A neighborhood statistics model for predicting stream pathogen indicator levels

Pramod K Pandey [1*], Gregory B Pasternack[2], Mahbubul Majumder[3], Michelle L Soupir[4], Mark S. Kaiser[3],

[1]Department of Population Health and Reproduction, University of California, Davis, California 95616, USA

[2]Department of Land, Air and Water Resources, University of California, Davis, California 95616, USA

[3]Department of Statistics & Statistical Laboratory, Iowa State University, Ames, IA 50014, USA

[4]Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50014, USA

[*]Corresponding author email: *pkpandey@ucdavis.edu*

Abstract

Because elevated levels of water borne *E. coli* in streams are a leading cause of water quality impairments in the U.S., water quality managers need tools for predicting aqueous *E. coli* levels. Presently *E. coli* levels may be predicted using complex mechanistic models that have a high degree of unchecked uncertainty or simpler statistical models. To assess spatio-temporal patterns of instream *E. coli* levels, herein we measured *E. coli*, a pathogen indicator, at 16 sites (at four different times) within the Squaw Creek watershed, Iowa, and subsequently the Markov Random Field model was exploited to develop a neighborhood statistics model for predicting instream *E. coli* levels. Two observed covariates, local water temperature (°C) and mean cross-sectional depth (m), were used as inputs to the model. Predictions of *E. coli* levels in the water column were compared with independent observational data collected from sixteen in-stream locations. The results revealed that spatio-temporal averages of predicted and observed *E. coli* levels were extremely close (all within factor of 2), while 66% of individual predicted *E. coli* concentrations were within a factor of 2 of the observed values. In only one event, difference between prediction and observation was beyond 1 order of magnitude. The mean of all predicted values at sixteen locations was approximately 1% higher than the mean of the observed values. The approach presented here will be useful while assessing instream contaminations such as pathogen/pathogen indicator levels at watershed scale.

Keywords: stream water; *E. coli*; neighborhood structures; Markov Random Field model

1. Introduction

Unsafe levels of pathogens in ambient water bodies such as streams, ground water, lakes and reservoirs, estuaries, and coastal waters are a major concern for the environment and pose a serious risk to public health (U.S. EPA 2012a). Predictive models have been developed to simulate watershed-scale hydrological processes and associated bacterial transport and interactions. In this study, we report spatio-temporal patterns of *E.coli* levels in a stream network and then introduce the use of a neighborhood statistics model for predicting stream pathogen indicator levels.

1.1 Research motivation

Water borne pathogens have been linked to various diseases, including diarrhea, malaria, yellow fever, dengue, hepatitis A, Hepatitis E, and typhoid fever. For example, approximately 37.5% of diarrhea cases in developing countries are due to contaminated water. Even in a developed country such as United States, approximately 60% of total diarrhea cases are attributable to unsafe water and poor hygiene. According to the World Health Organization (WHO) approximately 4% of the global disease burden is caused by contaminated water (WHO, 2010); improving water quality is a viable option for mitigating health risk.

One of the leading causes of stream water quality impairment in the U.S. is elevated levels of pathogens such as *E. coli*, which is also an indicator of the presence of other pathogens. According to the EPA's national summary of impaired waters, approximately 40,235 water bodies are impaired; approximately 15% of the total 71,363 impairments are due to water borne pathogens (EPA 2013). One major source of *E. coli* in streams is diffuse pollution (i.e., non-point source pollution from agriculture). For example, in Iowa, where approximately 75% of the watershed is dominated by cropping land and precipitation is a major source of water for agriculture, 69% of assessed streams are impaired, with 27.5% of those due to high levels of pathogens (U.S. EPA 2012a). In California, where approximately 43% of land is dominated by agriculture, but irrigation is the major source of agriculture water, approximately 89% of assessed streams are contaminated (U.S. EPA 2013). Currently, 15.8% of total streams in California are assessed, and 15% of assessed streams are impaired by pathogens.

1.2 Predicting bacterial concentrations

Evaluating public health risks caused by water borne pathogens requires predictions of pathogen levels in ambient water bodies such as streams. In turn, predicting instream pathogen concentrations requires understanding fate and transport of pathogens at watershed scale. Previously process based modelling approaches have been used extensively for predicting pathogen levels in streams (Hipsey et al. 2008; Rehmann and Soupir, 2009; Pandey et al., 2012a,b; Droppo et al. 2009; Jamieson et al. 2005; Schilling et al. 2009; Wilkes et al. 2011). Jamieson et al. (2005) used stream bed stresses and stream flow while computing stream water column *E. coli* levels. A study by Hipsey et al. (2008) emphasized sediment properties that potentially affects stream water column *E. coli* levels. Rehmann and Soupir (2009) used a one-dimensional approach to understand the impacts of interactions between water column and streambed sediment on *E. coli* concentrations in streams. Pandey et al. (2012a) calculated *E. coli* resuspension rate, while predicting water column *E. coli* levels at watershed scale. Similarly, Kim et al. (2010) embedded a resuspension of *E. coli* to the existing Soil Water Assessment Tool (SWAT) for predicting in stream *E. coli* levels, while Parajuli et al. (2009) used the SWAT model for predicting instream *E. coli* levels without adding resuspension process. While previous approaches considerably enhanced the understanding of bacteria fate and transport in streams, the development of relatively simpler approaches, such as the statistical model described herein, can be another option for predicting instream *E. coli* levels.

In addition to using process-based models, many previous studies implemented such models of instream *E. coli* levels within geographical information systems (GIS) taking advantage of geospatial data. For example, Pandey et al. (2012b) estimated watershed indexes considering undisturbed land cover (e.g., wetlands, vegetated streams) and disturbed land cover (e.g., crop land, crop land receiving animal manure, urban land) for identifying the relationships between in-stream *E. coli* levels and watershed characteristics. Studies by Rothwell et al. (2010a;b) exploited GIS tools to identify the relationships between water chemistry (e.g., pH, sulphate, cations, and nutrients) and a watershed's land cover, topography, soil, and hydrology. These studies reported that stream water quality is significantly linked to watershed characteristics.

Understanding how climate and land surface characteristics (e.g., land cover, soil, topography, and geology) interact at the watershed scale to generate runoff and transport materials is crucial for predicting and ultimately mitigating in-stream pathogen and pathogen-indicator levels. Watershed-scale models that account for these relationships to simulate processes and fluxes can help with development and implementation of a watershed management plan for improving in-stream water quality. For example, SWAT has been extensively used (Parajuli et al. 2009; Cho et al. 2010; Kim et al. 2010) to predict in-stream water *E. coli* levels. In the SWAT model, watershed characteristics such as cropland, grazing land, livestock density, decay of bacteria, and climate of the watershed (rainfall and temperature) are used as inputs for predicting bacteria levels in streams. Previous studies have shown that SWAT can help deriving suitable land management plans and guidelines supportive for mitigating instream pathogen levels.

Despite the potential opportunities for a model-based management approach, comparison between model predictions and observations of in-stream bacteria levels clearly indicates that considerable improvements in the existing models are required before their potential is reached (Nagels et al. 2002; Rehmann and Soupir 2009; Hipsey et al. 2008; Droppo et al. 2009; Dorner et al. 2006; Pachepsky and Shelton 2011). For instance, Dorner et al. (2006) developed a hydrological model (WATFLOOD model was augmented with a pathogen transport model) and found that daily predictions of *E. coli* levels varied from 1 to 4 orders of magnitude of observed values (more than 70 observations were compared with predicted values). Similarly Kim et al. (2010) predictions using SWAT model varied from 1 to 3 orders of magnitude of the observed values (more than 150 observations were compared with predicted values).

To address the underlying deficiencies, studies have suggested everything from adding more physical processes to improving statistical methods. For example, one idea has been to improve the formulations of in-stream processes such as resuspension of *E. coli* from streambed sediment to the water column in order to improve existing water quality models for bacteria predictions (Muirhead et al. 2004; Bai and Lung 2005; Jamieson et al. 2005). Another approach for improving in-stream *E. coli* predictions could be combining the capability of GIS data and statistics.

1.3 Research objectives

The overall goal of this study was to explore *E. coli* levels in a watershed stream network and test the value of a spatial neighborhood statistics model, the Markov Random Field model, to predict *E. coli* levels in streams. The model was formulated and tested for Iowa’s Squaw Creek Watershed, which is an agriculture-dominated watershed. Non-point source pollution is known to be the leading cause of bacterial contamination in the streams. This study builds upon the work of Kaiser (2010), who previously used this approach successfully to predict nitrate concentrations in the Des Moines River, Iowa prior to impoundment in Saylorville Reservoir. The study used stream flow and nitrate data (2954 observations from January, 1982 to December, 1996) from seven gaging stations along the Des Moines River from Boone to Pella (about 116 miles). The specific objectives of this study were to (i) observe and analyze how *E. coli* levels vary between four different times in relation to the covariates of water temperature and water depth, (ii) compare *E. coli* levels in tributaries versus the mainstem channel, and (iii) develop and assess the predictive prowess of a neighborhood statistics model.

2. METHODS

2.1 Field setting and observations

Squaw Creek passes through Story, Webster, Hamilton, and Boone Counties of Iowa (Figure 1). The Squaw Creek watershed, Hydrologic Unit Code (HUC) 10 (ID 0708010503), has a total drainage area of 592.4 sq km and average slope of 2%. The watershed's humid continental climate, Köppen climate classification *Dfa*, receives an average annual precipitation of 910 mm. In general, December and January are the coldest months (temperature variation from -1 to -10 ºC), and June and July are the warmest month (temperature variation from 30 to 35 ºC). The mainstem length (i.e., Squaw Creek) is 60.5 km and the total stream length (including tributaries) within the watershed is 346.7 km. There are 75 first order streams. Approximately 74% of the watershed is under agriculture: corn 41% and soybeans 33%. Forest cover is about 2.7%, and the total grassland is about 17% of the total watershed.

Corn and soybean are two major crops grown in the watershed. Planting and harvesting of corn in Iowa are done generally between April and October. Soybeans are usually planted in May after completing corn planning, with soybean harvesting in early-mid October. Corn is the major crop receiving liquid manure (mostly in fall) from confined animal feeding operations. Water samples (total 64 observations) in support of model development were collected at 16 locations along the stream on 27th June (t = 1), 6th July (t = 2), 17th July (t = 3), and 17th October 2009 (t = 4). Eight locations (1 – 8) were located in tributaries and another eight (9 – 16) were located along the mainstem (shown in Figure 1). Samples were collected using a Horizontal Polycarbonate Water Bottle Sampler (2.2 L, Forestry Suppliers Inc., Jackson, Mississippi City, USA) by lowering the instrument from a bridge into the center (≈ 15 cm below surface water) of the stream at the sampling location. After sample collections, samples were stored at 4 ^{0}C (in a cooler) immediately and were analyzed (triplicate) within 24 hours. Membrane filtration technique (US EPA method 1603) has been used for *E. coli* enumeration using modified mTEC agar (Difco^{TM}, Modified mTEC agar, Becton, Dickinson and Company, Sparks, MD, USA) (APHA 1999). In addition to *E. coli* enumeration, we also measured the stream water column depth (m) and temperature (^{0}C), while collecting water samples. Average stream water column depth along the transact at each sampling location was determined by marking off equal intervals of approximately 60 cm along the measuring string and then the mean of the water depths was used for analysis. Streamflow data was obtained for the U. S. Geological Survey gaging station (ID 05470500) at site 16 (Figure 1). Climate data, precipitation and temperature, were obtained for Ames City (lat 42.02, long – 93.77) using Iowa Mesonet, Iowa State University, Ames, Iowa (IEM 2012).

To address the first objective, we performed a comparative analysis of event based observation data of water column *E. coli* levels, stream water depths, and stream water temperatures that were collected at 16 locations along the stream at four different times. The second objective was addressed by exploiting the use of Mann-Whitney U test, a non-parametric test. The test was used to compare *E. coli* levels across all sites in tributaries and main stem at each time. Further, Pearson correlation coefficients were estimated to relate *E. coli* levels among sampling locations. The third objective was resolved by developing a statistics model that uses a neighborhood structure linking *E. coli* levels in downstream locations with upstream sampling locations. Subsequently model predictions were compared with observations to verify the model's predictability. In addition, EPA's water quality criteria of indicator organisms (*E. coli*) of fresh water were used as reference points, while comparing the model predictions and observations.

2.2 Neighborhood statistics model

To develop the model for the study area, we developed the conditionally specified model for Squaw Creek. In equation 1, Y is a random variable, and $s_i \equiv (l, t)$ where l is sampling locations (1–16), and t is sampling events (1 – 4).

$$\begin{aligned} Y &\equiv \left\{ Y\left(s_i\right) : i = 1, \dots 64 \right\} \\ &= \left\{ Y\left(l, t\right) : l = 1, \dots 16; t = 1, \dots 4 \right\} \end{aligned} \tag{1}$$

We assume that the temporal distributions of *E. coli* at a station, conditional on all stations upstream depends only on closest upstream stations. Based on sampling locations shown in Figure 1, neighborhood structures were developed, which are shown in Table 1. The criteria of neighbor selection were defined based on inflowing tributaries and sampling locations. For each sampling location, upstream tributaries, and immediate upstream and downstream locations were defined as neighbors. For example, location 10 has two tributaries just upstream, therefore these two tributaries (locations 1 and 2) are considered neighbors in addition to location 11 (immediately downstream).

We also assume that measurements of *E. coli* concentrations are independent in time. This leads us to define neighbors of *Y (s$_i$)* as:

$$N_i \equiv \left\{ s_j : s_j \in \left\{ (l-1,t),(l+1,t) \right\}; i = 1,\ldots,n \right\} \tag{2}$$

Then

$$\left[Y(s_i) \middle| \left\{ Y(s_j) : j \neq i \right\} \right] = \left[Y(s_i) \middle| Y(N_i) \right] \tag{3}$$

For *i* = 1, …., *n* let Y (*s$_i$*) have conditional density

$$f\left(y(s_i) \middle| y(N_i) \right) = Gau\left(\mu_i, \tau^2 \right) \tag{4}$$

where

$$\mu_i = \theta_i + \sum_{j \in N_i} c_{i,j} \left(y(s_j) - \theta_j \right) \tag{5}$$

subject to $c_{i,j} = c_j$ where

$\theta \equiv \left(\theta_1,...,\theta_n\right)^T$ $\theta = \{\theta\}$ is the parameter vector of marginal mean that incorporate the covariates X_i, i = 1, 2, .., p. We used two covariates (p = 2) temperature (^{o}C) and stream water depth (m). Thus we have

$$\theta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = X\beta . \tag{6}$$

The joint distribution (Besag 1974; Cressie 1993) is:

$$Y \approx Gau\left(\theta;\left(I - C\right)^{-1} I\tau^2\right) \tag{7}$$

where

$$C \equiv \left[c_{i,j}\right]_{N\times N} \; and \;\; c_{i,j} = 0 \, if \;\; j \notin N_i \tag{8}$$

For this model, C has the form

$$C = \eta I_n H \tag{9}$$

where H is a block diagonal matrix of size 64 × 64; and each block (size = 16 × 16) consists of the neighborhood structures based on inflowing tributaries and sampling locations. The neighborhood was defined as 1 if two locations are neighbor; and 0 otherwise.

To obtain the estimates of these parameters we apply the maximum likelihood approach (Kaiser and Nordman 2012; Kaiser 2010). The Log likelihood function for the above model is:

$$\begin{aligned} L\left(\beta,\tau^2,\eta\right) &= \left(1/2\right) Log\left(\left|I - C\right|\right) - \left(N/2\right) Log\left(2\pi\tau^2\right) \\ &- \left(1/\left(2\tau^2\right)\right)\left(y - X\beta\right)^T \left(1 - C\right)\left(y - X\beta\right). \end{aligned} \tag{10}$$

An advantage of this model specification is that for any given η the maximum likelihood estimate (MLE) β and τ^2 are the closed form solutions, which are given by:

$$\hat{\beta} = \left[X^T\left(I-C\right)X\right]^{-1}\left[X^T\left(1-C\right)y\right] \tag{11}$$

$$\hat{\tau}^2 = \frac{1}{n}\left(y - X\hat{\beta}\right)^T\left(I-C\right)\left(y - X\hat{\beta}\right) \tag{12}$$

Once we have MLE of β and τ^2 we can plug these values into Equation 9 to get the likelihood for η that gives us MLE for η as shown in Figure 2. Again plugging in the MLE of η in above two equations, we obtain the estimates of β and τ^2. We obtain the confidence intervals of the parameters using the maximum likelihood approach described by Kaiser (2010), Cressie (1993) and Besag (1974). The values of estimated parameters τ^2, η, $\beta_{0,}$ $\beta_{1,}$ β_2 are shown in Table 2. These values were used for predicting the *E. coli* concentrations at each sampling location.

3. RESULTS AND DISCUSSION

3.1 Event-based observations

Comparing the four sampling events, each one showed a different range of *E. coli* concentrations and there were identifiable factors explaining the observed differences, which are shown in Figure 3. As an example, *E. coli* levels varied from 144 to 944 CFU/100 mL in the first spatial sampling event (t = 1). During the second sampling event (t = 2), *E. coli* levels varied from 336 to 633 CFU/100 mL, which was a narrower range than for event 1. Prior to this sampling event, the watershed witnessed around 50 mm of cumulative rainfall in the first two weeks of June. Between the first two sampling events cumulative rainfall was less than 1 mm. The average of *E. coli* levels at 16 locations (shown in Figure 3) at t = 1 was 30% higher than that of at t = 2.

During t = 3, *E. coli* levels varied from 225 to 5467 CFU/100 mL. One location (sampling point 7 of Figure 1) showed the maximum large *E. coli* level. Though between t = 2 and t = 3, there was no additional rainfall, and streamflow was also identical to preceding sampling events, at t = 3, the average of *E. coli* levels at 16 locations was 78% and 132% higher than that of during t = 1 and t = 2, respectively.

During t = 4, *E. coli* levels in tributaries varied from 53 – 333 CFU/100 mL, while in mainstem variation was from 17 – 120 CFU/100 mL. *E. coli* levels during t = 4 were considerably low compared to t = 1, 2, 3. The average *E. coli* level at t = 4 was only 14% of the *E. coli* levels at t = 3. Between t = 3 and t = 4, the cumulative rainfall was only 6.5 mm; however, temperature was considerably lower. For instance, the minimum and maximum daily air temperatures at t = 3 were 11.3 and 19.2 °C, respectively, while at t = 4, these values were –0.5 and 7.4 °C, respectively. During this sampling event, stream flow was 0.13 m^3/s, which is about 90% lower than that during t = 3. Overall, event-scale results indicated that winter season (i.e., low temperature) could be the potential reason for low *E. coli* levels.

Stream water temperatures and stream water depths are shown in Figure 4. The average daily temperatures during t = 1, 2, 3, and 4 were 24.5, 22.9, 15.3, and 3.5 °C, while average stream water temperatures were 21.3 ± 2.5 °C, 24.2 ± 1.2 °C, 19.9 ±1.9 °C, and 12.2 ± 3 °C, respectively. The stream water depths during these sampling events (0.5, 0.5, 0.4, 0.3 m, respectively) were generally similar, but declined for the last two events, with the value at t = 4 being the lowest observed. A total rainfall of 146 mm occurred May 1 to June 27, 2009. The average streamflow for the same period was 9.6 (± 6.4) m^3/s with a range from 3 to 30.4 m^3/s.

3.2 Tributaries vs. mainstem *E. coli* level analyses

Results of Mann-Whitney U test indicated that there was no significant difference (significant level of 0.05) in *E. coli* levels among t = 1 and t = 2 (all 16 sampling locations). *E. coli* levels among t = 2 and t =3 were significantly different. There was also significant difference among *E. coli* levels of t = 3 and t = 4, and t = 2 and t = 4. Further, there was significant difference in *E. coli* levels among t = 1 and t = 4.

A Pearson correlation matrix relating *E. coli* observations among 16 sites are shown in Table 3. Analysis showed significant correlations among the sampling locations (Table 3). Out of 120 correlations of sixteen sampling locations, 65% have shown high correlation (r > 0.70; p = 0.05). Relatively greater levels of correlation existed among proximal locations, particularly along the Squaw Creek. For example, locations: 1, 2 and 10; and 12, 13, and 14; and 14, 15 and 16.

3.3 Model results

The neighborhood statistics model implemented to predict in-stream *E. coli* levels yielded values within the range observed. Whereas the model produced similar *E. coli* concentrations within a relatively narrow range or both tributaries (Figure 5A) and the mainstem stream (Figure 5B), sampling observations showed a wider range in both settings. As shown in the figure, the model was not able to predict very high and low values. Compared to low values, the model predictions were reasonable well for higher *E. coli* levels.

Time averages of observed and predicted *E. coli* levels for tributaries and mainstem were very similar, but those for predictions showed less spatial variability (Figure 5C,D). The spatio-temporal average of tributary observations was 341 CFU/100 ml, while that of tributary predictions was 343 CFU/100 ml. Similarly, that of mainstem observations and predictions was 337 and 339 CFU/100 ml, respectively. These averages are all extremely close.

Besides local water temperature and depth, many other local parameters of natural streams, such as channel geometry, nutrient concentrations, solar radiation, and dissolved oxygen also impact *E. coli* levels (Hipsey et al. 2008). In this model we use only two covariates stream water depth and temperature, which might be the reason for the relatively large difference (in few predictions) between measured and predicted values (Fig 5A & B); however, considering the uncertainties involved in predicting *E. coli* levels in natural streams, which is influenced by many factors such as grazing operations, livestock density, cropping land, and land management practices, this parsimonious model can be considered reasonably good for predicting instream *E. coli* concentrations. We anticipate availability of a larger dataset could improve the model results.

Comparing the predictions of this study with previous ones (Kim et al. 2010; Dorner et al. 2006), the model predictions fit reasonably well. For instance, in the referenced studies predictions were only within 1 – 4 orders of magnitude of the observations, while in this study the average of predicted values were within a factor of 2 of the observed values (Figs. 5C,D), which is substantially better. Figure 6 compares average observations with predictions of 16 sampling locations in reference to EPA guidelines (based on the 1986 RWQC) (U.S. EPA 2012b) that say geometric mean (GM) coliform density and statistical threshold value (STV) of indicator organisms for waters designated for primary contact recreation should be less than 126 CFU/100 mL and 410 CFU/100 mL, respectively. The figure showed that both average predictions and observations exceeded EPA's GM criteria. About 18% observations exceeded EPA's STV criteria, while all predictions were lower than the STV value indicating model's under predictions for few locations. Nevertheless, 81% of both predictions and observations were lower than the STV value indicating the model's suitability for assessing instream water quality.

In addition, the neighborhood statistics model proposed here does not requires intensive calibration, which is necessary in hydrological models while implementing for predicting in-stream *E. coli* levels at watershed scale. Even though it is not expected to fit the observations with predictions very well, while predicting in-stream bacteria levels (Dorner et al. 2006), advancing existing modelling approaches are necessary in order to derive/identify efficient watershed management plans for improving stream water quality. The approach we presented here requires further improvement, and we anticipate that using a larger observed dataset will potentially enhance the predictions. One major challenge in stream bacteria modeling is the availability of limited observed data. Therefore, future studies carrying out extensive monitoring as well as modeling based on the field observations will certainly improve the existing models.

4. CONCLUSIONS

To predict in-stream *E. coli* levels, we have developed a neighborhood statistics model, Markov Random Field model, which was implemented in the Squaw Creek watershed, Iowa. The model predictions were compared with the observed *E. coli* levels at 16 different locations. The two independent parameters water temperature (^{o}C) and stream water depth (m) were used for predicting the *E. coli* levels. Results indicated that the method used here is a potentially useful approach to predict instream *E. coli* levels at watershed scale with certain degree of predictability. The approach can be useful in understanding of the spatial variability of *E. coli* levels at watershed scale.

ACKNOWLEDGMENTS

The authors thank to Division of Agriculture and Natural Resources (ANR) and Veterinary Medicine Extension, University of California, Davis, and National Science Foundation (award No. CBET-0967845) for supporting this work.

References

American Public Health Association (APHA). (1999). Standard methods for the examination of water and wastewater, AWWA, Water Environment Federation.

Bai, S., & Lung, W. S. (2005). Modeling sediment impact on the transport of fecal bacteria. *Water Research, 39 (20),* 5232–5240.

Besag, J. (1974). Spatial integration and the statistical analysis of lattice systems. *Journal of Royal Statistical Society, 36 (2),* 192–236.

Cho, K. H., Pachepsky, Y. A. Kim, J. H., Guber, A. K., Shelton, D. R. & Rowland, R. (2010). Release of Escherichia coli from the bottom sediment in a first-order creek: Experiment and reach-specific modeling. *Journal of Hydrology, 391*, 322–332.

Cressie, N. (1993). Statistics for spatial data. John Wiley & Sons, New York.

Dorner, S.M., Anderson, W. B., Slawson, R. M., Kouwen, N., & Huck, P. M. (2006). Hydrologic modeling of pathogen fate and transport. *Environmental Science & Technology 40 (15),*4746–4753.

Droppo, I. G., Liss, S. N., Williams, D., Nelson, T., Jaskot, C., & Trapp, B. (2009). Dynamic existence of waterborne pathogens within river sediment compartments: Implications for water quality regulatory affairs. *Environmental Science & Technology, 43 (6),* 1737–1743.

Hipsey, M. R., Antenucci, J. P., & Brookes, J. D. (2008). A generic, process-based model of microbial pollution in aquatic systems. *Water Resources Research, 44 (7)*, W07408.

Iowa Environment Mesonet (IEM). (2012). Iowa AG Climate Network, Iowa State University. http://mesonet.agron.iastate.edu/agclimate/hist/hourlyRequest.php.

Jamieson, R.C., Joy, D. M., Lee, H., Kostaschuk, R., & Gordon, R. J. (2005). Resuspension of sediment-associated *Escherichia coli* in a natural stream. *Journal of Environmental Quality, 34 (2)*, 581–589.

Kaiser, M.S. (2010). Statistical methods for spatial data. Course lecture, Department of Statistics, Iowa State University.

Kaiser, M. S., & Nordman, D. J. (2012). Blockwise empirical likelihood for spatial Markov Model assessment. *Statistics and Its Interface, 0*,1–8.

Kim, J.W., Pachepsky, Y.A., Shelton, D. R., & Coppock, C. (2010). Effect of streambed bacteria release on *E. coli* concentrations: Monitoring and modeling with the modified SWAT. *Ecological Modelling, 221 (12)*,1592–1604.

Muirhead, R.W., Davies-Colley, R. J., Donnison, A. M. & Nagels, J. W. (2004). Fecal bacteria yields in artificial flood events: quantifying in-stream stores. *Water Research*, 38,1215–1224.

Nagels, J.W., Davies-Colley, R. J., Donnison, A. M., & Muirhead, R.W. (2002). Faecal contamination over flood events in a pastoral agricultural stream in New Zealand. *Water Science and Technology, 45 (12)*, 45–52.

Pachepsky, Y.A., & Shelton, D. R. (2011). *Escherichia coli* and fecal coliforms in freshwater and estuariene sedients. *Critical reviews in Environmental Science and technology, 41(12)*,1067–111.

Pandey, P. K., Soupir, M. L., & Rehmann, C. R. (2012a). Predicting resuspension of *Escherichia coli* from streambed sediments. *Water Research, 46*, 115–126.

Pandey, P. K., Soupir, M. L., Haddad, M., & Rothwell, J. J. (2012b). Assessing the impacts of watershed indexes and precipitation on spatial in-stream *E. coli* concentrations. *Ecological Indicator, 23*, 641–652.

Parajuli, P. B., Douglas-Mankin, K. R., Barnes, P. L., & Rossi, C.G. (2009). Fecal bacteria source characterization and sensitivity analysis of SWAT 2005. *Transaction of ASABE, 52 (6)*,1847–1858.

Rehmann, C.R., & Soupir, M.L. 2009. Importance of interactions between the water column and the sediment for microbial concentrations in streams. *Water Research, 43(18)*, 4579–4589.

Rothwell, J.J., Dise, N.B., Taylor, K.G., Allott, T.E.H., Scholefield, P., Davies, H., & Neal, C. (2010a). A spatial and seasonal assessment of river water chemistry across North West England. *Science of Total Environment, 408*, 841–855.

Rothwell, J.J., Dise, N.B., Taylor, K.G., Allott, T.E.H., Scholefield, P., Davies, H., & Neal, C. (2010b). Predicting river water quality across North West England using catchment characteristics. *Journal of Hydrology, 395(3-4)*, 153–162.

Schilling, K. E., Zhang, Y., Hill, D. R., Jones, C. S., & Wolter, C. F. (2009). Temporal variations of *E. coli* concentrations in a large Midwestern river. *Journal of hydrology, 365*,79–85.

U.S. Environmental Protection Agency (USEPA). (2012a). WATERS (Watershed Assessment, Tracking & Environmental ResultS). Washington, D.C. (accessed on 10/22/2012).

U.S. Environmental Protection Agency (U.S. EPA). (2012b). Recreational water quality criteria. http://water.epa.gov/scitech/swguidance/standards/criteria/health/recreation/index.cfm (accessed on 2/22/2013).

U.S. Environmental Protection Agency (USEPA). (2013). WATERS (Watershed Assessment, Tracking & Environmental ResultS). Washington, D.C. (accessed on 2/22/2013).

Wilkes, G., Edge, T. A., Gannon, V. P. J., Jokinen, C., Lyautey, E., Neumann, N. F., Ruecker, N., Scott, A., Sunohara, M., Topp, E., & Lapen, D. R. (2011). Associations among pathogenic bateria, parasites, and environmental and land use factors in multiple mixed-use watersheds. *Water Research*, *45(18)*, 5807–5825.

World Health Organization. (2010). Water Sanitation and Health. http://www.who.int/water_sanitation_health/diseases/en/

Table 1 Neighborhood structures of main channel

Sampling locations	Neighbors
1	10
2	10
3	11
4	12
5	12
6	13
7	13
8	16
9	16
10	1, 2,11
11	10, 3,12
12	11, 5, 4,13
13	12, 7, 6,14
14	13,15
15	14,16
16	15, 8, 9

Table 2 Parameter values of neighborhood structures

	τ^2	η	β_0	β_1	β_2
Estimate	4.0E+04	-0.02	-18.2	15.9	109.9
Lower limit	2.6E+04	-0.27	-219	5.3	-153
Upper limit	5.4E+04	0.23	183.5	26.7	373
p-value	7.8 E-09	0.56	0.42	0.001	0.20

Table 3 Correlation coefficients of *E. coli* levels at different locations

Locations	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1		0.97		0.98	0.93		0.83	0.92	0.82	0.96	1.0	0.75	0.67	0.78		
2				0.99	0.98		0.76	0.98	0.75	0.96	0.97	0.70	0.68	0.66		
3																
4					0.99		0.71	0.97	0.70	0.93	0.98					
5								0.99		0.90	0.93					
6							0.94		0.95	0.70		0.95	0.85	0.96	0.94	0.87
7									1.0	0.9	0.82	0.99	0.91	0.97	0.86	0.86
8										0.91	0.91					
9										0.89	0.82	0.99	0.90	0.98	0.87	0.86
10											0.96	0.87	0.84	0.81		0.70
11											0.75	0.66	0.78			
12													0.95	0.95	0.92	0.92
13														0.81	0.91	0.97
14															0.84	0.78
15																0.97
16																

NOTE: Only statistically significant numbers are shown in the table (p < 0.05).

Figure captions:

Fig. 1 Study Area (Squaw Creek Watershed, Iowa, U.S.A). Corn and soybean crops dominate the watershed

Fig. 2 Maximum likelihood of η

Fig. 3 Spatial observations of *E. coli* levels and climate of watershed. Top four figures shows *E. coli* levels along the stream, and bottom figure shows temperature, precipitation, and stream flow (stream flow was observed at the lowest end of the watershed i.e., location 16 of Figure 1)

Fig. 4 Observed stream water depth and water temperature at 16 locations of the watershed

Fig. 5 Comparison between observations and predictions of in-stream *E. coli* levels in Squaw Creek watershed

Fig. 6 Comparison between observed and predicted *E. coli* levels, and EPA's Geometric Mean (GM) and Statistical Threshold Value (STV) criteria of indicator organisms for fresh water. Average of observed values and predicted values of four sampling events are shown in the figure

Figure 1.

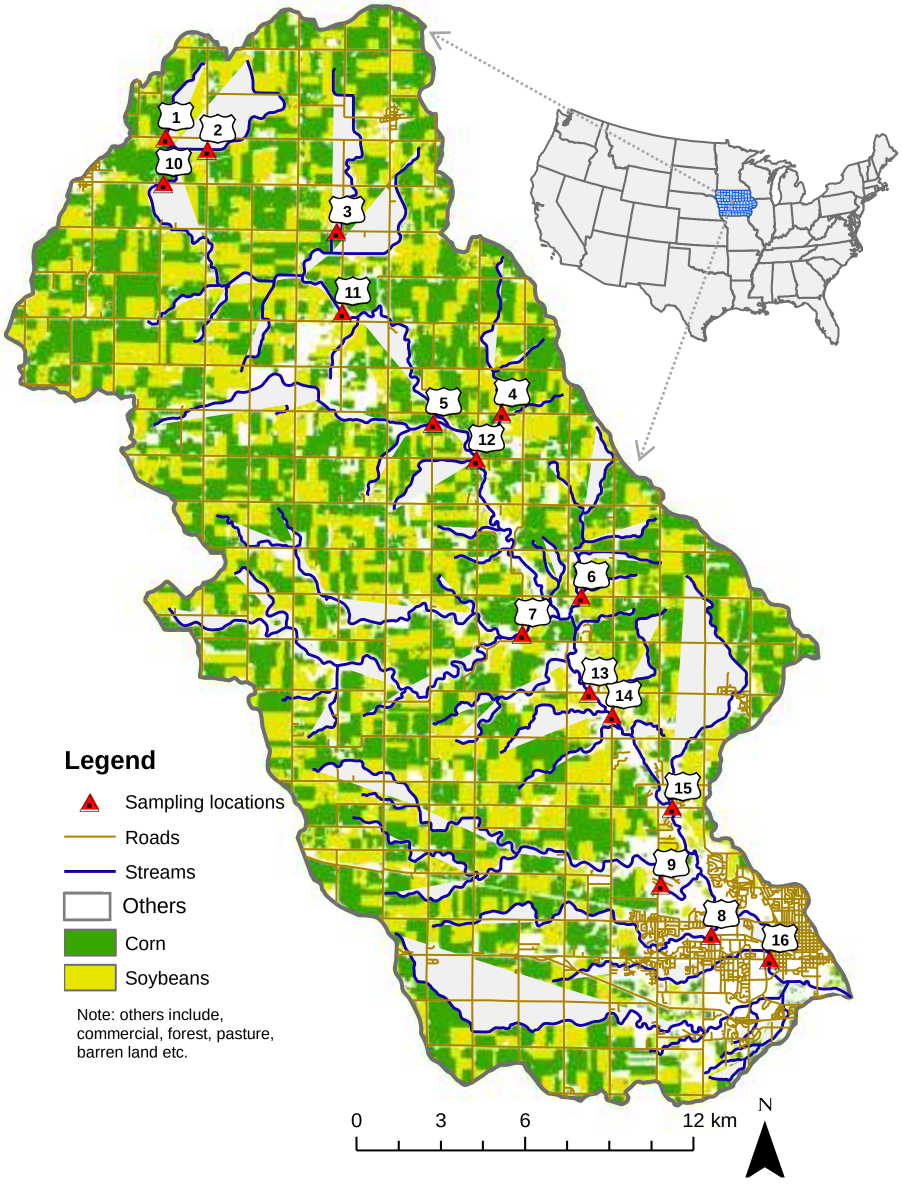

1
2
10
3
11
5
4
12
6
7
13
14
15
9
8
16
Legend
Sampling locations
Roads
Streams
Others
Corn
Soybeans
Note: others include, commercial, forest, pasture, barren land etc.
0
3
6
12 km
N

Figure 2

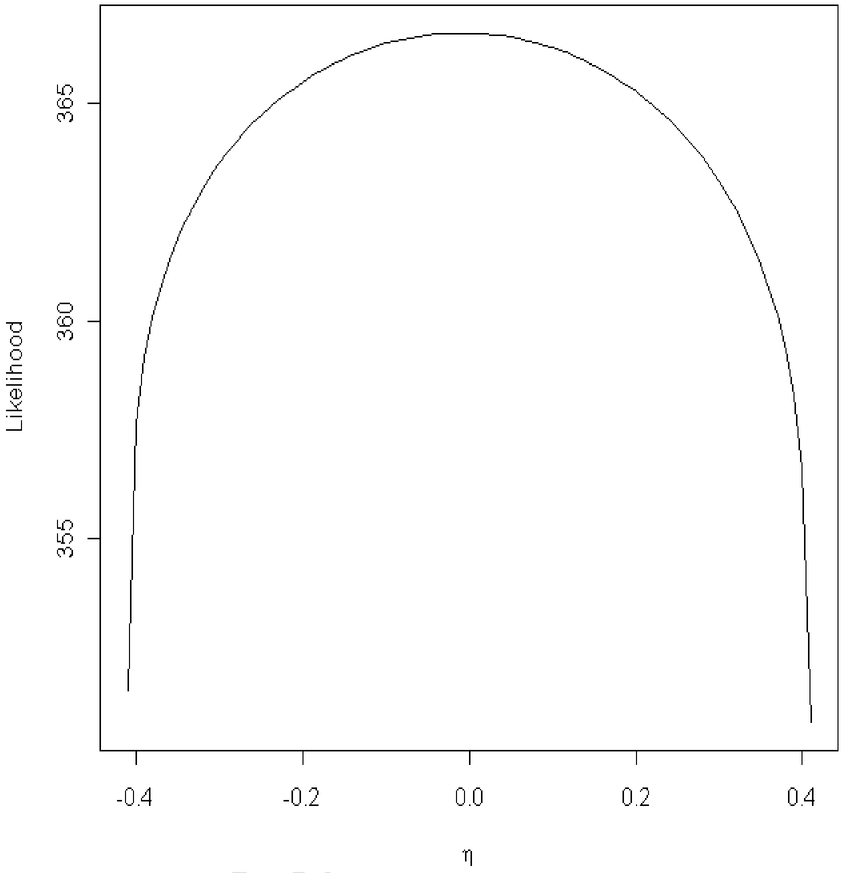
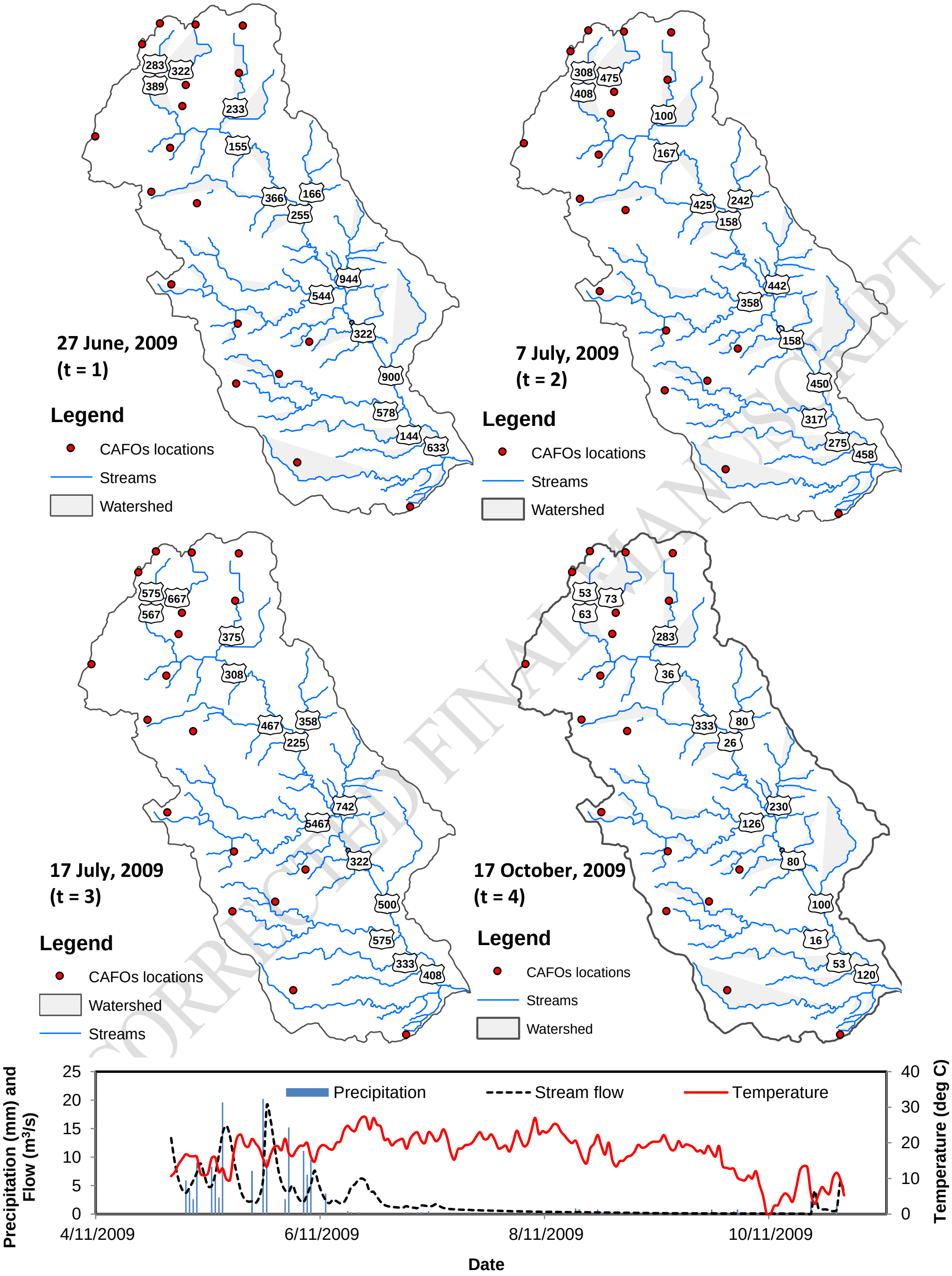


Figure 3



283
322
389
233
155
366
166
255
944
544
322
900
578
144
633
27 June, 2009
(t = 1)
Legend
● CAFOs locations
Streams
Watershed
308
475
408
100
167
425
242
158
442
358
158
450
317
275
458
7 July, 2009
(t = 2)
Legend
● CAFOs locations
Streams
Watershed
575
667
567
375
308
467
358
225
742
5467
322
500
575
333
408
17 July, 2009
(t = 3)
Legend
● CAFOs locations
Watershed
Streams
53
73
63
283
36
333
80
26
230
126
80
100
16
53
120
17 October, 2009
(t = 4)
Legend
● CAFOs locations
Streams
Watershed
Precipitation
Stream flow
Temperature
Precipitation (mm) and Flow (m^3/s)
25
20
15
10
5
0
Temperature (deg C)
40
30
20
10
0
4/11/2009
6/11/2009
8/11/2009
10/11/2009
Date

Figure 4

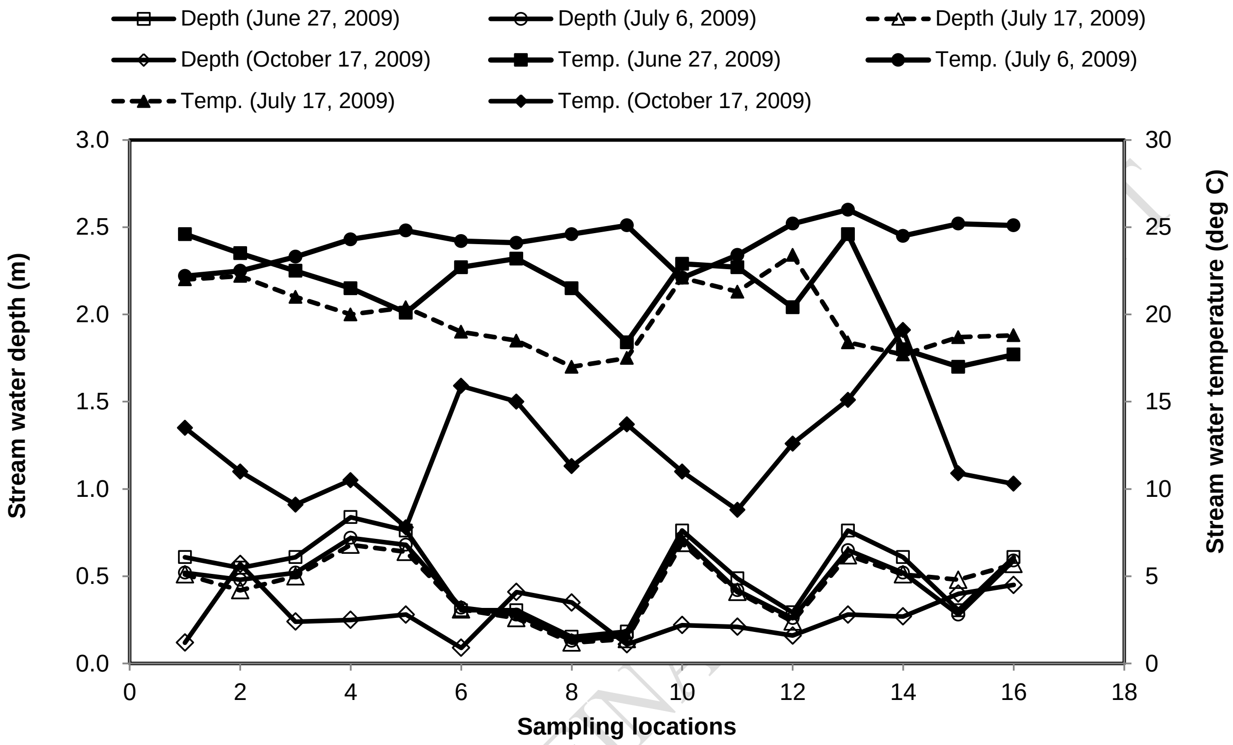

Depth (June 27, 2009)
Depth (July 6, 2009)
Depth (July 17, 2009)
Depth (October 17, 2009)
Temp. (June 27, 2009)
Temp. (July 6, 2009)
Temp. (July 17, 2009)
Temp. (October 17, 2009)
Stream water depth (m)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Stream water temperature (deg C)
30
25
20
15
10
5
0
0
2
4
6
8
10
12
14
16
18
Sampling locations

Figure 5

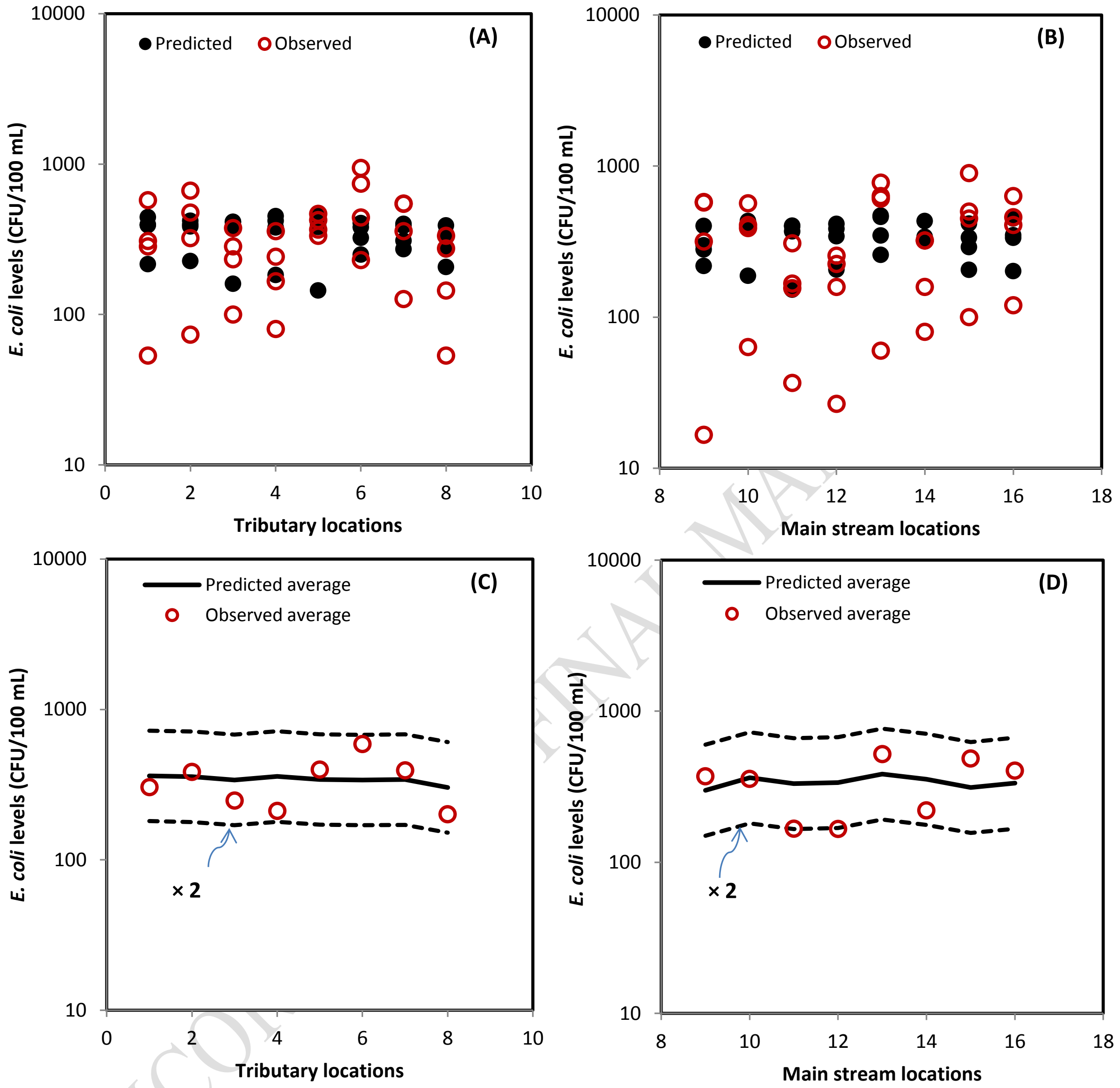
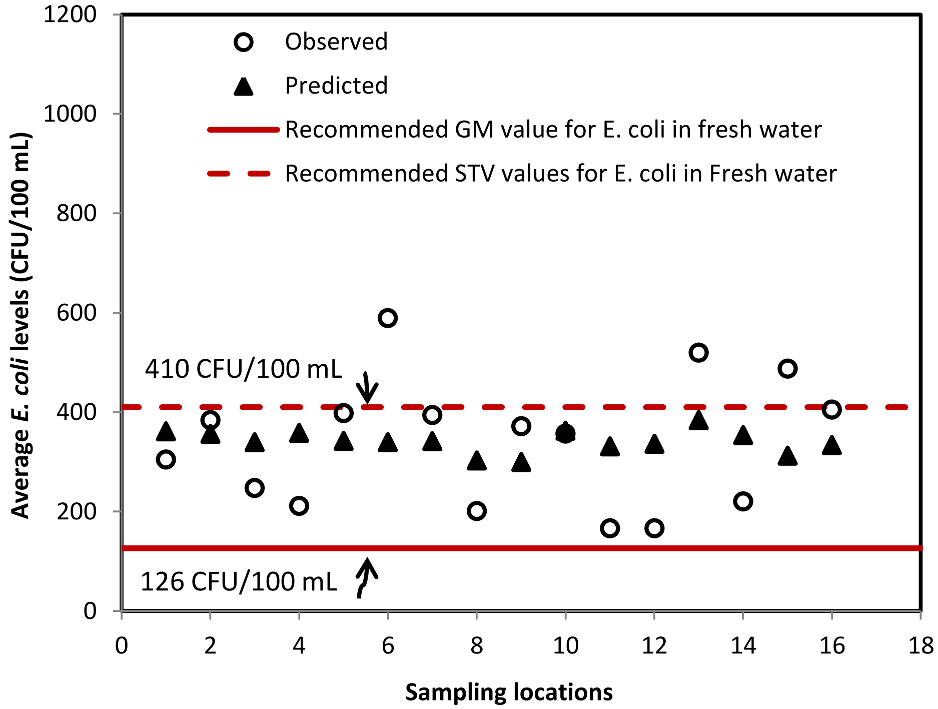

(A)
● Predicted
○ Observed
E. coli levels (CFU/100 mL)
10000
1000
100
10
0
2
4
6
8
10
Tributary locations
(B)
● Predicted
○ Observed
E. coli levels (CFU/100 mL)
10000
1000
100
10
8
10
12
14
16
18
Main stream locations
(C)
Predicted average
Observed average
× 2
E. coli levels (CFU/100 mL)
10000
1000
100
10
0
2
4
6
8
10
Tributary locations
(D)
Predicted average
Observed average
× 2
E. coli levels (CFU/100 mL)
10000
1000
100
10
8
10
12
14
16
18
Main stream locations

Figure 6.



Observed
Predicted
Recommended GM value for E. coli in fresh water
Recommended STV values for E. coli in Fresh water
410 CFU/100 mL
126 CFU/100 mL
Average *E. coli* levels (CFU/100 mL)
Sampling locations
0
200
400
600
800
1000
1200
0
2
4
6
8
10
12
14
16
18