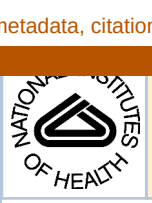

Published in final edited form as:
Environ Sci Technol. 2009 May 15; 43(10): 3736–3742.

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

Eric S. Money[1], **Gail P. Carter**[2], and **Marc L. Serre**[1,*]

[1]University of North Carolina – Chapel Hill, Dept. of Environmental Sciences and Engineering, Chapel Hill, NC 27599-7431

[2]New Jersey Dept. of Environmental Protection, Division of Science, Research, and Technology, P.O. Box 409, Trenton, NJ 08625-0409

Abstract

Escherichia coli (*E.coli*) is a widely used indicator of fecal contamination in water bodies. External contact and subsequent ingestion of bacteria coming from fecal contamination can lead to harmful health effects. Since *E.coli* data are sometimes limited, the objective of this study is to use secondary information in the form of turbidity to improve the assessment of *E.coli* at un-monitored locations. We obtained all *E.coli* and turbidity monitoring data available from existing monitoring networks for the 2000 – 2006 time period for the Raritan River Basin, New Jersey. Using collocated measurements we developed a predictive model of *E.coli* from turbidity data. Using this model, soft data are constructed for *E.coli* given turbidity measurements at 739 space/time locations where only turbidity was measured. Finally, the Bayesian Maximum Entropy (BME) method of modern space/time geostatistics was used for the data integration of monitored and predicted *E.coli* data to produce maps showing *E.coli* concentration estimated daily across the river basin. The addition of soft data in conjunction with the use of river distances reduced estimation error by about 30%. Furthermore, based on these maps, up to 35% of river miles in the Raritan Basin had a probability of *E.coli* impairment greater than 90% on the most polluted day of the study period.

Introduction

Fecal Indicator Bacteria in River Systems

Fecal indicator bacteria (FIB) provide important health and ecological information for many river basins. Although FIB's themselves are not harmful, their presence in streams suggests that pathogenic microorganisms might also be present, leading to possible human health risks. Diseases and illnesses that can be contracted in water with high fecal contamination include typhoid fever, hepatitis, gastroenteritis, and dysentery (1). The most commonly tested FIBs are

*Corresponding Author: Marc_serre@unc.edu, 919-966-7014 (phone), 919-966-7911 (fax).

Supporting Information Available:
A detailed description of the study area and *E.coli* and turbidity data is presented in figures S1, S2, table S1. The mathematical steps used to obtain the flow-connected covariance model are described, and the Ver Hoef et al. (2006) model is shown in figure S3 to be a limiting case of Eq. 5. Additional figures illustrate the *E.coli*/turbidity relationship and a summary of all estimation results. Movie S1 shows *E.coli* concentrations estimated every day for a subset of the study period. This information is available free of charge via the Internet at http://pubs.acs.org.

Brief:
This study combines *E.coli* and turbidity data in a river-based space/time geostatistical framework to provide a more accurate basin-wide assessment of fecal contamination.

total coliforms, fecal coliforms, *Escherichia coli* (*E.coli*), and enterococci. *E.coli* is a species of fecal coliform that is specific to fecal material from humans and other warm blooded animals. Based on studies conducted by the Environmental Protection Agency (EPA), *E.coli* is the best indicator of health risk from water contact in recreational waters (2). Therefore many states are now measuring *E.coli* instead of total coliforms to assess streams for fecal contamination. However, due to the limited scope of existing monitoring networks, budget limitations, and manpower constraints, it is difficult to assess all river miles. The purpose of this study is to examine the use of a modern spatiotemporal geostatistics technique, known as Bayesian Maximum Entropy (BME), to statistically assess *E.coli's* presence in both monitored and un-monitored streams using not only existing *E.coli* data but also integrating secondary information in the form of turbidity measurements to further improve the mapping of basin-scale fecal indicators.

Autocorrelation in E.coli

Geostatistical techniques such as kriging rely on the fact that many natural phenomenon exhibit spatial autocorrelation. Monitoring stations along the same stream, for example, tend to report similar physical and chemical characteristics. Kriging methods construct a regional model of correlation to estimate variables, such as *E.coli*, at un-sampled locations based on data from sampled locations (3–5). Cokriging, subsequently, uses not only the spatial correlation of a single variable, but also the correlations associated with other environmental variables. There have been numerous examples of cokriging for environmental variable estimation ranging from soil salinity, suspended sediment, and rainfall, to regional stream quality (6–9). It is most beneficial where the primary variable is under-sampled with respect to the secondary variable, as is the case for this study when examining *E.coli* and turbidity as secondary information. Generally the inclusion of secondary information results in more accurate local predictions than when considering a single variable alone (6,10).

A more general approach, and the approach used in this study, to estimating at un-sampled locations is the BME method of modern space/time geostatistics (11). This method accounts for both spatial and temporal correlations between data points. BME has been successfully applied to a variety of environmental issues, including water quality (12–13). As demonstrated in these studies, BME presents the flexibility of providing the space/time kriging methods as its linear limiting case, while it can be expanded to a non-linear estimator if other non-linear knowledge bases (e.g. soft data, non Gaussian distribution, etc.) need to be considered. In addition, the BME approach has recently been updated with river-based functionality to incorporate river distance instead of the typical Euclidean distance when dealing with river parameters. Several studies have noted that river distances might provide more appropriate models for the spatial autocorrelation of water quality along river networks (9,12). Therefore a major component of this study is to determine whether the use of river distances along with turbidity as a secondary variable, improves our estimation of *E.coli* for un-monitored stream reaches.

Turbidity and E.coli

Turbidity is the expression of the optical property that causes light to be scattered and absorbed rather than transmitted with no change in direction of flux level through the sample (14).. It is related to *E.coli* concentration in that research has shown that FIBs are oftentimes associated with particulate matter in the water column and transport of fecal bacteria via suspended sediments is an important aquatic mechanism (1,15). Numerous studies have examined the relationship between turbidity and *E.coli* and found significant correlation between both parameters (16–19). Our study area contained a larger number of measured turbidity values relative to *E.coli*, therefore turbidity was chosen as a secondary variable.

Experimental Section

Study Area and Data

The area under investigation is the Raritan River Basin in north-central New Jersey (Figure S1). The basin is 1100 square miles and consists of 36% urban, 19% agriculture, 27% forest, and approximately 17% wetland/water land uses. Approximately 1.2 million people live within this basin and both fecal coliforms and turbidity have been cited as major resource concerns (20). Water quality data for the Raritan Basin was obtained through the National Water Information System (NWIS), maintained by the United States Geological Survey (USGS) for the period January 1, 2000 – December 30th, 2007. A total of 44 monitoring stations provided 579 space/time data points for measured *E.coli* while 118 monitoring stations yielded 739 measurements of turbidity for the study period. *E.coli* data were log-normally distributed with a mean of 5.4 log-colony forming units (cfu)/100*mL*.

Generation of Soft Data

One of the primary goals of this research is to introduce a secondary variable, in the form of turbidity, to predict *E.coli* concentrations in areas where there are no direct *E.coli* measurements. These predicted values are referred to as 'soft' data because of the uncertainty associated with the predicted values. There are two types of soft data employed in this study, probabilistic and interval. To construct the probabilistic soft data, we used a total of 27 collocated samples of turbidity and *E.coli*. First, a simple linear regression was performed using log-transformed data to determine an initial correlation (r-squared = 0.54) which was consistent with other studies relating turbidity to *E.coli* or fecal coliform concentration (16–19). Because of the limited number of collocated points and relatively low values of turbidity represented, the final least squares predictive model for *E.coli* is a continuous piecewise function containing the linear relationship along with a polynomial model of order 2 to reduce overestimation of *E.coli* at extremely high turbidity values :

$$\text{Log}-E.\text{coli}=\begin{cases} 2.07z^2-0.02z+2.08 & z<0.6 \\ 2.05z+1.57 & z\geq 0.6 \end{cases} \quad (1)$$

where log-*E.coli* is expressed in log-cfu/100*mL*, and log-turbidity (*z*) is expressed in log-NTU. Using this relationship the log-*E.coli* prediction error variance was calculated using the mean of the squared differences between predicted and measured log-*E.coli* for a series of given windows of log-turbidity values. Finally, for every space/time point where log-turbidity (but not necessarily log-*E.coli*) was measured, a Gaussian probability distribution function (PDF) was constructed for log-*E.coli* with a mean given by (1) and a variance corresponding to the prediction error variance at the measured log-turbidity. This mean and variance were then used to construct soft log-*E.coli* data of Gaussian probabilistic type at 739 space/time points.

We also accounted for the uncertainty associated with the direct measurements of low levels of *E.coli*. The data downloaded from the USGS use the membrane filtration (m-Tec) method for bacteria enumeration and several intercalibration studies suggest ± 0.5 log as a working point to account for measurement error (21–22). Therefore, for any measured log-*E.coli* < 2 log-cfu/100mL in this study, interval soft data were introduced in the general form of equation (2), where a = measured log-*E.coli* − 0.5 and b = measured log-*E.coli* + 0.5. This resulted in an additional 15 soft data points.

$$\text{Prob}[a<\text{log}-E.\text{coli}<b]=1 \quad (2)$$

Bayesian Maximum Entropy Framework

To integrate the soft data with the measured log-*E.coli* values and then estimate at un-monitored locations, the BME method of modern space/time geostatistics is used. BME provides a rigorous mathematical framework to process a wide variety of knowledge bases characterizing the space/time distribution and monitoring data available for log-*E.coli*, and obtain a complete stochastic description at any un-monitored space/time point in terms of its posterior PDF. The BME method was introduced by Christakos (11), and a detailed description of the conceptual underpinnings of the BME framework are provided in Christakos (23–24), while its *BMElib* numerical implementation is described in Serre *et al.* (25), Serre and Christakos (26) and Christakos *et al* (27). *BMElib* version 2.0b with river functionality was used in this analysis. It was written using the MATLAB® R2000a programming platform. Details about the implementation of river distance calculations in *BMElib* are provided in Money *et al*. (28). The BME procedure consists of defining the general knowledge (i.e. covariance), site specific knowledge (i.e. monitoring data), and integrating the two to calculate a posterior PDF. Site specific knowledge includes both hard data (e.g. measured values) and soft data (i.e. log-*E.coli* predictions based on turbidity). By way of summary, BME uses the maximization of a Shannon measure of information entropy and an operational Bayesian updating rule to process the general and site specific knowledge bases, and obtain the posterior PDF describing log-*E.coli* concentration at any un-sampled point of the river network.

Covariance Model Selection

An important aspect of this work is to select a covariance model that uses river distances. We restricted our mapping analysis to rivers that can be represented by a directed tree consisting of a set of downstream-combining stream reaches (Fig 1), which is highly relevant for the Raritan River considered in this study. Each stream reach is identified by a unique stream reach index i, and we let V be the set of all stream reach indexes; $V=\{1,2,\ldots\}$. We define the longitudinal coordinate l of a point on the river network as the length of the continuous line connecting the river outlet to that point along the river network (by convention, negative l values represent fictitious locations downstream of the outlet). A point $\boldsymbol{r}=(\boldsymbol{s},l,i)$ on the river network is uniquely identified by either its spatial coordinate $\boldsymbol{s}=(s_1,s_2)$; or its river coordinate (l,i) identifying the longitudinal coordinate l and the reach index i where the point is located (Fig 1). Using this convention to define points along a river network, we consider two classes of covariance models that incorporate river distances.

The first class of models to consider are isotropic river covariance models, which can be expressed as a function of the distance between two points $\boldsymbol{r}$ and $\boldsymbol{r'}$, i.e. $\mathrm{cov}(\boldsymbol{r},\boldsymbol{r'})=c(d(\boldsymbol{r},\boldsymbol{r'}))$, where $d(\boldsymbol{r},\boldsymbol{r}')$ is a distance metric. An important member of this class is the isotropic exponential-power river covariance model introduced by Money *et al*. (28–29)

$$\mathrm{cov}(\boldsymbol{r},\boldsymbol{r}')=\exp(-(d_\alpha(\boldsymbol{r},\boldsymbol{r}')/a_r)^\beta),\quad 0\le\alpha\le 1 \text{ and } 0<\beta\le 2 \tag{3}$$

where $d_\alpha(\boldsymbol{r},\boldsymbol{r'})=\alpha d_R(\boldsymbol{r},\boldsymbol{r'})+(\alpha-1)d_E(\boldsymbol{r},\boldsymbol{r'})$ is a linear combination of the river distance (i.e. shortest length along the river connecting $\boldsymbol{r}$ and $\boldsymbol{r'}$) $d_R(\boldsymbol{r},\boldsymbol{r'})$ and the Euclidean distance (i.e. straight-line distance) $d_E(\boldsymbol{r},\boldsymbol{r'})$, and a_r is the overall spatial range. This covariance model is permissible for *any* directed tree river network for ($\alpha=0,\beta\in[0,2]$) or for ($\alpha=1,\beta=1$) (28–29).

Another important class of river covariance models are flow-connected covariance models, which are a function of both river distance *and* flow. We introduce here a novel (although modest) generalization of the flow-connected covariance model introduced by Ver Hoef et al. (30). We let $\omega(\boldsymbol{r})$ be a positive *density* function characterizing the flow entering the river per unit stream length along the river network, and we refer to its corresponding *flow* function Ω

$$\Omega(\boldsymbol{r})=\int_{\mathbf{u}\in U(\mathbf{r})} \mathrm{dl}(\mathbf{u})\omega(\mathbf{u})$$

(***r***) as its integral along all flow lines upstream of ***r***, i.e. , where is $U(\boldsymbol{r})$ is the set of points upstream of ***r***, and $l(\boldsymbol{u})$ is the longitudinal coordinate of point ***u***. The density function $\omega(\boldsymbol{r})$ may be obtained from overland flow discharge if that information is available, or from a proxy such as contributing watershed area, or it may be set to a constant value. When $\omega(\boldsymbol{r})$ is non-zero throughout the river network, the resulting flow function $\Omega(\boldsymbol{r})$ varies with ***r***, as illustrated in Fig 1. Combining ideas introduced in Ver Hoef *et al*. (30), de Fouquet and Bernard-Michel (31), and Cressie *et al*. (32) to construct permissible flow-connected covariance models, we propose here to define a spatial random field $X(\boldsymbol{r})$ as

$$X(\boldsymbol{r})=\int_{\mathbf{u}\in U(\mathbf{r})} \mathrm{dl}(\mathbf{u})\sqrt{\omega(\mathbf{u})/\Omega(\mathbf{r})}W(\mathbf{u})Y(l(\mathbf{r})) \quad (4)$$

where $W(\boldsymbol{u})$ is a white noise process, $Y(l)$ is a zero mean random process on R^1 with covariance $\mathrm{cov}(Y_i(l),Y_i(l'))=c_1(h)$, $h=|l-l'|$, and $c_1(h)$ may be any permissible covariance function in R^1. The covariance of $X(\boldsymbol{r})$ provides a permissible covariance model given by (see supporting information for detailed steps)

$$c_X(\boldsymbol{r},\boldsymbol{r}')=\sqrt{\Omega(\boldsymbol{r},\boldsymbol{r}')}c_1(h) \quad (5)$$

where $\Omega(\boldsymbol{r},\boldsymbol{r}')=\Omega(\boldsymbol{r})/\Omega(\boldsymbol{r}')$ if ***r*** is upstream of ***r'***, and $\Omega(\boldsymbol{r},\boldsymbol{r}')=0$ if ***r*** and ***r'*** are not flow-connected. $\Omega(\boldsymbol{r},\boldsymbol{r}')$ is a number between 0 and 1 that quantifies the flow connection between ***r*** and ***r'***. As shown in supporting information, the flow-connected covariance model introduced by Ver Hoef *et al*. (30) corresponds to the limiting case of Eq. (5) where the flow function is constant along each reach, i.e. $\Omega(\boldsymbol{r})=\Omega(i(\boldsymbol{r}))$, and is additive at each junction so that $\Omega(i(\mathbf{r}))=\sum_{j\in V_{\mathbf{r}}(\infty)}\Omega(j)\ \forall\,\mathbf{r}$, where $i(\boldsymbol{r})$ is the reach index of point $\boldsymbol{r}=(\boldsymbol{s},l,i)$, and $V_r(\infty)$ is the set of the indexes of the leaf reaches (i.e. stream reaches at the upstream ends of the river network) feeding into ***r***. Our covariance model (Eq. 5) adds the flexibility to consider flow functions $\Omega(\boldsymbol{r})$ that increase along any given stream reach. For example, as shown in Fig. 1, this added flexibility allows our flow-connected covariance model (Eq. 5) to account for overland flow discharge between points ***r'*** and ***r"***, as well as the flow contribution of the small river reach (shown in dotted lines) that was ignored in the representation of the river network. This is illustrated by the fact that $\Omega(\boldsymbol{r}')<\Omega(\boldsymbol{r}")$ in Fig. 1, even though ***r'*** and ***r"*** are on the same reach. This generalization of Ver Hoef *et al.* (30) covariance model is useful in situations where there are several monitoring data points along the same stream reach.

An obvious advantage to using flow-connected models is that they incorporate flow connectivity into the model of autocorrelation. However, as noted by Peterson and Urquhart (33), setting the covariance to zero when points are not flow-connected may be a hindrance if very few monitoring sites are flow-connected, leading to less informed estimation maps than those produced using an isotropic covariance model. In the case of log-*E.coli* in the Raritan Basin, considering a spatial range equal to the area of the basin itself, on average only 1.6 data points were flow-connected. Therefore an isotropic covariance model was chosen to estimate log-*E.coli* in the Raritan Basin. The final model used in this study for the space/time covariance of log-*E.coli* between space/time points $\boldsymbol{p}=(\boldsymbol{r},t)$ and $\boldsymbol{p}'=(\boldsymbol{r}',t')$ is

$$\mathrm{cov}(\boldsymbol{p},\boldsymbol{p}')=c_1\exp\left(\frac{-3h}{a_{r1}}\right)\exp\left(\frac{-3\tau}{a_{t1}}\right)+c_2\exp\left(\frac{-3h}{a_{r2}}\right)\exp\left(\frac{-3\tau}{a_{t2}}\right)+c_3\exp\left(\frac{-3h}{a_{r3}}\right)\exp\left(\frac{-3\tau}{a_{t3}}\right) \quad (6)$$

where t and t' are times, $h=d_\alpha(\boldsymbol{r},\boldsymbol{r'})$ and $\tau=|t-t'|$ are the spatial and temporal lags, respectively, and a_r and a_t refer to the overall spatial and temporal ranges, respectively. In this study we used either $\alpha=0$ (Euclidean distance) or $\alpha=1$ (river distance).

Comparing River and Euclidean Based Estimation

A comparison was made between estimations using river distance, as described above, and estimation using the typical Euclidean distance, alongside the incorporation of soft data from measured turbidity. Cross-validation tests were performed on three different scenarios to determine the best model for estimating basin-wide log-*E.coli*. Each data point was removed sequentially and re-estimated using the remaining space/time data points. The Mean Square Error (MSE) is calculated as the sum of the squared differences between re-estimated and measured values. Scenario 1 used the measured log-*E.coli* data (i.e. the 15 interval soft data points and all the hard data) with the Euclidean distance. Scenario 2 contained the same data as scenario 1 except the river distance was used. Scenario 3 built upon scenario 2 by adding in the turbidity data (incorporated as the soft Gaussian data constructed using Eq. 1). The method with the lowest MSE was then used in the assessment and estimation of *E.coli* for the entire Raritan Basin.

BME Estimation of Basin-wide E.coli

Using the selected distance metric within the BME framework we estimate *E.coli* at equidistant estimation points (i.e. distributed at a fixed interval of 0.1*km*) along the Raritan River Basin network. The network shapefiles were obtained from the NJDEP, and projected in a geographic coordinate system (NAD83) using decimal degrees. For visualization purposes, a small buffer (.01km) was overlaid using a geographic information system. For each estimation point we select the hard and soft log-*E.coli* data situated in its local space/time neighborhood, and calculate the corresponding BME posterior PDF describing log-*E.coli* at that estimation point. The variance of the BME posterior PDF provides an assessment of the estimation uncertainty, while the back-log transform of the mean of the BME posterior PDF is used as an approximation of the median estimator for *E.coli* concentrations. This is then used to produce chloropleth maps of estimated *E.coli* concentration, and delineate river miles that are more-likely-than-not impaired.

Assessing Impaired River Miles

In order to better understand the pattern of fecal contamination impairment and better quantify the probability of these impairments, a criterion-based space/time assessment framework is used to categorize the fraction of river miles meeting certain probability thresholds, as discussed in Akita *et al.* (12). These thresholds give us the ability to classify the probability of violation of a standard for any space/time estimation point based on the BME posterior PDF of log-*E.coli*. We set our standard for *E.coli* concentration at235cfu/100mL, which is the standard set by NJDEP for primary contact recreation. Using this standard, the probability of violation at space/time point $\boldsymbol{p}$ is defined as the probability that *E.coli*>235cfu/100mL, i.e.

$$\text{Prob.}[\text{Violation}, \boldsymbol{p}]=\text{Prob.}[\text{E.coli}(\boldsymbol{p})>235\text{cfu}/100\text{mL}] \quad (7)$$

The fraction of river miles impaired on any given day of the study period is then obtained by calculating the fraction of equidistant estimation points for which the probability of violation (Eq. 7) is in excess of some pre-selected probability threshold (e.g. 90%).

Results and Discussion

Covariance Analysis

Figure 2 shows the covariance $c_X(h,\tau)$ of log-*E.coli* obtained for the Raritan Basin. The top panel displays $c_X(h,\tau=0)$ which shows how the covariance varies as a function of spatial lag h for a temporal lag τ equal to 0, while the bottom panel displays $c_X(h=0,\tau)$ which shows how the covariance varies as a function of temporal lag for a zero spatial lag. Experimental covariance values estimated from data are shown with markers, while the covariance models obtained by fitting Eq. 6 to the markers are shown with lines. The covariance was calculated and modeled using both a Euclidean distance (dashed line) and river distance (plain line). The covariance model parameters obtained with the Euclidean and river distances are summarized in table 1. The first structure of the covariance model (with parameters c_{01}, a_{r1} and a_{t1}) is similar for both Euclidean and river distance-based models, with 50% of the total variability of log-*E.coli* being characterized by a fairly short range of 30–40*km* in space and 80 *days* in time. This could be due to variability we would expect from point-like sources of *E.coli* pollution that are not constant and therefore may dissipate over a few months. The second and third structures of both Euclidean and river covariance models indicate that the remaining 50% of variability in log-*E.coli* levels is autocorrelated over longer distances and durations. As noted before, *E.coli*, and fecal bacteria in general, is oftentimes associated with suspended sediment in the water column. Because of this association it is hypothesized that *E.coli* associated with suspended sediment remains in the water at high levels for a longer period of time than free bacteria because sediments are retained along a stream network for long distances (15). This phenomenon is captured in the longer spatial and temporal ranges of the covariance models. In the Euclidean based model, the longer range was between 100–200*km* in space and 200–500 days in time. Interestingly, for the river based-model, the spatial ranges were anywhere from 1.5 to 2 times longer (300–400*km*), suggesting that by accounting for the river connections between points, *E.coli* concentrations may remain correlated over much longer distances than previously considered.

Cross-Validation Analysis

The cross-validation analysis outlined above resulted in mean square errors of MSE_1=2.87 (log-cfu/100*mL*)^2 for scenario 1, MSE_2=2.57(log-cfu/100*mL*)^2 for scenario 2, and MSE_3=1.99 (log-cfu/100*mL*)^2 for scenario 3. Comparing scenario 1 to scenario 2 we see that by using river distances instead of Euclidean distances we reduce the estimation error by about 10%, which is similar to the reduction found in a previous study examining dissolved oxygen in the Raritan Basin (28). If we then add in soft log-*E.coli* data derived from measured turbidity (scenario 3), there is an additional 24% decrease in estimation error. Therefore by incorporating river distances along with soft data from turbidity, the estimation error was reduced by 31% when compared to log-*E.coli* estimation using the typical Euclidean distance and no secondary information. This is one of the first instances in a space/time context that river distances and secondary information have been combined to significantly reduce estimation error. As a result, the river-based covariance model was deemed to be the most accurate representation of *E.coli* in the Raritan Basin, and was used in the subsequent basin-wide estimation and mapping of fecal contamination.

Fecal Contamination in the Raritan Basin

Median estimates of *E.coli* concentration were calculated for every day of the study period between 2000–2007. A movie showing changes in these estimated concentrations over time and space can be viewed in supporting information. Figure 3 depicts the *E.coli* concentration on May 4th, 2002 and is representative of many of the days in this study. The squares indicate locations of monitoring stations with measured *E.coli* values and the chloropleth map shows areas where the concentration exceeds the single sample standard of 235cfu/100mL. One can

see from this map and the animation that extremely high *E.coli* concentrations (> 600cfu/100mL) can be found along the eastern side of the basin in the North and South Branch and Lower Raritan watershed management areas (WMA). Over the study period the Lower Raritan WMA remained consistently contaminated with *E.coli* well above the state standard for contact recreation. A large proportion of the Lower Raritan is urban while the NS Branch is a mix of agricultural/forest/urban (figure S1). Urban areas have a large concentration of potential E.coli sources, while forested and agricultural areas have fewer controls on surface water inputs, therefore higher E.coli concentrations may be expected in these areas. In addition, several hot spots could be identified in the upper Millstone WMA that would appear and then dissipate, suggesting the occurrence of acute point source contamination in those areas. Figure S1 illustrates the locations of sewage receiving plants that discharge into surface water. A large cluster of these plants in combination with a high percentage of agricultural land in the Millstone WMA could potentially explain these hot spots, but further investigation is needed to identify specific sources and is beyond the scope of this work. It should also be noted that high *E.coli* concentrations were estimated in many areas where no monitoring stations existed. In these areas, additional monitoring strategies may be needed to capture potential harmful levels of *E.coli*. On a basin-wide scale, the average daily E.coli concentration was highly variable, with exceedances spiking in 2003 (Figure S5).

It is also important to assess the confidence in these estimations and describe the probability that a particular river mile is impaired for *E.coli*. This information is important for decision-makers and environmental managers when deciding how to allocate resources and devise public warnings of fecal contamination. Using the log-*E.coli* posterior PDF calculated at regularly spaced estimation points along the Raritan, we calculated for each day of the study period the percentage of river miles with a probability of impairment (Eq. 7) greater than 90%. Figure 4 depicts these results for a 300 day window of the study period. The x-axis is the day of estimation and the y-axis is the percentage of river miles in the Raritan Basin that exceeded the standard with 90% confidence. The fraction of river miles having a >90% probability of being impaired was highly variable from one day to another, and reached a maximum of 35% on the most polluted day of this time period. Figure 4 and Figure S5 show a similar trend toward high *E.coli* concentrations in late 2003/early 2004, suggesting larger than normal influences on *E.coli* emissions. Vidon *et al.* (18) suggest that discharge and precipitation are the best indicators of *E.coli* loading, they also found loading tended to be higher in the winter/spring, which may help to account for the higher than normal readings estimated in the Raritan basin for this time period. Further analysis of precipitation and discharge patterns would help to quantify these potential influences in the Raritan. In addition, a majority of *E.coli* measurements were taken in the summer months, while turbidity measurements occurred throughout the year, which provides crucial secondary information during the high *E.coli* winter months and may further explain the accuracy improvement that we obtained.

It should be noted that our geostatistical model can be applied to a variety of basins with varying data density; however, too few data can affect the covariance calculations. A minimum of 10–50 data points should exist to construct a correlation model, and depending on the size of the watershed, more points may be necessary. A larger watershed with large datasets may also be numerically challenging, something that will be expanded upon in future work. Overall this study provides a unique spatiotemporal framework for incorporating river distances and secondary information into the basin-wide assessment of water quality. Accuracy has been improved by over 30% when combining river distances and turbidity as an indicator of *E.coli* concentration. By constructing our model in this way, we are better able to estimate *E.coli* along un-monitored stream segments, thereby increasing the overall number of river miles assessed and providing environmental managers with accurate maps that not only show the spatial and temporal distribution of *E.coli* but that can also highlight areas of concern, which

can be useful when evaluating future monitoring strategies and allocating state and local resources.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful for the assistance of staff for the state of New Jersey. This work was supported by the New Jersey Dep. of Environmental Protection (contracts SR03-046, SR04-062 and SR05-050) and grants from the National Inst. of Environmental Health Sciences (grants no. 5 P42 ES05948, P30ES10126 and T32 ES007018).

References

1. Mallin MA, Williams KE, Esham EC, Lowe RP. Effect of human development on bacteriological water quality in coastal watersheds. Ecological Applications 2000;10(4):1047–1056.
2. U.S. Environmental Protection Agency. Guide to Monitoring Water Quality. Section 5.11 Fecal Bacteria. Washington, DC: EPA; 2000.
3. Delhomme JP. Kriging in the hydrosciences. Advances in Water Res 1978;1(5):251–266.
4. Cressie N. The origins of kriging. Math. Geol 1990;22(3):239–252.
5. Stein, ML. Interpolation of Spatial Data: Some Theory for Kriging. New York: Springer; 1999.
6. Darwish, KhM; Kotb, MM.; Ali, R. Mapping soil salinity using collocated cokriging in Bayariya, Oasis, Egypt. Proceedings of the 5th International Symposium on Spatial Data Quality, ITC Enschede; June 13–15; The Netherlands. 2007.
7. Li Z, Zhang Y-K, Schilling K, Skopec M. Cokriging estimation of daily suspended sediment loads. Journal of Hydro 2006;327:389–398.
8. Seo DJ, Krajewski WF, Azimi-Zonooz A, Bowles DS. Stochastic interpolation of rainfall data from rain gauges and radar using cokriging. Water Resources Res 1990;26(5):915–924.
9. Jager HI, Sale MJ, Schmoyer RL. Cokriging to assess regional stream quality in the southern blue ridge province. Water Resources Res 1990;26(7):1401–1412.
10. Goovaerts, P. Geostatistics for Natural Resources Evaluation. London: Oxford University Press; 1997.
11. Christakos G, Li X. Bayesian maximum entropy analysis and mapping: A farewell to kriging estimators. Math. Geol 1998;30:435–462.
12. Akita Y, Carter G, Serre ML. Spatiotemporal nonattainment assessment of surface water Tetrachloroethene in New Jersey. Journal of Env. Qual 2007;36(2):508–520.
13. LoBuglio JN, Characklis GW, Serre ML. Cost-effective water quality assessment through the integration of monitoring data and modeling results. Water Resources Res 2007;43
14. American Public Health Association. Standard Methods for the Examination of Water and Wastewater. Washington, DC: APHA; 1992.
15. Sayler GS, Nelson JD, Justice A, Colwell RR. Distribution and significance of fecal indicator organisms in the upper Chesapeake Bay. Applied Micro 1975;30:625–638.
16. Adams, PD.; Hollabaugh, CL.; Harris, RR. Environmental assessment of the Chattahoochee River in west Georgia: relationships between flow and sediment and bacteria. Geological Society of America Annual Meeting; October 28–31; Denver, CO. 2007.
17. Dorner SM, Anderson WB, Huck PM, Gaulin T, Candon HL, Slawson RM, Payment P. Pathogen and indicator variability in a heavily impacted watershed. Journal of Water & Health 2007;5(2):241–257. [PubMed: 17674573]
18. Vidon P, Tedesco LP, Wilson J, Campbell MA, Casey LR. Direct and indirect hydrological controls on *E.coli* concentration and loading in midwestern streams. Journal of Env. Qual 2008;37:1761–1768.
19. Reeves RL, Grant SB, Mrse RD, Oancea CMC, Sanders BF, Bochm AB. Scaling and management of fecal indicator bacteria in runoff from coastal urban watershed in southern California. Environ. Sci. Technol 2004;38:2637–2648. [PubMed: 15180060]

20. N.J. Department of Environmental Protection. Trenton, NJ: NJDEP; 2002. Raritan Basin: Portrait of a Watershed.
21. Noble RT, Weisberg SB, Leekaster MK, Mcgee CD, Ritter K, Walker KO, Vainik PM. Comparison of beach water quality indicator measurement methods. Environ. Modeling and Assessment 2003;81:301–312.
22. Griffith JF, Aumand LA, Lee IM, Mcgee CD, Othman LL, Ritter KJ, Walker OK, Weisberg SB. Comparison and verification of bacterial water quality indicator measurement methods using ambient coastal water samples. Environ. Modeling and Assessment 2003;116:335–344.
23. Christakos G. A Bayesian/maximum-entropy view on the spatial estimation problem. Math. Geol 1990;22(7):763–776.
24. Christakos, G. Modern Spatiotemporal Geostatistics. Vol. 2nd Edition. New York: Oxford University Press; 2000.
25. Serre, ML.; Bogaert, P.; Christakos, G. Computational investigations of bayesian maximum entropy spatiotemporal mapping. In: Buccianti, A.; Nardi, G.; Potenza, R.; De Frede, Editors, editors. 4th Annual Conference of the International Association of Mathematical Geology; Naples, Italy: 1998. p. 117-122.
26. Serre ML, Christakos G. Modern geostatistics: computational BME in light of uncertain physical knowledge—the equus beds study. Stoch. Environ. Res. and Risk Assessment 1999;13(1):1–26.
27. Christakos, G.; Bogaert, P.; Serre, ML. Temporal GIS: Advanced Functions for Field Based Applications. New York: Springer; 2002. p. 217
28. Money E, Carter GP, Serre ML. Using river distance in the space/ time estimation of dissolved oxygen along two impaired river networks in New Jersey. Water Research. In Press
29. Money, E.; Carter, GP.; Serre, ML. Department of Environmental Sciences and Engineering. UNC-BMElab Technical Report 2008-08. Chapel Hill, NC, USA: University of North Carolina; 2008. Covariance models for directed tree river networks; 18 p.
30. Ver Hoef JM, Peterson EE, Theobold D. Spatial statistical models that use flow and stream distance. Environ. and Ecol. Statistics 2006;13:449–464.
31. de Fouquet C, Bernard-Michel C. Modèles géostatistiques de concentrations ou de débits le long des cours d'eau. Comptes Rendus Geosciences 2006;338(3):307–318.
32. Cressie N, Frey J, Harch B, Smith M. Spatial prediction on a river network. Journal of Agricultural, Biological, and Environ. Stat 2006;11:127–150.
33. Peterson EE, Urquhart NS. Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: A case study in Maryland. Environ. Monitoring and Assessment 2006;121:615–638.

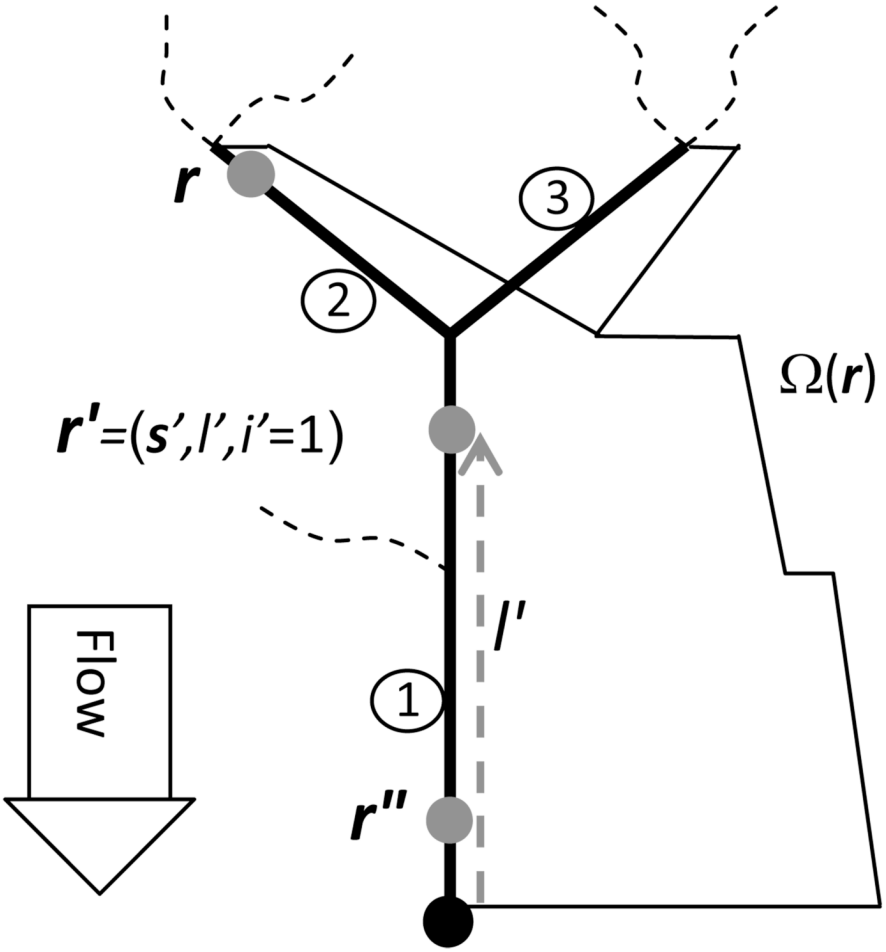

Figure 1.
Directed tree river network represented by 3 stream reaches numbered in circles. The small stream reaches shown in dotted lines have been ignored in this representation. Points $\boldsymbol{r}'=(\boldsymbol{s}',l',i'=1)$ and $\boldsymbol{r}''=(\boldsymbol{s}'',l'',i''=1)$ are on reach 1, and point $\boldsymbol{r}=(\boldsymbol{s},l,i=2)$ is on reach 2. The flow function $\Omega(\boldsymbol{r})$ is shown varying as a function of $\boldsymbol{r}$.

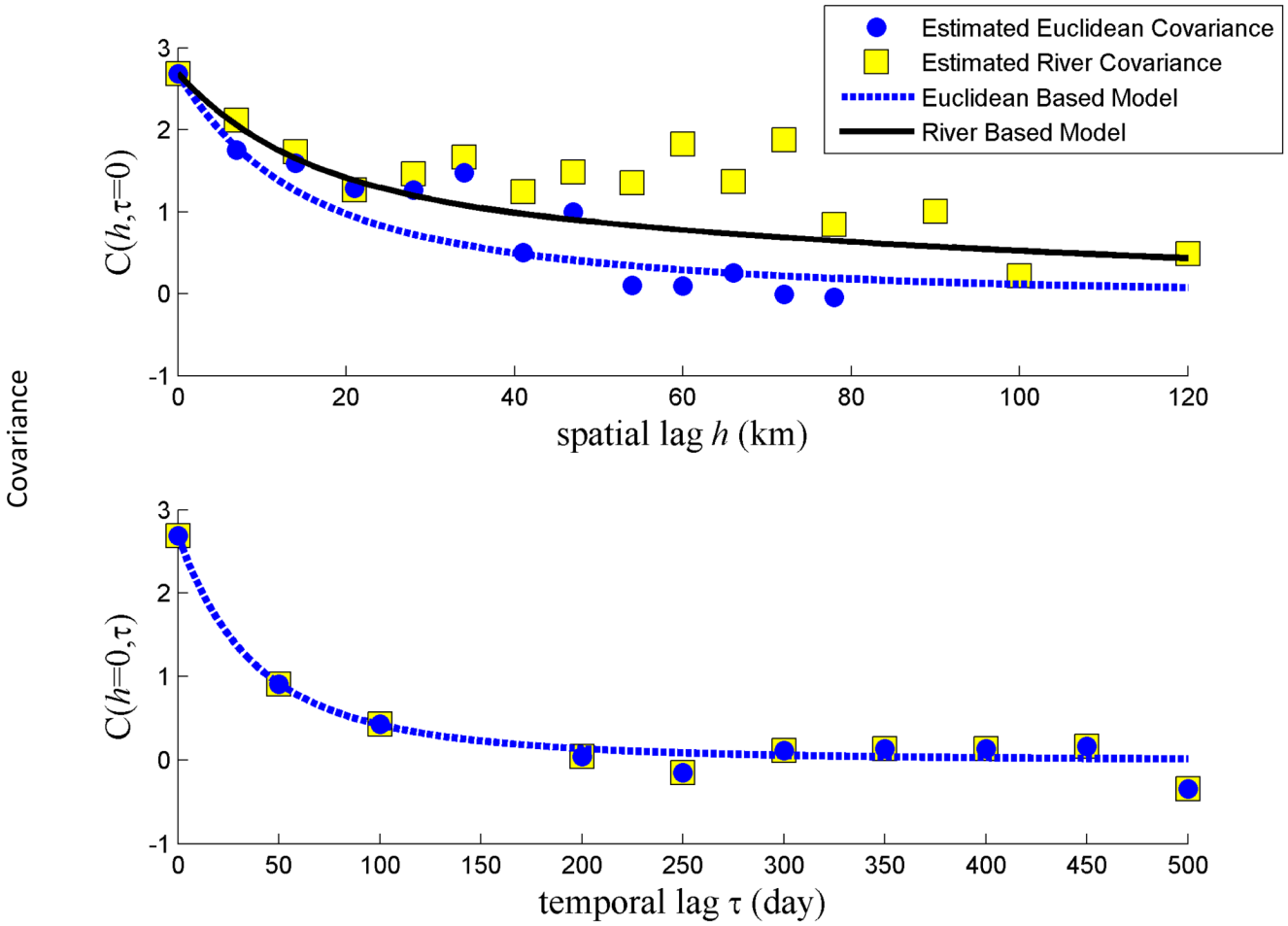

Figure 2.
Spatial (top) and temporal (bottom) covariance for *E.coli* in the Raritan Basin

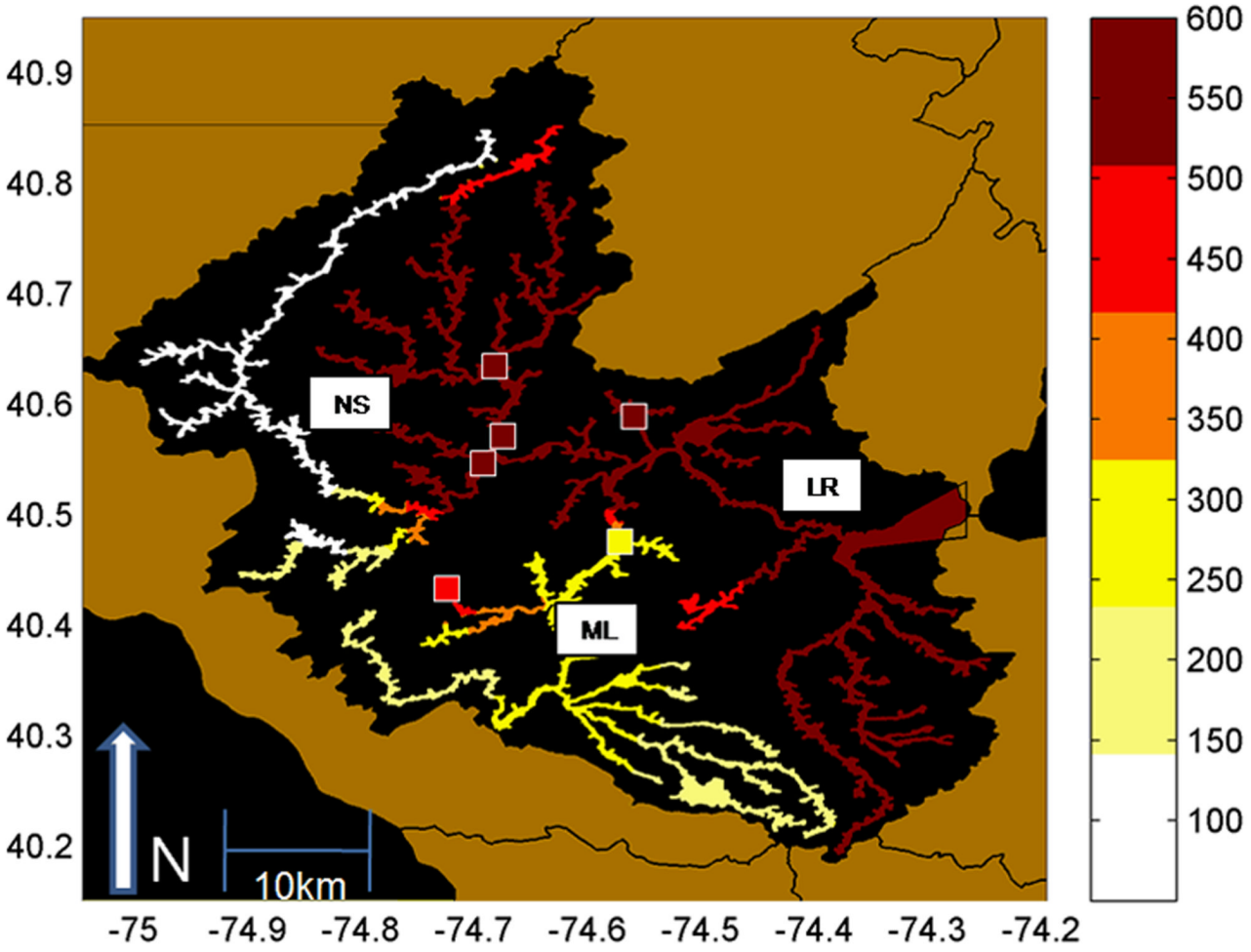

Figure 3.
Estimation of *E.coli*(cfu/100mL) on May 4, 2002 in the Raritan Basin. The NJ standard for *E.coli* is 235cfu/100mL. NS=North/South Branch WMA; ML=Millstone WMA; LR=Lower Raritan WMA.

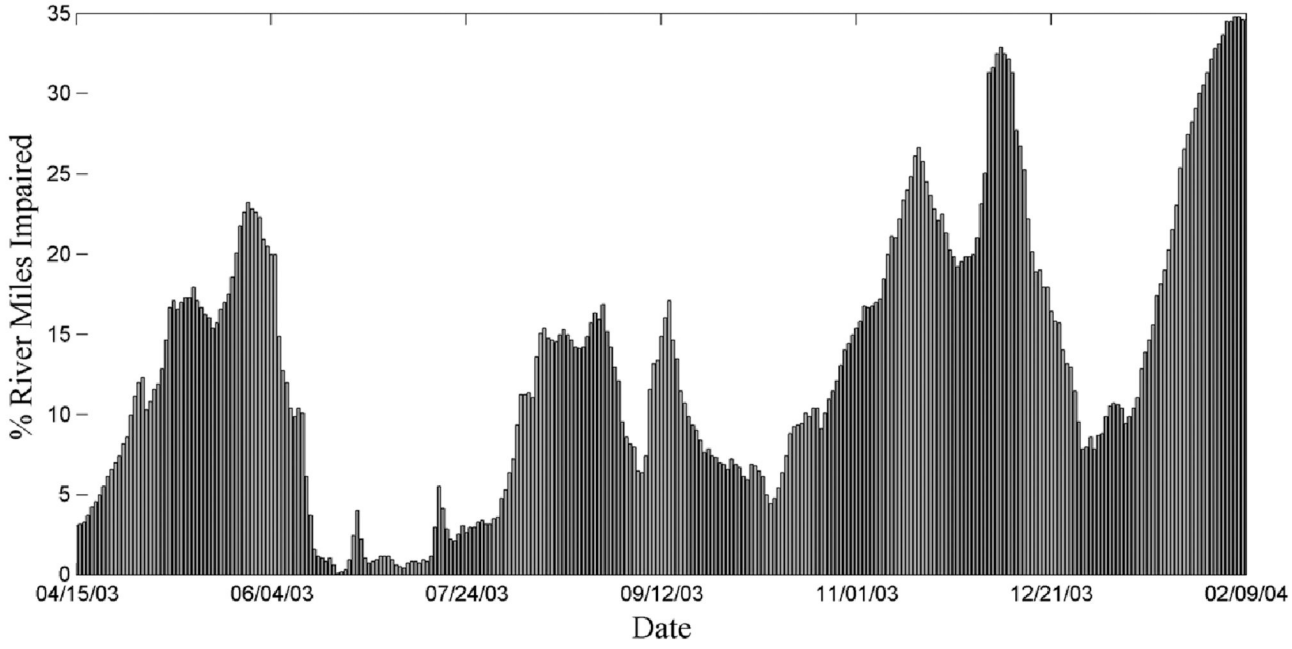

Figure 4.
Percentage of river miles with a probability of impairment > 90%. Impairment = exceeding the single sample standard of 235 cfu/100mL.

Table 1

E.coli Space/Time Covariance Model Parameters

	c_1 (*)	a_{r1} (*km*)	a_{t1} (*days*)	c_2 (*)	a_{r2} (*km*)	a_{t2} (*days*)	c_3 (*)	a_{r3} (*km*)	a_{t3} (*days*)
Euclidean	1.35	30	80	1.08	100	200	0.27	200	500
River	1.35	40	80	1.08	300	200	0.27	400	500

(*) c_1, c_2, c_3 are expressed in (log-cfu/100*mL*)^2