UC Davis

UC Davis Previously Published Works

Title

Explaining and modeling the concentration and loading of Escherichia coli in a stream-A case study.

Permalink

<https://escholarship.org/uc/item/5tf7c6gj>

Authors

Wang, Chaozi
Schneider, Rebecca L
Parlange, Jean-Yves
et al.

Publication Date

2018-09-01

DOI

10.1016/j.scitotenv.2018.04.036

Peer reviewed

Explaining and modeling the concentration and loading of *Escherichia coli* in a stream—A case study

Chaozi Wang^{a,b}, Rebecca L. Schneider^c, Jean-Yves Parlange^a, Helen E. Dahlke^b, M. Todd Walter^{a,*}

^a Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY 14853, USA

^b Department of Land, Air, and Water Resources, UC Davis, Davis, CA 95616, USA

^c Department of Natural Resources, Cornell University, Ithaca, NY 14853, USA

HIGHLIGHTS

- Identified two new variables indicating instream *E. coli* levels
- Found three critical processes driving the *E. coli* fate and transport in watersheds
- Built a PLS regression model that can predict hazardous *E. coli* levels well
- Provided suggestions for management practices to reduce instream *E. coli*

GRAPHICAL ABSTRACT

ARTICLE INFO

Article history:

Received 27 November 2017

Received in revised form 12 March 2018

Accepted 4 April 2018

Available online xxx

Editor: D. Barcelo

Keywords:

Pathogen

Fate

Transport

Water quality

Watershed

Multivariate data analysis

ABSTRACT

Escherichia coli (*E. coli*) level in streams is a public health indicator. Therefore, being able to explain why *E. coli* levels are sometimes high and sometimes low is important. Using citizen science data from Fall Creek in central NY we found that complementarily using principal component analysis (PCA) and partial least squares (PLS) regression provided insights into the drivers of *E. coli* and a mechanism for predicting *E. coli* levels, respectively. We found that stormwater, temperature/season and shallow subsurface flow are the three dominant processes driving the fate and transport of *E. coli*. PLS regression modeling provided very good predictions under stormwater conditions ($R^2 = 0.85$ for log (*E. coli* concentration) and $R^2 = 0.90$ for log (*E. coli* loading)); predictions under baseflow conditions were less robust. But, in our case, both *E. coli* concentration and *E. coli* loading were significantly higher under stormwater condition, so it is probably more important to predict high-flow *E. coli* hazards than low-flow conditions. Besides previously reported good indicators of in-stream *E. coli* level, nitrate-/nitrite-nitrogen and soluble reactive phosphorus were also found to be good indicators of in-stream *E. coli* levels. These findings suggest management practices to reduce *E. coli* concentrations and loads in-streams and, eventually, reduce the risk of waterborne disease outbreak.

© 2018 Published by Elsevier B.V.

1. Introduction

Pathogens in surface waters raise a serious international concern (Ahmed et al., 2009; Ferguson et al., 2003; Hofstra, 2011; Marcheggiani et al., 2015; Oliver et al., 2005; Schets et al., 2008;

* Corresponding author.

E-mail address: mtw5@cornell.edu (M.T. Walter).

Schriewer et al., 2010; Wilkes et al., 2013). *Escherichia coli* (*E. coli*) is a commonly used indicator organism for pathogens in freshwater (Brooks et al., 2013; Dwivedi et al., 2013; Vidon et al., 2008). Understanding the factors controlling instream *E. coli* levels is critical to developing strategies to reduce disease risks due to contaminated drinking and recreational waters.

Spatially distributed watershed models, for example, Soil and Water Assessment Tool (SWAT), requires a lot of spatially distributed information, much of which is not readily available, e.g., field management schedules such as manure spreading, wildlife populations, and septic system distribution (Baffaut and Sadeghi, 2010; Benham et al., 2006; Cho et al., 2012; Frey et al., 2013). Frey et al. (2013) needed to combine SWAT with host-associated bacteroidales marker to predict the presence and absence of pathogen well. However, microbial source tracking data are sparse and hard to obtain. Cho et al. (2012) needed to modify SWAT model to include solar radiation intensity and beaver activity to predict fecal coliform concentrations in a forest dominated watershed. Walker and Stedinger (1999) required information on manure “freshness,” in addition to land use distributions, to adequately model pathogen concentrations.

Our work is focused on instream predictors of *E. coli*, therefore spatially distributed watershed models are not necessary for our purpose. Climatic/meteorological, hydrological, and water quality factors have all been shown to affect concentrations and/or loadings of *E. coli* in streams. Whitman et al. (2008) found that sunlight, season, snowmelt, and storms can influence in-stream. Vidon et al. (2008) stated that stream flow conditions (high-flow vs. baseflow), discharge, precipitation (especially seven day antecedent precipitation), turbidity, location (i.e. headwater or lower reaches) and temperature all showed significant correlations with either concentrations or loadings of *E. coli* in streams. Dwivedi et al. (2013) investigated the correlation between 13 water quality factors with *E. coli* loading and concluded that dissolved oxygen, phosphate, ammonia, chlorophyll, suspended solids, and temperature, were most correlated. Diurnal variability (Meays et al., 2006) and dissolved carbon dioxide (Gray, 1975) have also been found to be relevant.

However, current models that do not require much spatially distributed information cannot both predict and explain *E. coli* concentrations and loadings well. Artificial neural network models predict *E. coli* level well (Basant et al., 2010; Dwivedi et al., 2013), but these types of analyses do not clearly elucidate the underlying processes influencing *E. coli* transport. Statistical models offer some advantages over other approaches in that they do not require full a priori concepts of the sources, sinks, transport processes, etc. (Chen and Chang, 2014). Unfortunately, multiple linear regression (MLR) models cannot avoid co-linearity (Nevers and Whitman, 2005), which is common in water quality data. Models like LOADEST estimate constituent loads utilizing only streamflow and time (Dwivedi et al., 2013; Runkel et al., 2004); however, as summarized above, *E. coli* levels depend not only on streamflow, but also on other, perhaps intercorrelated, factors.

Principal component analysis (PCA) is a different statistical tool, which has been successfully used for concentrating the most important information in a dataset (Esbensen et al., 2002). A few researchers have used it to identify factors correlated to *E. coli* concentrations in field samples (Bech et al., 2014; Dwivedi et al., 2013). Dwivedi et al. (2013) used 6 variables in a PCA, and found that the first principal component (PC1) corresponded mostly to physical factors, such as dissolved oxygen and temperature, and the second principal component (PC2) corresponded primarily to chemical and biological factors, including phosphate, ammonia, suspended solids, and chlorophyll. Using PCA, Bech et al. (2014) found that concentrations of leached fecal bacteria from well-structured agricultural field into drainage water were negatively correlated with the number of days after slurry-manure application, which was consistent with the finding of Falbo et al. (2013). However, contrary to other studies, Bech et al. (2014) found that leached fecal bacteria was negatively-correlated with soil water

content and that leached fecal bacteria was not correlated with precipitation.

Partial least squares (PLS) regression is a regression informed by PCA, and is used to build predictive models with highly inter-correlated predictor variables (Esbensen et al., 2002). Thus, PLS is widely used in water quality modeling (Aguilera et al., 2000; Basant et al., 2010; Ortiz-Estarellas et al., 2001a; Ortiz-Estarellas et al., 2001b; Singh et al., 2007). Carroll et al. (2009) built PLS regression models to predict the corresponding sinks to confirm or challenge possible source-sink relationships and showed that the *E. coli* contamination of natural water resources in urban areas are highly correlated with human fecal contamination sources. Brooks et al. (2013) successfully used PLS regression to predict fecal indicator bacteria on Great Lakes beaches using a large suite of possible predictors. Some of these predictors are specific to beaches, like wave height, wind speed and wind direction; some predictors are also applicable to streams, such as turbidity, air temperature, antecedent rainfall (24 h and 48 h), and variables related to season (Julian date and month). Like many other statistical models, PLS regression is not necessarily reliable for revealing the underlying processes controlling the *E. coli* levels (Brooks et al., 2013).

However, to our knowledge, no previous studies have used PCA and PLS regression in combination to understand and model concentrations or loadings of *E. coli* in streams. The objective of this study was to test the applicability of combining PCA and PLS regression to model stream *E. coli* concentrations and loadings to both predict the *E. coli* level and provide insights into controlling mechanisms.

2. Site description and data availability

Fall Creek watershed (Fig. 1) is a 326 km² in central upstate New York that drains into Cayuga Lake, with 41% agricultural land, 52% forest and the rest developed area (Johnson et al., 2007). The elevation of the watershed ranges from 270 to 600 m (Johnson et al., 2007). Year-round-manure-fertilized pastures or hay fields takes up 15% of the watershed and row crop fields, which receive manure in April, May, October, or November, take up 17% of the watershed (Archibald, 2015).

We used water quality data accrued by the Community Science Institute (CSI, 2014). We chose the three CSI monitoring sites with longest time series of *E. coli* data: Fall Creek at Freeville (FCF), Freese Road Bridge (FRB), and Cayuga Street Bridge (CSB) (Fig. 1). FCF is 19 km from stream mouth and located immediately upstream of the confluence of Fall Creek and Virgil Creek (CSI, 2014). FRB is 6 km from the stream mouth and located upstream of Cornell water treatment plant intake and downstream of eroding stream banks at Monkey Run Road (CSI, 2014). The CSB is 1 km from stream mouth, which is close to the mouth without any lake backwater interference- and it is located downstream Ithaca, i.e., an urban area (CSI, 2014).

Water quality data were obtained from the CSI (2014) database for period 2004–2013. A total of 13 water quality analytes were selected for use in our analyses, including: *E. coli* concentration (colonies/100 mL), total coliform concentration (colonies/100 mL), alkalinity (mg CaCO₃/L), chloride (mg/L), total hardness (mg CaCO₃/L), pH, total Kjeldahl nitrogen (mg-N/L), nitrate-nitrite-nitrogen (as mg-N/L), soluble reactive phosphorus (µg-P/L), total phosphorus (µg-P/L), specific conductance (µS/cm), total suspended solids (mg/L), and turbidity (NTU). The CSI (2014) water quality samples were labeled either stormwater or baseflow. Stormwater originates directly from precipitation or melting snow while baseflow recharges from groundwater (Isensee et al., 1989).

The instantaneous discharges at each of the three water sampling locations was estimated with the drainage-area ratio method, a standard method to estimate streamflow when discharge at a site is unavailable but discharge at a nearby site is available (Emerson et al., 2005). Specifically, we scale the instantaneous discharge at USGS gauge (#04234000) (USGS, 2007, 2014) on Fall Creek with the drainage area corresponding to the other three water quality monitoring sites (Fig. 1).

Fig. 1. Map of the Fall Creek Watershed. The circles with a dot indicate the CSI sampling sites used in our analyses: FCF, FRB, and CSB. The open circle indicates the watershed outlet. The dotted triangle indicates the USGS gauge. And the open triangles indicate the weather stations. The inset map also shows the location of the Fall Creek Watershed in the New York State.

Since about half of the water quality data points did not have water temperature, we used instantaneous air temperature instead. Both instantaneous air temperature and hourly precipitation data were from weather stations closest to the sampling sites (Fig. 1): mostly from Ithaca 13 E (I13E) and from one of the two weather stations associated with the Ithaca Tompkins Regional Airport (ITRA and ITR) when I13E was missing data (NOAA, 2014). The seven day antecedent precipitations were acquired by summing up all the hourly precipitations in the 7×24 h preceding each water quality sampling time.

We eliminated the data on days that discharge or climate data were unavailable, and the final total number of samples was 102 over 10 years. The number of data points used was: $n = 32$ at FRB, $n = 39$ at CSB, and $n = 34$ at FCF. Three outliers of extremely high reported

E. coli concentrations were excluded, because all the associated water quality analytes were not correspondingly extremely high or low.

3. Methodology

A total of 8 categorical variables and 17 continuous variables were used in the analyses (Table 1). We performed a one-way ANOVA to see if there were any significantly different groups of data that should be treated separately in subsequent analyses. We also calculated Pearson correlation coefficients for comparison between *E. coli* (concentration and loading) and the other variables in order to provide basic information about correlations within the dataset.

Table 1
Notations.

Category	Variable	Meaning (unit)
Sites	CSB	1 for data collected at Cayuga Street Bridge (CSB), and 0 for data collected at the other two sites
	FRB	1 for data collected at Freese Road Bridge (FRB), and 0 for data collected at the other two sites
	FCF	Fall Creek Freeville sampling site, default site option
Season	W	1 for data collected during winter (Nov 15th to Feb 28th or 29th), and 0 for data collected during other seasons
	Sp	1 for data collected during spring (March 1st to May 31st), and 0 for data collected during other seasons
	Su	1 for data collected during summer (June 1st to Aug 31st), and 0 for data collected during other seasons
	F	Fall, Sep 1st to Nov 14th, default season option
Hydrology and climate/meteorology	Season	Used in PCA, 1 for winter (W), 2 for spring (Sp), 3 for fall (F), and 4 for summer (Su)
	B	1 for baseflow, and 0 for stormwater
Water quality	Discharge	Instantaneous discharge scaled by drainage area (cfs)
	7dP	Seven day antecedent precipitation (mm)
	airT	Instantaneous air temperature (°C)
	Alkalinity	Alkalinity (mg CaCO ₃ /L)
	Cl	Chloride concentration (mg/L)
	Ecoli	<i>E. coli</i> concentration (colonies/100 mL)
	Ecoliloat	<i>E. coli</i> loading (colonies/s), product of Ecoli and discharge
	Hardness	Total hardness (mg CaCO ₃ /L)
	KNT	Total Kjeldahl nitrogen (as N, mg/L)
	NNN	Nitrate-nitrite-nitrogen (as N, mg/L)
	pH	pH
	SC	Specific conductance (μS/cm)
	SRP	Soluble reactive phosphorus (as P, μg/L)
	Tcoliform	Total coliform concentration (colonies/100 mL)
	TP	Total phosphorus (as P, μg/L)
	TSS	Total suspended solids (mg/L)
	Turbidity	Turbidity (NTU)

Since the water quality analytes, discharge and weather variables have different units, before PCA and PLS regression, each variable was centered to its mean and scaled by its standard deviation, i.e. autoscaled, according to [Esbensen et al. \(2002\)](#).

MATLAB was used to run both the PLS regression and PCA. PCA was used to identify independent driving processes of the patterns of *E. coli* observed in the stream, and, further, to provide better understanding of the variation of *E. coli* level in the stream together with PLS regression.

In order to obtain the most important processes controlling the *E. coli*, we used as few variables as possible in the PCA. First, the four seasons were combined into one variable, Season (1 = winter, 2 = spring, 3 = fall, 4 = summer). Second, the spatial difference was neglected as the three sites were not far from each other and all located at lower reaches of the Fall Creek Watershed ([Fig. 1](#)). Third, hardness and alkalinity were deleted, as they can be represented by pH; and turbidity was deleted as it can be represented by total suspended solids. Last, *E. coli* loading was not used, as it was the product of *E. coli* concentration and discharge.

Each principal component (PC) is a linear combination of all the input variables. We had a total of 14 variables in the PCA. Thus, each PC was a 14-dimensional vector. The PC coefficients for one PC are the coefficients of the linear combination. In other words, each PC is the correlation coefficient between one pair of variables and the PC. However, the *p*-values corresponding to the correlation coefficients could not be obtained. Therefore, we developed our own way to determine the significance of the correlations. We calculated the mean of the absolute

values of the coefficients for each PC. Those variables whose PC coefficients were larger than the mean were regarded as significantly positively correlated with the corresponding PC; those variables whose principal component coefficients were smaller than negative mean were regarded as significantly negatively correlated with the corresponding PC; those variables whose PC coefficients were close to zero, i.e. between -0.05 and 0.05 , were regarded as uncorrelated with the corresponding PC.

PLS regression uses Y matrix (dependent variables) as a starting point to iteratively find the optimal principal components (PCs) to project the X matrix (independent variables) and Y matrix interdependently ([Esbensen et al., 2002](#)). In our case, the *E. coli* concentrations and *E. coli* loadings were the dependent variables, and all the other variables, except for concentration of total coliform, were regarded as independent variables. Note that fall is the default season category and FCF is the default site category. We used segmented cross validation ([Esbensen et al., 2002](#)), i.e. dividing all the data points into segments, using one segment as the test set each time, until each segment having been test set once. A total of 13 segments were used for data collected under stormwater conditions and a total of 6 segments were used for data collected under baseflow conditions.

After trial and error, and comprehensively considering R^2 , adjusted R^2 , root mean square error of prediction (RMSEP) and residual ([Esbensen et al., 2002](#)), we found that the models using the log of *E. coli* concentration and *E. coli* loading, and using the first seven principal components (sorted by the explained variance) provided the best results. In addition, independent variables were deleted from the corresponding model, if its regression coefficient was in the range of -0.002 to 0.002 .

In addition, we performed PLS regression using only the driving factors identified by PCA analysis and compared with the corresponding PLS regression model using all the variables.

4. Results

4.1. Basic information

A one-way ANOVA of instream *E. coli* level indicated that there were significant differences between the stormwater and baseflow conditions: stormwater conditions had significantly higher average *E. coli* concentrations, discharge, and *E. coli* loadings than the baseflow conditions ([Fig. 2a–c](#)). Whereas, the *E. coli* concentrations were not significantly different among the three sites ($n = 31, 38$ and 33 ; $p = 0.71$) ([Fig. 2d](#)). Therefore, the total dataset was divided into two groups: stormwater and baseflow datasets. Note that for one-way ANOVA, the means used are arithmetic means, but for easier comparison with [Vidon et al. \(2008\)](#), we also calculated geometric means of the variables ([Table 2](#)).

One-way ANOVA of all the continuous variables indicated that, except for chloride, nitrate-nitrite-nitrogen, air temperature, and seven day antecedent precipitation, all the other variables were significantly different between the two flow conditions ([Table 2](#)). Alkalinity, total hardness, pH and specific conductance were significantly higher for baseflow conditions than in stormwater conditions. In contrast, total Kjeldahl nitrogen, total suspended solids, soluble reactive phosphorus, total coliform, total phosphorus and turbidity, in addition to the previously mentioned *E. coli* concentrations, discharge and *E. coli* loadings, were significantly higher for stormwater conditions relative to baseflow conditions.

E. coli was a good indicator of total coliform under both baseflow and stormwater conditions ([Tables 3 and 4](#)). We found that *E. coli* concentration and *E. coli* loading were significantly-correlated with the same variables ([Table 3](#)). Most notably, neither of them was significantly correlated with any season or any sampling site, but both of them were significantly negatively correlated with baseflow ([Table 3](#)). However, separating the data set into two subgroups—baseflow and stormwater

Fig. 2. Comparison between baseflow and stormwater conditions in terms of (a) *E. coli* concentration, (b) discharge, and (c) *E. coli* loading. Comparison among the three sites in terms of *E. coli* concentration (d). The bars are arithmetic means and the error bars are standard errors.

provided more useful insights into the relationships between these variables (Table 4). *E. coli* level (both concentration and loading) became correlated to summer and sampling sites, but not correlated to variables related to pH and seven day antecedent precipitation (Table 4).

4.2. Principal component analysis (PCA)

The first three PCs explained 65% of the total variance (Table 5). The first principal component (PC1) explained nearly two fifth of the total variance (Table 5). *E. coli* concentration, total Kjeldahl nitrogen (N), total suspended solids (TSS), soluble reactive phosphorus (P), total P and discharge were significantly positively correlated with PC1; while baseflow and specific conductance were significantly negatively correlated with it (Table 5).

The second principal component (PC2) explained about one fifth of the total variance (Table 5). Only nitrate-nitrite-N was significantly positively correlated with PC2; while season, *E. coli* concentration, total Kjeldahl N, specific conductance and air temperature were significantly negatively correlated with it; and baseflow was nearly uncorrelated (Table 5).

The third principal component (PC3) explained about one tenth of the total variance (Table 5). *E. coli* concentration, nitrate-nitrite-N, and seven day antecedent precipitation were significantly positively correlated with PC3; while chloride, total suspended solids, total P and discharge were significantly negatively correlated with it; and specific conductance and total Kjeldahl N were uncorrelated with it (Table 5).

The rest principal components either explained less than 5% of the total variance or were not significantly correlated with *E. coli*.

4.3. Partial least squares (PLS) regression

Under stormwater conditions, the predicted vs. measured *E. coli* concentrations and loadings were close to 1:1 line, well-distributed over the whole range, and the residuals had no clear, systematic bias (Fig. 3e–h). In addition, the predicted *E. coli* concentrations / loadings had relatively large R^2 and adjusted R^2 values, and relatively small RMSEP (Fig. 3e, g). In contrast, under baseflow conditions, it was hard to precisely predict either *E. coli* concentrations or loadings, as both of the models had relatively small R^2 and adjusted R^2 , relatively large RMSEP, and the residuals showed clear systematic patterns (Fig. 3a–d).

The regression coefficients of the four PLS regression models were tabulated in Table 6. Each dependent variable was predicted by linear regression consisting of all the independent variables and their corresponding regression coefficients, plus an intercept that was 0 for each of the four models in this study.

Additionally, the significance of fitting and regression coefficients of the PLS regression model using only driving factors identified by PCA are shown in Fig. S1 and Table S1, respectively.

5. Discussion

5.1. Data distribution compared to previous research

Our data were consistent with Vidon et al. (2008) in that baseflow had significantly higher pH and specific conductance compared to stormwater (Table 2); and had significantly lower turbidity, *E. coli*

Table 2

Count or average value for each variable in stormwater and baseflow conditions, and ANOVA results.

	Count for the categorical variables		<i>p</i>	Geometric mean	
	Baseflow	Stormwater		Baseflow	Stormwater
	Mean for baseflow	Mean for stormwater		Baseflow	Stormwater
W	13	15			
Sp	14	6			
S	20	9			
F	21	4			
FRB	21	10			
CSB	24	14			
FCF	23	10			
Alkalinity (mg CaCO ₃ /L)	133	107	9 × 10⁻⁶	132	101
Cl (mg/L)	25.8	23.6	0.1243	25.0	22.6
Ecoli (colonies/100 mL)	223	1524	2 × 10⁻⁵	98.6	548
Hardness (mg CaCO ₃ /L)	159	128	2 × 10⁻⁶	157	122
pH	7.95	7.75	0.0005	7.94	7.74
KNT (as N, mg/L)	0.405	0.592	8 × 10⁻⁶	0.386	0.542
NNN (as N, mg/L)	0.805	0.931	0.1637	0.618	0.882
TSS (mg/L)	8.16	75.4	5 × 10⁻⁶	4.35	25.0
SRP (as P, µg/L)	14.0	22.0	4 × 10⁻⁵	12.2	19.2
SC (µS/cm)	357	316	0.0005	355	308
Tcoliform (colonies/100 mL)	9696	36,964	4 × 10⁻⁶	5802	21,065
TP (as P, µg/L)	32.1	103	1 × 10⁻⁶	28.5	65.4
Turbidity (NTU)	7.34	67.8	2 × 10⁻⁶	5.38	25.0
Discharge (cfs)	117	451	6 × 10⁻⁷	77.3	243
airT (°C)	9.38	6.99	0.1551	n.a.	n.a.
7dP (mm)	22.3	26.8	0.2367	n.a.	21.5
Ecoliload (colonies/s)	6.02 × 10 ⁶	1.71 × 10 ⁸	9 × 10⁻⁵	2.16 × 10 ⁶	3.77 × 10 ⁸

Bold: significantly different.

n.a.: values contain negative number or 0, geometric mean could not be calculated.

concentration, discharge and *E. coli* loading compared to stormwater (Fig. 2 and Table 2).

Under baseflow conditions, we had similar geometric means of pH and turbidity as those of Vidon et al. (2008). However, compared to Vidon et al. (2008), we saw lower specific conductance and temperature, and had much lower *E. coli* concentration (10 to 20 fold) and much higher discharge (10 to 20 fold), resulting in similar *E. coli* loading (Table 2).

Under stormwater conditions, we had similar geometric means of pH and turbidity as those of Vidon et al. (2008). However, compared to (Vidon et al., 2008), we had lower specific conductance and temperature, and had much lower *E. coli* concentration (5 to 8 folds) and much higher discharge (10 to 20 folds), resulting in higher *E. coli* loading (Table 2).

5.2. Interpretation of one-way ANOVA results and the Pearson correlation coefficients

Those variables that were significantly higher under baseflow conditions than under stormwater conditions (Table 2) were due to groundwater base-cation inputs and dilution of base-cations by stormwater. Whereas those variables that were significantly higher under

Table 3

Pearson correlation coefficients between *E. coli* concentration and loading and other variables for all the data points (*n* = 102).

	<i>E. coli</i> concentration (MPN/100 mL)	<i>E. coli</i> loading (MPN/s)
B	−0.41***	−0.38***
W	0.08	0.10
Sp	−0.14	0.003
Su	0.08	−0.14
F	0.04	0.03
FRB	−0.05	−0.04
CSB	0.08	0.16
FCF	−0.04	−0.12
Alkalinity	−0.30**	−0.44***
Cl	−0.22*	−0.28**
Hardness	−0.30**	−0.46***
pH	−0.27**	−0.22*
KNT	0.74***	0.67***
NNN	−0.13	−0.09
TSS	0.52***	0.59***
SRP	0.58***	0.59***
SC	−0.26**	−0.46***
Tcoliform	0.82***	0.71***
TP	0.61***	0.64***
Turbidity	0.61***	0.69***
Discharge	0.13	n/a
airT	0.12	−0.02
7dP	0.45***	0.50***

* Significant at 0.05.

** Significant at 0.01.

*** Significant at 0.001.

stormwater conditions than under baseflow conditions (Table 2), together with *E. coli* concentration and loading (Fig. 2), suggested that the transport paths of organic N, sediments (represented by total suspended solids and turbidity), P, and coliform share common features with the transport paths of *E. coli* (i.e., surface/near-surface pathways). The findings that both *E. coli* concentrations and *E. coli* loadings were significantly lower under baseflow conditions compared to stormwater conditions (Fig. 2) and that the *E. coli* concentration and loading only

Table 4

Pearson correlation coefficients between *E. coli* concentration/loading and other variables under baseflow conditions (*n* = 68), and under stormwater conditions (*n* = 34).

	<i>E. coli</i> concentration (MPN/100 mL)	<i>E. coli</i> loading (MPN/s)	<i>E. coli</i> concentration (MPN/100 mL)	<i>E. coli</i> loading (MPN/s)
	Baseflow		Stormwater	
W	−0.11	0.11	−0.02	0.006
Sp	−0.21	−0.02	−0.20	0.03
Su	0.43***	0.16	0.08	−0.25
F	−0.15	−0.23	0.16	0.29
FRB	−0.30*	−0.19	−0.008	−0.06
CSB	0.25*	0.35**	0.05	0.23
FCF	0.04	−0.16	−0.04	−0.18
Alkalinity	0.21	−0.10	−0.23	−0.43*
Cl	0.09	0.19	−0.32	−0.40*
Hardness	0.19	−0.03	−0.22	−0.45**
pH	−0.18	−0.20	−0.19	−0.14
KNT	0.72***	0.33**	0.77***	0.76***
NNN	−0.06	0.09	−0.52**	−0.41*
TSS	0.68***	0.40***	0.41*	0.52**
SRP	0.27*	0.22	0.66***	0.72***
SC	0.27*	0.07	−0.21	−0.47**
Tcoliform	0.46***	0.60***	0.81***	0.67***
TP	0.64***	0.39**	0.52**	0.58***
Turbidity	0.69***	0.46***	0.51**	0.64***
Discharge	−0.16	n.a.	−0.07	n.a.
airT	0.26*	−0.07	0.34*	0.09
7dP	−0.02	0.11	0.74***	0.80***

n.a.: not independent.

* Significant at 0.05.

** Significant at 0.01.

*** Significant at 0.001.

Table 5

Principal component coefficients, i.e. the correlation coefficients between the first three PCs and the variables. The mean of the absolute value of the coefficients for the three PCs were 0.25, 0.24 and 0.21, respectively.

	PC1	PC2	PC3
Explained (%)	35.1	20.6	9.1
B	−0.29	−0.01	0.11
Season	−0.17	−0.44	0.14
Cl	−0.14	−0.18	−0.57
EColi	0.31	−0.26	0.21
pH	−0.21	−0.13	−0.09
KNT	0.36	−0.26	0.00
NNN	0.11	0.39	0.29
TSS	0.35	−0.14	−0.32
SRP	0.29	−0.21	0.21
SC	−0.30	−0.29	0.04
TP	0.37	−0.18	−0.25
Discharge	0.27	0.21	−0.23
airT	−0.13	−0.45	0.07
7dP	0.18	−0.13	0.49

Bold: significantly positively correlated;

Bold and italic: significantly negatively correlated; and.

Underline: nearly not correlated.

showed spatial differences under baseflow conditions (Tables 3 and 4) were consistent with the findings of Vidon et al. (2008). The only significant correlation related to season that we found was under baseflow conditions: *E. coli* concentration was significantly higher in summer (Tables 3 and 4).

In terms of correlations between *E. coli* (either *E. coli* concentration or *E. coli* loading, or both) and the continuous variables, our findings were consistent with Vidon et al. (2008), i.e., pH (pH, alkalinity, and hardness) were significantly correlated with *E. coli* level when combining all the data points (Table 3), while not significantly correlated when the data were separated by flow conditions (Table 4). And we also found that total Kjeldahl N and soluble reactive P were significantly correlated with *E. coli* (Tables 3 and 4). Contrary to Vidon et al. (2008), who found that turbidity was not significantly correlated with *E. coli* level during high flow conditions, we found that sediment (total suspended solid, total P, and turbidity) was always significantly correlated with *E. coli* regardless of flow (Tables 3 and 4).

5.3. Interpretation of the meaning of the PCs from PCA results

When interpreting the PCA results, we followed the procedure: 1) interpreting the meaning of each PC from its coefficients, and 2) use the meaning of each PC to interpret its influence on instream *E. coli* level and other factors.

5.3.1. PC1

The negative correlation between PC1 and the baseflow variable (Table 5) meant that PC1 was positively correlated with stormwater. And under stormwater conditions, discharge should be higher, the variables related to sediment erosion and resuspension (total suspended solids and total phosphorus) should be higher compared to baseflow conditions. Whereas, under stormwater conditions, instream ions should be diluted, i.e. lower specific conductance. When the total P increased, the soluble reactive P increased accordingly, due to solubility equilibrium. And the increase of total Kjeldahl N, which consisted of ammonia and organic N, might due to the erosion of manure on the land surface or in the top soil. So, it is probable that PC1 appears to be the stormwater component.

As *E. coli* was found to be eroded from the top soil (Wang, 2015; Wang et al., 2017; Wang et al., 2018) and resuspended from the bottom of the stream together with sediments (Ferguson et al., 2003; Jamieson et al., 2005), it was expected that *E. coli* concentrations were significantly positively correlated with PC1. As has been documented previously, *E. coli* concentrations were highly correlated with sediment

concentrations and related variables, including total P (Gentry et al., 2006; Vidon et al., 2008). According to previous findings (Wang, 2015; Wang et al., 2017; Wang et al., 2018), one reason was that transport of *E. coli* and sediment can be explained by the same mechanisms, especially when stormwater dominates. The high correlation between the increase of *E. coli* level and the occurrence of storm and the increase of sediments (Table 5) was generally explained by surface erosion and the co-transport of *E. coli* both alone and bound to eroding sediments (Muirhead et al., 2006a; Oliver et al., 2005). Surface erosion is a common process in agricultural lands, especially where no cover crops or conservation tillage are used. Therefore, it was expected to contribute to the largest PCA factor. Collins and Rutherford (2004) and Vidon et al. (2008) also found that at high flow, *E. coli* was transported by overland flow during precipitation.

5.3.2. PC2

When the air temperature was low, the microbial activity was low, leading to decreased denitrification and increased nitrate-nitrite-N (Wilhelm et al., 1994). So, we speculate that PC2 was the temperature/season component.

Tate (1995) reported that it was not freezing but thawing that killed the microbes. Therefore, the low microbial activity not only resulted from the low enzyme activity at low temperature, but could also result from freezing, or the dying off of microbes due to thawing. At low temperature, the growth rate of *E. coli* could not off-set the frozen and dying off rate. And the significantly negative correlation between season and PC2 (Table 5) further confirmed this: winter and spring were the times when freezing and thawing happens. The increase of discharge might due to snowmelt in spring, and the increased discharge diluted the stream water and led to decreased total Kjeldahl N and specific conductance.

Interestingly, the second largest driver of *E. coli* concentration appeared to be a seasonal influence on temperature and associated microbial activity. We are not the first who find the importance of season and temperature on *E. coli*. Vidon et al. (2008) found that winter/spring and summer/fall constituted two identifiably different conditions for *E. coli*. Brooks et al. (2013) found that variables related to season, like Julian date and month, affected *E. coli*, and that air temperature was important. Whitman et al. (2008) found that season was important, and, they also found that sunlight and snowmelt affected *E. coli* populations. Chen and Chang (2014) found that *E. coli* level was higher during summer when the temperature was suitable for *E. coli* survival and growth. However, we are the first who find that nitrate-nitrite-N, a previously ignored factor, is also a good *E. coli* indicator, as it indicates microbial activity, which generally follows seasonal and temperature trends.

5.3.3. PC3

The concurrence of large seven day antecedent precipitation and low discharge (Table 5) led us to propose that PC3 was the component of shallow subsurface flow, as antecedent precipitation controlled soil moisture and shallow subsurface flow (Takagi and Lin, 2011). The significantly negative correlation of total suspended solids and total phosphorus with PC3 (Table 5) supports our inference, as the shallow subsurface flow imported water with low sediment concentration into the stream.

The significantly positive correlation of *E. coli* concentration, soluble reactive P and nitrate-nitrite-N with PC3 (Table 5) was explainable. The increase of shallow subsurface flow provided more liquid water in the soil, so that the saturation of the soil was increased. Therefore, those *E. coli* that previously had been strained by soil or adsorbed to soil had more chances to be remobilized (Wan and Wilson, 1994; Zhuang et al., 2007); and P that previously had been adsorbed to soil or had precipitated also had chances to be remobilized (Domagalski and Johnson, 2011; Harman et al., 1996); those newly produced nitrate-nitrite-N from other forms of N in the soil by mineralization and oxidation had more chances to get dissolved (Harman et al., 1996; Wilhelm et al.,

Fig. 3. The results of PLS regression using all the variables. The left column showed predicted vs. measured $\log(E. coli)$ concentration under baseflow conditions (a), $\log(E. coli)$ loading under baseflow conditions (c), $\log(E. coli)$ concentration under stormwater conditions (e), and $\log(E. coli)$ loading under stormwater conditions (g). The right column showed residual vs. measured $\log(E. coli)$ concentration under baseflow conditions (b), $\log(E. coli)$ loading under baseflow conditions (d), $\log(E. coli)$ concentration under stormwater conditions (f), and $\log(E. coli)$ loading under stormwater conditions (h).

Table 6

Regression coefficients corresponding to the variables for the four PLS models using all the variables. The intercepts for the four models were all zero.

	Baseflow		Stormwater	
	log (<i>E. coli</i> concentration)	log (<i>E. coli</i> loading)	log (<i>E. coli</i> concentration)	log (<i>E. coli</i> loading)
W	−0.0072	0.031	0.25	0.31
Sp	−0.13	−0.11	0.11	0.14
Su	0.037	0.082	0.11	0.067
FRB	−0.39	−0.19	−0.24	−0.076
CSB	−0.012	0.22	−0.35	−0.16
Alkalinity	0.041	−0.11	−0.13	−0.22
Cl	0.088	Del.	0.19	0.19
Hardness	0.035	0.037	−0.012	−0.092
pH	−0.059	−0.033	0.066	0.031
KNT	0.56	0.51	0.76	0.57
NNN	−0.10	0.14	−0.17	−0.033
TSS	−0.24	−0.25	−0.059	−0.021
SRP	−0.0068	Del.	−0.090	−0.10
SC	0.21	0.25	0.17	0.10
TP	−0.0073	0.036	0.068	Del.
Turbidity	0.092	0.049	−0.20	−0.14
Discharge	0.24	0.51	0.14	0.39
airT	−0.0085	−0.034	0.18	0.13
7dP	−0.17	−0.15	0.33	0.37

Del.: corresponding variable deleted from the model due to a less than 0.002 absolute value of regression coefficient.

1994). In addition, when the saturation of the soil increased, *E. coli*, dissolved soluble reactive P and dissolved nitrate-nitrite-N had more pathways to streams. In other words, when shallow subsurface flow increased, the dilution effect of increased soil water content was exceeded by the increased mobility through vadose zone for these three kinds of contaminants.

Gentry et al. (2006) and Vidon et al. (2008) reported that the seven day antecedent precipitation was one of the best indicators of *E. coli* under baseflow conditions. However, to our knowledge, soluble reactive P is here correlated with *E. coli* for the first time. *E. coli* is prone to straining and sorption in soil (Oliver et al., 2005) and P is prone to sorption in soil (McGeachan and Lewis, 2002). So they shared similar transport pathways from soil through shallow subsurface flow into stream.

The significantly negative correlation of chloride with PC3 (Table 5) might due to continuous dilution by the antecedent rainfall, as rainwater has low chloride concentration.

5.4. Interpretation of the PLS regression results

For the PLS regression models using all the variables, both *E. coli* concentration and *E. coli* loading were more precisely predicted under stormwater conditions compared to baseflow conditions (Fig. 3), i.e., we were good at predicting the most hazardous events. In effect, a large relative error in predicting small values was perhaps not relevant from a human health perspective.

Comparing the PLS regression model using only the driving factors identified by PCA (Fig. S1) with the above model (Fig. 3(e) and (f)), we see that 1) using all the variables achieved better RMSEP, however, 2) using only PCA-identified driving factors achieved better adjusted R^2 . Although it is, then, hard to identify the best model, it appears we were justified in using PCA to identify the primary driving factors and that PLS regression model using only these factors can provide good predictions.

5.5. Suggestions on management practices

Since stormwater was the largest driver for the *E. coli*, management practices that can reduce surface erosion may help mitigate *E. coli* pollution. First, avoiding applying manure on hydrological sensitive areas may reduce the probability of *E. coli* getting into surface runoff

(Walter et al., 2000). Second, covering the soil surface with vegetation or crop residue could reduce the erosion of *E. coli* on or near soil surface (Wang et al., 2018). Third, applying vegetation buffer strips (Qi and Altinakar, 2011; Roodsari et al., 2005) on fields close to streams can filter *E. coli* in overland flow and largely prevent *E. coli* reaching streams by overland flow, the fastest pathway. Forth, since roadside ditches provide fast path for *E. coli* to reach streams (Falbo et al., 2013), applying vegetation buffer strips on fields close to roadside ditches may also help. During summer, when the microbial activity is high, the plant growth is also high, which means that the vegetation buffer should have best effect at worst situation.

6. Conclusions

In conclusion, we identified two new variables (nitrate-nitrite-N and soluble reactive P) that can indicate the level of *E. coli* in a stream. We identified three important processes driving the *E. coli* fate and transport in a watershed (stormwater, temperature/season and shallow subsurface flow). We also demonstrated that the strategy of applying both PCA and PLS regression can provided a useful, complementary and relatively simple way for evaluating datasets with lots of collinearity, both temporal and spatial. The PCA suggested possible underlying biophysical mechanisms driving the variation of the stream water quality. PCA could not predict the actual concentrations / loadings; whereas, the PLS regression was useful, once the significantly different groups of data (stormwater and baseflow in this study) were separated, to create predictive equations of the actual concentrations / loadings.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.scitotenv.2018.04.036>.

Acknowledgement

The authors want to thank Rebecca Marjerison for assistance with data acquisition and map production, and thank Françoise Vermeylen and Jacob Bien for assistance with statistical analysis. We also thank Community Science Institute (CSI) for providing the online water quality dataset. Chaozi Wang thanks China Scholarship Council for providing her a four-year scholarship.

References

- Aguilera, P.A., Garrido Frenich, A., Castro, H., Martinez Vidal, J.L., 2000. PLS and PCR methods in the assessment of coastal water quality. *Environ. Monit. Assess.* 62, 193–204.
- Ahmed, W., Sawant, S., Huygens, F., Goonetilleke, A., Gardner, T., 2009. Prevalence and occurrence of zoonotic bacterial pathogens in surface waters determined by quantitative PCR. *Water Res.* 43, 4918–4928.
- Archibald, J., 2015. Modeling Dissolved Phosphorus Transport From Agricultural Watersheds In Central New York. Cornell University, Ithaca, NY, 14853.
- Baffaut, C., Sadeghi, A., 2010. Bacteria modeling with SWAT for assessment and remediation studies: a review. *Trans. ASABE* 53, 1585–1594.
- Basant, N., Gupta, S., Malik, A., Singh, K.P., 2010. Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water — a case study. *Chemom. Intell. Lab. Syst.* 104, 172–180.
- Bech, T.B., Rosenbom, A.E., Kjaer, J., Amin, M.G.M., Olsen, P., Jacobsen, C.S., 2014. Factors influencing the survival and leaching of tetracycline-resistant bacteria and *Escherichia coli* through structured agricultural fields. *Agric. Ecosyst. Environ.* 195, 10–17.
- Benham, B.L., Baffaut, C., Zeckoski, R.W., Mankin, K.R., Pachepsky, Y.A., Sadeghi, A.M., Brannan, K.M., Soupir, M.L., Habersack, M.J., 2006. Modeling bacteria fate and transport in watersheds to support TMDLs. *Trans. ASABE* 49, 987–1002.
- Brooks, W.R., Fienen, M.N., Corsi, S.R., 2013. Partial least squares for efficient models of fecal indicator bacteria on Great Lakes beaches. *J. Environ. Manag.* 114, 470–475.
- Carroll, S.P., Dawes, L., Hargreaves, M., Goonetilleke, A., 2009. Faecal pollution source identification in an urbanising catchment using antibiotic resistance profiling, discriminant analysis and partial least squares regression. *Water Res.* 43, 1237–1246.
- Chen, H., Chang, H., 2014. Response of discharge, TSS, and *E. coli* to rainfall events in urban, suburban, and rural watersheds. *Environ. Sci.: Processes Impacts* 16, 2313–2324.
- Cho, K.H., Pachepsky, Y.A., Kim, J.H., Kim, J.-W., Park, M.-H., 2012. The modified SWAT model for predicting fecal coliforms in the Wachusett Reservoir Watershed, USA. *Water Res.* 46, 4750–4760.
- Collins, R., Rutherford, K., 2004. Modelling bacterial water quality in streams draining pastoral land. *Water Res.* 38, 700–712.

- Community Science Institute, 2014. CSI database. <http://communityscience.org> (last accessed July 1, 2014).
- Domagalski, J.L., Johnson, H.M., 2011. Subsurface transport of orthophosphate in five agricultural watersheds, USA. *J. Hydrol.* 409, 157–171.
- Dwivedi, D., Mohanty, B.P., Lesikar, B.J., 2013. Estimating *Escherichia coli* loads in streams based on various physical, chemical, and biological factors. *Water Resour. Res.* 49, 2896–2906.
- Emerson, D.G., Vecchia, A.V., Dahl, A.L., 2005. Evaluation of drainage-area ratio method used to estimate streamflow for the Red River of the North Basin, North Dakota and Minnesota.
- Esbensen, K.H., Guyot, D., Westad, F., Houmoller, L.P., 2002. *Multivariate Data Analysis-In Practice: An Introduction to Multivariate Data Analysis and Experimental Design*. 5th ed. CAMO Process AS, Oslo, Norway.
- Falbo, K., Schneider, R.L., Buckley, D.H., Walter, M.T., Bergholz, P.W., Buchanan, B.P., 2013. Roadside ditches as conduits of fecal indicator organisms and sediment: implications for water quality management. *J. Environ. Manag.* 128, 1050–1059.
- Ferguson, C., Husman, A.M.d.R., Altavilla, N., Deere, D., Ashbolt, N., 2003. Fate and transport of surface water pathogens in watersheds. *Environ. Sci. Technol.* 33, 299–361.
- Frey, S.K., Topp, E., Edge, T., Fall, C., Gannon, V., Jokinen, C., Marti, R., Neumann, N., Ruecker, N., Wilkes, G., Lapen, D.R., 2013. Using SWAT, Bacteroidales microbial source tracking markers, and fecal indicator bacteria to predict waterborne pathogen occurrence in an agricultural watershed. *Water Res.* 47, 6326–6337.
- Gentry, R.W., McCarthy, J., Layton, A., McKay, L.D., Williams, D., Koirala, S.R., Saylor, G.S., 2006. *Escherichia coli* loading at or near base flow in a mixed-use watershed. *J. Environ. Qual.* 35, 2244–2249.
- Gray, E., 1975. Survival of *Escherichia coli* in stream water in relation to carbon dioxide and plant photosynthesis. *J. Appl. Bacteriol.* 39, 47–54.
- Harman, J., Robertson, W.D., Cherry, J.A., Zanini, L., 1996. Impacts on a sand aquifer from an old septic system: nitrate and phosphate. *Ground Water* 34, 1105–1114.
- Hofstra, N., 2011. Quantifying the impact of climate change on enteric waterborne pathogen concentrations in surface water. *Curr. Opin. Environ. Sustain.* 3, 471–479.
- Isensee, A.R., Johnson, L., Thornhill, J., Nicholson, T.J., Meyer, G., Vecchioli, J., Laney, R., 1989. *Subsurface-water Flow and Solute Transport: Federal Glossary of Selected Terms*. Reston, VA, p. 46.
- Jamieson, R.C., Joy, D.M., Lee, H., Kostaschuk, R., Gordon, R.J., 2005. Resuspension of sediment-associated *Escherichia coli* in a natural stream. *J. Environ. Qual.* 34, 581–589.
- Johnson, M.S., Woodbury, P.B., Pell, A.N., Lehmann, J., 2007. Land-use change and stream water fluxes: decadal dynamics in watershed nitrate exports. *Ecosyst.* 10, 1182–1196.
- Marcheggiani, S., D'Ugo, E., Puccinelli, C., Giuseppetti, R., D'Angelo, A.M., Gualerzi, C.O., Spurio, R., Medlin, L.K., Guillebault, D., Weigel, W., Helmi, K., Mancini, L., 2015. Detection of emerging and re-emerging pathogens in surface waters close to an urban area. *Int. J. Environ. Res. Public Health* 12, 5505–5527.
- McGeachan, M.B., Lewis, D.R., 2002. Transport of particulate and colloid-sorbed contaminants through soil, part 1: general principles. *Biosyst. Eng.* 83, 255–273.
- Meays, C.L., Broersma, K., Nordin, R., Mazumder, A., Samadpour, M., 2006. Diurnal variability in concentrations and sources of *Escherichia coli* in three streams. *Can. J. Microbiol.* 52, 1130–1135.
- Muirhead, R.W., Collins, R.P., Bremer, P.J., 2006. The association of *E. coli* and soil particles in overland flow. *Water Sci. Technol.* 54, 153–159.
- Nevers, M.B., Whitman, R.L., 2005. Nowcast modeling of *Escherichia coli* concentrations at multiple urban beaches of southern Lake Michigan. *Water Res.* 39, 5250–5260.
- NOAA, 2014. NOAA National Centers for Environmental Information: Climate Data Online. <http://gis.ncdc.noaa.gov/map/viewer/#app=cdo&cfg=lcd&theme=lcd> (last accessed July 22, 2014).
- Oliver, D.M., Clegg, C.D., Haygarth, P.M., Heathwaite, A.L., 2005. Assessing the potential for pathogen transfer from grassland soils to surface waters. *Adv. Agron.* 85, 125–180.
- Ortiz-Estarellas, O., Martin-Biosca, Y., Medina-Hernandez, M.J., Sagrado, S., Bonet-Domingo, E., 2001a. Multivariate data analysis of quality parameters in drinking water. *Analyst* 126, 91–96.
- Ortiz-Estarellas, O., Martín-Biosca, Y., Medina-Hernández, M.J., Sagrado, S., Bonet-Domingo, E., 2001b. On the internal multivariate quality control of analytical laboratories. A case study: the quality of drinking water. *Chemom. Intell. Lab. Syst.* 56, 93–103.
- Qi, H., Altinakar, M.S., 2011. Vegetation buffer strips design using an optimization approach for non-point source pollutant control of an agricultural watershed. *Water Resour. Manag.* 25, 565–578.
- Roodsari, R.M., Shelton, D.R., Shirmohammadi, A., Pachepsky, Y.A., Sadeghi, A.M., Starr, J.L., 2005. Fecal coliform transport as affected by surface condition. *Trans. ASAE* 48, 1055–1061.
- Runkel, R.L., Crawford, C.G., Cohn, T.A., 2004. Load Estimator (LOADEST): A Fortran Program for Estimating Constituent Loads in Streams and Rivers. US Department of the Interior, US Geological Survey.
- Schets, F.A., van Wijnen, J.H., Schijven, J.F., Schoon, A., Husmant, A., 2008. Monitoring of waterborne pathogens in surface waters in Amsterdam, The Netherlands, and the potential health risk associated with exposure to *Cryptosporidium* and *Giardia* in these waters. *Appl. Environ. Microbiol.* 74, 2069–2078.
- Schriewer, A., Miller, W.A., Byrne, B.A., Miller, M.A., Oates, S., Conrad, P.A., Hardin, D., Yang, H.H., Chouicha, N., Melli, A., Jessup, D., Dominik, C., Wuertz, S., 2010. Presence of bacteroidales as a predictor of pathogens in surface waters of the Central California coast. *Appl. Environ. Microbiol.* 76, 5802–5814.
- Singh, K.P., Malik, A., Basant, N., Saxena, P., 2007. Multi-way partial least squares modeling of water quality data. *Anal. Chim. Acta* 584, 385–396.
- Takagi, K., Lin, H.S., 2011. Temporal dynamics of soil moisture spatial variability in the Shale Hills critical zone observatory. *Vadose Zone J.* 10, 832–842.
- Tate, R.L., 1995. *Soil Microbiology*. 2nd ed. John Wiley & Sons, Inc., New York, USA.
- Vidon, P., Tedesco, L.P., Wilson, J., Campbell, M.A., Casey, L.R., Gray, M., 2008. Direct and indirect hydrological controls on *E. coli* concentration and loading in Midwestern streams. *J. Environ. Qual.* 37, 1761–1768.
- Walker, F.R., Stedinger, J.R., 1999. Fate and transport model of *Cryptosporidium*. *J. Environ. Eng.* 125, 325–333.
- Walter, M.T., Walter, M.F., Brooks, E.S., Steenhuis, T.S., Boll, J., Weiler, K., 2000. Hydrologically sensitive areas: variable source area hydrology implications for water quality risk assessment. *J. Soil Water Conserv.* 55, 277–284.
- Wan, J., Wilson, J.L., 1994. Colloid transport in unsaturated porous media. *Water Resour. Res.* 30, 857–864.
- Wang, C., 2015. *Escherichia coli* Transport Modeling at Soil Column Scale and Watershed Scale. Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY.
- USGS, 2007. USGS Instantaneous Data Archive - IDA. http://ida.water.usgs.gov/ida/available_records.cfm (last accessed July 22, 2014).
- USGS, 2014. USGS Surface-Water Historical Instantaneous Data for the Nation. <http://nwis.waterdata.usgs.gov/nwis/uv/> (last accessed July 20, 2014).
- Wang, C., Parlange, J.-Y., Rasmussen, E.W., Wang, X., Chen, M., Dahlke, H.E., Walter, M.T., 2017. Modeling the release of *Escherichia coli* from soil into overland flow under raindrop impact. *Adv. Water Resour.* 106, 144–153.
- Wang, C., Parlange, J.-Y., Schneider, R.L., Rasmussen, E.W., Wang, X., Chen, M., Dahlke, H.E., Truhlar, A.M., Walter, M.T., 2018. Release of *Escherichia coli* under raindrop impact: the role of clay. *Adv. Water Resour.* 111, 1–5.
- Whitman, R.L., Przybyla-Kelly, K., Shively, D.A., Nevers, M.B., Byappanahalli, M.N., 2008. Sunlight, season, snowmelt, storm, and source affect *E. coli* populations in an artificially ponded stream. *Sci. Total Environ.* 390, 448–455.
- Wilhelm, S.R., Schiff, S.L., Cherry, J.A., 1994. Biogeochemical evolution of domestic wastewater in septic systems.1. Conceptual-model. *Ground Water* 32, 905–916.
- Wilkes, G., Ruecker, N.J., Neumann, N.F., Gannon, V.P.J., Jokinen, C., Sunohara, M., Topp, E., Pintar, K.D.M., Edge, T.A., Lapen, D.R., 2013. Spatiotemporal analysis of *Cryptosporidium* species/genotypes and relationships with other zoonotic pathogens in surface water from mixed-use watersheds. *Appl. Environ. Microbiol.* 79, 434–448.
- Zhuang, J., McCarthy, J.F., Tyner, J.S., Perfect, E., Flury, M., 2007. In situ colloid mobilization in Hanford sediments under unsaturated transient flow conditions: effect of irrigation pattern. *Environ. Sci. Technol.* 41, 3199–3204.