn subbasins in Maryland. Baltimore, Ellicott, and Columbia are tree major cities in these subbasins.

2.2. Left-Censored Data

Left-censored chloride data correspond to data for which the true log-concentration is known only to be below a censoring limit (CL) of interest. Censoring data is a common practice when measured values are below the detection limit (DL) of an instrument. The BME approach has recently been shown to rigorously process left-censored data (Messier et al., 2012). Briefly, the maximum-likelihood estimation (MLE) method is used to estimate the mean (μ) and standard deviation (σ) of stream chloride concentrations by finding the μ and σ values that maximizes the MLE likelihood function (Helsel, 2005 and Messier et al., 2012)

$$\mathcal{L}(z|\mu,\sigma) = \left\{\prod_{z_i|z_i \geq CL_i} f_{\mu,\sigma}(z_i)\right\} * \left\{\prod_{z_i|z_i \leq CL_i} F_{\mu,\sigma}(CL_i)\right\} \quad (1)$$

where $f_{\mu,\sigma}(z_i)$ denotes the normal probability distribution function (PDF) of observed chloride log-concentrations, z_i, with population mean (μ) and standard deviation (σ), and $F_{\mu,\sigma}(CL_i)$ denotes the CDF of the distribution taken at the log of the censoring limit (CL_i). The uncertainty associated with a left-censored data with CL_i is then fully characterized by the Truncated Gaussian PDF (TGPDF) obtained by truncating a Gaussian PDF above CL_i. The TGPDF(μ, σ, CL_i) has a mean<μ because of the truncation.

2.3. Space/time BME Geostatistical Framework for Mapping Analysis

BME, a space/time geostatistical estimation framework grounded in epistemic principles, reduces to the kriging methods as its linear limiting case. BME theory and its numerical implementation details are given elsewhere (Christakos, 1990, Serre and Christakos, 1999, Christakos and Serre, 2000, and Patrick Bogaert, 2001). Details about the application of BME to river networks are given elsewhere (Money et al., 2009).

Our notation to describe a space/time random field (S/TRF) will consist of denoting a single random variable Z in capital letters, its realization, z, in lower case; and vectors in bold faces (e.g., $\boldsymbol{z} = [z_1,..., z_n]^T$). Let $\boldsymbol{z}_d$ be the vector of log-concentrations observed at locations $\boldsymbol{p}_d$, let $o_z(\boldsymbol{p})$ be an known offset function (Messier et al., 2015), where $\boldsymbol{p} = (\boldsymbol{s}, t)$, $\boldsymbol{s}$ is the space coordinate and t is time, and let $\boldsymbol{x}_d = \boldsymbol{z}_d - o_z(\boldsymbol{p}_d)$ be the vector of offset removed log-concentrations. The suffix d in $\boldsymbol{p}_d$ is used to specify a location where data is available (i.e. a data point), whereas $\boldsymbol{p}$ without suffix d specify any location in the study domain. We define $X(\boldsymbol{p})$ as a homogenous/stationary S/TRF with realization $\boldsymbol{x}_d$, and we let

$$Z(\boldsymbol{p}) = X(\boldsymbol{p}) + o_z(\boldsymbol{p}). \quad (2)$$

be the S/TRF representing the distribution of stream chloride log-concentrations.

The total knowledge base *K* characterizing the S/TRF $X(\boldsymbol{p})$ can be divided in the general knowledge base (G-KB) and the site-specific knowledge base (S-KB). The G-KB describes general characteristics of the S/TRF including its mean $m_x(\boldsymbol{p}) = E[X(\boldsymbol{p})]$ and covariance function

$$c_x(\boldsymbol{p}, \boldsymbol{p}') = E\left[\left(X(\boldsymbol{p}) - m_x(\boldsymbol{p})\right)\ \left(X(\boldsymbol{p}') - m_x(\boldsymbol{p}')\right)\right], \quad (3)$$

where E[.] is the stochastic expectation operator. The S-KB refers to the sampling data $\boldsymbol{x}_d$, including both the hard (above detect) data $\boldsymbol{x}_h$ collected at $\boldsymbol{p}_h$, and the soft (left-censored) data $\boldsymbol{x}_s$ collected at $\boldsymbol{p}_s$ with an uncertainty expressed in terms of the PDF $f_s(\boldsymbol{x_s})$ (e.g. TGPDF(μ, σ, CL_i)).

We briefly describe here the main stages of the BME analysis used to estimate chloride log-concentration at unsampled locations $\boldsymbol{p}_k$ along the river network. At the prior stage, the $G - KB = \{E[X(\boldsymbol{p})], C_X(\boldsymbol{p}, \boldsymbol{p}')\}$ is examined to obtain the prior PDF $f_G(.)$ describing the S/TRF $X(\boldsymbol{p})$ at mapping points of interest. At the integration stage, the prior PDF is updated using Bayesian epistemic conditionalization on $S - KB = \{\boldsymbol{x}_h, f_s(\boldsymbol{x}_s),\}$, leading to the BME posterior PDF

$$f_K(x_k) = A^{-1} \int d\boldsymbol{x_s} f_G(\boldsymbol{x_h}, \boldsymbol{x_s}, x_k) f_S(\boldsymbol{x_s}) \quad (4)$$

where x_k is a value of X_k=$X(\boldsymbol{p}_k)$, $f_G(\boldsymbol{x}_h, \boldsymbol{x}_s, x_k)$ is the multivariate Gaussian PDF for $(\boldsymbol{x}_h, \boldsymbol{x}_s, x_k)$ with mean and variance-covariance given by the G-KB, and $A = \int dx_k \int d\boldsymbol{x}_s f_G(\boldsymbol{x}_h, \boldsymbol{x}_s, x_k) f_S(\boldsymbol{x}_s)$ is a normalization coefficient. At the interpretive stage, the relation $Z_k = X_k + o_z(\boldsymbol{p}_k)$ is used together with $f_K(x_k)$ to obtain the BME mean and variance log-concentration at the estimation points, which are then used to produce maps describing the estimated chloride log-concentration and associated estimation uncertainty at space/time locations of interest.

Several approaches exist to calculate an offset function $o_Z(\boldsymbol{p})$. In this work we use the approach described in Akita et al. (2007) and Money et al. (2009), where $o_Z(\boldsymbol{p}) = o_Z(\boldsymbol{s}, t)$ is the sum of a spatial component $o_{Z,s}(\boldsymbol{s})$ and a temporal component $o_{Z,t}(t)$ that are calculated using an exponential kernel smoothing of the time-averaged and spatially averaged data, respectively. Specifically, the spatial component at a given location $\boldsymbol{s}$ is given by

$$o_{Z,s}(\boldsymbol{s}) = \sum_i w(\boldsymbol{s}, \boldsymbol{s}_i)\, \overline{z(\boldsymbol{s}_i)} \tag{5}$$

where $\overline{z(\boldsymbol{s}_i)}$ is the time-averaged log-concentration at location $\boldsymbol{s}_i$, $w(\boldsymbol{s}, \boldsymbol{s}_i)$ is an exponential kernel weight given by

$$w(\boldsymbol{s}, \boldsymbol{s}_i) = B^{-1} \exp(-3d(\boldsymbol{s}, \boldsymbol{s}_i)/k_r), \tag{6}$$

$d(\boldsymbol{s}, \boldsymbol{s}_i)$ is the distance between $\boldsymbol{s}$ and $\boldsymbol{s}_i$, k_r is the spatial exponential smoothing range, and $B = \sum_i \exp(-3d(\boldsymbol{s}, \boldsymbol{s}_i)/k_r)$ is a normalization coefficient calculated so that the sum of weights equals 1. In the previous water quality studies (Akita et al., 2007 and Money et al., 2009) the distance $d(\boldsymbol{s}, \boldsymbol{s}_i)$ in Eq. (6) is based on an Euclidean metric. In this work we extend past works by calculating that distance based on either an Euclidean or a river distance metric, i.e.

$$d(\boldsymbol{s}, \boldsymbol{s}') = \begin{cases} d_E(\boldsymbol{s}, \boldsymbol{s}') & \text{Euclidean distance} \\ d_R(\boldsymbol{s}, \boldsymbol{s}') & \text{River distance} \end{cases} \quad (7)$$

To the best of our knowledge, this is the first study implementing an offset calculated using a kernel smoothing based on a river metric, hence the river offset presented here is novel. Note that the calculation of the temporal component $\boldsymbol{o}_{Z,t}(\boldsymbol{t})$ is done as described in Akita et al. (2007), i.e. by replacing the spatial distance in Eq (6) with the corresponding time difference. As shown in the SI, the offset function described here captures well the broad spatial and temporal trends in chloride log-concentrations, indicating that this offset function is suitable in this study area.

Alternatively, the offset function can be calculated using a Land Use Regression (LUR) as described in Messier et al., (2012), and Reyes and Serre (2014), where the LUR uses land imperviousness as a predictor, since it has been found to be a predictor of stream water quality degradation (Brabec et al., 2002 and King et al., 2011)

The $c_x(\boldsymbol{p}, \boldsymbol{p}')$ function describing the covariance of the homogeneous/stationary S/TRF $X(\boldsymbol{p})$ can be expressed as an exponential function of the spatial distance and time difference between space/time points $\boldsymbol{p}$=($\boldsymbol{s}$,t) and $\boldsymbol{p}$'=($\boldsymbol{s}$',t'), i.e.

$$\boldsymbol{c_x}\big((\boldsymbol{s}, \boldsymbol{t}), (\boldsymbol{s}', \boldsymbol{t}')\big) = \boldsymbol{c_0}\,\mathbf{exp}(-\mathbf{3}\,\boldsymbol{d}(\boldsymbol{s}, \boldsymbol{s}')/\boldsymbol{a_r})\mathbf{exp}(-\mathbf{3}\,|\boldsymbol{t} - \boldsymbol{t}'|/\boldsymbol{a_t}) \qquad \textbf{(8)}$$

where $\boldsymbol{c_0}$, a_r and a_t are the variance, spatial covariance range, and temporal covariance range, respectively, of the S/TRF $X(p)$, and $d(s,s')$ can again be either the Euclidean or river distance (equation 7). In this work we choose an exponential covariance model because it has been shown to be permissible for any river networks (Ver Hoef et al., 2006; Peterson and Urquhart, 2006 and Money et al., 2009) and to our knowledge no other covariance model has been shown to fulfill that same property.

To quantify the impact of using either the Euclidean or river distance (eq. 7) in the offset (eq. 6) and covariance (eq. 8), we implement all combinations of offset and covariance models (i.e. Euclidean offset/Euclidean covariance, Euclidean offset/River covariance, River offset/Euclidean covariance, and River offset/River covariance models) and we compare their mapping accuracy.

Another alternative for the covariance model is using a WAHD covariance model (Peterson and Urquhart, 2006, and Money et al., 2009), however, we excluded it from detailed analysis because we found it has a lower mapping accuracy than the Euclidean covariance model, which is consistent with what Peterson and Urquhart (2006) found for DOC using the MBSS data.

2.4. Comparison of BME using River versus Euclidean Distance

The DL for our MBSS chloride data is very low (0.01 mg/l), and all 390 measured values are above DL. In that case the BME method treats all the data as hard, and no soft data are used. In this baseline case the effect of using a river versus Euclidean distance in the BME estimation method was assessed by performing a leave-one-out cross-validation (LOOCV) whereby each chloride log-concentration value z_j was removed one at a time, and re-estimated using only the remaining data. For a given estimation method (m) that uses either the river or Euclidean distance, the overall estimation error was quantified using the Mean Squared Error, $MSE^{(m)} = \frac{1}{n}\sum_{j=1}^{n}\left(z_j^{*(m)} - z_j\right)^2$, the consistent estimation error (i.e. the bias) was quantified using the Mean Error $ME^{(m)} = \frac{1}{n}\sum_{j=1}^{n}\left(z_j^{*(m)} - z_j\right)$, and the random error (i.e. lack of precision) was quantified using the squared Pearson coefficient, $R^2 = 1 - \sum_{j=1}^{n}\left(z_j^{*(m)} - z_j\right)^2 / \sum_{j=1}^{n}\left(z_j^{*(m)}\right)^2$, where $z_j^{*(m)}$ is the re-estimation of z_j. This cross validation analysis was used to quantify the gain in mapping accuracy when the Euclidean distance is replaced with the river distance in the covariance model, and then in the offset model. This results in four baseline approaches (Euclidean offset/Euclidean covariance, Euclidean offset/river covariance, river offset/Euclidean covariance, and river offset/ river covariance) which are all mathematically permissible regardless of their physical meaningfulness.

2.5. Sensitivity Analysis with respect to the Proportion of Left Censored Data

Methods are needed to deal with situations where there is a large proportion of left censored data. This can happen for cost effectiveness purposes when low-cost data is used (LoBuglio et al., 2007), or when measuring toxic compounds that are difficult to detect.

The usual approaches used to deal with left censored data have been to delete them, or to fabricate numbers for them (equal to half of the CL, or equal to the CL), which are flawed approaches that can introduce a strong bias in mean and standard deviation (Singh and Nocerino, 2002).

On the other hand, the BME approach has recently been shown to rigorously process left-censored data (Messier et al., 2012). However few studies have investigated the loss of accuracy associated with left-censored data (Helsel, 2005, and Messier et al., 2012), and this study provides a unique opportunity to do that. As stated earlier, all 390 measured values are above the DL, which provided us an opportunity to investigate the sensitivity of the loss in mapping accuracy with respect to the proportion of censored data. This sensitivity analysis consisted in left censoring a proportion of the data, and comparing the cross validation statistics of the following three methods: (a) BME rigorously modeling the censored data using the TGPDF, (b) kriging replacing the censored data with half the CL, and (c) kriging replacing the censored data with the CL. Comparison of the loss in the mapping accuracy of these three methods revealed whether BME (methods a) better handles left-censored data than its kriging limiting cases (methods b and c).

2.6. Assessment of Impaired River Miles

The space/time distribution of chloride is governed by complex natural and physical processes. Imperfect knowledge about these complex processes result in a significant uncertainty in chloride estimation. Not accounting for estimation uncertainty in impairment assessment may lead to a wrong conclusion and hence accounting for uncertainty is considered to be an essential aspect of any decision making framework. Our river BME method is a geostatistical approach and as such its advantage is that it provides not only concentration estimates but also the probability that chloride exceeds a specific regulatory level. Using river BME, we calculated the probability that chloride exceeds the EPA guideline level of 230 mg/l along each of the 6018 river miles in the study area from 2005 to 2014, and we classified a given river reach as impaired if the average probability of exceedance of the EPA guideline level along that river reach is greater than 90 %, as non-assessed if that probability is between 10% and 90%, and clean if that probability is less than 10%. The average probability of exceedance along a river reach is calculated as the arithmetic average of the probability of exceedance calculated at equidistant points along that river reach.

3. Results and Discussion

3.1. Covariance Models of Offset-Removed Chloride log-Concentrations

Details about LUR analysis (R=0.6), the three offset models (Euclidean, river and LUR), and the weighted least square covariance fitting procedure used to obtain the covariance parameters for each offset model are available in the SI. The sill (i.e. variance) and the spatial covariance range for the Euclidean offset removed chloride log-concentrations are c_o = 0.41 (log-mg/l) ^{2} and a_r = 19 km (across land) for the Euclidean covariance model, and c_o = 0.41 (log-mg/l)

^{2} and a_r = 28 km (along rivers) for the river covariance model. For the river offset removed chloride log-concentrations, c_o = 0.25 (log-mg/l)^2 and a_r = 28 km (across land) for the Euclidean covariance model, and c_o = 0.25 (log-mg/l)^2 and a_r =36 km (along rivers) for the river covariance model. For the LUR offset removed chloride log-concentrations, c_o = 0.61 (log-mg/l)^2 and a_r = 58 km (across land) for the Euclidean covariance model, and c_o = 0.61 (log-mg/l)^2 and a_r =96 km (along rivers) for the river covariance model. The temporal range is a_t = 12 years for all covariance models.

3.2. Cross-Validation Results Contrasting the Euclidean versus River Covariance models

The cross validation results (Table 1) obtained in the baseline case (where none of the 390 values are censored) show that using an Euclidean offset (first row of Table 1), space/time BME using a river covariance better predicts chloride (R^2=0.711) than when using an Euclidean covariance (R^2=0.638), corresponding to an 11.44% percent change (PC) in R^2. This work is the first to demonstrate that the river covariance model is better than the Euclidean covariance model for chloride estimation in these subbasins. This means that the autocorrelation of chloride is best described using distances measured along the river network, which indicates that processes that are distributed along river networks (e.g. highways -a known source of chloride, vegetation buffers -a known attenuation process, etc.), are important drivers of the distribution of chloride along rivers.

Table 2.1: Leave-one-out cross-validation statistics obtained using the BME method with different offset and covariance models for the estimation of chloride log-concentration (*)

	Euclidean Covariance			**River Covariance**		
	MSE (log-mg/l)^2	**ME** (log-mg/l)	**R^2**	**MSE** (log-mg/l)^2	**ME** (log-mg/l)	**R^2**
Euclidean Offset	0.343	0.002	0.638	0.264	0.002	0.711
River Offset	0.224	0.003	0.760	0.194	0.018	0.789

(*) The Euclidean covariance and river covariance models use the Euclidean and river distance metrics, respectively. The Euclidean offset and the river offset use the Euclidean and river distance metrics, respectively; MSE is the mean squared error; ME is the mean error; R^2 is the squared coefficient of determination between observed and estimated values.

3.3. Cross-Validation Results Contrasting Euclidean versus River Offsets

Since we conclude in the baseline case that the covariance should be based on the river distance rather than the Euclidean distance, then the next question is whether the offset should also be calculated based on the river distance rather than the Euclidean distance. To answer that question we implemented space/time BME using our novel river offset (second row of Table 1). The only difference between the first and second row of table 1 is the introduction of the river offset, and by comparing these two rows we find that the river offset consistently outperforms the Euclidean offset. For example when using a river covariance (second column of Table 1), space/time BME using the river offset better predicts chloride (R^2 =0.789) than when using the Euclidean offset (R^2=0.711), corresponding to a 10.97% PC in R^2. Our work is the first to introduce the river offset and to demonstrate that it leads to an appreciable improvement over the Euclidean offset used in previous works.(Akita et al., 2007, and Money et al., 2011) The implication of this finding is that the river network topology should be taken into account for both the offset and covariance models. Doing so results in an overall PC in R^2 of 23.67 %, which considerably improves our ability to accurately predict chloride across space and time.

3.4. Sensitivity Analysis Results with respect to Censoring Limit

To assess sensitivity analysis of the estimation accuracy of the river BME and kriging methods with respect the proportion of censored data, we performed a cross validation analysis for 6 different proportions of censored data ranging from 0% (baseline case) to 46.2% of the overall data (figure 2). Each censored dataset was generated by selecting a CL, censoring all values below the CL and only providing the CL value. River BME rigorously models the uncertainty contained in censored data using the TGPDF, while the kriging methods simply treat them as data without any uncertainty since these data are replaced with half the CL, or with the CL. As expected, the estimation accuracy degrades with increasing proportion of censored data. However figure 2 clearly demonstrates that the rate of deterioration in estimation accuracy is lower for river BME (method a) than for its kriging linear limiting cases (method b and c). This trend can also be seen from the cross validation R^2 which indicates that BME improves the R^2 by a factor of about 2 to 7.5 over kriging (with censored data replaced by half the CL) when the proportion of censored data ranges from 13.6% to 46.2% (see SI for more details). Overall these results indicate that when a dataset includes censored data, then the BME method used in this work is consistently more efficient than the kriging method at extracting the information contained in these censored data.

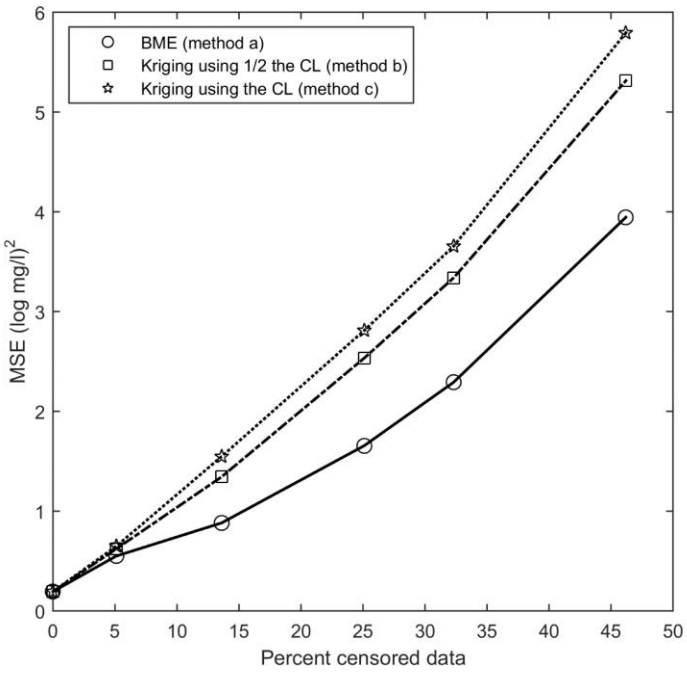


Figure 2.2: Cross validation MSE for river BME and its kriging linear limiting cases shown with respect to the proportion of censored data. BME (method a) rigorously models the uncertainty in the censored data using the TGPDF, while kriging treats them as data with no uncertainty by simply replacing them with half of the CL (method b) or by the CL (method c).

3.5. Cross Validation Results Contrasting the River and LUR Offsets

The LUR offset is obtained based on the average imperviousness in HUC12 subwatersheds, which is a weak predictor of chloride in our study area (R=0.6, see SI for more details). LUR is an integral part of many water quality models and is an attractive method because it takes advantage of seemingly free data (e.g. imperviousness calculated for other purposes), but in practice its implementation require dedicated modelers to preprocess these data, which can be time consuming for local regulatory agencies. The cross-validation statistics MSE increases from 0.194 (log-mg/l)^2 for the river offset BME method to 0.313 (log-mg/l)^2 for the LUR offset BME method, and the corresponding R^2 drops from 0.789 to 0.660. These cross-validation statistics indicates that using a LUR offset fails to produce better results than using the

river offset presented in this work, however LUR models with river buffers and temporally varying imperviousness maps may improve the LUR based approach.

3.6. Difference in the Maps Produced Using Euclidean versus River BME

To the best of our knowledge, previous studies have not compared, and quantified, the difference in estimated levels obtained using an Euclidean versus river BME methods in that situation. To address this question, we provide here a comparison of the Euclidean versus river BME maps in area B and area C (figure 1 depicts where areas B and C are located). The purpose of this comparison is purely to emphasize the difference in chloride estimates using Euclidean versus river BME along unsampled river reaches. These maps are not meant to compare the estimation accuracy of the Euclidean and river BME methods at unsampled locations.

The Euclidean BME and river BME maps for area B are shown in figures 3(a) and 3(b), respectively. In that area we are interested in the assessment of Bynum Run, which lacks monitoring data, and runs parallel to Winters Run where monitoring data are available. Figures 3(a) and 3(b) show that in this area major highways (a known source of chloride) are aligned along the river network. The river distance between the monitoring stations on Winter Run and estimation points on Bynum Run are long, resulting in a low autocorrelation in chloride measurements. The situation for the Euclidean BME model is the converse, the estimated values along Bynum Run are strongly affected by what's measured in Winters Run. Figures 3(a) and 3(b) show this difference in estimated chloride, and reveal that the chloride levels along Bynum Run are substantially higher in the Euclidean BME map (figure 3(a)) than in the river BME map (figure 3(b)). To quantify this difference, we calculate the number of river miles with estimates exceeding two thresholds of interest: 230 mg/l (an ambient water quality criteria for chloride

defined by the U.S. EPA (U.S. Environmental Protection Agency. Ambient water quality criteria for chloride., 1988)), and 145 mg/l (a concentration level at which declines in survival of salamanders have been documented (Stranko et al., 2013)). We find that according to Euclidean BME, 14% of Bynum Run river miles North of US 40 exceed 230 mg/l, and 62% of these river miles exceed 145 mg/l, whereas none of these river miles exceed either threshold limits according to river BME.

Similarly, the river BME estimates along the Grays and Cranberry Runs (figure 3(d)) are low as opposed to the high chloride estimates obtained with Euclidean BME (figure 3(c)). According to Euclidean BME, 9% of river miles along the Grays and Cranberry Runs exceed 230 mg/l, and 52% of these river miles exceed 145 mg/l, while none of these river miles exceed either threshold limits according to river BME.

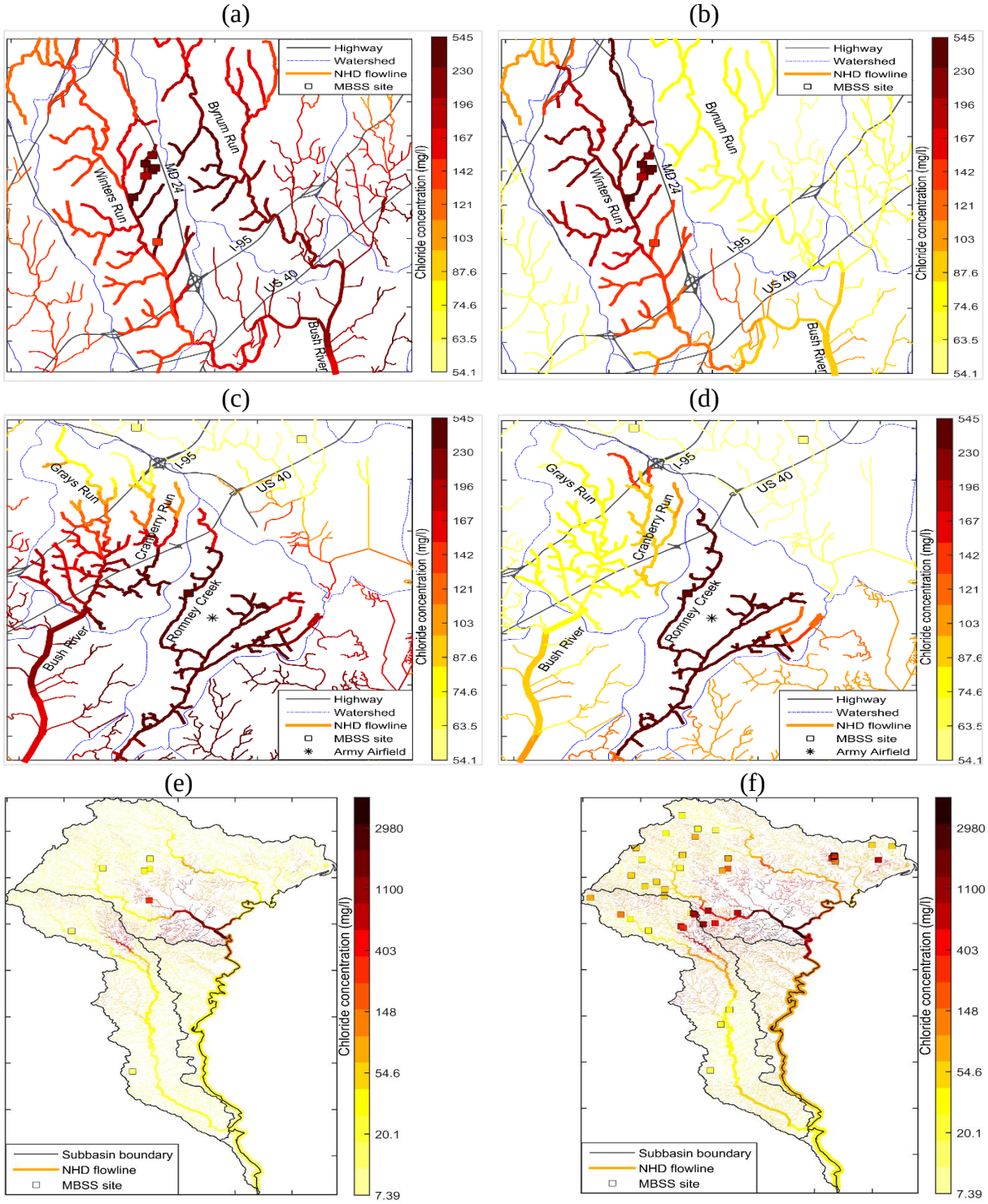


Figure 2.3: Maps of the BME mean estimate of chloride concentrations in 2014. The maps on the left panels are estimated using Euclidean BME, the maps on the right are estimated using river BME. Panel (a), (c) and (e) show the Euclidean BME estimate of chloride in area B, area C, and the study domain, respectively. The corresponding river BME maps are in the Panels (b), (d) and (g), respectively. The flow lines in panels (a), (b), (c), and (d) are highlighted (increased width) for better visual appearances of segments compared for estimation accuracy. The width of the flow lines in panels (e) and (f) correspond to their cumulative river miles.

These results demonstrate that there can be big differences in the estimated chloride concentration using Euclidean BME and river BME, which may lead to substantial differences in the assessment of whether a river reach is impaired. For example using the Euclidean approach one might conclude that Bynum Run and the Grays and Cranberry Runs are in need of remedial action, while using the river approach one might conclude that remedial action is less needed and added monitoring is desired. The implication of this finding is that using the proper approach does matter, and therefore one should use the river BME approach introduced in this work rather than the classical Euclidean approach when estimating chloride along unmonitored river miles. Another implication of this finding is that using river BME, one will delineate impaired areas that are confined along river reaches, as opposed to spread isotropically across land, which may be easier to remediate because resources will be targeted to a specific subwatershed, rather than spread across multiple subwatersheds.

3.7. Space/time Patterns in Chloride Contamination

The rate of urban development, changes in road salt application practices, and changing climate conditions may drive a variety of spatial and temporal patterns in chloride concentrations (Corsi et al., 2015). Accurate estimation of chloride is crucial to understand these patterns, to improve our understanding of the extent and nature of chloride contamination, and to design effective measures to control the chloride pollution. A series of chloride concentration maps from 2005 to 2014 are constructed using the space/time river BME method introduced in this study. The maps obtained for 2014 are shown in figure 3, while maps for other years are in SI. These maps provide the first representation of chloride distribution that fully integrates information about space/time variability and river network topology.

In the study area, the high population density area is made up of Baltimore and Columbia-Ellicott cities, which have a high concentration of impervious surfaces and are separated by a narrow green buffer along the Patapsco River. Conversely the surrounding area is generally green with localized concentrations of impervious surfaces where small towns are located.

Our river BME maps of chloride concentrations reveal that there are two distinct cores of chloride contamination corresponding to Baltimore and Columbia-Ellicott cities, which are persistently contaminated from 2005 to 2014. This indicates that once an area is contaminated it remains contaminated for a long time, which is consistent with what has been reported in previous studies (Harte and Trowbridge, 2010 and Perera et al., 2013) These two core areas are initially separated by a clean buffer along the Patapsco River. This buffer is revealed by the river BME estimation method as it accounts for river network topology. These two core areas are expanding outwards at a low rate during 2005-2009, resulting in a narrowing and eventual loss of the green buffer separating Baltimore and Columbia-Ellicott cities. There is a stagnation in 2010 and 2011, followed by an accelerated rate of outward expansion of the two core areas during 2012-2014 up until they coalesce in 2014, resulting in significant contamination over the whole Baltimore-Columbia--Ellicott urban area. Major factors for this significant urban-wide contamination may include increased rate of salt application, as well as the loss of green buffer separating Baltimore and Columbia--Ellicott cities.

Our river BME maps further reveal that at the beginning of the study period (2005) the concentration of chloride is low or inexistent in the streams located outside of the Baltimore-Columbia-Ellicott urban area. However in that area several pockets of high chloride concentration emerge in 2005-2009 and remain contaminated till the end of the study period

(2014). Each of these pockets can be visually detected using river BME because they are confined along distinct river branches, whereas it is more difficult to see them when using an Euclidean approach that averages out concentration across river branches. These pockets of contamination illustrate the usefulness of river BME to identify such areas so that they can be targeted for monitoring.

3.8. Probabilistic Assessment of Impaired River Miles

The probabilistic assessment of impaired river miles indicates that there are two distinguishable time periods (2005-2009 and 2011-2014) during which the fraction of unassessed and impaired river miles increased (figure 4). In the first time period the impaired river miles increased from 1.3% in 2005 to 3.5% in 2009, corresponding to a 0.55% rate of increase in impaired river miles per year. In second time period, the impaired river miles increased from 2.3% to 6%, corresponding to a 1.23% rate of increase in impaired river miles per year. These results demonstrate that there is a marked acceleration of the impairment of the study area, with a greater than two fold increase in the rate at which river miles become impaired. As stated earlier mechanisms causing this acceleration of impairment include the loss of buffer along the Patapsco River, the coalescence of core impaired areas, and the increased rate of chloride application. The implication of this finding is that there is sufficient evidence of increased impairment to justify taking strong measures to control chloride applications in these watersheds.

Interestingly, there is an even stronger acceleration in the unassessed river miles. There is a 1.05% and 3.17% rate of increase in unassessed river miles per year during the 2005-2009 and 2011-2014 periods, respectively. This dramatic acceleration of the rate of increase of unassessed river miles indicates that the monitoring effort, which in 2005 was sufficient to differentiate

between clean and impaired river miles, is becoming insufficient to fulfill its task, and increased monitoring is needed while chloride levels are rising. Hence the overall finding of our work is that there is an urgent need for increased monitoring in areas where chloride is unassessed, and these unassessed areas can efficiently be identified using the river BME approach.

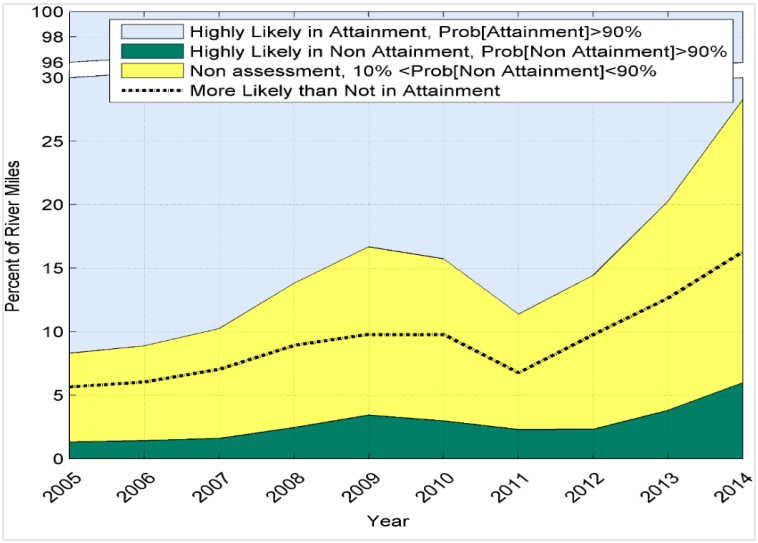


Figure 2.4: Time series of average fraction of river miles in Gunpowder-Patapsco, Patuxent, and Severn subbasins in Maryland that are highly likely in non-attainment (the probability of exceedance of the EPA guideline (230 mg/l) is greater than 90 %), non-assessed (probability between 10% and 90%), and highly likely in attainment (probability less than 10%) from 2005 to 2014. See Supplementary Information for maps showing for each year from 2005 to 2014 the spatial distribution of the probability that chloride exceeds 230 (mg/l).

4. Conclusions

This work is making an important methodological contribution for the assessment of water quality along rivers. It consists in the introduction of a river kernel smoothing function used to capture large distance scale variability in water quality. We find that when combined with geostatistical estimation of offset-removed concentrations, the river kernel smoothing is more accurate than earlier approaches that used Euclidean kernel smoothing.

This is because river kernel smoothing better captures river topology than Euclidean kernel smoothing. To our knowledge, this work is the first to perform a mapping analysis using the river kernel smoothing described here in a river geostatistical framework, and to demonstrate that it substantially improves mapping accuracy over an Euclidean approach. This approach is a contribution to the field of river geostatistics, and will be applicable to the estimation a wide range of river water quality parameters.

Another important contribution is our analysis of the mapping efficiency of the BME method of modern geostatistics when dealing with dataset with left censored data, as is the case when measurements are below the DL. We demonstrate that when a proportion of data is left censored, then BME always outperforms its kriging linear limiting case. This is a widely applicable finding of our work because there are many instances where environmental agencies have to measure trace level toxic constituents that have concentrations less than the DL of the measuring instruments. In such cases we recommend that these agencies use the full non-linear and non-Gaussian BME approach rather than arbitrarily setting the left censored data to half the CL or to the CL value.

Turning to the analysis of river chloride in Maryland, we find that there are big differences in the estimated chloride concentration using Euclidean BME versus river BME, particularly along unmonitored river reaches that run parallel to a river reach with monitoring data. We demonstrate that the differences in estimated chloride concentrations lead to substantial differences in the assessment of whether a river reach is impaired. Hence, an appropriate estimation method is important as estimates change the outcome of regulatory or policy decisions and the remediation strategy selected.

Using the river BME approach we find that chloride contamination in Maryland is characterized by wide contamination throughout Baltimore and Columbia-Ellicott cities, the disappearance of a clean buffer separating these two large urban areas, and the emergence of multiple localized pockets of contamination in surrounding areas. The number of impaired river miles increased by 0.55% per year in 2005-2009 and by 1.23% per year in 2011-2014, corresponding to a marked acceleration of the rate of impairment that justify taking strong measures to control chloride applications in these watersheds. We also find that the number of unassessed river miles has increased even more drastically over these periods, indicating the need of increased monitoring required as large clean areas become fragmented with pockets with persistently high chloride concentration. These unassessed pockets areas can efficiently be identified using the river BME approach for optimal sampling design for targeted monitoring. Since the river BME approach accounts for river network topology, the areas identified as unassessed are confined along specific river reaches, which will make regulatory effort more targeted and efficient.

Acknowledgements

This work was supported in part by grant number P42ES005948 of the National Institute of Environmental Health Sciences, and by NSF grant 1316318 as part of the joint NSF-NIH-USDA Ecology and Evolution of Infectious Diseases program.

REFERENCES

(1) Akita, Y., Carter, G., Serre, M.L., 2007. Spatiotemporal nonattainment assessment of surface water tetrachloroethylene in New Jersey. J. Environ. Qual. 36, 508–520. doi:10.2134/jeq2005.0426

(2) Brabec, E., Schulte, S., Richards, P.L., 2002. Impervious Surfaces and Water Quality: A Review of Current Literature and Its Implications for Watershed Planning. J. Plan. Lit. 16, 499–514. doi:10.1177/088541202400903563

(3) Chen, Y.C., Yeh, H.C., Wei, C., 2012. Estimation of river pollution index in a tidal stream using kriging analysis. Int. J. Environ. Res. Public Health 9, 3085–3100. doi:10.3390/ijerph9093085

(4) Christakos, G., 1990. A Bayesian/maximum-entropy view to the spatial estimation problem. Math. Geol. 22, 763–777. doi:10.1007/BF00890661

(5) Christakos, G., Li, X., 1998. Bayesian Maximum Entropy Analysis and Mapping: A Farewell to Kriging Estimators? Math. Geol. 30, 435–462. doi:10.1023/A:1021748324917

(6) Christakos, G., Serre, M.L., 2000. BME analysis of spatiotemporal particulate matter distributions in North Carolina. Atmos. Environ. 34, 3393–3406. doi:10.1016/S1352-2310(00)00080-7

(7) Church, P.E., And, Friesz, P.J., 1995. Delineation of a road-salt plume in ground water, and traveltime measurements for estimating hydraulic conductivity by use of borehole-induction logs, in: 5th Symposium, Minerals and Geotechnical Logging Society

(8) Corsi, S.R., De Cicco, L.A., Lutz, M.A., Hirsch, R.M., 2015. River chloride trends in snow-affected urban watersheds: Increasing concentrations outpace urban growth rate and are common among all seasons. Sci. Total Environ. 508, 488–497. doi:10.1016/j.scitotenv.2014.12.012

(9) Ganio, L.M., Torgersen, C.E., Gresswell, R.E., 2009. Geostatistical Approach in Stream Pattern for Describing Networks 3, 138–144. doi:10.1890/1540-9295(2005)

(10) Gardner, B., Sullivan, P.J., Lembo, Jr., A.J., 2003. Predicting stream temperatures:

geostatistical model comparison using alternative distance metrics. Can. J. Fish. Aquat. Sci. 60, 344–351. doi:10.1139/f03-025

(11) George Christakos, Patrick Bogaert, and M.S., 2001. Temporal Geographical Information Systems: Advanced Functions for Field-Based Applications, 2001 editi. ed. Springer.

(12) Harte, P.T., Trowbridge, P.R., 2010. Mapping of road-salt-contaminated groundwater discharge and estimation of chloride load to a small stream in southern New Hampshire, USA. Hydrol. Process. 24, 2349–2368. doi:10.1002/hyp.7645

(13) Helsel, D.R., 2005. Nondetects and data analysis. Statistics for censored environmental data.

(14) Kanevski, M., 2008. Advanced Mapping of Environmental Data. Wiley.

(15) Kelly, V.R., Lovett, G.M., Weathers, K.C., Findlay, S.E.G., Strayer, D.L., Burns, D.J., Likens, G.E., 2008. Long-term sodium chloride retention in a rural watershed: Legacy effects of road salt on streamwater concentration. Environ. Sci. Technol. 42, 410–415. doi:10.1021/es071391l

(16) LoBuglio, J.N., Characklis, G.W., Serre, M.L., 2007. Cost-effective water quality assessment through the integration of monitoring data and modeling results. Water Resour. Res. 43, 1–16. doi:10.1029/2006WR005020

(17) Messier, K.P., Akita, Y., Serre, M.L., 2012. Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. Environ. Sci. Technol. 46, 2772–2780. doi:10.1021/es203152a

(18) Messier, K.P., Campbell, T., Bradley, P.J., Serre, M.L., 2015. Estimation of Groundwater Radon in North Carolina Using Land Use Regression and Bayesian Maximum Entropy. Environ. Sci. Technol. 49, 9817–9825. doi:10.1021/acs.est.5b01503

(19) Money, E., Carter, G.P., Serre, M.L., 2010. Using River Distance in the Space/Time Estimation of Dissolved Oxygen Along Two Impaired River Networks in New Jersey. Water 43, 1948–1958. doi:10.1016/j.watres.2009.01.034

(20) Money, E.S., Carter, G.P., Serre, M.L., 2009. Modern space/time geostatistics using river distances: Data integration of turbidity and E. coli measurements to assess fecal contamination along the Raritan River in New Jersey. Environ. Sci. Technol. 43, 3736–3742. doi:10.1021/es803236j

(21) Money, E.S., Sackett, D.K., Aday, D.D., Serre, M.L., 2011. Using river distance and existing hydrography data can improve the geostatistical estimation of fish tissue mercury at unsampled locations. Environ. Sci. Technol. 45, 7746–7753. doi:10.1021/es2003827

(22) Perera, N., Gharabaghi, B., Howard, K., 2013. Groundwater chloride response in the Highland Creek watershed due to road salt application: A re-assessment after 20years. J. Hydrol. 479, 159–168. doi:10.1016/j.jhydrol.2012.11.057

(23) Peterson, E.E., Urquhart, S.N., 2006. Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: A case study in Maryland. Environ. Monit. Assess. 121, 613–636. doi:10.1007/s10661-005-9163-8

(24) Reyes, J.M., Serre, M.L., 2014. An LUR/BME framework to estimate PM2.5 explained by on road mobile and stationary sources. Environ. Sci. Technol. 48, 1736–1744. doi:10.1021/es4040528

(25) Rose, N., Cowie, C., Gillett, R., Marks, G.B., 2011. Validation of a spatiotemporal land use regression model incorporating fixed site monitors. Environ. Sci. Technol. 45, 294–299. doi:10.1021/es100683t

(26) Savelieva, E., Demyanov, V., Kanevski, M., Serre, M., Christakos, G., 2005. BME-based uncertainty assessment of the Chernobyl fallout. Geoderma 128, 312–324. doi:10.1016/j.geoderma.2005.04.011

(27) Singh, A., Nocerino, J., 2002. Robust estimation of mean and variance using environmental data sets with below detection limit observations. Chemom. Intell. Lab. Syst. 60, 69–86. doi:10.1016/S0169-7439(01)00186-1

(28) Stranko, S., Bourquin, R., Zimmerman, J., Kashiwagi, M., McGinty, M., Kauda, R., 2013. Do Road Salts Cause Environmental Impacts ? Maryl. DNR.

(29) Taylor-rogers, S., 1997. State of the Streams.

(30) U.S. Environmental Protection Agency. Ambient water quality criteria for chloride., 1988. Washington D.C. doi:EPA 440/5-88-001

(31) USGS Hydrography data, 2015. URL http://nhd.usgs.gov/data.html (accessed 10.6.15).

(32) Ver Hoef, J.M., Peterson, E., Theobald, D., 2006. Spatial statistical models that use flow and stream distance. Environ. Ecol. Stat. 13, 449–464. doi:10.1007/s10651-006-0022-8

(33) Yang, X., Jin, W., 2010. GIS-based spatial regression and prediction of water quality in river networks: A case study in Iowa. J. Environ. Manage. 91, 1943–1951. doi:10.1016/j.jenvman.2010.04.011

CHAPTER 3 (PAPER 2): A NOVEL GEOSTATISTICAL APPROACH COMBINING EUCLIDEAN AND GRADUAL-FLOW COVARIANCE MODELS TO ESTIMATE FECAL COLIFORM ALONG THE HAW AND DEEP RIVERS IN NORTH CAROLINA[2]

1. Introduction

Assessing water quality along rivers is vital for watershed management[1] and to protect public health. Geostatistical studies estimate river water quality using covariance models that characterize the spatial variability in surface waters[2 , 3 ,4]. Euclidean covariance models are successful in describing autocorrelation driven by terrestrial sources using the Euclidean (straight line) distance between points.[5,6,7,1] River covariance models use the river distance between points to account for the river network topology and have also been successful in many studies. [8 ,9,10 ,11,12,13] Flow-weighted covariance models[14,15,5,1] referred herein simply as flow covariance models, add physical meaningfulness by using both river distance and the ratio of flow between points. More specifically in the model introduced in 2006 by Ver Hoef et al. (2006),[14] Cressie et al.(2006)[15], de Fouquet and Bernard-Michel (2006),[16] and Bernard-Michel and de Fouquet (2006),[17] the flow function is constant along any river reach, and it is additive where two reaches combine. We will refer to this as the pipe-flow covariance model. In 2009, Money et al. (2009)[10] introduced a generalization of the flow covariance model that is based on a gradually varying flow along each river reach, which we will refer to as the gradual-flow covariance model.

[2] This chapter is under review in the Journal of Environmental Sciences and Technology. Jat P. and M.L. Serre, 2016. A Novel Geostatistical Approach Combining Euclidean and Gradual-Flow Covariance Models to Estimate Fecal Coliform along the Haw and Deep Rivers in North Carolina. . (Submitted to ES&T)

Surprisingly; very few studies have demonstrated an improvement in estimation accuracy using the pipe-flow covariance model[14,18], and to the best of our knowledge, no study has implemented the gradual-flow model.

The lack of success in using flow covariance models may have been due to conceptual limitations (for example because, for some water quality parameters, both Euclidean distance and flow connectivity are important factors), or to implementation challenges (for example in obtaining cumulated areas as a proxy for flow, or to calculate a pipe flow approximation of the underlying gradual flow). To address these issues we introduce here a novel hybrid Euclidean/Gradual-flow covariance model, and we demonstrate its use by performing a spatiotemporal estimation of fecal coliform along the Haw and Deep river in North Carolina from 2006 to 2010.

Fecal coliform, the most common microbiological contaminants of surface waters, are considered to be indicator organisms for the potential presence of disease-causing organisms that pose human health risks[19,20] . The EPA has proposed a guideline level of 200 CFU/100ml to limit risk of swimming-associated gastrointestinal illness[21]. Fecal coliform have been reported as exceeding the EPA guideline level in many river reaches of the Haw and Deep rivers in North Carolina,[22] and it is therefore critical to assess fecal coliform in the Haw and Deep rivers to better protect public health.

The distribution of fecal coliform in the Haw and Deep river is driven both by terrestrial sources (in areas with a high percentage of impervious surface in the headwater of this river system) and hydrological transport downstream of these sources. We therefore hypothesize that the hybrid Euclidean/Gradual-flow model will improve the estimation of fecal coliform compared to a purely Euclidean, purely river, or purely flow model. Furthermore, we

hypothesize that the Euclidean/Gradual-flow covariance model will better capture the effect of flow on spatial variability than the Euclidean/Pipe-flow model whenever using a coarse representation of the underlying river system.

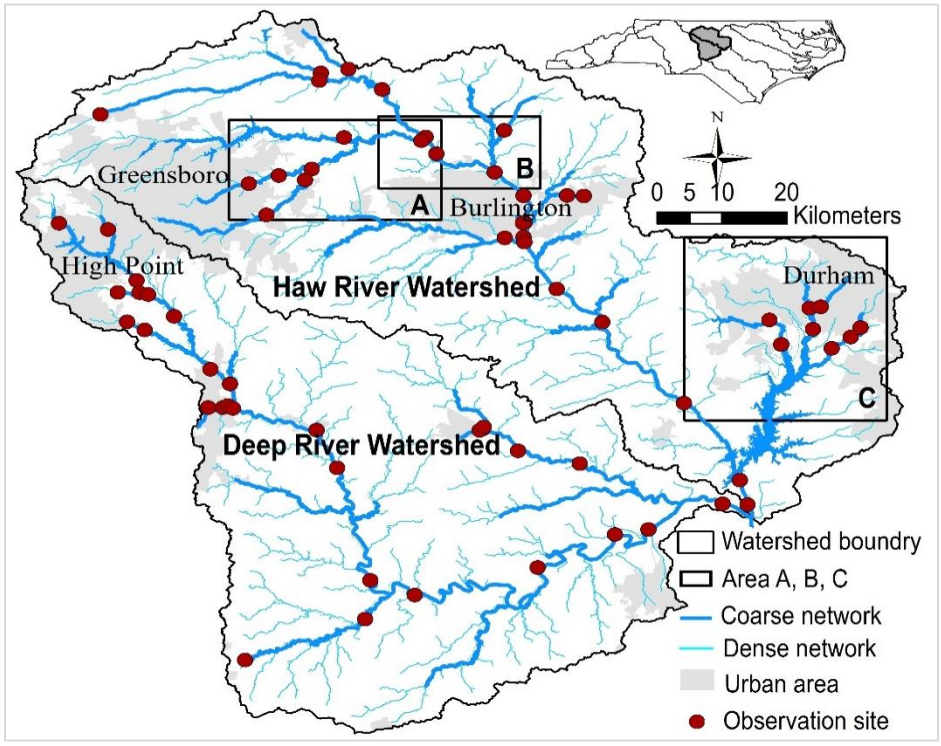
2. Materials and Methods

2.1. Fecal coliform and hydrography Data

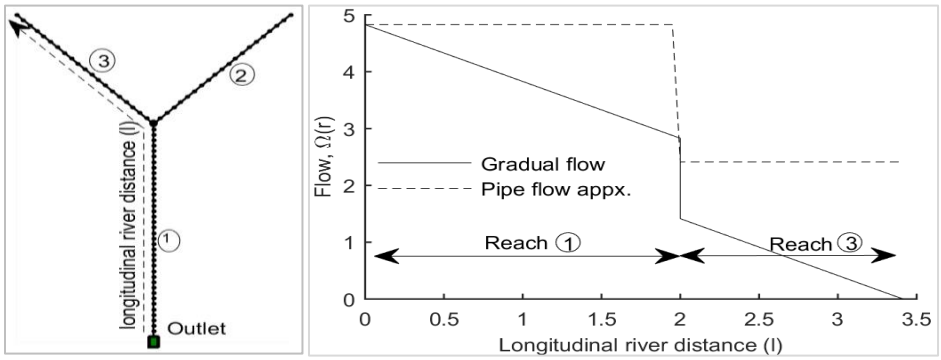
The fecal coliform concentration data for the Haw and Deep rivers were obtained from an existing monitoring network (managed by Cape Fear River Basin Monitoring Coalition) over a period of 2006−2010. Over this period there were a total of 69 unique observation sites (figure 1a) that were sampled for fecal coliform, resulting in 3848 space/time fecal coliform measurements ranging from 1 to 12500 CFU/100ml, with mean 723 CFU/100ml and standard deviation 2062 CFU/100ml (see table S1in Supplemental Information (SI) for additional statistics).

The Haw river (with a 1,526 square miles watershed area) and the Deep river (with a 1,441 square miles watershed area) are at the headwaters of the Cape Fear River basin, which is the largest watershed basin in North Carolina and discharges into the Atlantic Ocean. The river network along the Haw and Deep rivers is described based on stream lines (figure 1a) obtained from the USGS National Hydrography Data ("USGS Hydrography data," 2014, see SI for more details).

(a)


Greensboro
A
B
Burlington
High Point
Haw River Watershed
Durham
C
Deep River Watershed
N
0 5 10 20
Kilometers
Watershed boundry
Area A, B, C
Coarse network
Dense network
Urban area
Observation site

(b) (c)


③
②
①
longitudinal river distance (l)
Outlet
Flow, Ω(r)
Gradual flow
Pipe flow appx.
Reach ①
Reach ③
Longitudinal river distance (l)

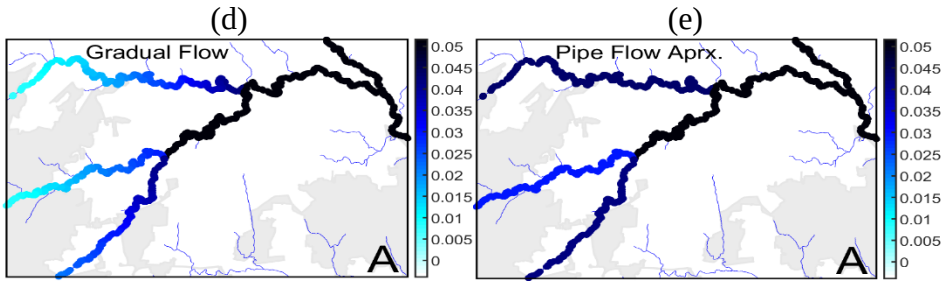


Figure 3.1: Panel (a) shows a map of the study area depicting the fecal coliform observation sites located in the Haw river and the Deep river watersheds. The thick stream lines represent the coarse river network, consisting mainly of the river reaches where monitoring sites are located, as well as their downstream reaches. The thin river lines show the additional upstream stream lines making up the dense river network. Panel (b) shows a fictitious coarse river network and panel (c) shows that its gradual flow along reaches 1 and 3 is poorly approximated by the corresponding pipe flow. Likewise, the gradual flow along the coarse river network shown in panel (d) for area A is poorly approximated by its corresponding pipe flow shown in panel (e). In particular the pipe flow along the upstream branches of area A are not able to reproduce well the gradual flow in these reaches.

2.2. Space/time Bayesian Maximum Entropy geostatistical framework

Our notation will consist in denoting a single random variable Z in capital letters, its realization, z, in lower case; and vectors in bold faces (e.g., $\boldsymbol{z} = [z_1,..., z_n]^T$). Using this notation we represent a space/time random field (S/TRF) as $Z(\boldsymbol{p})$, where $\boldsymbol{p} = (\boldsymbol{r}, t)$ is a space/time point, $\boldsymbol{r}$ is the spatial coordinate along the river network and t is time. We use Bayesian Maximum Entropy (BME), a space/time geostatistical estimation framework grounded in epistemic principles and its linear limiting case, kriging, to estimate fecal coliform log-concentration along rivers.[24,25,26,27]

The general BME framework used to estimate water quality at un-sampled location has been defined in several recent BME studies.[28,13] In brief, let $\boldsymbol{z}_d$ be the vector of log-concentrations observed at locations $\boldsymbol{p}_d$, let o_z be an known constant offset value[29] and let $\boldsymbol{x}_d = \boldsymbol{z}_d - o_z$ be the vector of offset removed log-concentrations. We define $X(\boldsymbol{p})$ as a homogenous/stationary S/TRF with realization $\boldsymbol{x}_d$, and we let $Z(\boldsymbol{p}) = X(\boldsymbol{p}) + o_z$.be the S/TRF

representing the distribution of fecal coliform log-concentrations. The knowledge base characterizing the S/TRF $X(\boldsymbol{p})$ includes its mean $m_x(\boldsymbol{p}) = E[X(\boldsymbol{p})]$, , where E[.] is the stochastic expectation operator, its covariance function $c_x(\boldsymbol{p}, \boldsymbol{p}') = E\left[\left(X(\boldsymbol{p}) - m_x(\boldsymbol{p})\right)\ \left(X(\boldsymbol{p}') - m_x(\boldsymbol{p}')\right)\right]$, and the data $\boldsymbol{x}_d$.

In this work we estimate fecal coliform log concentration at un-sampled locations using the space/time ordinary kriging limiting case of BME implemented in the *BMElib* numerical library[27] where the mean $m_x(\boldsymbol{p}) = m_x$ is assumed constant within the local estimation neighborhood, and the covariance is the product of its spatial and temporal components, i.e. $c_x(\boldsymbol{p}, \boldsymbol{p}') = c_x\left((\boldsymbol{r}, t), (\boldsymbol{r}', t')\right) = c_{spatial}(\boldsymbol{r}, \boldsymbol{r}')c_{temporal}(t, t')$. For the temporal component we use the stationary exponential model that is a function of time lag, i.e. $c_{temporal}(t, t') = \exp(-3\tau/a_t)$ where $\tau = |t - t'|$ is the time lag and a_t is the temporal covariance range. Choosing a spatial covariance model that is permissible for river networks is more intricate and several models are presented next.

2.3. Euclidean covariance model

Euclidean covariance models are a function of the Euclidean (or straight-line) distance between points, i.e. $c_E(\boldsymbol{r}, \boldsymbol{r}') = c_1(d_E(\boldsymbol{r}, \boldsymbol{r}'))$, where $d_E(\boldsymbol{r}, \boldsymbol{r}')$ is the Euclidean distance between locations $\boldsymbol{r}$ and $\boldsymbol{r}'$, and $c_1(.)$ can be any permissible covariance model for one dimensional fields. Euclidean covariance models can adequately describe the autocorrelation in water quality when contamination is from terrestrial sources distributed over long Euclidean distances and hydrological transport along the river is comparatively negligible.

There are many functions $c_1(.)$ that are permissible in R^1, such as the exponential, power, Gaussian, spherical, etc. In this work we use the exponential model for its high interpretability, for consistency with previous studies, and to facilitate comparison with other models, i.e. we use

$$c_E(\boldsymbol{r}, \boldsymbol{r}') = \sigma^2 \exp(-3\, d_E(\boldsymbol{r}, \boldsymbol{r}')/a_E) \quad (1)$$

where σ^2 is the variance and a_E is the Euclidean covariance range of the random field.

2.4. River covariance model

River covariance models are a function of the river distance $d_R(\boldsymbol{r}, \boldsymbol{r}')$ between any two locations ***r*** and ***r'***, i.e. the distance traveled along the river between these two locations. River covariance models incorporate the river network topology in their description of water quality autocorrelation, and they are adequate when pollution source is autocorrelated along rivers, such as when it is coming from elongated agricultural fields or roads that happen to follow the river topography.

However, simply replacing the Euclidean distance with the river distance in a covariance function can lead to a non-permissible covariance model. Therefore, the geostatistical modeling assumption of positive-definiteness needs to be assessed to ensure the validity of spatial covariance models using the river distance. It has been shown that the exponential covariance model using river distances is permissible for river network.[14,10] Hence river covariance models are of the form

$$c_R(\boldsymbol{r}, \boldsymbol{r}') = \sigma^2 \exp(-3\, d_R(\boldsymbol{r}, \boldsymbol{r}')/a_R) \quad (2)$$

where σ^2 is the variance and a_R is the river covariance range of the random field.

2.5. Flow-weighted covariance model using pipe flow

Flow-weighted covariance models account for both river topology and flow connectivity by incorporating river distance and flow in the covariance model. This kind of covariance models are useful to describe autocorrelation for persistent pollutants with autocorrelation driven by longitudinal transport over long downstream distances along rivers, and therefore for which dilution of the pollutant along a river is an important driver for the autocorrelation exhibited by that pollutant.

The flow-weighted covariance model is derived by first defining a spatial random field and then calculating its covariance. Let us identify a point ***r***=(***s***,*l*,*i*) on the river network either by its Euclidean coordinate ***s***=*{longitude, latitude}*; or by its river coordinate (*l*,*i*) consisting of the longitudinal coordinate (*l*) corresponding to the length of the continuous line connecting the river outlet to ***s*** along the river network (by convention, negative *l* values represent fictitious locations downstream of the outlet), and the reach index (*i*) uniquely defining the river reach where ***s*** is located. Ver Hoef et al. (2006)[14] and Cressie et al. (2006)[15] define the spatial random field $X(l, i)$ as the moving-average of a white noise random process, while de Fouquet and Bernard-Michel (2006)[16] and Bernard-Michel and de Fouquet (2006)[17] define $X(l, i)$ as the sum of uncorrelated one dimensional fields along each flow line (see SI for details). In both cases it can be shown that the covariance between $X(l, i)$ and $X(l', i')$ is zero when *i* and *i'* are not flow connected, and

when *i* is upstream of *i'* it is given by $cov(\boldsymbol{r}, \boldsymbol{r}') = \sqrt{\Omega(i(\boldsymbol{r}))/\Omega(i'(\boldsymbol{r}'))}\ c_1(d_R(\boldsymbol{r}, \boldsymbol{r}'))$, where $c_1(.)$ is the class of permissible covariance models in R^1, and the flow function $\Omega(i(\boldsymbol{r}))= \Omega(i)$ is constant along the reach *i* where $\boldsymbol{r}$ is located, and additive when two reaches combine (i.e. if two reaches *i* and *i*'' combine into a downstream reach *i'*, then $\Omega(i) + \Omega(i'') = \Omega(i')$). We refer to this flow as pipe flow since the flow at the river network outlet is the sum of the flows in all inlet (leaf) reaches, and we set $c_1(.)$ equal to the exponential model for interpretability and consistency with the other covariance models, so that the pipe-flow covariance model is given by

$$c_P(\boldsymbol{r}, \boldsymbol{r}') = \sigma^2\sqrt{\Omega(i(\boldsymbol{r}))/\Omega(i'(\boldsymbol{r}'))}\ \exp(-3\ d_R(\boldsymbol{r}, \boldsymbol{r}')/a_P) \quad (3)$$

where the subscript *P* emphasizes that pipe flow is used.

The flow ratio $\Omega(i)/\Omega(i')$ is a number between 0 and 1 expressing the proportion of flow in reach *i'* that is coming from its upstream reach *i*. This flow ratio captures the effect of dilution from side river reaches contributing side flow between reaches *i* and *i'*. However a limitation of this model is that it assumes that no flow is gradually added along a given river reach. In truth the flow gradually increases along each river reach, and therefore pipe flows are only an approximation of the underlying gradually varying flow.

Several approaches were proposed to calculate pipe flows. Cressie et al. (2006)[15] proposed calculating the flow using stream order. According to their method, all leaf reaches in the river network are set to equal to 1 and then stream orders are added when streams merge at confluence nodes. This leads to an additive pipe flow that is easy to calculate; however it does not provide an approximation of the underlying gradual flow. Ver Hoef et al. (2006)[14] proposed a

method based on the reach proportional influence (PI), which they define as the proportion of the flow that a given reach contributes at its downstream confluence node. Money et al. (2009)[10] showed that PIs can be used to approximate the underlying gradual flow by setting the pipe flow for any reach equal to the multiplication of the PI of that reach and those downstream of it. For example in figure 1(b), if reach 3 contributes 50% of the flow at its downstream end, then its PI is 0.5, and its pipe flow is half of that in reach 1.

In this work we will use the Ver Hoef et al. (2006)[14] pipe flow because it provides an (almost perfect) approximation of the underlying gradual flow when the river network is dense enough so that the sum of flows at the leaf reaches is almost equal to the outlet flow. However, it has two limitations: (1) the approximation breaks down for coarse river networks such as that shown in figure 1(b) where obviously the pipe flow (dashed line in figure 1(c)) is a rather poor approximation of the underlying gradual flow (plain line), and (2) perhaps more importantly, the calculation of PIs requires some expertise in river topology and flow connectivity, which can be a problem when the computer language script used to calculate PIs becomes outdated. Therefore we seek an alternative that can truly accommodate gradual flows and is easy to implement.

2.6. Flow-weighted covariance model using gradual flow

Money et al. (2009)[10] introduced a generalization of the flow-weighted covariance model that rigorously accounts for flows that gradually increase along river reaches. They defined $\omega(\mathbf{r})$ as a positive density function characterizing the flow entering the river per unit length along the river network. Then, the flow function $\Omega(\boldsymbol{r})$ can simply be obtained by integrating the flow density along all river reaches upstream of $\boldsymbol{r}$, i.e.

$$\Omega(\boldsymbol{r}) = \int_{u \in U(r)} dl(\boldsymbol{u})\, \omega(\boldsymbol{u}) \tag{4}$$

where $U(\boldsymbol{r})$ is the set of points upstream of $\boldsymbol{r}$, and $l(\boldsymbol{u})$ is the longitudinal coordinate of point $\boldsymbol{u}$. The $\omega(\boldsymbol{r})$ is usually nonzero and positive throughout the river network, and as a result the flow function $\Omega(\boldsymbol{r})$ gradually increases with $\boldsymbol{r}$ in the direction of flow, as opposed to the pipe flow approximation where flow density is zero and the flow along any given river reach remains constant.

Based on the flow density $\omega(\boldsymbol{r})$ and corresponding flow $\Omega(\boldsymbol{r})$ Money et al (2009) [10] define the spatial random field $X(\boldsymbol{r})$ as

$$X(\boldsymbol{r}) = \int_{u \in U(r)} dl(\boldsymbol{u}) \sqrt{\omega(\boldsymbol{u})/\Omega(\boldsymbol{r})}\; W(\boldsymbol{u})\, Y(l(\boldsymbol{r})) \tag{5}$$

where $W(\boldsymbol{u})$ is a white noise process, $Y(l(\boldsymbol{r})$ is a zero mean random process with covariance $c_1(h),\ h = |l - l'|$ is the river distance, and $c_1(h)$ can be any permissible covariance function, which as noted earlier we set equal to the exponential model. Then the covariance between $X(\boldsymbol{r})$ and $X(\boldsymbol{r}')$ is zero when *i* and *i'* are not flow connected, and when $\boldsymbol{r}$ is upstream of $\boldsymbol{r}'$ it is given by (see SI for details)

$$c_G(\boldsymbol{r}, \boldsymbol{r}') = \sigma^2 \sqrt{\Omega(\boldsymbol{r})/\Omega(\boldsymbol{r}')}\; \exp(-3\, d_R(\boldsymbol{r}, \boldsymbol{r}')/a_G) \tag{6}$$

where the subscript *G* emphasizes that the flow ratio $\Omega(\boldsymbol{r})/\Omega(\boldsymbol{r}')$ gradually varies with $\boldsymbol{r}$ and $\boldsymbol{r}'$.

This generalization is physically realistic and meaningful because it lets the flow function $\Omega(\boldsymbol{r})$ gradually increase along each river reach in the direction of flow. Several gradually varying functions can be used for $\Omega(\boldsymbol{r})$, including historical flow, cumulated area, cumulated river length, etc. Since our goal is to test a function that gradually varies along river reaches but is also easy to implement, we chose to use cumulated river length because this is the easiest function to implement (by simply setting $\omega(\boldsymbol{r})$ to 1). If we find that using Eq 6 with $\Omega(\boldsymbol{r})$ equal to the cumulated river length improves estimation; then this will be of tremendous benefit to practitioners, because this completely eliminates the laborious task of estimating historical flows or processing a digital terrain model to calculate cumulated areas that match a user's river network.

2.7. Hybrid Euclidean-flow covariance model

The Euclidean covariance model better describes the effect of terrestrial sources processes while the flow covariance model better describes the effect of longitudinal hydrological transport. When both processes act simultaneously a hybrid Euclidean and flow covariance models may be most appropriate. Mathematically, a combination of two permissible covariance models is also permissible. Hence in this work we define the hybrid Euclidean/gradual-flow covariance model as the linear combination of the Euclidean and gradual-flow covariance models, respectively, i.e.

$$c_{EG}(\boldsymbol{r},\boldsymbol{r}') = \alpha_E\sigma^2 \exp(-3\, d_E(\boldsymbol{r},\boldsymbol{r}')/a_E) + \alpha_G\sigma^2\sqrt{\Omega(\boldsymbol{r})/\Omega(\boldsymbol{r}')}\, \exp(-3\, d_R(\boldsymbol{r},\boldsymbol{r}')/a_G) \quad (7)$$

where α_E and α_G are the proportions of contribution from the Euclidean and gradual-flow covariance models, respectively, such that $\alpha_E + \alpha_G = 1$. The same rules are followed to define the Euclidean/Pipe-flow covariance model $c_{EP}(\boldsymbol{r},\boldsymbol{r}')$.

2.8. Calculating experimental covariance values

The covariance of the offset removed fecal coliform log-concentration S/TRF $X(\boldsymbol{r},t)$ is modeled by first calculating experimental covariance values $\hat{c}_X$ based on the measurement data $\boldsymbol{x}_d = [x_1, \ldots, x_n\,]$ of the S/TRF, and then fitting a covariance model to these experimental covariance values. In this work the experimental covariance $\hat{c}_X$ between $X(\boldsymbol{r},t)$ and $X(\boldsymbol{r}',t')$ is potentially of a function of the Euclidean lag $d_E(\boldsymbol{r},\boldsymbol{r}')$, river lag $d_R(\boldsymbol{r},\boldsymbol{r}')$, flow ratio $f = \Omega(\boldsymbol{r})/\Omega(\boldsymbol{r}')$, and time lag $\tau = |t - t'|$, hence the experimental covariance for a given Euclidean lag d_E, river lag d_R, flow ratio f, and time lag τ is calculated as

$$\hat{c}_X(d_E, d_R, f, \tau) = \frac{1}{N(d_E,d_R,f,\tau)} \sum_{i=1}^{N(d_E,d_R,f,\tau)} x_{head,i}\, x_{tail,i} - {m_X}^2 \quad (8)$$

where $N(d_E, d_R, f, \tau)$ is the number of pairs of values $(x_{head,i}x_{tail,i})$ separated by a Euclidean lag d_E, river lag d_R, flow ratio f and time lag τ, and m_X is the mean of the $\boldsymbol{x}_d$ data.

The space/time covariance model that we use is space/time separable. We first model its temporal component by plotting $\hat{c}_X(0,0,0,\tau)$ as a function of the temporal lag τ, and we then do

a least-square fitting of the exponential temporal covariance model to obtain the temporal covariance range a_t. We can then focus on how the spatial component of the covariance varies with Euclidean lag, river lag, and flow ratio, which is the primary focus of this work. To do this it is useful to plot $\hat{c}_X(d_E, d_R, f, 0)$ as a function of Euclidean lag d_E for fixed values of the river lag and flow ratio, and then as a function the flow ratio f for fixed Euclidean and river lags, in order to understand the relative contribution of the Euclidean and flow covariance models to the overall spatial variability of the offset removed fecal coliform log-concentration. This type of exploratory covariance analysis is, to our knowledge, novel, and widely applicable not only to our study but to any other river water quality studies. Finally, we can obtain the parameters of any of the candidate spatial covariance model (Euclidean, river, flow and Euclidean/flow) by doing a least square fitting of that model with the spatial experimental covariance values.

2.9. Model performance evaluation and assessment of river miles with high fecal coliform

Model performance is evaluated by conducting a leave-one-out cross-validation (LOOCV) for each covariance model, calculating the Mean Square Error (MSE), Mean Error (ME) and R^2, and selecting the covariance model with the smallest MSE (see SI for details). In order to contrast the pipe flow and gradual flow models, this LOOCV is conducted on a coarse river network consisting mainly of the river reaches where monitoring sites are located, as well as their downstream reaches.

Assessment of river miles with high fecal coliform is done in two stages. First we calculate the BME mean and variance of fecal coliform concentration at equidistant estimation points along all river reaches. Second, we assess an estimation point as having high fecal

coliform if the probability that fecal coliform exceeds 200 CFU/100ml is greater than 90%, i.e. *Prob*[*FC*>200CFU/100ml]>90%.[6,13]

3. Results and Discussion

3.1. The hybrid Euclidean/Gradual-flow estimates are more accurate than those obtained using a purely Euclidean or purely flow-weighted covariance model

Cross-validation results were obtained for the coarse river network using the Euclidean, River, Gradual-flow, Pipe-flow, Euclidean/Gradual-flow, and Euclidean/Pipe-flow covariance models (Table 1). The cross validation statistics (MSE, ME and R^2) indicate that fecal coliform estimates obtained using the Euclidean covariance model are more accurate than estimates obtained using the river covariance and flow-weighted covariance models, suggesting that terrestrial sources are the dominant factor in the fecal contamination along rivers.

However, the estimation is improved when using a hybrid Euclidean/Gradual-flow covariance consisting 70% of the Euclidean covariance model and 30% of the flow-weighted covariance model, as shown by the 12.4% reduction in MSE achieved by the Euclidean/Gradual-flow model compared to that of the Euclidean model. This indicates that in fact both terrestrial source and hydrological transport play an important role in the distribution of fecal contamination along rivers, and therefore the best way to incorporate flow in a geostatistical estimator is through a hybrid Euclidean/flow model rather than a purely Euclidean or purely flow-weighted covariance model. To the best of our knowledge, this is the first case study demonstrating improved estimation accuracy using a hybrid Euclidean/Gradual-flow covariance model, which is widely applicable to many surface water quality studies.

Table3.1: Leave-one-out cross-validation statistics and corresponding covariance parameter values obtained using the (E) Euclidean, (R) River, (G) Gradual-flow, (P) Pipe-flow, (EG) Euclidean/Gradual-flow, and (EP) Euclidean/Pipe-flow covariance models. All results were obtained based on the coarse river network. For the E, R, G and P models, $Sill_1$ and $Range_1$ are the covariance sill (σ^2) and range (a_E, a_R, a_G and a_P for the E, R, G and P model respectively) obtained through least square fitting. For the EG and EP models $Sill_1 = \alpha_E\sigma^2$ and $Range_1 = a_E$ are the covariance sill and range of the Euclidean model, and $Sill_2$ and $Range_2$ are the covariance sill and range of the flow covariance model (e.g. $Sill_2 = \alpha_G\sigma^2$ and $Range_2 = a_G$ for the EG model). For the EG and EP models, α_E and α_G (or α_P) are obtained by selecting the α_E that minimizes the cross-validation MSE (See SI for more details on covariance modeling). For both the EP and EG models, the α_E=70%. In all models, the temporal range is $a_t = 30$ days.

Covariance type	MSE*	ME**	R^2	$Sill_1$*	$Range_1$†	$Sill_2$*	$Range_2$†
Euclidean (E)	1.716	0.007	0.446	2.940	88	-	
River (R)	1.760	0.037	0.434	2.940	155	-	
Gradual-flow (G)	2.243	0.112	0.283	2.940	1554	-	
Pipe-flow (P)	2.185	0.025	0.299	2.940	1554	-	
EG (70%/30%)	1.504	0.018	0.494	2.058	164	0.882	155
EP (70%/30%)	1.573	0.014	0.477	2.058	162	0.882	155

* (CFU/100 ml) ^{2}
** CFU/100 ml
† (km)

3.2. When using a coarse river network, the Euclidean/gradual-flow estimates are more accurate than the Euclidean/pipe-flow estimates

The pipe flow was found to be an almost perfect approximation of the gradual flow for the dense river network. However, this approximation significantly deteriorates for a coarse river network, as seen in area A on figure 1d-e (see SI for additional details), indicating that the pipe flow is a poor approximation of the gradual flow for coarse river networks.

To investigate the effect of this finding, we calculated experimental covariance values for various classes of Euclidean lags, river lags, and flow ratios (see SI for details), and we found that the increase in covariance with respect to flow ratio is statistically significant (p-value<0.05)

when using gradual flow, but this increase is not significant (p-value>0.05) when using pipe flow. This finding demonstrates that the gradual-flow covariance model better captures and integrates the effect of flow on the spatial variability of fecal coliform along rivers than the pipe-flow covariance model when a coarse network is used.

The significant increase in covariance with respect to gradual flow ratio leads to a better estimation of fecal coliform concentration using the Euclidean/gradual-flow covariance model (MSE=1.504) compared to the Euclidean/pipe-flow covariance model (MSE=1.573). The implication of this finding is that gradual flow should be used instead of pipe flow whenever estimating fecal coliform along rivers.

3.3. Fecal coliform concentrations vary over long spatial distances and short time scales

We calculated fecal coliform along the dense river network for each observation day from 2006 to 2010. Maps of the estimates obtained on 12-Jun-2006 using the Euclidean and Euclidean/Gradual-flow models are shown in figures 2a and 2b, respectively. Maps and animations for other days are shown in SI. As a whole, these maps show that the fecal coliform concentrations vary over long spatial distances that cover a significant portion of the study domain, but that can change in the matter of a few days, as is apparent in the spatial and temporal covariance ranges shown in table 1. This indicates that contamination events are sporadic in time, but when they occur they are wide spread across space, and therefore the maps provide an effective tool to target areas where measures are needed to protect the public health.

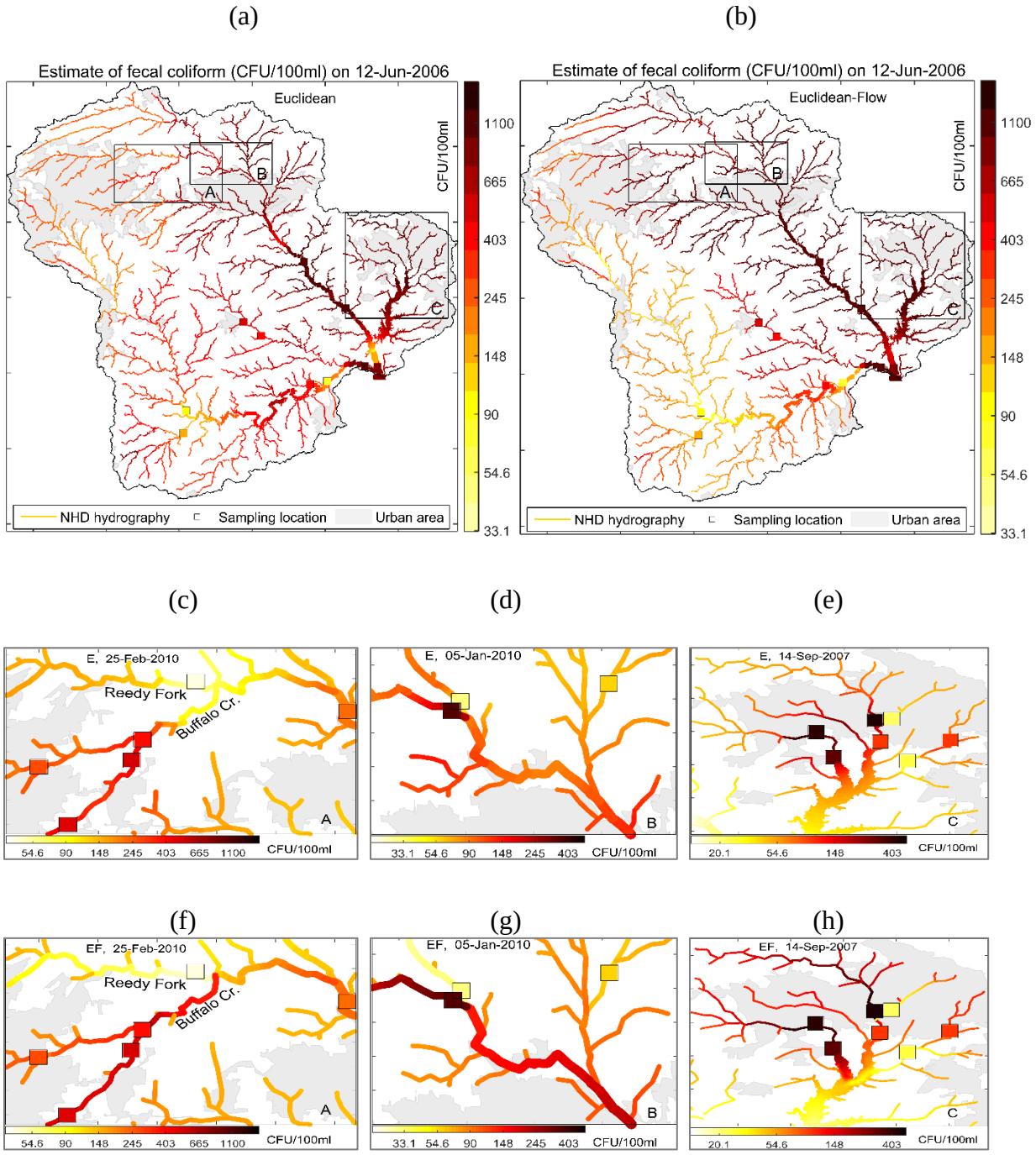


Figure 3.2: Maps of fecal coliform estimates (CFU/100ml) obtained on 12-Jun-2006 across the study area are shown in panels (a) and (b), those obtained on 25-Feb-2010 over area A are shown in panels (c) and (f), those obtained on 05-Jan-2010 over area B are shown in panels (d) and (g), and those obtained on 14-Sep-2007 over area C are shown in panels (e) and (h). Estimates obtained using the Euclidean covariance model are shown in panels (a), (c), (d) and (e) while those obtained using the Euclidean/Gradual-flow covariance model are shown in panels (b), (f), (g), and (h).

3.4. Euclidean/Gradual-flow estimates reveal that fecal contamination is more watershed specific and covers more river miles than traditionally thought

Our novel Euclidean/Gradual-flow map for 12-Jun-2006 (figure 2b) shows that the contaminated area is more watershed specific (i.e. contamination stays within a watershed) than what is seen in the Euclidean map (figure 2a). The watershed specific nature of contamination is physically meaningful due to hydrologic transport, and supported by the monitoring data. Indeed, as seen in figures 2a-b, monitoring data in a watershed are likewise values, and differ from those in a different watershed (this is also seen for other dates, see SI).

By accounting for the watershed specific nature of contamination, the Euclidean/Gradual-flow model reveals that contamination remains autocorrelated over much longer distances (covariance ranges a_E=164km and a_G=155km) than what is estimated based on the Euclidean model (a_E=88km). This means that the Euclidean/Gradual-flow model is able to capture contamination over more river miles within a specific watershed than the Euclidean model is able to do.

3.5. Euclidean/Gradual-flow estimates capture hydrological transport

The Euclidean/Gradual-flow estimates along Buffalo Creek in area A (figure 2f) are high because they are flow connected with 4 high measured values observed upstream of Buffalo Creek. This shows that the Euclidean/Gradual-flow model captures the hydrological transport of fecal contamination along this river reach, whereas the Euclidean model fails to do so (figure 2c). Furthermore, the Euclidean/Gradual-flow estimate abruptly changes at the confluence node where Buffalo Creek merges with Reedy fork, thereby capturing the effect of dilution past the

confluence node. This effect happens at all confluence nodes, and we also show it in areas B (figures 2d and 2g) and C (figures 2e and 2h) for demonstration purposes.

This illustrates how the Euclidean/Gradual-flow model better captures the effect of hydrological transport than the Euclidean model, and explains why, given the same monitoring data, more river miles will be identified as contaminated by the Euclidean/Gradual-flow model than by the Euclidean model.

3.6. The Euclidean/Gradual-flow model substantially increases sensitivity in the detection of fecal impairment

There are 573 sampling days in 2006-2010, during which at least one sample was collected. For each river mile, we counted the number of sampling days assessed as having high fecal coliform (i.e. with $Prob[FC>200CFU/100ml]>90\%$). We found that across the study area, and more specifically across areas A, B and C located in areas with high percent of impervious surface, the number of sampling days assessed as having high fecal coliform was consistently greater for the Euclidean/Gradual-flow estimates compared to the Euclidean estimates (figure 3). We furthermore assessed a river mile as being impaired if it had more than 60 sampling days assessed as having high fecal coliform out of a total of 573 sampling days. We found that 96 river miles were detected as being impaired according to the Euclidean/Gradual-flow method, which is more than twice than the 39 river miles found according to the Euclidean estimate (see SI for additional similar results). This demonstrates that the Euclidean/Gradual-flow model more than doubles the sensitivity in the detection of fecal impairment in the Haw and Deep rivers, and our map reveals that this impairment occurs primarily in the headwaters of the river system where a high percentage of surface is impervious.

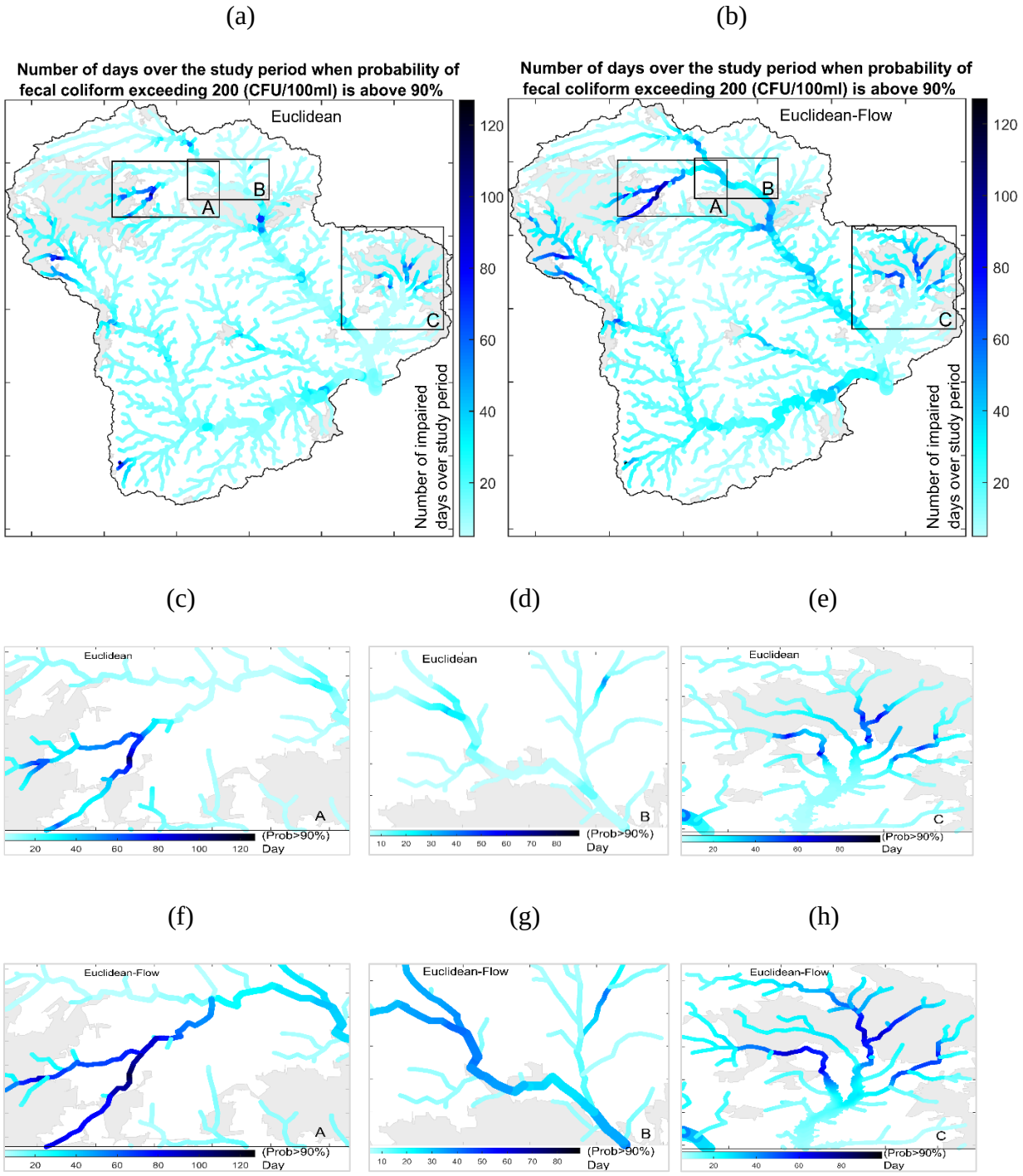


Figure 3.3 : These maps show, for each river mile, the number of sampling days (out of a total of 573 sampling days in 2006-2010) assessed as having high fecal coliform (i.e. with Prob[FC>200CFU/100ml]>90%). The study area is shown in panels (a) and (b), area A is shown in panels (c) and (f), area B is shown in panels (d) and (g), and area C is shown in panels (e) and (h). Estimates obtained using the Euclidean covariance model are shown in panels (a), (c), (d) and (e) while those obtained using the Euclidean/Gradual-flow covariance model are shown in panels (b), (f), (g), and (h).

3.7. Concluding remarks and future works

There have been very few studies that successfully used the flow covariance model.[14 1] This may be for a variety of reasons, including implementation difficulties. Our novel approach fills a critical need because it is the first case study to implement the gradual flow covariance model, to demonstrate improved estimation accuracy using a hybrid Euclidean/Gradual-flow covariance model, and, more critically, it removes several barriers in implementation by (a) using cumulated river length as a proxy for flow (which removes the cumbersome processing of digital terrain models to calculate cumulated areas), (b) it uses gradual flow (which removes the need to calculate a pipe flow approximation), and (c) it performs well regardless of the coarseness of the network used to model the river system.

Using our novel approach, we created the first geostatistical maps of fecal coliform that capture variability associated with both terrestrial sources and hydrological transport and that increase the number of river miles where fecal impairment is detected in the Haw and Deep rivers. These maps provide a critical tool to assess fecal impairment and to take measures to protect the public health.

Future works include the application of our novel model to other river systems and pollutants, the investigation of the tradeoffs in using various proxies for flow, and the integration of land use and weather variables in the estimation framework.

Acknowledgements

This work was supported in part by grant number P42ES005948 of the National Institute of Environmental Health Sciences, and by NSF grant 1316318 as part of the joint NSF-NIH-USDA Ecology and Evolution of Infectious Diseases program.

REFERENCES

(1) Isaak, D. J.; Peterson, E. E.; Ver Hoef, J. M.; Wenger, S. J.; Falke, J. A.; Torgersen, C. E.; Sowder, C.; Steel, E. A.; Fortin, M.-J.; Jordan, C. E.; et al. Applications of spatial statistical network models to stream data. *WIREs Water* 2014, *1* (June), 277–294.

(2) Delhomme, J. P. Kriging in the hydrosciences. *Adv. Water Resour.* 1978, *1* (5), 251–266.

(3) Jager, H. I.; Sale, M. J.; Schmoyer, R. L. Cokriging to assess regional stream quality in the Southern Blue Ridge Province. *Water Resour. Res.* 1990, *26* (7), 1401–1412.

(4) Curriero, F. C. On the use of non-euclidean distance measures in geostatistics. *Math. Geol.* 2006, *38* (8), 907–926.

(5) Peterson, E. E.; Urquhart, S. N. Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: A case study in Maryland. *Environ. Monit. Assess.* 2006, *121* (1–3), 613–636.

(6) Akita, Y.; Carter, G.; Serre, M. L. Spatiotemporal nonattainment assessment of surface water tetrachloroethylene in New Jersey. *J. Environ. Qual.* 2007, *36* (2), 508–520.

(7) LoBuglio, J. N.; Characklis, G. W.; Serre, M. L. Cost-effective water quality assessment through the integration of monitoring data and modeling results. *Water Resour. Res.* 2007, *43* (3), 1–16.

(8) Gardner, B.; Sullivan, P. J.; Lembo, Jr., A. J. Predicting stream temperatures: geostatistical model comparison using alternative distance metrics. *Can. J. Fish. Aquat. Sci.* 2003, *60* (3), 344–351.

(9) Money, E.; Carter, G. P.; Serre, M. L. Using River Distances in the Space/Time Estimation of Dissolved Oxygen Along Two Impaired River Networks in New Jersey. *Water Res.* 2009, *43* (7), 1948–1958.

(10) Money, E. S.; Carter, G. P.; Serre, M. L. Modern space/time geostatistics using river distances: Data integration of turbidity and E. coli measurements to assess fecal contamination along the Raritan River in New Jersey. *Environ. Sci. Technol.* 2009, *43* (10), 3736–3742.

(11) Money, E. S.; Sackett, D. K.; Aday, D. D.; Serre, M. L. Using river distance and existing hydrography data can improve the geostatistical estimation of fish tissue mercury at unsampled locations. *Environ. Sci. Technol.* 2011, *45* (18), 7746–7753.

(12) Yang, X.; Jin, W. GIS-based spatial regression and prediction of water quality in river networks: A case study in Iowa. *J. Environ. Manage.* 2010, *91* (10), 1943–1951.

(13) Jat, P.; Serre, M. L. Bayesian Maximum Entropy space/time estimation of surface water chloride in Maryland using river distances. *Environ. Pollut.* 2016.

(14) Ver Hoef, J. M.; Peterson, E.; Theobald, D. Spatial statistical models that use flow and stream distance. *Environ. Ecol. Stat.* 2006, *13* (4), 449–464.

(15) Cressie, N.; Frey, J.; Harch, B.; Smith, M. Spatial prediction on a river network. *J. Agric. Biol. Environ. Stat.* 2006, *11* (2), 127–150.

(16) de Fouquet, C.; Bernard-Michel, C. Modeles geostatistiques de concentrations ou de debits le long des cours d'eau. *Comptes Rendus - Geosci.* 2006, *338* (5), 307–318.

(17) C. Bernard-Michel and C. de Fouquet1. Construction of valid covariances along a hydrographic network . Application to specific water discharge on the Moselle Basin. 2006, No. 1, 1–4.

(18) Peterson, E. E.; Merton, A. A.; Theobald, D. M.; Urquhart, N. S. Patterns of spatial autocorrelation in stream water chemistry. *Environ. Monit. Assess.* 2006, *121* (1–3), 569–594.

(19) Soller, J. A.; Schoen, M. E.; Bartrand, T.; Ravenscroft, J. E.; Ashbolt, N. J. Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. *Water Res.* 2010, *44* (16), 4674–4691.

(20) Heaney, C. D.; Sams, E.; Dufour, A. P.; Brenner, K. P.; Haugland, A.; Chern, E.; Wing, S.; Marshall, S.; Love, D. C.; Noble, R.; et al. Fecal indicators in sand, sand contact, and risk of enteric illness among beachgoers. *Epidemiology* 2012, *23* (1), 95–106.

(21) USEPA. *Ambient water quality criteria for bacteria*; 1986.

(22) NC Department of Environment and Natural Resources Division. *North Carolina Water Quality Assessment and Impaired Waters List (2006 Integrated 305(b) and 303(d) Report)*; 2006; Vol. 305.

(23) USGS Hydrography data http://nhd.usgs.gov/data.html (accessed Oct 6, 2015).

(24) Christakos, G. *Modern Spatiotemporal Geostatistics. Oxford University Press*; Oxford University Press: New York, 1990.

(25) Serre, M. L.; Christakos, G. Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge - the Equus Beds study. *Stoch. Environ. Res. Risk Assess.* 1999, *13* (1–2), 1–26.

(26) Christakos, G.; Serre, M. L. BME analysis of spatiotemporal particulate matter distributions in North Carolina. *Atmos. Environ.* 2000, *34* (20), 3393–3406.

(27) George Christakos, Patrick Bogaert, and M. S. *Temporal Geographical Information Systems: Advanced Functions for Field-Based Applications*, 2001 editi.; Springer, 2001.

(28) Messier, K. P.; Akita, Y.; Serre, M. L. Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ. Sci. Technol.* 2012, *46* (5), 2772–2780.

(29) Messier, K. P.; Campbell, T.; Bradley, P. J.; Serre, M. L. Estimation of Groundwater Radon in North Carolina Using Land Use Regression and Bayesian Maximum Entropy. *Environ. Sci. Technol.* 2015, *49* (16), 9817–9825.

CHAPTER 4 (PAPER 3): SPACE/TIME ESTIMATION OF DISSOLVE ORGANIC CARBON ALONG RIVERS IN MARYLAND USING A COMBINATION OF EUCLIDEAN AND FLOW-WEIGHTED COVARIANCE MODELS[3]

1. Introduction

Dissolved organic carbon (DOC) is operationally defined as organic molecules that pass through a filter, most often 0.45 µm. DOC is an important constituent of water quality because it affects the physical, chemical, and biological condition of freshwater ecosystems. DOC is a significant energy source for aquatic life in stream and river waters (Wetzel et al., 1995). It absorbs biologically harmful ultraviolet rays (Williamson et al., 1996). DOC acts as a weak acid and binds dissolved substances, such as metals, making them temporarily less bioavailable (Driscoll et al., 1995) (Prusha and Clements, 2004). However, excess DOC can release pesticides from particulate agricultural residue matter in suspension (Worrall et al. 1997) and form harmful by-products with disinfectants during drinking water treatment processes (Chu et al., 2002). Therefore high DOC levels may be harmful and it is important to estimate DOC along all river miles to assess where levels may be in exceedance of safe levels.

The distribution of DOC across a river network is influenced by two major processes. First the concentration of DOC is strongly influenced by the terrestrial sources of DOC. Soil, groundwater, and dead terrestrial plant material are major sources of DOC (Wetzel et al., 1995,

[3] This chapter is under manuscript preparation for the Journal of Water research. Jat P. and M.L. Serre, 2016. Space/Time Estimation of Dissolve Organic Carbon along rivers in Maryland using a Combination of Euclidean and Flow-weighted Covariance models. (In preparation: Water Research)

and overland flow through wetlands and organic soil layers contributes significant DOC concentrations into proximal streams and rivers (Mulholland et al., 2008). This process results in a spatial distribution of stream DOC levels that follows the terrestrial landscape. Second, once DOC reaches streams it is transported downstream over distances that may be non-negligible. Several studies have reported that river DOC concentrations typically increase with increasing flow discharges (Hobbie and Likens, 1973). Volk et al. (Volk et al., 2002) found that DOC concentration could increase by as much as 3 fold when discharge also increases by 3 fold in a small stream. High DOC concentrations at high discharge provide conditions under which hydrological transport may occur over some distance downstream of areas where DOC is released in the stream waters. Hence the spatial variability of DOC is governed by both terrestrial sources and longitudinal transport.

Estimating DOC concentration along all river miles of the Gunpowder-Patapsco, Patuxent, and Severn sub-basins in Maryland is vital to assess impairment of this river system. Assessing river impairment is critical in informing watershed management and in taking appropriate measures where DOC levels are high. For example it is important for water utilities using surface waters to know where and when levels are in excess of the 3mg/L advisory level because high DOC may lead to the formation of carcinogenic disinfectants by-products in the treated water.

Monitoring all river miles is not feasible because it is too costly and too time consuming for environmental agencies. In practice only limited monitoring data are available, and geostatistical methods provide the most cost effective methodological approach to assess all river miles based solely on limited monitoring data. The key defining feature of geostatistical methods is the covariance model used to describe the variability of surface water quality along the river

system. The covariance models that have been successfully used in previous surface water quality studies are the Euclidean model (Peterson and Urquhart, 2006)·(Akita et al., 2007)·(LoBuglio et al., 2007)·(Isaak et al., 2014) and the river covariance model (Gardner et al., 2003) ·(E. Money et al., 2009)·(E. S. Money et al., 2009) ·(Money et al., 2011)·(Yang and Jin, 2010)·(Jat and Serre, 2016), which are based on the Euclidean and river distances, respectively. However since hydrologic transport is an important factor governing the spatial distribution of DOC, it would make sense to use a covariance model that accounts for flow when estimating DOC, else important characteristics of the spatial distribution of DOC may be misrepresented. In 2006 Ver Hoef et al (Ver Hoef et al., 2006) and others (Cressie et al., 2006)) introduced a covariance model that uses flow. The introduction of this flow covariance model was a breakthrough; however, surprisingly, very few studies have been successful in implementing that model and demonstrating an improvement in estimation accuracy (Ver Hoef et al., 2006) (Peterson et al., 2006). In fact Peterson and Urquhart (Peterson and Urquhart, 2006) compared the Euclidean and the flow covariance models in the estimation of DOC in Maryland, and they found the Euclidean model estimates were more accurate than those obtained with the flow covariance model. Hence currently the best available method available to assess DOC in our study area is the Euclidean model, however this model does not account for hydrological transport and therefore lacks physical meaningfulness.

The goal of this work is to address this critical issue by implementing a spatiotemporal geostatistical approach that will incorporate flow in the geostatistical estimation of DOC across our study domain over multiple years. To do this we will use the Euclidean/Gradual-flow approach recently presented in Jat and Serre (in review), which uses a hybrid covariance model that includes both Euclidean distance and flow in the estimation process. Our hypothesis is that

this novel approach will result in maps that are more accurate and physically meaningful than past maps.

2. Materials and Methods

2.1. DOC and hydrography data

A total of 391 space/time TOC concentration values were obtained from the Maryland Biological Stream Survey (MBSS) dataset from 2005 to 2014 in stream waters located in the Gunpowder-Patapsco, Severn, and Patuxent sub-basins (figure 1). The concentration values ranged from 0.192 mg/l to 19.034 mg/l, with mean 1.7272 mg/l and standard deviation 1.7440 mg/l. The river network in the Gunpowder-Patapsco, Severn, and Patuxent sub-basins is described based on stream lines (figure 1) obtained from the USGS National Hydrography Data ("USGS Hydrography data," 2014)

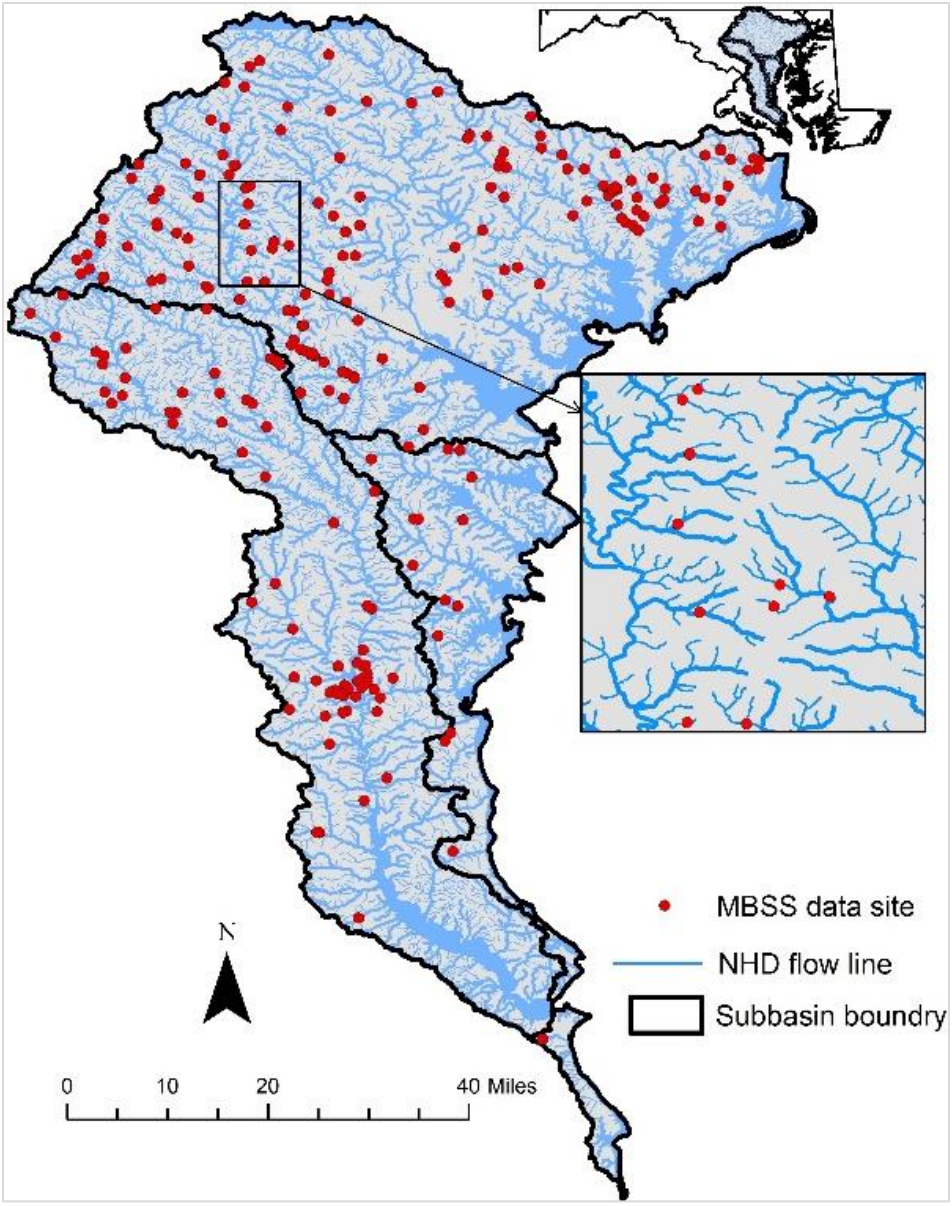


Figure 4.1: Map of the study area depicting Maryland Biological Stream Survey (MBSS) monitoring sites in the Gunpowder-Patapsco, Patuxent, and Severn sub-basins in Maryland.

2.2. Space/time Bayesian Maximum Entropy framework

Estimation of TOC log concentrations is made using the ordinary kriging limiting case of the Bayesian Maximum Entropy (BME) method and its *BMElib* numerical implementation (Serre and Christakos, 1999), (Christakos and Serre, 2000), (George Christakos, Patrick Bogaert,

2001). We provide here a short description of the implementation of BME to estimate TOC log concentrations, and more details is available elsewhere (Messier et al., 2012), (Jat and Serre, 2016), Jat and Serre (in review).

As explained in Jat and Serre (in review) we denote a random variable Z in capital letters, its realization, z, in lower case; and vectors in bold faces (e.g., $\boldsymbol{z} = [z_1,..., z_n]^T$). We denote a space/time random field (S/TRF) as $Z(\boldsymbol{p})$, where $\boldsymbol{p} = (\boldsymbol{r}, t)$ is a space/time point, $\boldsymbol{r}$ is the spatial coordinate along the river network and t is time. Let $\boldsymbol{z}_d$ be the vector of log-concentrations observed at locations $\boldsymbol{p}_d$, let o_z be an known constant offset value (Messier et al., 2015) and let $\boldsymbol{x}_d = \boldsymbol{z}_d - o_z$ be the vector of offset removed log-concentrations. We define $X(\boldsymbol{p})$ as a homogenous/stationary S/TRF with realization $\boldsymbol{x}_d$, and we let $Z(\boldsymbol{p}) = X(\boldsymbol{p}) + o_z$.be the S/TRF representing the distribution of fecal coliform log-concentrations. The knowledge base characterizing the S/TRF $X(\boldsymbol{p})$ includes its mean $m_x(\boldsymbol{p}) = E[X(\boldsymbol{p})]$, where E[.] is the stochastic expectation operator, its covariance function $c_x(\boldsymbol{p}, \boldsymbol{p}') = E\big[\big(X(\boldsymbol{p}) - m_x(\boldsymbol{p})\big)\ \big(X(\boldsymbol{p}') - m_x(\boldsymbol{p}')\big)\big]$, and the data $\boldsymbol{x}_d$.

In ordinary kriging the mean $m_x(\boldsymbol{p}) = m_x$ is assumed constant within the local estimation neighborhood. Following Jat and Serre (in review) we select a space/time covariance model equal to the product of a purely spatial and purely temporal components, i.e. $c_x(\boldsymbol{p}, \boldsymbol{p}') = c_x\big((\boldsymbol{r}, t), (\boldsymbol{r}', t')\big) = c_{spatial}(\boldsymbol{r}, \boldsymbol{r}')c_{temporal}(t, t')$. For the temporal component we use the stationary exponential model that is a function of time lag, i.e. $c_{temporal}(t, t') = \exp(-3\tau/a_t)$ where $\tau = |t - t'|$ is the time lag and a_t is the temporal covariance range. The spatial covariance model deserves special attention and is described next.

2.3. Spatial covariance model

In this work we primarily implement the Euclidean, River, Gradual-flow, and the Euclidean/Gradual-flow exponential covariance models described in details in Jat and Serre (in review).

In brief, the Euclidean and river covariance models use the Euclidean and river distances, respectively, and they are defined as

$$c_E(\boldsymbol{r},\boldsymbol{r}') = \sigma^2 \exp(-3\, d_E(\boldsymbol{r},\boldsymbol{r}')/a_E) \quad (1)$$

and

$$c_R(\boldsymbol{r},\boldsymbol{r}') = \sigma^2 \exp(-3\, d_R(\boldsymbol{r},\boldsymbol{r}')/a_R) \quad (2)$$

where $c_E(\boldsymbol{r},\boldsymbol{r}')$ and $c_R(\boldsymbol{r},\boldsymbol{r}')$ are the Euclidean and river covariance models, respectively, σ^2 is the variance, $d_E(\boldsymbol{r},\boldsymbol{r}')$ and $d_R(\boldsymbol{r},\boldsymbol{r}')$ are the Euclidean and river distances, respectively, and a_E is a_R are the Euclidean and river covariance ranges, respectively.

The Gradual-flow covariance model (E. S. Money et al., 2009) uses both river distance and a flow function $\Omega(\boldsymbol{r})$ that gradually increases in the direction of flow. Following Jat and Serre, in review we use the upstream cumulated length as a proxy for the gradual flow because it is easy to obtain, which greatly facilitates the implementation of this model by practitioners. The Gradual-flow covariance between $X(\boldsymbol{r})$ and $X(\boldsymbol{r}')$ is zero when $\boldsymbol{r}$ and $\boldsymbol{r}'$ are not flow connected, and when $\boldsymbol{r}$ is upstream of $\boldsymbol{r}'$ it is given by

$$c_G(\boldsymbol{r},\boldsymbol{r}') = \sigma^2 \sqrt{\Omega(\boldsymbol{r})/\Omega(\boldsymbol{r}')}\, \exp(-3\, d_R(\boldsymbol{r},\boldsymbol{r}')/a_G) \quad (3)$$

where the flow ratio $\Omega(\boldsymbol{r})/\Omega(\boldsymbol{r}')$ quantifies the proportion of the downstream flow that is coming from the upstream point, and the Gradual-flow covariance range a_G is the distance over which

TOC is autocorrelated along a reach when the flow ratio is one (i.e. in the absence of dilution from tributaries), which is indicative of the distance that TOC travels downstream from a source.

The hybrid Euclidean/gradual-flow covariance model (Jat and Serre, in review) is the linear combination of the Euclidean and gradual-flow covariance models, i.e.

$$c_{EG}(\boldsymbol{r},\boldsymbol{r}') = \alpha_E\sigma^2 \exp(-3\, d_E(\boldsymbol{r},\boldsymbol{r}')/a_E) + \alpha_G\sigma^2\sqrt{\Omega(\boldsymbol{r})/\Omega(\boldsymbol{r}')}\, \exp(-3\, d_R(\boldsymbol{r},\boldsymbol{r}')/a_G) \quad (4)$$

where α_E and α_G are the proportions of contribution from the Euclidean and gradual-flow covariance models, respectively, such that $\alpha_E + \alpha_G = 1$.

2.4. Calculating experimental covariance values and selecting covariance parameters

The experimental covariance value $\hat{c}_X$ between $X(\boldsymbol{r}, t)$ and $X(\boldsymbol{r}', t')$, where $\boldsymbol{r}$ and $\boldsymbol{r}'$ are separated by the Euclidean lag $d_E(\boldsymbol{r},\boldsymbol{r}')$, river lag $d_R(\boldsymbol{r},\boldsymbol{r}')$, flow ratio $f = \Omega(\boldsymbol{r})/\Omega(\boldsymbol{r}')$, and time lag $\tau = |t - t'|$ is calculated using the equation

$$\hat{c}_X(d_E, d_R, f, \tau) = \frac{1}{N(d_E,d_R,f,\tau)} \sum_{i=1}^{N(d_E,d_R,f,\tau)} x_{head,i}\, x_{tail,i} - {m_X}^2 \quad (5)$$

where $N(d_E, d_R, f, \tau)$ is the number of pairs of offset-removed log-concentration TOC values ($x_{head,i}x_{tail,i}$) separated by a Euclidean lag d_E, river lag d_R, flow ratio f and time lag τ, and m_X is the mean of the offset-removed log-concentration TOC data.

The parameters (sill and range) of the Euclidean, river, and Gradual-flow spatial covariance models are then obtained by doing a least square fitting of these covariance models onto experimental covariance values $\hat{c}_X(d_E, d_R, f, 0)$ obtained for various values of d_E, d_R, and f, and a temporal lag τ equal to zero.

In the case of the hybrid Euclidean/Gradual flow spatial covariance model, we fix α_E to a value between 0 and 1, we set α_G to $1 - \alpha_E$, and we then obtain the corresponding sill, Euclidean range and Gradual-flow range by least square fitting of the hybrid model onto the experimental covariance values $\hat{c}_X(d_E, d_R, f, 0)$. We then need to decide what is proper value for α_E. For that we simply select the α_E that results in the lowest mean square error in a leave-one-out cross-validation.

Finally the temporal covariance range is obtained by fitting the temporal component of the space/time covariance model to experimental covariance values $\hat{c}_X(0,0,0,\tau)$ obtained for pairs of offset-removed log-concentration TOC values $(x_{head,i}x_{tail,i})$ that are spatially collocated and separated by various termporal lags τ.

2.5. Accuracy of model estimates and probabilistic assessment of DOC impaired river miles

The accuracy of given estimation model is evaluated by doing a leave-one-out cross-validation (LOOCV) analysis consisting in removing each DOC measured value, and re-estimating that values from the remaining data. The model with the lowest Mean Square Error (MSE) is the most accurate model. Other useful validation statistics are the Mean Error (ME) characterizing consistent bias, and the R^2 characterizing precision.

The most accurate model is used to perform a probabilistic assessment of DOC impairment at each river mile along the river system. The ordinary kriging estimate of DOC and its corresponding estimation error variance are calculated at equidistant estimation points along all river reaches. A given river mile is then identified as impaired if at that river mile the

probability that the DOC concentration exceeds 3 mg/l is greater than 90%, and as unassessed if that probability is between 0.1 and 0.9 (Akita et al., 2007).

3. Results and Discussion

3.1. The Euclidean model is more accurate than the river and the flow models, indicating that terrestrial sources is the primary driver of DOC variability along rivers

The LOOCV statistics (MSE, ME and R^2) obtained for the Euclidean (E), River (R), Gradual-flow (G) models and tabulated in table 1. These cross-validation results show that the DOC estimates obtained using the Euclidean covariance model are more accurate than estimates obtained using the purely river covariance or purely flow-weighted covariance model. The Euclidean estimates explains approximately 60.7% of space/time variability in DOC concentrations whereas the purely river and the purely Gradual-flow estimates explain only 58.2% and 41.3% that variability, respectively. This finding indicates that terrestrial sources of DOC are the primary factor driving the spatial variability of DOC along rivers. This result is in agreement with the Peterson et al. (2006)'s finding that the Euclidean covariance model better predicts the spatial distribution of DOC along rivers compared to the flow-weighted covariance model in a purely spatial analysis for only one year, and extends that result in the context of a spatiotemporal analysis conducted over 10 years.

Table 4. 1: Leave-one-out cross-validation statistics and corresponding covariance parameter values obtained using the (E) Euclidean, (R) River, (G) Gradual-flow, (EG) Euclidean/Gradual-flow covariance models. For the E, R, and G models, $Sill_1$ and $Range_1$ are the covariance sill (σ^2) and range (a_E, a_R, a_G for the E, R, G model respectively) obtained through least square fitting. For the EG model $Sill_1 = \alpha_E\sigma^2$ and $Range_1 = a_E$ are the covariance sill and range of the Euclidean model, and $Sill_2 = \alpha_G\sigma^2$ and $Range_2 = a_G$ are the covariance sill and range of the flow covariance model. For the EG model, α_E and α_G are obtained by selecting the α_E that minimizes the cross-validation MSE, resulting in α_E=80% and α_G=20%. In all models, the temporal range is a_t =7 years

Covariance type	MSE*	ME**	R^2	$Sill_1$*	$Range_1$†	$Sill_2$*	$Range_2$†
Euclidean (E)	0.273	0.036	0.607	0.677	36.3		
River (R)	0.289	0.014	0.582	0.677	98.2		
Gradual flow (G)	0.422	-0.001	0.413	0.677	981.8		
EG (80%/20%)	0.226	0.021	0.676	0.542	43.6	0.135	981.8

* (log ml/l) ^{2}
** (log ml/l)
† (km)

3.2. The hybrid Euclidean/Gradual-flow model is the most accurate model, indicating that flow plays a role in the distribution of DOC along rivers

The α_E value for the Euclidean/Gradual-flow (EG) covariance model was determined by setting α_E to a fixed value chosen from 0 to 1 by increment of 0.05, performing a LOOCV analysis to obtain the corresponding MSE, and selecting the α_E with the smallest MSE. As shown in figure 2a the MSE clearly changes with α_E, and the minimum MSE of 0.226(log ml/l)^2 is obtained for α_E=80% and $\alpha_G = 1 - \alpha_E$ =20%.

The covariance parameter values we obtained for α_E=80% and α_G =20% are $a_E = 36.3\ km$ and $a_F = 981.2\ km$ (Table 1), and the corresponding covariance model is shown in figure 2b as a function Euclidean lag for a fixed river lag and fixed flow ratios, and in figure 2c as a function of flow ratio for fixed Euclidean and river lags.

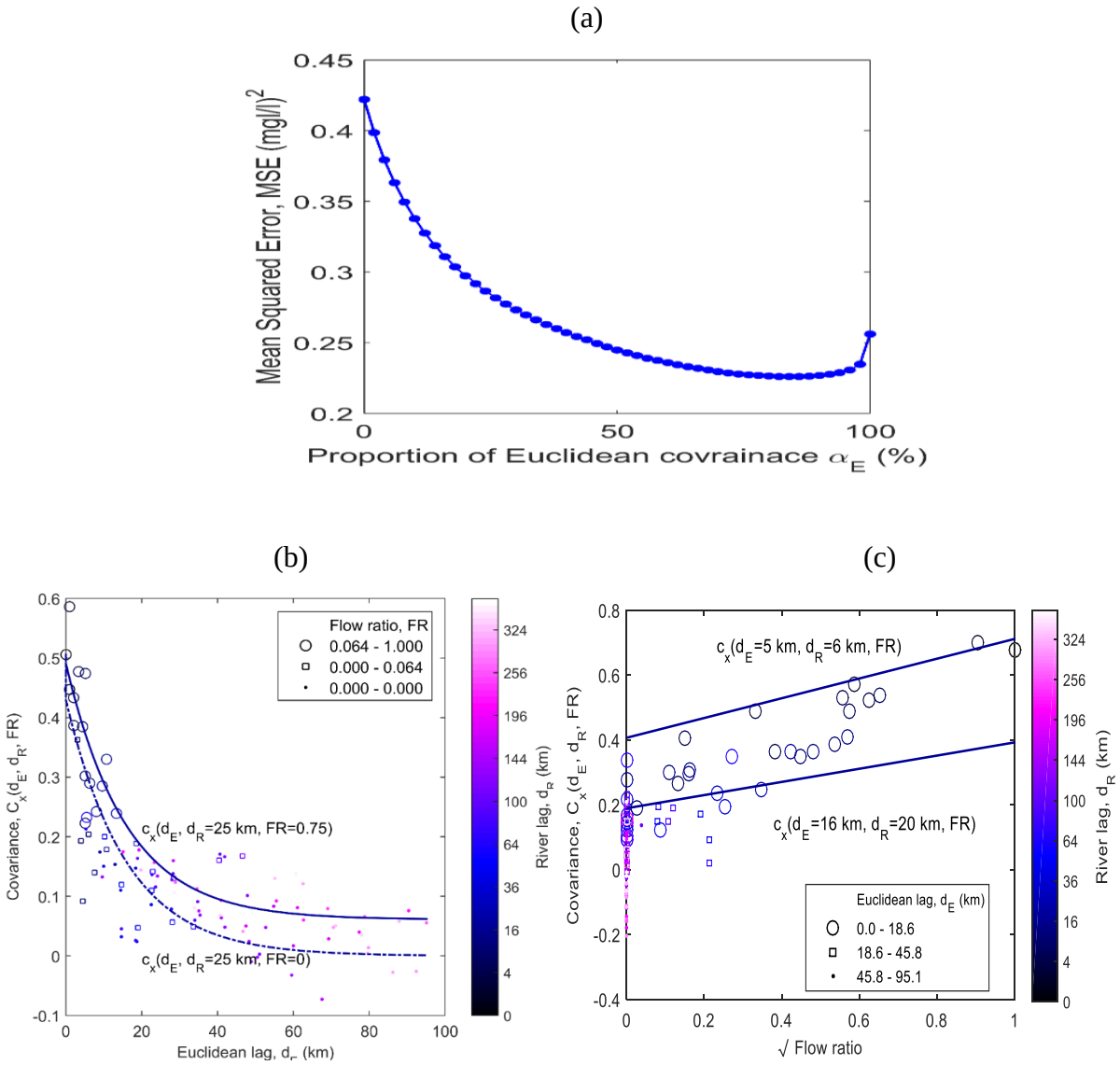


Figure 4.2: (a) Plot of the MSE as a function of α_E, the proportion of the Euclidean component in the hybrid Euclidean/Gradual-low covariance model. Experimental covariance values (markers) and Euclidean/Gradual-flow covariance model (lines) shown as a function of (b) Euclidean lag for a fixed river lag and fixed flow ratios, and (c) as a function of flow ratio for fixed Euclidean and river lags.

The LOOCV statistics obtained for the Euclidean/Gradual-flow (EG) covariance model with α_E=80% are added alongside those of the purely E, R and G models in Table 1. These cross-validation results demonstrate that the hybrid Euclidean/Gradual-flow estimates are the most accurate amongst all models. The hybrid Euclidean/Gradual-flow model explains 67.6% of the space/time variability in DOC concentrations as opposed to the 60.7% explained by the

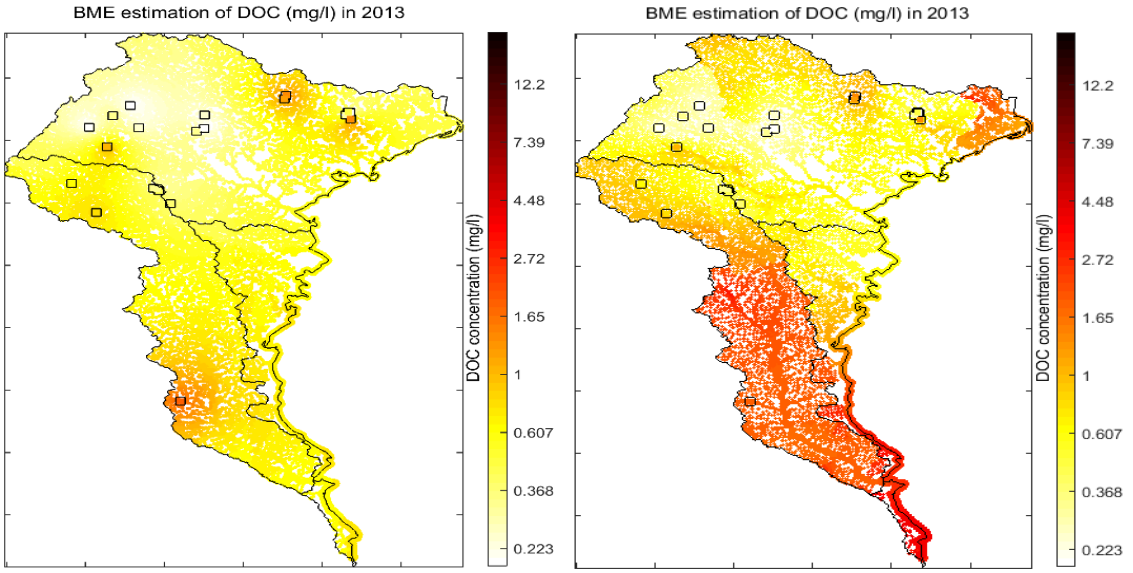
Euclidean model. The MSE of the Euclidean/Gradual-flow model is 17% lower than that of the Euclidean model. This appreciable decrease in estimation error indicates that in fact both terrestrial source and hydrological transport play an important role in the distribution of DOC along rivers, and therefore the best way to incorporate flow in a geostatistical estimator is through a hybrid Euclidean/flow model rather than a purely Euclidean or purely flow-weighted covariance model.

The implication of this finding is that a hybrid covariance model should be used instead of purely Euclidean or purely flow covariance model whenever estimating DOC along rivers.

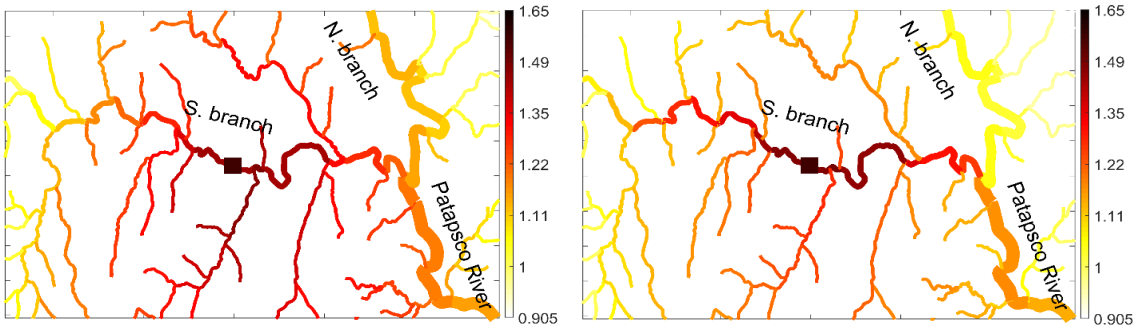
3.3. The domain wide variability of DOC is watershed specific

DOC estimates in 2013 are obtained using the Euclidean (figure 3a) and Euclidean/Gradual-flow (figure 3b) models. These estimates show that the areas of high DOC concentrations stay within a watershed when using the Euclidean/Flow covariance model as opposed to extending across watersheds when the traditional Euclidean covariance model is used. The watershed specific nature of DOC concentrations is physically meaningful as watershed characteristics such as topography, land use, hydrologic cycles, and many other natural processes are similar within a given watershed, and vary across watersheds. The Euclidean/Gradual-flow covariance model better captures the influence of the river network topology and reveals that DOC concentration within watershed remains autocorrelated over much longer distances (covariance ranges a_E=48km and a_G=982km) than what is estimated based on the Euclidean model (a_E=36km). This means that the Euclidean/Gradual-flow model is able to estimate DOC concentrations over more river miles within a specific watershed than the Euclidean model is able to do.

(a) (b)


BME estimation of DOC (mg/l) in 2013
DOC concentration (mg/l)
12.2
7.39
4.48
2.72
1.65
1
0.607
0.368
0.223
BME estimation of DOC (mg/l) in 2013
DOC concentration (mg/l)
12.2
7.39
4.48
2.72
1.65
1
0.607
0.368
0.223

(c) (d)


N. branch
S. branch
Patapsco River
1.65
1.49
1.35
1.22
1.11
1
0.905
N. branch
S. branch
Patapsco River
1.65
1.49
1.35
1.22
1.11
1
0.905

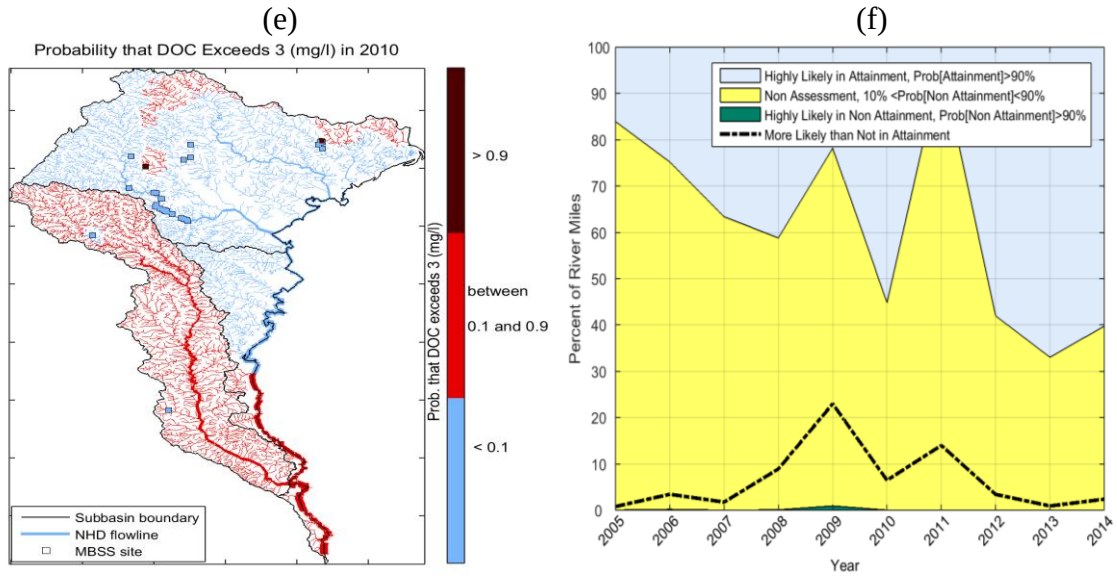


Figure 4.3: Panels (a) and (b) show the maps depicting the spatial distribution of DOC (mg/l) across the study domain in 2013 obtained using Euclidean and Euclidean/Gradual-flow models, respectively. Panels (c) and (d) are maps of estimates obtained using Euclidean and Euclidean/Flow models, respectively, showing the spatial distribution of DOC in 2008 near the confluence of the North and South branches of the Patapsco River. The map depicting the probability that DOC exceeds 3mg/l in 2010 is shown in panel (e), and the probabilistic assessment of DOC impairment over the study domain from 2005 to 2014 is shown in panel (f). Both panel (e) and (f) were obtained using Euclidean/Gradual-flow covariance model.

3.4. The fine scale variability of DOC is influenced by hydrological transport along individual river reaches and by dilution at confluence points

There are noticeable differences between the Euclidean (figure 3c) and Euclidean/Gradual-flow estimates of DOC in 2008 near the confluence of the North and South branches of the Patapsco River. There is a monitoring site on South branch that recorded a high DOC concentration in 2008. The Euclidean estimates of DOC are continuously changing downstream of that monitoring site along South branch and along the Patapsco River, without exhibiting an abrupt change at the confluence of South and North branch, nor at the confluence of any tributaries that flow into South branch downstream of the monitoring site. This indicates that the Euclidean model is not able to account for the fact that waters in these tributaries are not flow connected to the monitoring sites. On the other hand the Euclidean/Gradual-flow estimates on South branch exhibit an abrupt change where South branch merges with North branch to form

the Patapsco River. An abrupt change in estimated DOC is also seen at the confluence point between South branch and each of its tributaries. This is because the monitoring site is not flow connected with tributaries that merge with South branch downstream of the monitoring site. As a result the concentration estimated in these tributaries are distinct from the concentration estimated on South branch. Hence the Euclidean/Gradual-flow model depicts fine scale variability of DOC concentration that is governed by hydrologic transport along each river reach, and by dilution at confluence nodes. The dilution effect can for example be seen at the confluence of South and North branch, where the concentration at the downstream end of South branch (1.27 mg/l) is different than that at the downstream end of north branch (0.96 mg/l), resulting in a new concentration past the confluence point that is in between those two upstream concentrations. These differences in concentrations demonstrate how the Euclidean/Gradual-flow model accounts for dilution. The Euclidean model estimates values are exactly the same directly before and past the confluence point, demonstrating that this model fails to account for the dilution that occurs at confluence points.

The implication of this finding that is the Euclidean/Gradual-flow model provides a fine scale representation of the spatial distribution of DOC concentrations that is substantially more physically meaningful than that of the Euclidean model.

3.5. There is a small fraction of impaired river miles but a large fraction of unassessed river miles

The space/time distribution of DOC is governed by complex natural and physical processes. Imperfect knowledge about these complex processes may result in a significant uncertainty in geostatistical estimation of DOC concentrations, and hence not accounting for

estimation uncertainty in impairment assessment may lead to a wrong conclusion. Using the Euclidean/Gradual-flow covariance model we not only obtained DOC estimates but also the probability that DOC exceeds a specific threshold level. Maps of non-attainment probability at any threshold level of interest can provide important insight for policy guideline and watershed management.

Figure 3(e) shows the probability that the DOC concentration exceeds the 3 mg/l threshold. This map clearly shows that the Patuxent sub-basin is almost entirely unassessed in 2010, whereas the probability that DOC concentration exceeds the 3 mg/l in the other two sub-basins is highly unlikely. This indicates that there is sufficient monitoring in these two watersheds to assess that the water is below 3 mg/L, but this is not the case in the Patuxent sub-basin, where more monitoring is needed in order to know whether DOC is below or above the 3 mg/L threshold level. Hence the probabilistic assessment map for 2010 indicates that a substantial fraction of the study area is unassessed.

In order to determine whether this finding extends to other years, we tabulated for each year the fraction of river miles that were assessed as impaired (i.e. assessed as being above 3 m/L) versus unassessed (figure 3f). We find that while very few river miles are assessed as impaired for DOC, there is a large fraction of river miles that are unassessed. The maps of unassessed river miles produced by the Euclidean/Gradual-flow therefore provide critical new information indicating which river miles require more monitoring in order to determine the ecological health of the river system in these areas, and whether utilities using surface waters in these areas need to treat water in a way that avoids the formation of carcinogenic disinfectant by products.

REFERENCES

(1) Akita, Y., Carter, G., Serre, M.L., 2007. Spatiotemporal nonattainment assessment of surface water tetrachloroethylene in New Jersey. J. Environ. Qual. 36, 508–520. doi:10.2134/jeq2005.0426

(2) Christakos, G., Serre, M.L., 2000. BME analysis of spatiotemporal particulate matter distributions in North Carolina. Atmos. Environ. 34, 3393–3406. doi:10.1016/S1352-2310(00)00080-7

(3) Chu, H.P., Wong, J.H.C., Li, X.Y., 2002. Trihalomethane formation potentials of organic pollutants in wastewater discharge. Water Sci. Technol. 46, 401–406.

(4) Cressie, N., Frey, J., Harch, B., Smith, M., 2006. Spatial prediction on a river network. J. Agric. Biol. Environ. Stat. 11, 127–150. doi:10.1198/108571106X110649

(5) Gardner, B., Sullivan, P.J., Lembo, Jr., A.J., 2003. Predicting stream temperatures: geostatistical model comparison using alternative distance metrics. Can. J. Fish. Aquat. Sci. 60, 344–351. doi:10.1139/f03-025

(6) George Christakos, Patrick Bogaert, and M.S., 2001. Temporal Geographical Information Systems: Advanced Functions for Field-Based Applications, 2001 editi. ed. Springer.

(7) Hobbie, J., Likens, G.E., 1973. ORGANIC CARBON , AND FINE PARTICULATE CARBON and of the interactions between nutrient and. Limnol. Oceanogr. 18, 734–742.

(8) Isaak, D.J., Peterson, E.E., Ver Hoef, J.M., Wenger, S.J., Falke, J.A., Torgersen, C.E., Sowder, C., Steel, E.A., Fortin, M.-J., Jordan, C.E., Ruesch, A.S., Som, N., Monestiez, P., 2014. Applications of spatial statistical network models to stream data. WIREs Water 1, 277–294. doi:10.1002/wat2.1023

(9) Jat, P., Serre, M.L., 2016. Bayesian Maximum Entropy space/time estimation of surface water chloride in Maryland using river distances. Environ. Pollut. doi:http://dx.doi.org/10.1016/j.envpol.2016.09.020

(10) LoBuglio, J.N., Characklis, G.W., Serre, M.L., 2007. Cost-effective water quality assessment through the integration of monitoring data and modeling results. Water Resour.

Res. 43, 1–16. doi:10.1029/2006WR005020

(11) Messier, K.P., Akita, Y., Serre, M.L., 2012. Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. Environ. Sci. Technol. 46, 2772–2780. doi:10.1021/es203152a

(12) Messier, K.P., Campbell, T., Bradley, P.J., Serre, M.L., 2015. Estimation of Groundwater Radon in North Carolina Using Land Use Regression and Bayesian Maximum Entropy. Environ. Sci. Technol. 49, 9817–9825. doi:10.1021/acs.est.5b01503

(13) Money, E., Carter, G.P., Serre, M.L., 2009. Using River Distances in the Space/Time Estimation of Dissolved Oxygen Along Two Impaired River Networks in New Jersey. Water Res. 43, 1948–1958. doi:10.1016/j.watres.2009.01.034.Using

(14) Money, E.S., Carter, G.P., Serre, M.L., 2009. Modern space/time geostatistics using river distances: Data integration of turbidity and E. coli measurements to assess fecal contamination along the Raritan River in New Jersey. Environ. Sci. Technol. 43, 3736–3742. doi:10.1021/es803236j

(15) Money, E.S., Sackett, D.K., Aday, D.D., Serre, M.L., 2011. Using river distance and existing hydrography data can improve the geostatistical estimation of fish tissue mercury at unsampled locations. Environ. Sci. Technol. 45, 7746–7753. doi:10.1021/es2003827

(16) Mulholland, P.J., Helton, A.M., Poole, G.C., Hall, R.O., Hamilton, S.K., Peterson, B.J., Tank, J.L., Ashkenas, L.R., Cooper, L.W., Dahm, C.N., Dodds, W.K., Findlay, S.E.G., Gregory, S. V, Grimm, N.B., Johnson, S.L., McDowell, W.H., Meyer, J.L., Valett, H.M., Webster, J.R., Arango, C.P., Beaulieu, J.J., Bernot, M.J., Burgin, A.J., Crenshaw, C.L., Johnson, L.T., Niederlehner, B.R., O'Brien, J.M., Potter, J.D., Sheibley, R.W., Sobota, D.J., Thomas, S.M., 2008. Stream denitrification across biomes and its response to anthropogenic nitrate loading. Nature 452, 202–205. doi:10.1038/nature06686

(17) Peterson, E.E., Merton, A.A., Theobald, D.M., Urquhart, N.S., 2006. Patterns of spatial autocorrelation in stream water chemistry. Environ. Monit. Assess. 121, 569–594. doi:10.1007/s10661-005-9156-7

(18) Peterson, E.E., Urquhart, S.N., 2006. Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: A case study in Maryland. Environ. Monit. Assess. 121, 613–636. doi:10.1007/s10661-005-9163-8

(19) Prusha, B. a., Clements, W.H., 2004. Landscape attributes, dissolved organic C, and metal bioaccumulation in aquatic macroinvertebrates (Arkansas River Basin, Colorado). J. North Am. Benthol. Soc. 23, 327–339. doi:10.1899/0887-3593(2004)023<0327:LADOCA>2.0.CO2

(20) Serre, M.L., Christakos, G., 1999. Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge - the Equus Beds study. Stoch. Environ. Res. Risk Assess. 13, 1–26. doi:10.1007/s004770050029

(21) USGS Hydrography data [WWW Document], 2015. URL http://nhd.usgs.gov/data.html (accessed 10.6.15).

(22) Ver Hoef, J.M., Peterson, E., Theobald, D., 2006. Spatial statistical models that use flow and stream distance. Environ. Ecol. Stat. 13, 449–464. doi:10.1007/s10651-006-0022-8

(23) Volk, C., Wood, L., Johnson, B., Robinson, J., Zhu, H.W., Kaplan, L., 2002. Monitoring dissolved organic carbon in surface and drinking waters. J. Environ. Monit. - JEM 4, 43–47. doi:10.1039/b107768f

(24) Wetzel, R.G., Hatcher, P.G., Bianchi, T.S., 1995. Natural photolysis by ultraviolet irradiance of recalcitrant dissolved organic matter to simple substrates for rapid bacterial metabolism. Limnol. Oceanogr. 40, 1369–1380. doi:10.4319/lo.1995.40.8.1369

(25) Williamson, C.E., Stemberger, R.S., Morris, D.P., Frost, T.M., Paulsen, S.G., 1996. Ultraviolet radiation in North American lakes: Attenuation estimates from DOC measurements and implications for plankton communities. Limnol. Oceanogr. 41, 1024–1034. doi:10.4319/lo.1996.41.5.1024

(26) Yang, X., Jin, W., 2010. GIS-based spatial regression and prediction of water quality in river networks: A case study in Iowa. J. Environ. Manage. 91, 1943–1951. doi:10.1016/j.jenvman.2010.04.011

CHAPTER 5: CONCLUSIONS

The primary goal of this work was to implement and test a mixture of river and flow covariance models to estimate water quality parameters along river networks. The Bayesian Maximum Entropy (BME) method of modern space/time geostatistics was extended to better account for the river metric in its mean trend functionality and to better incorporate a mixture of Euclidean distance, river distance, and flow connectivity in its covariance functionality. This creates a rich set of new flow-based BME functionalities in the *BMElib* numerical implementation of the BME framework. These new functionalities can be used in any surface water quality studies, providing practitioners with new tools for the mapping analysis of surface water quality that can work on a wide range of river network topology characteristics. These tools were useful for the case studies considered in this work; and they are widely applicable and generalizable to many other surface water quality studies.

Three real world case studies were presented, which provides a broad range of applications demonstrating the use of the river covariance model as well as a mixture of Euclidean/flow covariance models. For each case study an exhaustive range of covariance models were tested, using Euclidean and/or river distances and their combinations, in order to assess which model worked best for each case study. It was hypothesized that in the case of a pollutant (Chloride) for which the sources are along roads that follow the river network the river covariance model would be the best, while for water quality parameters such as fecal coliform and DOC where both terrestrial source and hydrological transport are important the best model

would be a hybrid Euclidean/flow covariance model. These hypotheses were tested using cross validation to see which model results in the most noticeable improvement in estimation accuracy compared to the other covariance models.

The results of the three case studies confirmed our hypotheses, as all three resulted in a 12% to 24% improvement in estimation accuracy. This large range of accuracy improvement is due to a number of factors including the number and density of monitoring stations, the river network resolution and complexity, and the variability and autocorrelation characterizing water quality in each case study. The highest improvement in estimation accuracy was obtained in first case study on Chloride, where we observed that a river covariance model improved the cross-validation R2 by 23.67% compared to an Euclidean covariance model, and where we found that river BME maps were significantly different than the Euclidean BME maps, indicating that a covariance modeling choice can significantly impact the conclusions drawn from these maps for remediation and targeted monitoring. In case study two (fecal coliform) and three (DOC) we observed a 12% and a 17% improvement in estimation accuracy when using a hybrid Euclidean/flow covariance model compared to a purely Euclidean model, and we again found that the maps of water quality obtained with the Euclidean/flow model are significantly different, and generate new findings, compared to the maps obtained using an Euclidean model. Overall the BME framework with the newly introduced flow functions was able to significantly improve water quality estimation along a variety of river networks and for a host of pollutants. There are limitations, however, to this approach. First, the covariance functions used in the analysis were restricted to the exponential function, which is permissible for any river networks. However, there are a variety of other possible covariance functions that need to be examined for permissibility before they can be used in a river and flow covariance functions. Secondly, soft

data or secondary information from physical or water quality models (i.e. Qual2, SWAT) can be used to even further improve the estimation accuracy of water quality.

Future research directions should investigate the tradeoffs in using various proxies for flow (i.e. watershed area, actual volumetric flow, cumulative upstream length), and integrate land use and weather variables in the estimation framework. The river and flow BME functions developed here are general tools that set the stage for a multitude of research regarding spatiotemporal trends in water quality along river networks. It will provide local, state, and federal environmental managers a sound modeling framework for better allocating resources, targeted monitoring, and informing the public when water quality impairments put the public at risk of adverse health impacts.

APPENDIX A: SUPPLEMENTARY INFORMATION FOR 'BAYESIAN MAXIMUM ENTROPY SPACE/TIME ESTIMATION OF SURFACE WATER CHLORIDE IN MARYLAND USING RIVER DISTANCES' PAPER

NHD Flowlines in Subbasins

Our study domain is made up of three of the subbasins defined by the United States Geological Services (USGS) using Hydrology Unit Codes (HUC) with 8 digits. These three HUC8 subbasins are located in an area that drains to the Chesapeake Bay, and they consist of the Gunpowder-Patapsco subbasin, the Severn subbasin and the Patuxent subbasin. The Gunpowder-Patapsco subbasin area is 98.9% in Maryland and 1.1% in Pennsylvania, whereas the Severn and Patuxent subbasins are 100% in Maryland.

The river hydrographic network is defined based on flow lines obtained from the USGS national hydrography dataset (U.S. Geological Survey, National Hydrographic Dataset, http://nhd.usgs.gov/data.html). The number of NHD flow lines and size of the subbasins in our study domain are reported in table S1.

Table A.S1: Subbasin name, number of NHD flowlines, stream length, and area of the subbasin in our study domain (Source: ArcView analysis- 1:24,000 scale NHD hydrography dataset[1,])

Subbasin name	HUC8 code (unitless)	No. of NHD flow lines (unitless)	Total stream length (mile)	Area (mile ^{2})
Gunpowder-Patapsco	02060003	9342	3002	1417
Severn	02060004	2835	1006	369
Patuxent	02060006	6321	2010	927

The previous study of dissolve organic carbon (DOC) by Peterson and Urquhart (2006) used 3083 stream segments throughout Maryland (12407 mile^2), which corresponds to an

average density of 0.25 stream segment per mile^2. This is substantially less than the average of 6.8 NHD flow lines per mile^2 used in our study. This indicates that our work refined the resolution of the river network by a factor of about 27.

Impervious Surface Data

Impervious surfaces are manmade hard areas that are essentially impenetrable to water. Urbanization is a key factor of increasing the imperviousness of watersheds as it adds roads, rooftops, parking lots, sidewalks etc.

Percent developed imperviousness layers were retrieved from the National Land Cover Database (NLCD 2011) published by the Multi-Resolution Land Characteristics Consortium for the conterminous United States, and then they were cropped to our study domain as shown in figure S1. These Multi-Resolution Land Characteristics Consortium based percent developed imperviousness layers provide the imperviousness (%) for each 30m by 30m pixel in our study domain. This fine resolution description of imperviousness was then aggregated to provide the impervious percentage for each HUC with 12 digits (HUC12) subwatersheds delineated in figure S1. The aggregation was performed using ArcGIS (ArcGIS 10.3 version).

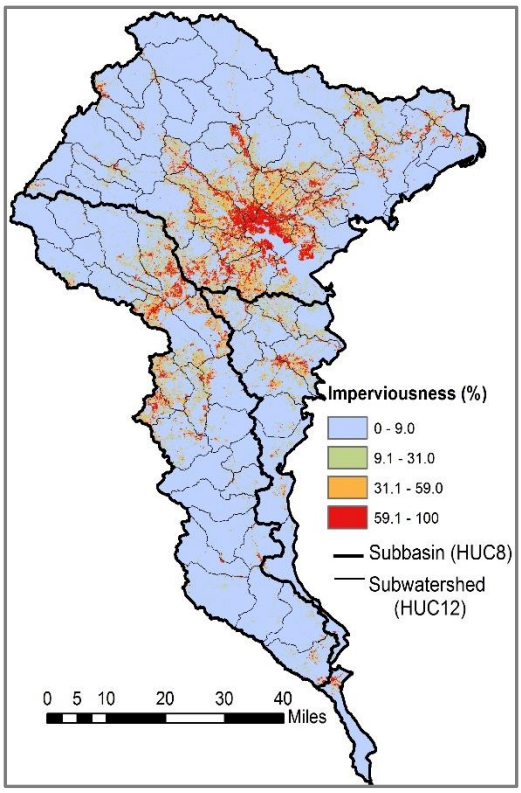


Figure A.S1: Figure A.S1: Multi-Resolution Land Characteristics based percent developed imperviousness layers in 2011.

Land Use Regression (LUR) Model

A strong link between percent imperviousness and water quality degradation in a watershed has been reported by several studies (James 1965, Klein 1979, Demers and Sage, 1990, Kaushal et al., 2005 and Morgan et al., 2007). In the winter, roads and sidewalks are treated with deicing salts. As snow melts, imperviousness physically limits the infiltration of melted snow and most of the chloride in road deicing salts is directly transported to the surface waters, which strongly influences the water chemistry of streams and rivers.

A land use regression (LUR) model for the linear relationship between subwatershed percent imperviousness and chloride log-concentration is developed, which helps elucidate the road salt contribution to elevated chloride concentrations across our study domain, as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{(S1)}$$

where Y_i is the natural log transform of chloride concentration at point i, X_i is the percent imperviousness of the HUC12 subwatershed containing point i, β_1 is its source regression coefficient, and ε_i is an error term.

Figure S2 shows the regression plot of log-chloride concentrations versus subwatershed impervious percentages. We found that the coefficients of regression β_0 and β_1 are equal to 3.2 (log-mg/l) and 0.074 (log-mg/l) per percent impervious surface. The Pearson correlation coefficient (R), a measure of the linear correlation between subwatershed percent imperviousness and chloride log-concentrations, is 0.6.

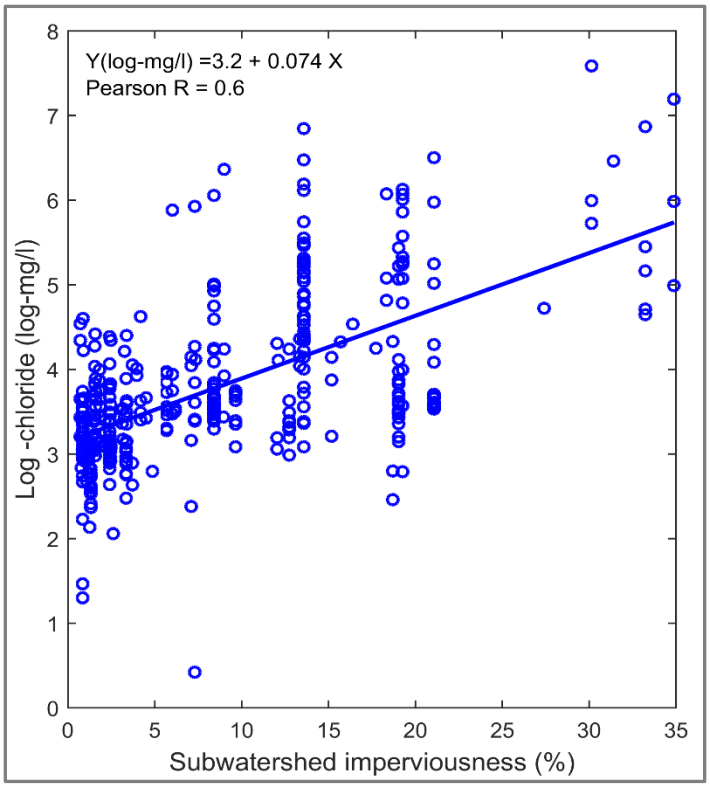


Figure A.S2: Regression plot of log –chloride versus subwatershed imperviousness percentage

Offset models

As described in the main paper, three global offset models are considered in this work. The first offset model is described in previous studies (Akita et al., 2007, Money et al., 2009, and

Money et al., 2011). It consists of the sum of a spatial component (Fig S3a) that is obtained by smoothing time averaged chloride log-concentrations using an exponential kernel filter calculated using Euclidean distances with a spatial exponential smoothing range k_r=75 km (across land), and a temporal component (Fig S4) obtained by smoothing spatially averaged log-concentrations using an exponential kernel filter calculated using time differences with a temporal exponential smoothing range k_t=5 years. The spatial component exhibits spatial trend that varies isotropically across land and across unconnected river branches.

The second global offset model is similar to the first offset model, with the only difference being that its spatial component (Fig S3b) is obtained using an exponential kernel filter based on *river* distances (instead of Euclidean distances) with a spatial exponential smoothing range k_r=75 km (along rivers). This spatial component exhibits spatial trends that varies along the river network (as opposed to across land), and therefore unconnected river branches display non similar concentrations.

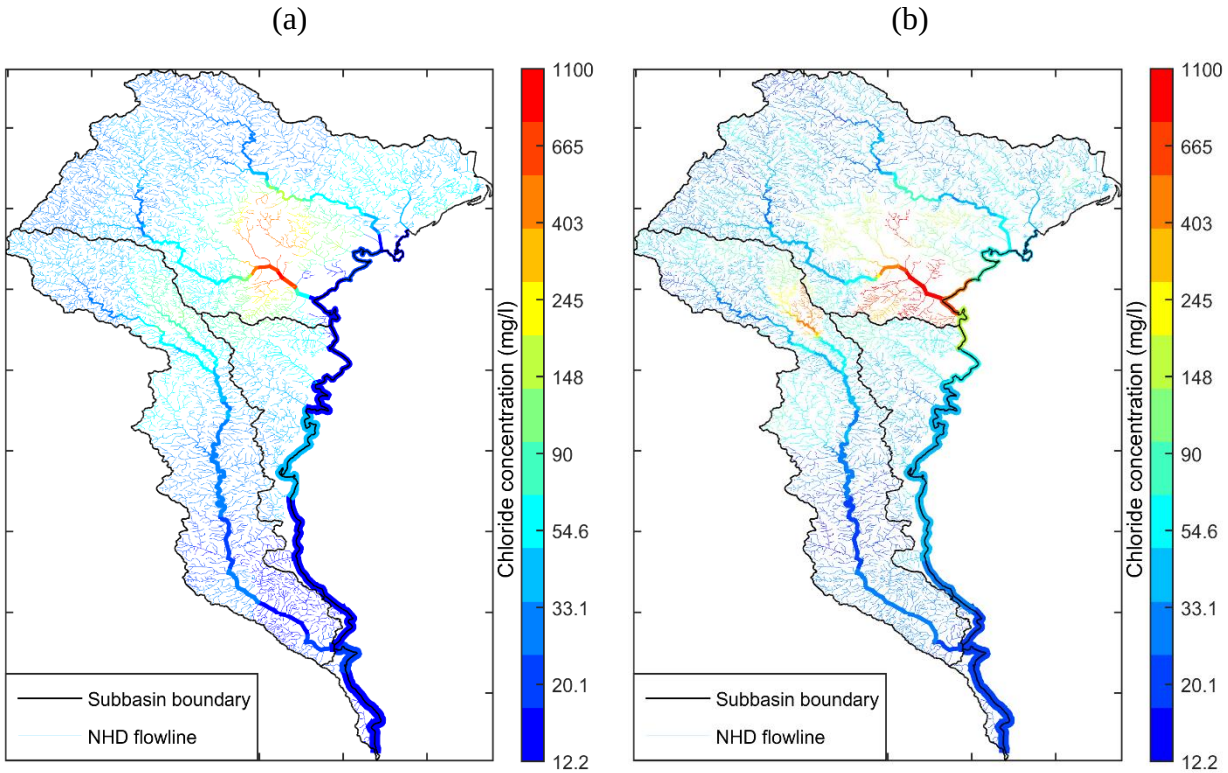


Figure A.S3: Spatial component of the global offset calculated using kernel smoothing of time averaged chloride concentration measurements using an exponential kernel function based on (a) Euclidean distances and (b) river distances.

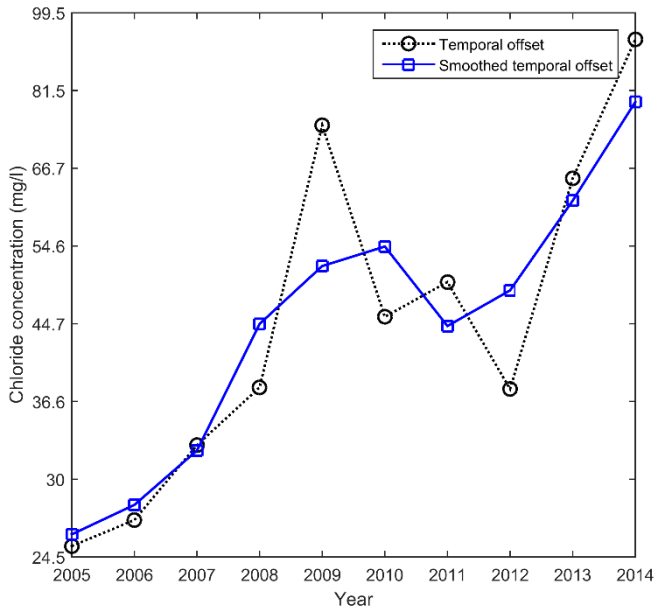


Figure A.S4: Temporal component of the offset, obtained using an exponential kernel smoothing of spatially averaged chloride log-concentrations

The third global offset model is the LUR estimate (Figure S5) calculated using the linear regression line (Figure S2) between log-chloride concentrations and HUC12 subwatershed imperviousness percentages obtained from the Multi-Resolution Land Characteristics based percent developed imperviousness layers (Figure S1). The LUR offset does not change with time.

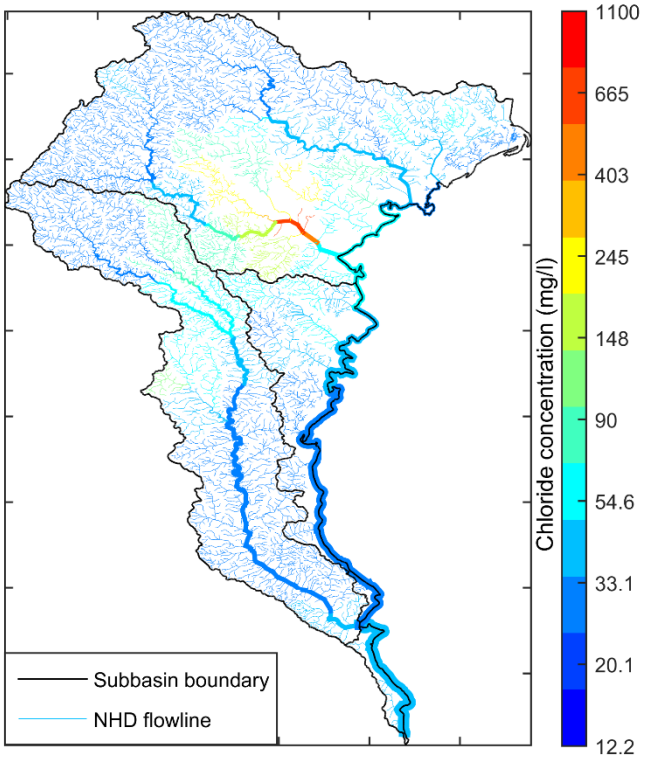


Figure A.S5: Offset of chloride concentration calculated as the LUR estimate obtained based on a linear regression between chloride log-concentrations and HEC12 subwatershed imperviousness percentages.

Weighted Least Square Covariance Fitting Procedure

We define the random field $X(\boldsymbol{p})$, where $\boldsymbol{p}=(\boldsymbol{s}, t)$ is the space time coordinate, as a spatially homogeneous and temporally stationary space/time random field for which the set of offset-removed chloride log-concentrations is one realization. There always exist such a space/time random field, and its space/time covariance function will capture the variability of the offset-removed log concentrations. In this work we consider three offset models, which each

exhibit its own space/time variability, and therefore needs its own Euclidean and river covariance models.

For each offset model we calculate the corresponding offset-removed log concentrations, and from those we calculate the experimental covariance value corresponding to pairs of offset-removed log concentrations measured at points $\boldsymbol{p}=(\boldsymbol{s}, t)$ and $\boldsymbol{p}'=(\boldsymbol{s}', t')$ separated by a spatial lag $r=\mathrm{d}(\boldsymbol{s},\boldsymbol{s}')$ and a temporal lag $\tau = |t-t'|$ of interest. The spatial distance $\mathrm{d}(\boldsymbol{s},\boldsymbol{s}')$ is calculated either using an Euclidean distance or a river distance.

Experimental covariance values obtained for various spatial and temporal lags were then used to fit an exponential space/time covariance model given by

$$c_x(r,\tau) = c_0 \exp\left(-\frac{3r}{\mathrm{a_r}}\right)\exp\left(-\frac{3\tau}{\mathrm{a_t}}\right) \qquad \text{(S2)}$$

where c_o is the variance, a_r is the spatial covariance range (measured as a straight line for the Euclidean covariance model, and along the river for the river covariance model), and a_t is the temporal covariance range.

The covariance fitting was performed using a weighted least square (WLS) approach that finds the covariance parameters $\boldsymbol{\theta}= (c_o, a_r, a_t)$ which minimizes the weighted sum of squares given by

$$\mathrm{WSS}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w_i\left(\hat{\mathrm{c}}(\boldsymbol{h}_i) - \mathrm{c}(\boldsymbol{h}_i;\boldsymbol{\theta})\right)^2 \qquad (3)$$

where $\hat{\mathrm{c}}(\boldsymbol{h}_i)$ is the i-th experimental covariance value calculated for space/time lag $\boldsymbol{h}_i=(r_i, \tau_i)$, w_i is the weight of space/time lag $\boldsymbol{h}_i$ corresponding in this work to number of pairs of observations separated by that lag, and $\mathrm{c}(\boldsymbol{h}_i;\boldsymbol{\theta})$ denotes the covariance model value calculated for space/time lag $\boldsymbol{h}_i$ using the parameter value $\boldsymbol{\theta}$.

Sensitivity Analysis with Respect to the Proportion of Left Censored Data

A sensitivity analysis was conducted with respect to proportion of left censored data. This analysis consisted in left censoring a proportion of the data, and comparing the cross validation mean square error (MSE) and R^2 of the following three methods: (a) BME rigorously modeling the censored data using the TGPDF, (b) kriging replacing the censored data with half the CL, and (c) kriging replacing the censored data with the CL. Table S2 shows the cross-validation statistics obtained in this sensitivity analysis. The MSE value increases and the R^2 decreases as the proportion of left-censored data increases, indicating a loss in estimation accuracy that was expected since more censored data means less information. However Table S2 clearly demonstrates that the rate of deterioration in estimation accuracy is lower for river BME (method a) than for its kriging linear limiting cases (method b and c), and as a result the BME method consistently outperforms the kriging methods. Focusing on the comparison between BME (method a) and kriging using half the censoring limit (method b), we see as expected that the R^2 is the same between method a (BME) and method b (kriging) when the proportion of censored data is zero (because in that case there is no censored data). When there is 5% of censored data, the R^2 is 0.412 for kriging and 0.448 for BME, corresponding to a percent change in R^2 (PC in R^2) of 9%. This means that BME improves the R^2 by 9% over kriging when 5% of the data is left censored. Interestingly, the PC in R^2 is 109%, 480%, 658% and 133%, respectively, when the proportion of censored data is 13.6%, 25.1%, 32.3% and 46.2%, respectively. This means that BME improves the R^2 by a factor of about 2 to 7.5 over kriging when the proportion of censored data ranges from 13.6% to 46.2%.

Table A.S2: Sensitivity analysis of the estimation accuracy of the river BME and kriging methods with respect to the proportion of left censored data

		BME (method a)		Kriging using ½ the CL (method		Kriging using the CL (method	
Censoring limit (mg/l)	**Proportion of censored**	**MSE***	**R^2**	**MSE***	**R^2**	**MSE***	**R^2**
0	0.0	0.194	0.789	0.194	0.789	0.194	0.789
15	5.1	0.550	0.448	0.622	0.412	0.648	0.402
20	13.6	0.882	0.340	1.344	0.163	1.548	0.113
25	25.1	1.656	0.174	2.534	0.030	2.811	0.100
30	32.3	2.291	0.091	3.337	0.012	3.654	0.010
35	46.2	3.945	0.007	5.314	0.003	5.794	0.007

* $(mg/l)^2$

Maps and Movies

BME Estimate of chloride concentration

Chloride concentration was estimated along each river mile in our study, and a series of concentration maps from 2005 to 2014 were constructed and posted at the following website: http://www.unc.edu/depts/case/BMElab/studies/PJ_ClMD/

The estimation was performed using either the river BME or Euclidean BME method, the maps of estimated Chloride concentrations are shown either over the study domain or over region A, and concentration values are shown using either a continuous colors or bicolor, resulting in the following eight sets of maps:

River BME estimate of chloride concentration

Maps and movie shown in continuous color over the study domain

Maps and movie shown in continuous color over area A

Maps and movie shown in bicolor over the study domain

Maps and movie shown in bicolor over the area A

Euclidean BME estimate of chloride concentration

Maps and movie shown in continuous color over the study domain

Maps and movie shown in continuous color over area A

Maps and movie shown in bicolor over the study domain

Maps and movie shown in bicolor over the area A

BME estimate of the probability that chloride exceeds 230 (mg/l)

A comparison of estimated Chloride concentration with the EPA guideline level of 230 mg/l was visualized by calculating and mapping the probability that the Chloride concentration is above 230 mg/l. Maps showing the spatial distribution of the probability that Chloride exceeds 230 mg/l along the rivers of our study domain for each year from 2005 to 2014 are posted at the following website: http://www.unc.edu/depts/case/BMElab/studies/PJ_ClMD/

The probability that Chloride exceeds 230 mg/l was estimated using either the river BME or Euclidean BME method, and the maps of probabilities are shown either over the study domain or over region A, resulting in the following four sets of maps:

River BME estimate of the probability that chloride exceeds 230 (mg/l)

Maps and movie shown over the study domain

Maps and movie shown over area A

Euclidean BME estimate of the probability that chloride exceeds 230 (mg/l)

Maps and movie shown over the study domain

Maps and movie shown over area A

REFERENCES

(1) Akita, Y., Carter, G., Serre, M.L., 2007. Spatiotemporal nonattainment assessment of surface water tetrachloroethylene in New Jersey. J. Environ. Qual. 36, 508–520. doi:10.2134/jeq2005.0426

(2) Demers, C.L., Sage, R.W., 1990. Effects of road deicing salt on chloride levels in four adirondack streams. Water, Air, Soil Pollut. 49, 369–373. doi:10.1007/BF00507076

(3) Kaushal, S.S., Groffman, P.M., Likens, G.E., Belt, K.T., Stack, W.P., Kelly, V.R., Band, L.E., Fisher, G.T., 2005. Increased salinization of fresh water in the northeastern United States. Proc. Natl. Acad. Sci. U. S. A. 102, 13517–13520. doi:10.1073/pnas.0506414102

(4) Money, E.S., Carter, G.P., Serre, M.L., 2009. Modern space/time geostatistics using river distances: Data integration of turbidity and E. coli measurements to assess fecal contamination along the Raritan River in New Jersey. Environ. Sci. Technol. 43, 3736–3742. doi:10.1021/es803236j

(5) Money, E.S., Sackett, D.K., Aday, D.D., Serre, M.L., 2011. Using river distance and existing hydrography data can improve the geostatistical estimation of fish tissue mercury at unsampled locations. Environ. Sci. Technol. 45, 7746–7753. doi:10.1021/es2003827

(6) Morgan, R.P., Kline, K.M., Cushman, S.F., 2007. Relationships among nutrients, chloride and biological indices in urban Maryland streams. Urban Ecosyst. 10, 153–166. doi:10.1007/s11252-006-0016-1

(7) Peterson, E.E., Urquhart, S.N., 2006. Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: A case study in Maryland. Environ. Monit. Assess. 121, 613–636. doi:10.1007/s10661-005-9163-8

APPENDIX B: SUPPLEMENTAL INFORMATION FOR 'A NOVEL GEOSTATISTICAL APPROACH COMBINING EUCLIDEAN AND GRADUAL-FLOW COVARIANCE MODELS TO ESTIMATE FECAL COLIFORM ALONG THE HAW AND DEEP RIVERS IN NORTH CAROLINA' PAPER

Details on the fecal coliform and hydrography data

The fecal coliform concentration data for the Haw and Deep rivers in North Carolina were obtained from the Cape Fear River Basin Monitoring Coalition's water quality data (http://lcfrp.uncw.edu/riverdatabase/). A query of this database was performed in September 11, 2014, to download all the fecal coliform data in the years 2006−2010, which resulted in a dataset with 3869 entries, including 9 missing values. After removing these 9 missing values, snapping sampling locations to the nearest point on the river network, and averaging 12 duplicate values, the dataset consisted in 3848 space/time fecal coliform concentrations located at 69 unique observations sites. Descriptive statistics of these 3848 fecal coliform concentrations observed along the Haw and Deep rivers from 2006-2010 are tabulated in table S1.

Table B.S1: Descriptive statistics of fecal coliform concentrations observed along the Haw and Deep rivers in North Carolina from 2006-2010

Parameter	**Haw river**	**Deep river**	**Haw-Deep rivers**
Number of fecal coliform observations	2378	1470	3848
Unique geographical observation sites	39	30	69
Minimum conc. (CFU/100 ml)	1	2	1
Maximum conc. (CFU/100 ml)	12500	12000	12500
Mean conc. (CFU/100 ml)	633	867	723
Median conc. (CFU/100 ml)	105	87	100
Standard deviation (CFU/100 ml)	1839	2373	2062

The river network along the Haw and Deep rivers is described based on flow lines obtained from the medium resolution USGS National Hydrography Data (NHD). The medium resolution USGS NHD flow lines were obtained on October 11, 2014 by going to the USGS website (http://nhd.usgs.gov/), selecting 'Get Data' → 'Go to NHD extract by States' → 'MediumResolution' → 'Shape' and selecting the 'NHD_M_37_North_Carolina_ST.zip' compressed file and extracting the 'NHDflowline.shp' shapefile containing the flow lines for all the rivers in North Carolina. The USGS defines its medium resolution NHD data as being at the scale of 1:100,000, which in our study provides a fine resolution description of all the river

reaches where observation sites are located (shown in plain line in figure 1a), as well as their named upstream reaches (shown in fine lines in figure 1a).

Details on the flow-weighted covariance models using pipe flow

The pipe flow-weighted covariance model is derived by first defining a spatial random field $X(l,i)$ and then calculating its covariance. Ver Hoef et al. (2006)[1] and Cressie et al. (2006)[2] define $X(l,i)$ as the moving-average of a white noise random process, while de Fouquet and Bernard-Michel (2006)[3] and Bernard-Michel and de Fouquet (2006)[4] define $X(l,i)$ as the sum of uncorrelated one dimensional fields along each flow line. We provide here the mathematical expression of $X(l,i)$ for these two approaches, and we refer the readers to their papers and to Money et al (2009)[5] for an in-depth derivation of how the pipe flow covariance model is derived from $X(l,i)$.

Let us identify a point $\boldsymbol{r}=(\boldsymbol{s},l,i)$ on the river network either by its Euclidean coordinate $\boldsymbol{s}=\{longitude, latitude\}$; or by its river coordinate (l,i) consisting of the longitudinal coordinate (l) corresponding to the length of the continuous line connecting the river outlet to $\boldsymbol{s}$ along the river network, and the reach index (i) uniquely defining the river reach where $\boldsymbol{s}$ is located. Ver Hoef et al. (2006)[1] and Cressie et al. (2006)[2] define the spatial random field $X(l,i)$ as

$$X(l,i) = \int_l^\infty d\boldsymbol{u} \sum_{j \in V_i(\boldsymbol{u})} \sqrt{\Omega(j)/\Omega(i)}\, g(\boldsymbol{u}-l)\, W(\boldsymbol{u},l) \qquad \text{(S1)}$$

where $V_i(\boldsymbol{u})$ is the set of river is reaches at longitudinal coordinate $\boldsymbol{u}$ upstream of reach i, $g(\boldsymbol{u}-l)$ is a moving average function that lead to a valid covariance function, $W(u,l)$ is a white noise process with mean zero i.e. $E\big(W(\boldsymbol{u},l)\big)=0$, $\Omega(j)/\Omega(i)$ is a real number between 0

and 1 expressing the amount of flow shared between reach *i* and reach *j* such that $\sum_{j\in V_i(\boldsymbol{u})}\Omega(i,j)=1$, $\Omega(i)$ and $\Omega(j)$ are flow additive functions that increase in the direction of flow (i.e. if two reaches *i'* and *i*'' combine into a downstream reach *i*, then according to flow additivity $\Omega(i')+\Omega(i'')=\Omega(i)$).

On the other hand de Fouquet & Bernard-Michel (2006)[3] and Bernard-Michel and de Fouquet (2006)[4] define $X(l,i)$ as

$$X(l,i)=\sum_{j\in V_i(\infty)}\sqrt{\Omega(j)/\Omega(i)}\;\; Y_j(l) \qquad \text{(S2)}$$

where $V_i(\infty)$ is the set of flow-connected leaf reaches upstream of reach *i*, and $Y_j(l)$ are independent zero mean random processes on R^1 i.e. $E[Y_j(l)]=0$ with covariance $cov(Y_i(l),Y_i(l'))=c_1(h), h=|l-l'|$, $c_1(h)$ may be any permissible covariance function in R^1.

As explained above, the covariance of the spatial random field given in equations S1 or S2 is the pipe flow covariance model given in Eq. 3.

Derivation of the flow-weighted covariance model using gradual flow

As explained in the main paper, Money et al. (2009)[5] introduced a generalization of the flow-weighted covariance model (Eq. 6) that rigorously accounts for flows that gradually increase along river reaches. Here we summarize the derivation of this covariance model.

As detailed in the main paper, let's define the flow density ω(**r**) as a positive density function and let's define the flow function Ω(***r***) as its integral upstream of ***r***, i.e. $\Omega(\boldsymbol{r})=\int_{u\in U(r)} dl(\boldsymbol{u})\;\omega(\boldsymbol{u})$, where *U*(***r***) is the set of points upstream of ***r***, and *l*(***u***) is the longitudinal coordinate of point ***u***.

Money et al. (2009)[5] define the SRF $X(\boldsymbol{r}) = \int_{u\in U(r)} dl(\boldsymbol{u})\,\sqrt{\omega(\boldsymbol{u})/\Omega(\boldsymbol{r})}\; W(\boldsymbol{u})\, Y(l(\boldsymbol{r}))$, where $W(\boldsymbol{u})$ is a white noise process, $Y(l(\boldsymbol{r})$ is a zero mean random process with covariance $c_1(h),\ h = |l - l'|$ is the river distance, and $c_1(h)$ can be any permissible covariance function.

When $\boldsymbol{r}$ is upstream of $\boldsymbol{r}$', the covariance between $X(\boldsymbol{r})$ and $X(\boldsymbol{r}')$ is derived using the following steps (Money et al., 2009)[5]

$$
\begin{aligned}
c_X(\boldsymbol{r},\boldsymbol{r}') &= E[X(\boldsymbol{r})X(\boldsymbol{r}')] \\
&= \int_{\boldsymbol{u}\in U(\boldsymbol{r})} dl(\boldsymbol{u}) \int_{\boldsymbol{u}'\in U(\boldsymbol{r}')} dl(\boldsymbol{u}')\; \sqrt{\frac{\omega(\boldsymbol{u})\omega(\boldsymbol{u}')}{\Omega(\boldsymbol{r})\Omega(\boldsymbol{r}')}}\; E[W(\boldsymbol{u})\, W(\boldsymbol{u}')\, Y(l(\boldsymbol{r})Y(l'(\boldsymbol{r}'))] \\
&= \int_{\boldsymbol{u}\in U(\boldsymbol{r})} dl(\boldsymbol{u}) \int_{\boldsymbol{u}'\in U(\boldsymbol{r}')} dl(\boldsymbol{u}')\; \sqrt{\frac{\omega(\boldsymbol{u})\omega(\boldsymbol{u}')}{\Omega(\boldsymbol{r})\Omega(\boldsymbol{r}')}}\; \delta(\boldsymbol{u}-\boldsymbol{u}')c_1(h = |l - l'|) \\
&= \int_{\boldsymbol{u}\in U(\boldsymbol{r})} dl(\boldsymbol{u})\; \frac{\omega(\boldsymbol{u})}{\sqrt{\Omega(\boldsymbol{r})\Omega(\boldsymbol{r}')}}\; c_1(h) \\
&= \sqrt{\Omega(\boldsymbol{r})/\Omega(\boldsymbol{r}')}\; c_1(h) \qquad \text{(S3)}
\end{aligned}
$$

where $\delta(\boldsymbol{u}-\boldsymbol{u}')$ is a Dirac function with property $\int d\boldsymbol{u}' f(\boldsymbol{u}')\,\delta(\boldsymbol{u}-\boldsymbol{u}') = f(\boldsymbol{u})$ for sufficiently smooth functions $f(\boldsymbol{u})$.

Eq. S3 leads to Eq. 6 when $c_1(h)=\sigma^2 \exp(-3\, d_R(\boldsymbol{r},\boldsymbol{r}')/a_G)$.

Leave-One-Out Cross-Validation (LOOCV) statistics

To assess the accuracy of the BME estimation of fecal coliform using the Euclidean, river, flow, and Euclidean/flow covariance models, a leave-one-out cross-validation (LOOCV) analysis was performed. Each fecal coliform measured value z_j was removed one at a time, and

re-estimated using only the remaining data. This method was repeated again for each monitoring station.

For a given estimation method (m) that uses any of the covariance models stated above, the overall estimation error was quantified using the Mean Squared Error, $MSE^{(m)} = \frac{1}{n}\sum_{j=1}^{n}\left(z_j^{*(m)} - z_j\right)^2$, the consistent estimation error (i.e. the bias) was quantified using the Mean Error $ME^{(m)} = \frac{1}{n}\sum_{j=1}^{n}\left(z_j^{*(m)} - z_j\right)$, and the random error (i.e. lack of precision) was quantified using the squared Pearson coefficient, $R^2 = 1 - \sum_{j=1}^{n}\left(z_j^{*(m)} - z_j\right)^2 / \sum_{j=1}^{n}\left(z_j^{*(m)}\right)^2$, where $z_j^{*(m)}$ is the re-estimation of z_j.

Pipe flow is a poor approximation of gradual flow along a coarse river network representation of the Haw and Deep rivers

A dense river network representing river reaches in our Haw and Deep river study area is depicted with thin lines in figure 1a. We calculated the gradual and pipe flow along this dense river network and found that they are almost perfectly similar, indicating that Ver Hoef et al. (2006)[1] proportional influence calculation leads to an almost perfect approximation of gradual flow if the river network is dense. This dense river network was used for our assessment of water quality along all river miles.

We also restricted the dense river network to a coarse river network consisting mostly of the river reaches where monitoring sites are located, as well as their downstream reaches. The gradual flow in area A located in the headwaters of the Haw river (figure 1d) is significantly different from its pipe flow approximation (figure 1e). The gradual and pipe flows are also significantly different in areas B and C (figures not shown). These results show that there can be

significant differences between gradual and pipe flows at the headwaters and along long river reaches of a coarse river network.

These results are amongst the first to clearly visualize that pipe flow is a good approximation of gradual flow for a dense river network; and that this deteriorates for a coarse river network. The implication of this finding is that the estimation accuracy between gradual versus pipe flow should be the same for dense river network, but the estimation accuracy between gradual versus pipe flow might be different whenever a coarse river network is used.

Modeling the spatial covariance using Euclidean distance, river distance, and flow ratio

Experimental covariance values with a zero temporal lag ($\tau = 0$) are obtained by selecting all the pairs ($x_{head,i}$, $x_{tail,i}$) of offset removed fecal coliform log concentration measurements that were measured at the same time, stratifying them into classes based on Euclidean lag d_E , river lag d_R, and flow ratio f, and calculating the experimental covariance for each class of pairs using Eq. 8. The experimental covariance values obtained on the coarse river network using gradual flow are shown in figures S1(a) and (b), and those obtained using the coarse network with pipe flow are shown in figure S1(c). In figure S1(a) the experimental covariance values are shown as a function of Euclidean lag on the *x*-axis and displayed with a marker that is based on its flow ratio and with a color that is based on its river lag. As can be seen in this figure, the experimental covariance value decreases with Euclidean lag for fixed river lag and flow ratio, as indicated by the general downward trend in covariance values along the *x*-axis. Furthermore, the experimental covariance values decrease with decreasing flow ratio, as indicated by the fact that the circles are generally above the squares. These experimental covariance values provide visual evidence that both the Euclidean lag and flow ratio play an

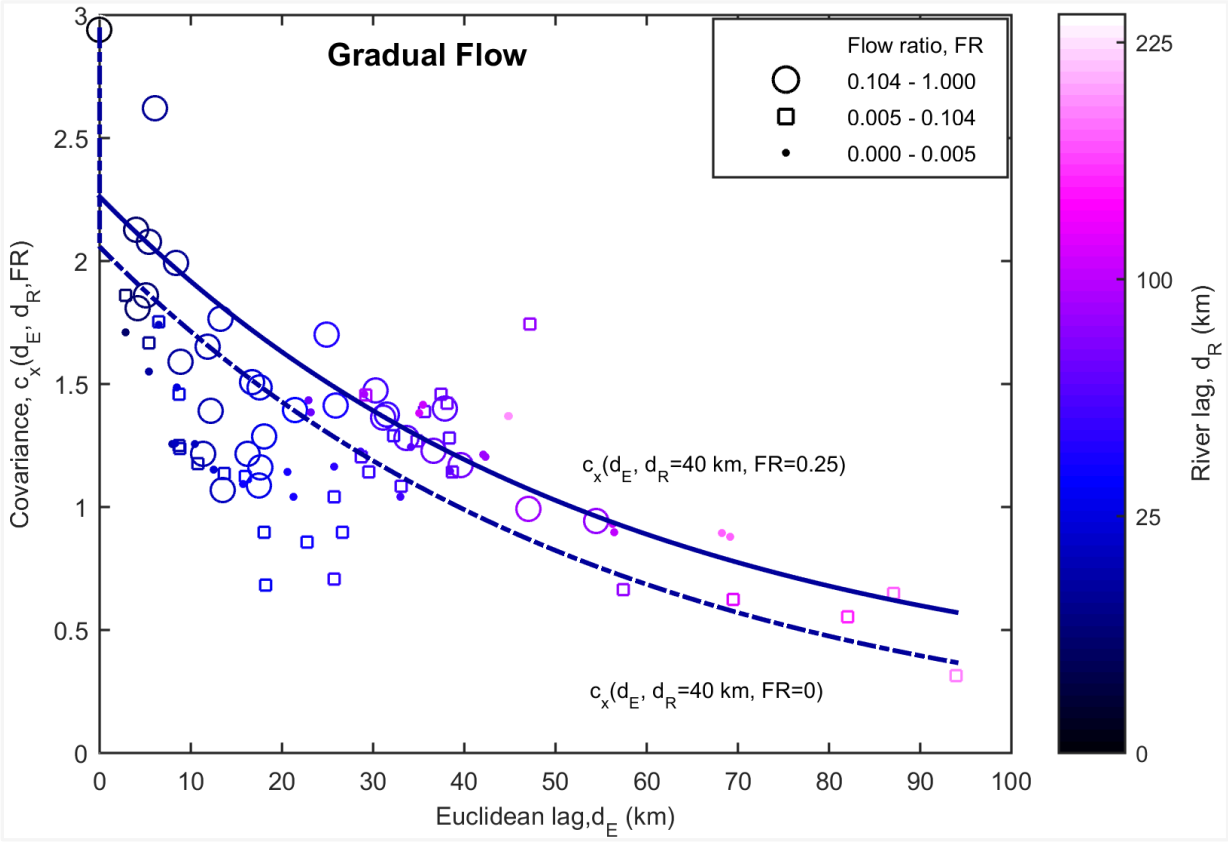
important role in describing the spatial variability of fecal coliform. As a result, the Euclidean/gradual flow covariance model (Eq 7) is the most suitable choice for modeling the variability of fecal coliform.

For each possible combination of α_E and α_G, for α_E going from 0 to 1 by increment of 0.05, we did a least square regression of the Euclidean/gradual flow covariance model onto the experimental covariance values to obtain the covariance range parameters a_E and a_F, and using these range parameter values we performed a cross validation analysis. We then selected the α_E and α_G values and associated parameters a_E and a_F that minimized the cross-validation MSE. The covariance parameter values we obtained are α_E=0.7, $\alpha_G = 0.3$, $a_E = 164\ km$ and $a_F = 155\ km$ (Table 1), and the corresponding covariance model is shown in figure S1(a) as a function Euclidean lag for a fixed river lag of 40*km* and a fixed flow ratio of 0.25 (plain line), and for the same river lag but a fixed flow ratio of 0 (dashed line). This model captures the decrease in covariance with respect to increasing Euclidean lag and decreasing flow ratio. According to this covariance model, the autocorrelation in fecal coliform in stream waters comes from terrestrial source contaminating land over distance ranges of approximately 164*km*, and from the subsequent longitudinal transport in suspended solid over distances of up to approximately 155*km* along river reaches where no dilution occurs.

In order to compare the impact that gradual and pipe flows have on modeling spatial variability, we repeated the calculation of experimental covariance values using pipe flow. We show side by side how experimental covariance values change with respect to flow ratio when gradual flow is used (figure S1(b)) and when pipe flow is used (figure S1(c)). As can be seen from these figures, the increase in experimental covariance values with respect to flow ratio is better captured when gradual flow is used. To ascertain this finding, we performed a statistical

test to quantify whether the increase in covariance with respect to flow ratio is significant for three different sets of Euclidean and river lags. We found that the p-value is less than 0.05 for each of these sets when gradual flow is used (figure S1(b)), whereas the p-value is greater than 0.05 for each of these sets when pipe flow is used (figure S1(c)). This finding indicates that when using a coarse river network, then the gradual flow covariance model better captures spatial variability of fecal coliform than the pipe flow covariance model.

(a)



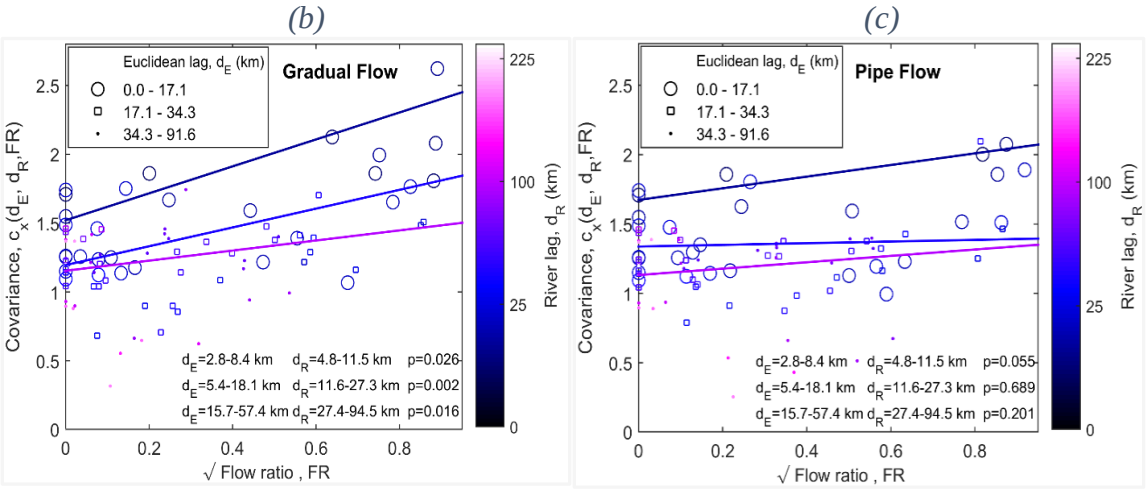


Figure B.S1: Experimental covariance values obtained using gradual flow and shown as a function of (a) Euclidean lag for fixed river lags and flow ratios, and (b) as a function of flow ratio for fixed Euclidean and river lags. The experimental covariance values obtained with pipe flow are shown in (c) with respect to flow ratio.

Daily estimates of fecal coliform across the study area

Maps of fecal coliform estimates obtained using the Euclidean covariance model and using the Euclidean/Gradual-flow covariance model are shown in figure S2 for May 08 and May 09 of 2006. The data on May-09-2006 is watershed specific, with values between distinctly different in the Haw river compared to those on the Deep river. The Euclidean/Gradual-flow estimates follow the same pattern on both May-09-2006 (when there are data in each watershed) as the day prior, May-08-2006. The Euclidean estimates are also watershed specific on May-09-2006, but fail to be watershed specific on May-08-2006, leading to a substantial difference between the Euclidean and the Euclidean/Gradual-flow estimates along the Deep river on May-08-2006.

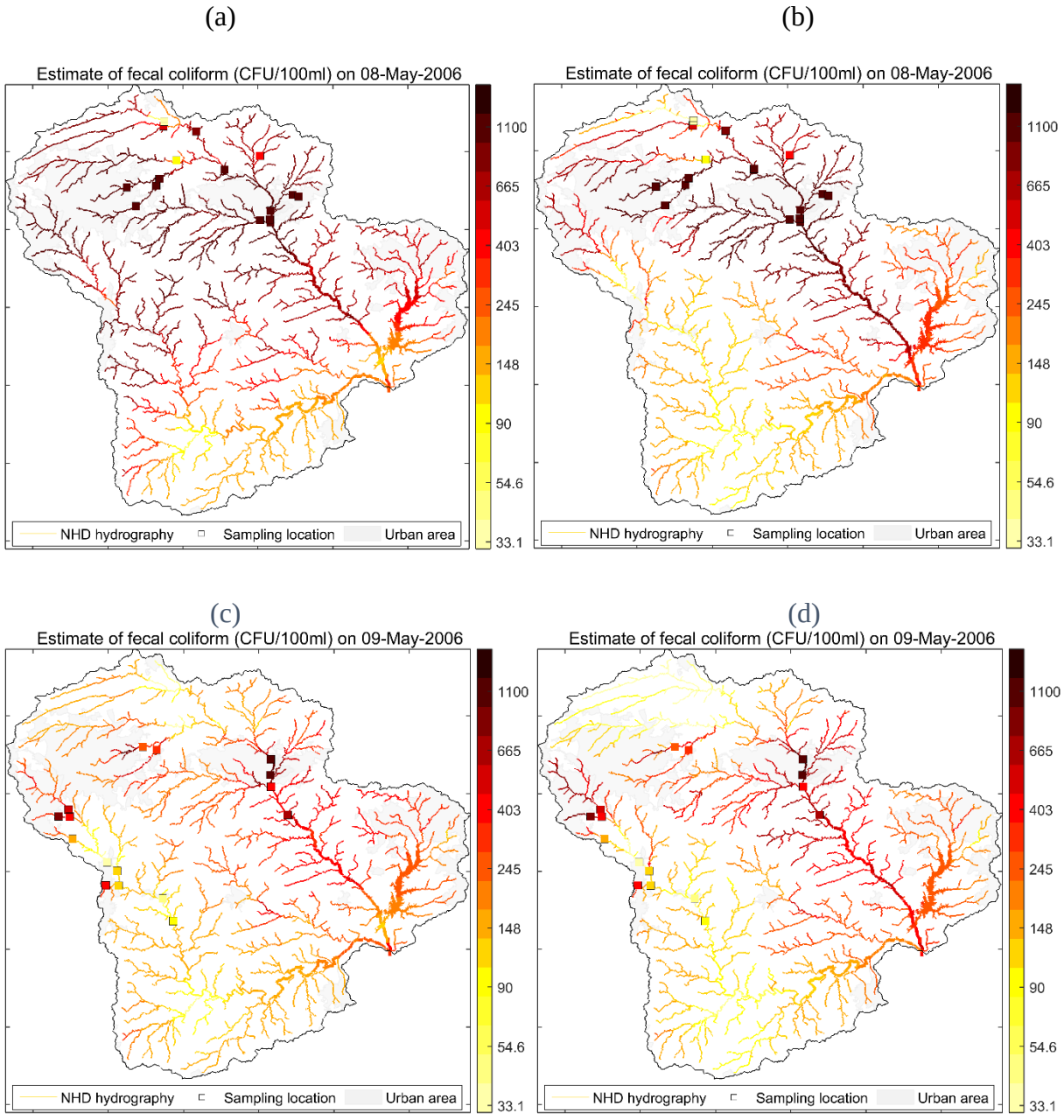


Figure B.S2: Maps of fecal coliform estimates (CFU/100ml) obtained using the Euclidean covariance on 08-May-2006 (panel a) and 09-May-2006 (panel c). Estimates obtained using the Euclidean/Gradual-flow covariance model are shown in panels (b) and (d).

Maps of fecal coliform estimates obtained using the Euclidean covariance model and using the Euclidean/Gradual-flow covariance model are shown in figure S3 for December 28 of

2006. These maps show again that the Euclidean/Gradual-flow estimates are watershed specific while the Euclidean estimates are not.

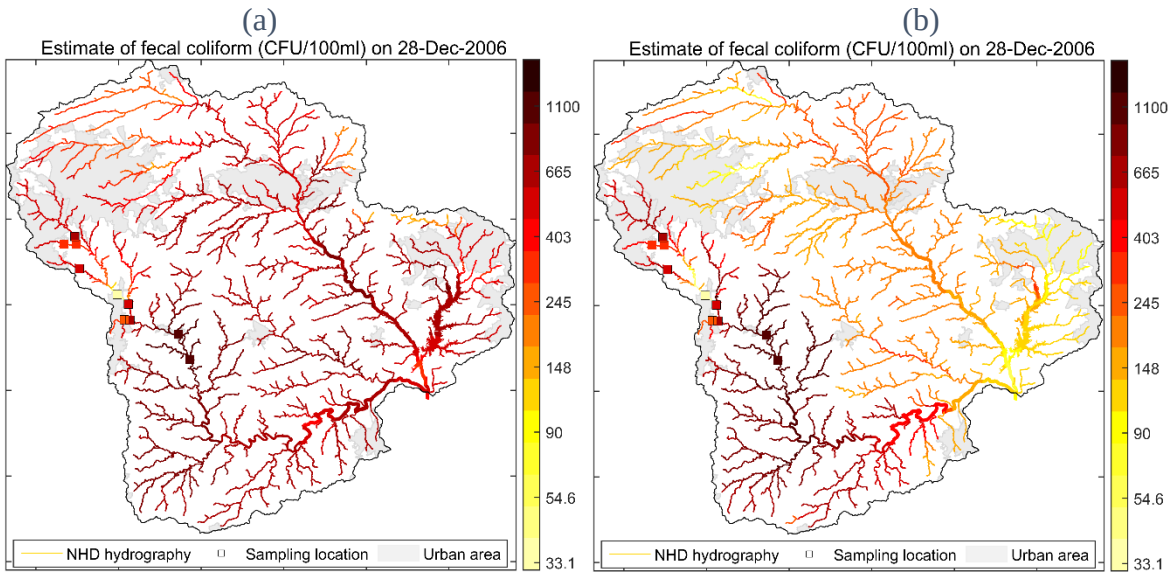


Figure B.S3: Maps of fecal coliform estimates (CFU/100ml) obtained on 28-Dec-2006 using the Euclidean covariance (panel a) and the Euclidean/Gradual-flow covariance model panel (b).

Movies showing maps of fecal coliform along the Haw and Deep rivers on specific days from 2006 to 2010 are available at http://www.unc.edu/depts/case/BMElab/studies/PJ_FC_NC/. Specifically movies are available for the following days:

- Sampling days from Jan-2006 to Jun-2008
- Sampling days from Jul-2008 to Dec-2010
- Consecutive days from Aug-31-2009 to Oct-30-2009

Number of impaired river miles

For each river mile and each sampling day in the 2006-2010 study period, we used the Euclidean covariance model versus the Euclidean/Gradual-flow covariance model to calculate

the probability that fecal coliform concentration exceeds 200 CFU/100ml on that day. A river mile was then assessed as impaired if there are more than *N* days over the 2006-2010 study period during which the probability that fecal coliform exceeds 200 CFU/100ml is greater than *P*. Table S2 reports the number of impaired river miles calculated for *P*=68% and 90%, and for *N*=60days, 90 days and 120 days. The results show that the number of impaired river miles estimated using the Euclidean/Gradual-flow covariance model is consistently greater than that estimated with the Euclidean covariance model.

Table B.S2: Number of impaired river miles estimated using the Euclidean covariance model versus the Euclidean/Gradual-flow covariance model

Threshold probability of impairment, *P* (%)	**Threshold number of impairment days, *N***	**Number of impaired river miles(*)**	
		Euclidean covariance model	**Euclidean/Gradual-Flow covariance model**
90	60	39	96
90	90	7	8
90	120	1	1
68	60	969	1537
68	90	196	312
68	120	64	92

(*) a river mile is impaired if there are more than *N* days over the 2006-2010 study period during which the probability that fecal coliform exceeds 200 CFU/100ml is greater than *P*

Furthermore, the increase in impaired river miles is more visible in the headwaters of the Haw and Deep rivers, as illustrated in areas A, B and C shown in figure 3. This is supported by the maps showing the average of fecal coliform estimates across the 2006-2010 study period (figure S4). This map shows that large urban centers, including High Point, Greensboro,

Burlington and Durham, are located at the headwaters of the Haw and Deep rivers. These urban areas have a large percentage of impervious surface, which increases runoff and fecal contamination of rivers in these areas. As a result, these areas are more frequently impaired. As fecal coliforms are then transported downstream from these urban areas, they have time to die off or to settle down, thereby resulting in the lower average concentrations seen in figure S4 at the downstream end of the river network.

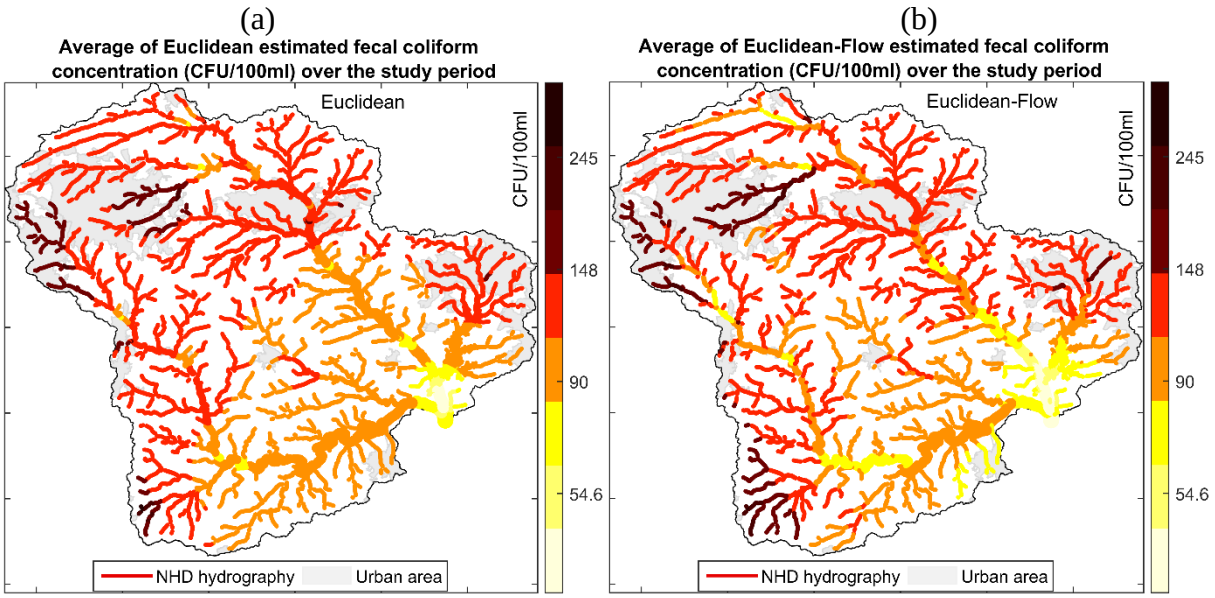


Figure B.S4: Maps of the fecal coliform estimates averaged across the 2006-2010 study period are shown in panel (a) using estimates obtained with Euclidean covariance and in panel (b) using estimates obtained with Euclidean/Gradual-flow covariance model.

REFERENCES

(1) Ver Hoef, J. M.; Peterson, E.; Theobald, D. Spatial statistical models that use flow and stream distance. *Environ. Ecol. Stat.* 2006, *13* (4), 449–464.

(2) Cressie, N.; Frey, J.; Harch, B.; Smith, M. Spatial prediction on a river network. *J. Agric. Biol. Environ. Stat.* 2006, *11* (2), 127–150.

(3) de Fouquet, C.; Bernard-Michel, C. Modeles geostatistiques de concentrations ou de debits le long des cours d'eau. *Comptes Rendus - Geosci.* 2006, *338* (5), 307–318.

(4) C. Bernard-Michel and C. de Fouquet1. Construction of valid covariances along a hydrographic network . Application to specific water discharge on the Moselle Basin. 2006, No. 1, 1–4.

(5) Money, E. S.; Carter, G. P.; Serre, M. L. Modern space/time geostatistics using river distances: Data integration of turbidity and E. coli measurements to assess fecal contamination along the Raritan River in New Jersey. *Environ. Sci. Technol.* 2009, *43* (10), 3736–3742.