Published in final edited form as:
J Am Water Resour Assoc. 2017 August ; 53(4): 944–960. doi:10.1111/1752-1688.12543.

IMPROVING PREDICTIVE MODELS OF IN-STREAM PHOSPHORUS CONCENTRATION BASED ON NATIONALLY-AVAILABLE SPATIAL DATA COVERAGES

Murray W. Scown, **Michael G. McManus**, **John H. Carson Jr.**, and **Christopher T. Nietch**
Formerly, ORISE Postdoctoral Research Participant, c/o Office of Research and Development, U.S. Environmental Protection Agency, currently Postdoctoral Research Fellow (Scown), Lund University Centre for Sustainability Studies, Lund, Sweden 22362; Ecologist (McManus), National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, Ohio 45268; formerly, Senior Statistician, CB&I Federal Services, currently Director (Carson), P&J Carson Consulting, LLC, Findlay, Ohio 45840; Ecologist (Nietch), National Risk Management Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, Ohio 45268

Abstract

Spatial data are playing an increasingly important role in watershed science and management. Large investments have been made by government agencies to provide nationally-available spatial databases; however, their relevance and suitability for local watershed applications is largely unscrutinized. We investigated how goodness of fit and predictive accuracy of total phosphorus (TP) concentration models developed from nationally-available spatial data could be improved by including local watershed-specific data in the East Fork of the Little Miami River, Ohio, a 1290 km^2 watershed. We also determined whether a spatial stream network (SSN) modeling approach improved on multiple linear regression (nonspatial) models. Goodness of fit and predictive accuracy were highest for the SSN model that included local covariates, and lowest for the nonspatial model developed from national data. Septic systems and point source TP loads were significant covariates in the local models. These local data not only improved the models but enabled a more explicit interpretation of the processes affecting TP concentrations than more generic national covariates. The results suggest that SSN modeling greatly improves prediction and should be applied when using national covariates. Including local covariates further increases the accuracy of TP predictions throughout the studied watershed; such variables should be included in future national databases, particularly the locations of septic systems.

Keywords

Spatial data; stream networks; statistical modeling; phosphorus; autocorrelation

---

Supporting Information
Additional supporting information may be found online under the Supporting Information tab for this article: S1. Final .ssn file objects (.zip); S2. R code and outputs for entire analysis (.zip); and S3. High resolution Figure 3 (.png).

INTRODUCTION

Spatial data and analyses are playing an increasingly important role in natural resource management, particularly in a watershed context. Spatial analysis of watersheds is continually being developed and implemented to identify impacts on water quality or ecosystem health (Strayer *et al.*, 2003), prioritize conservation and restoration (Flotemersch *et al.*, 2015), and predict the potential impacts of climate change and disturbance on watersheds (Isaak *et al.*, 2010). Such studies are often supported by data in geographic information systems (GIS) produced by government natural resource or environmental agencies at the national or international level. Examples of such databases include the National Hydrography Dataset Version 2 (NHDPlus v2), a geospatial framework of streamlines, waterbodies, catchments, and associated attributes throughout the United States (McKay *et al.*, 2012); the equivalent Australian Hydrologic Geospatial Fabric, or Geofabric (Bureau of Meteorology, 2015); and the National Land Cover Database (NLCD), which provides descriptive gridded spatial data in thematic classes such as urban, forest, and agriculture throughout the United States (Homer and Fry, 2012). Some of these data have been further refined and summarized to increase their compatibility with other analysis tools or packages; for example, the National Stream Internet (NSI; Nagel *et al.*, 2015) and the StreamCat database (Hill *et al.*, 2015). The large investment in these national and international spatial databases is aimed at providing standardized, readily available data coverages across large spatial extents. Spatial data and GIS are often used in watershed analyses for three broad types of studies: description of the watershed; estimation of relationships between a sampled response variable and land cover or land use covariates; and spatial prediction of a response variable at unsampled locations along the stream network. Within each of these types of studies numerous watershed aspects have been described and mapped, such as hydrogeomorphic patches (Williams *et al.*, 2013) or watershed integrity (Flotemersch *et al.*, 2015); and equally numerous response variables have been modeled and predicted, such as water chemistry (Johnson *et al.*, 1997; Peterson *et al.*, 2006; Zampella *et al.*, 2007), temperature (Isaak *et al.*, 2010), or biotic condition (Frieden *et al.*, 2014). Such studies often require intensive geoprocessing and statistical analyses; thus, having nationally- or internationally-available spatial data ready to use in these endeavors can save considerable time and money. Additionally, standardization of spatial data across databases and custodians enables valid comparisons among such studies to be made. However, national and international spatial databases often contain relatively general sets of variables (e.g., land use, elevation, soil type), and local watershed applications, such as nutrient modeling, can be improved by including additional covariates specifically available within that watershed (e.g., septic systems and other point source nutrient loads; Sferratore *et al.*, 2005).

With nationally- and internationally-available spatial databases being used for an increasingly broad range of objectives, their relevance and effectiveness for specific applications requires some scrutiny. While readily available spatial data coverages empower users across all levels of research and governance, they must be used with some caution in specific applications. In studies of lakes throughout the United States, for example, lake-specific variables are known to produce significantly improved predictive models of water quality and trophic state than models based on nationally-available spatial covariates alone

(Read *et al.*, 2015; Hollister *et al.*, 2016). Thus, while national databases enable prediction of response variables in lakes lacking in situ data, these predictions can be greatly improved with additional data not currently available nationally. In a watershed context, however, the effectiveness of nationally-available spatial databases for nutrient modeling in stream networks remains largely unscrutinized.

Improvements to traditional statistical modeling and prediction of response variables in watersheds have recently been reported through the use of spatial stream network (SSN) models (Frieden *et al.*, 2014; Isaak *et al.*, 2014). SSN models incorporate covariates along with spatial autocovariance in the response variable to potentially improve upon multiple linear regression in stream networks (Ver Hoef *et al.*, 2006). SSN models can also reduce prediction errors at unsampled locations by incorporating nearby, correlated observations into the prediction (Ver Hoef and Peterson, 2010). SSN modeling has been applied to a variety of response variables observed through monitoring programs. Examples of biological response variables used in SSN modeling have included proportion of native fish expected and macroinvertebrate indices (Peterson and Ver Hoef, 2010; Frieden *et al.*, 2014). Stream water temperature and stream chemistry, such as pH, conductivity, concentrations of nitrate, sulfate, and dissolved organic carbon have also been modeled (Peterson *et al.*, 2006; Garreta *et al.*, 2010; Isaak *et al.*, 2010; Ver Hoef and Peterson, 2010). Spatial autocorrelation has been shown to exist in in-stream phosphorus concentration (Dent and Grimm, 1999; McGuire *et al.*, 2014), suggesting that SSN modeling would also be useful in this context (Hagy, 2015).

Several variables in nationally-available spatial databases have proved to be significant covariates or predictor variables in SSN models. For example, the covariates used by Peterson *et al.* (2006) included watershed area and percentages of high intensity urban, low intensity urban, row crop, and coalmine, with those percentages of land cover derived from Multi-Resolution Land Characterization (Mercurio *et al.*, 1999). Besides land cover, other covariates from nationally-developed databases used in SSN models have included percentages of particular rock types in a watershed, mean slope, and categorization of sites based on Ecosystem Health Monitoring Program regions or ecoregions (Peterson *et al.*, 2006; Peterson and Ver Hoef, 2010; Ver Hoef and Peterson, 2010). However, the relevance and effectiveness as model covariates of spatial data from national and international databases largely depends upon the specific response variable of interest. While nationally-available covariates have proved effective for modeling certain physical parameters and biotic indices, national databases do not currently contain complete data on point source covariates (e.g., septic systems and waste water treatment plants) known to affect stream nutrient concentrations, particularly phosphorus (Sferratore *et al.*, 2005; Withers and Jarvie, 2008). Although, summaries of reported pollutant discharges from some permitted facilities are publicly available throughout the United States (e.g., the U.S. EPA's Discharge Monitoring Report Pollutant Loading Tool; U.S. EPA, 2016a).

Along with increasingly available spatial data coverages and a push for more spatially explicit modeling in watersheds, more and more observed response data from a broad range of monitoring programs are also becoming available. The response data can come from existing monitoring programs (Peterson *et al.*, 2006; Isaak *et al.*, 2010) or studies specifically designed for SSN modeling (Frieden *et al.*, 2014; Som *et al.*, 2014). Because of the spatial dependencies among monitoring sites inherently necessary for SSN modeling, the sampling design of such studies influences the statistical analysis approaches that can be adopted (McDonnell *et al.*, 2015) and possibly the validity of the inferences made from their results. Using or combining data from probabilistic and targeted (non-probabilistic) surveys can further complicate analyses and inferences (Maas-Hebner *et al.*, 2015). Thus, careful exploratory data analysis and study design are imperative in SSN modeling applications.

The convergence of national spatial data sets, emerging analytical tools, and increasing amounts of monitoring data is at the forefront of widespread spatial investigations of watersheds. While these developments afford novel research avenues to scientists, uncertainty exists around their effectiveness for watershed-specific applications, and the implications of using national spatial data sets and new techniques for stream nutrient modeling should be examined. Our objective is to determine whether improvements to a predictive model of in-stream phosphorus concentration using nationally-available spatial covariates can be achieved by including additional locally-derived covariates and adopting an SSN modeling approach. We examine natural and anthropogenic influences on our response variable of median total phosphorus (TP) concentration in streams, and develop models using covariate coverages that are nationally-available throughout the United States and free of charge, as well as using additional covariate coverages that are highly specific to our study area and were costly to assemble. In particular, we ask two research questions: 1) Do models derived from national covariate coverages predict TP concentrations as well as those that include additional local coverages? 2) Do models derived using the SSN approach predict TP concentrations better than those based on a more traditional multiple linear regression approach? We also discuss approaches to SSN study design and validation of SSN models, as well as their effect on inferences gleaned from the results. Finally, we make a brief comparison to SPAtially Referenced Regression On Watershed attributes (SPARROW) that has also been used to make predictions of TP concentration in this watershed.

STUDY AREA

This study was conducted using data obtained from streams in the East Fork of the Little Miami River watershed in southwestern Ohio, USA. The East Fork is a major tributary of the Little Miami River, which in turn discharges into the Ohio River approximately 6 km east of downtown Cincinnati (Fig 1). The East Fork watershed is approximately 1,290 km^2 in area and the river has a mean annual discharge of 16.3 m^3/s at its mouth. The highest elevation in the East Fork watershed is 365 meters above sea level (m.a.s.l.), while the river's confluence with the Little Miami is at 149 m.a.s.l. The East Fork watershed is comprised of two Level IV Ecoregions (U.S. EPA, 2016b): the Loamy High Lime Till Plain and the Pre-Wisconsonian Drift Plain. Soils in the till plain, in the uppermost part of the watershed (Fig 1), are more permeable and less erodible than those in the drift plain, which is known to affect stream nutrient concentrations in this area (Daniel *et al.*, 2010). Till plain soil types are abundant in the till plain itself but also extend into the drift plain, mainly along valley bottoms (Fig 1). Land use in the East Fork watershed is dominated by row crop agriculture, which occupies 55% of the total area, primarily in the upper two-thirds of the watershed.

Deciduous forest occupies 32% of the watershed area, mainly in a transition zone between the agriculturally-dominated upper part and the urban-dominated lower part. Urban development occupies 12% of the watershed. Approximately 17,400 septic systems are also known in the East Fork watershed. One major reservoir exists along the main stem of the East Fork—Harsha Lake—with a surface area of 8.7 km^2 and maximum depth of 34 m. Harsha Lake is used for recreation and as a water source for a Clermont County-operated drinking water treatment plant (Karcher *et al.*, 2013). Two smaller reservoirs—Stonelick Lake and Lake Lorelei—also occur on tributaries of the East Fork. There are 28 pollutant discharge permits authorized under the National Pollutant Discharge Elimination System (NPDES) permit program in the East Fork watershed (Fig 1). NPDES permits are distributed to industrial, municipal, and other facilities that are permitted to discharge pollutants from discrete point sources directly into surface waters.

Methods

Study design and geoprocessing

Median total phosphorus concentration (TP) was calculated at 105 monitoring sites throughout the East Fork stream network from multiple sampling visits (median of 5 visits with a minimum of 3 and maximum of 88) between June 26th and September 11th, 2012. Monitoring of these sites was conducted by either the U.S. or Ohio Environmental Protection Agency (EPA), or both, following each agency's standard phosphorus sampling protocols (Nietch, 2006; Ohio EPA, 2009). The U.S. EPA established a nutrient monitoring program in the East Fork beginning in 2006 as part of a case study for watershed management research and development. The U.S. EPA routine stream monitoring sites range from headwaters to the main stem, and were established to capture land use variation and account for spatial nesting within tributary networks. The sites are sampled year-round with some sites visited daily, others weekly, and others every three weeks. The goal of the U.S. EPA monitoring is to assess long-term trends in nutrient chemistry at a system scale. The OHEPA sites, on the other hand, were established as part of the 2012 East Fork Watershed assessment for the State's required 303D reporting. The OHEPA watershed water quality assessment targets the low-flow conditions of streams in the region, corresponding to the conditions during which WWTPs in the system could be having the greatest impact on water quality. To combine the information collected from the two programs we limited the data obtained from the U.S. EPA program to only the period when OHEPA was sampling in the system. The sampling schemes of both programs more readily capture base flow conditions. Eighty-five of these sites were used in the construction of statistical models (called modeling sites in Fig 1), while 20 sites were selected using a spatially-balanced random sample (Olsen *et al.*, 2012) and withheld from the models for validation (called validation sites in Fig 1). A further 779 sites were included in the study for prediction, which were obtained from the NSI data set (Nagel *et al.*, 2015).

Two sets of covariates were used for statistical modeling of median TP concentration. The first, from here on referred to as the 'national' data set, included 14 covariates for which spatial data coverages are freely available throughout the entire conterminous United States. These included nine landscape variables, four land use variables, and one point source variable, which were hypothesized to affect TP in streams throughout the watershed (Table 1). Watershed area was used as a surrogate for discharge because discharge data were not available at every sampling and prediction site used in the study. In addition to the national landscape and land use variables, three additional covariate coverages (referred to as 'local') were available in the study area. These were the location of septic systems throughout the watershed, as well as the TP load and average TP concentration of releases from all wastewater treatment plants (WWTP) in the watershed during 2012. A layer of septic locations throughout the watershed was created from GIS data and parcel numbers obtained from the five County Health Departments within the watershed.

The Discharge Monitoring Report (DMR; U.S. EPA, 2016a) Pollutant Loading Tool is a national database that does provide TP concentration, and other water quality measurements, from many NPDES permit holders, such as waste water treatment plants. However, given our local knowledge of the dischargers in the watershed we were aware that several of the WWTPs monitored TP concentration as frequently as weekly, but the DMR tool at best contains monthly values or averages. To access the most data available on WWTP discharges we asked a partner at the Ohio EPA to request a data retrieval on our behalf from an in-house electronic DMR database (Paul Gledhill, Modeler, Surface Water Division, Ohio EPA, March 18, 2014). The delivered data was the same base information as is contained in the national DMR tool, but instead of monthly averages for several of the plants the data from the Ohio EPA included the weekly values reported at select WWTPs. Using the data from Ohio EPA, we were able to handle multiple permitting requirements to obtain daily WWTP TP concentrations and loads. All WWTPs in the watershed are required to report discharge, ammonia concentration, and total suspended solids (TSS); however, several plants do not have a reporting requirement for TP concentration, while others do. For the plants that do have to monitor TP concentration, we used generalized linear modeling (GLM) to interpolate daily TP concentrations as a function of ammonia, discharge, and TSS. We also tested for a seasonal effect, which was significant, and, therefore, also included it in the GLM. The model used to interpolate daily TP data was then used in a predictive mode to obtain values for WWTPs that don't have a TP reporting requirement or only are required to report TP on a monthly or quarterly basis. The daily interpolated or estimated data was required to parameterize a watershed loading model as part of another project (Karcher *et al.*, 2012). For this study, the daily TP loads were summed for each point source to get an annual TP load for 2012. This was divided by the summed daily discharges to get an average TP concentration for each point source for 2012. WWTP TP loads and concentrations were calculated for the entire year (2012) rather than for the sampling period (June–September) in order to gain some measure of the continued long-term input of TP from wastewater, which likely has lagged effects on in-stream TP concentrations. We chose to include only the location of NPDES discharge permits in the national model because using data available on the DMR tool would introduce caveats regarding the completeness of this database nationally, although we acknowledge its potential utility in other studies. Thus, the point source variables differed between our national and local data sets in that only the location of a discharge permit was included for the national, whereas actual release loads and concentrations for all WWTPs were included for the local.

Each areal covariate was hypothesized to influence TP in the watershed in a particular spatial manner (Diebel *et al.*, 2009; Frieden *et al.*, 2014). Thus, the spatial treatment of each areal covariate was selected a priori (Table 1), following initial exploratory spatial data analysis. Those covariates that were widespread throughout the watershed and hypothesized to have cumulative effects downstream were treated in a ‘cumulative watershed’ manner; that is, the total area or amount of that particular covariate in the subwatershed contributing to any reach on the stream network. Those covariates that were patchily distributed throughout the watershed and hypothesized to have more localized effects were treated in a ‘proportion of Reach Contributing Area (RCA)’ manner; that is, the proportion of the adjacent catchment area contributing directly to an individual reach (RCA) (Peterson and Ver Hoef, 2014), and not the entire subwatershed above the reach. Reaches were delineated according to stream segments in the NHDPlus v2 (McKay *et al.*, 2012). Septic systems were treated in a cumulative watershed manner and as a density in the cumulative watershed, because of their hypothesized potential to influence TP concentration in a diffuse manner (Arnscheidt *et al.*, 2007; Withers and Jarvie, 2008), as well as the known effects of septic densities on other water quality parameters in the East Fork watershed (Peed *et al.*, 2011; Schenck *et al.*, 2015). The septic data available were areal, as opposed to point data, with each septic system occupying one or several 10 m^2 raster grid cells, depending upon its size. Because of this, our septic density covariate has the units of cumulative septic area (km^2) divided by cumulative watershed area (km^2). Till plain soils in the upper part of the watershed, as well as soils occurring predominantly in valley bottoms, were treated as presence/absence because of the large number of zero values throughout the watershed, which makes transformation of the distribution of continuous variables to approximate symmetry difficult. Discharge permits were also treated as presence/absence in the national data set. Waste water treatment plant TP loads and concentrations in the local data set were accumulated from their point source downstream. Release loads were ‘reset’ below the Harsha Lake reservoir; that is, WWTP TP releases upstream of the Harsha Lake were accumulated to the reservoir but did not continue to accumulate downstream of the dam. This was done to account for the nutrient ‘sink’ effect of the reservoir. TP load and average concentration released from the Harsha Lake dam were known and included as the initial values for point source accumulation downstream of the dam, to account for dam releases as a ‘source’ of nutrients. Thus, Harsha Lake was considered to act as both a source and sink of TP in the watershed.

A total of 1,311 km of digital stream network was analyzed throughout the East Fork watershed. The stream network was initially reconditioned from the NHDPlus v2 to ensure only a single streamline existed for each reach (users can now download such reconditioned streamlines directly from the NSI (Nagel *et al.*, 2015) for the entire conterminous United States). The stream network was then converted to a topologically constructed ‘landscape network’ (LSN) using the STARS 2.0.1 toolbox (Peterson and Ver Hoef, 2014) in ArcGIS 10.2.2 (ESRI, 2014). This type of LSN is a geodatabase that contains the topological relationship information among all segments in a stream network, including flow direction, via a number of relationship tables (Peterson and Ver Hoef, 2014). All national and local covariates were then attached to their respective LSN along with the 85 observation sites, 20 validation sites, and 779 prediction sites, according to the methods outlined by Peterson

(Peterson, 2014). The NHDPlus v2 provides additional watershed attributes for each stream segment; however, the local covariates used in this study required manual processing, so the national covariates were also manually processed for consistency. Once the geoprocessing was completed, each LSN was exported as a .ssn file object, using the STARS 2.0.1 toolbox, for use in the SSN package in R Statistical Software version 3.2.0 (Ver Hoef *et al.*, 2014; R Core Team, 2015). The final .ssn file objects used in this study are provided as supporting information (S1). All exploratory analysis and model selection and evaluation subsequent to the geoprocessing was performed in R and the outputs are provided as supporting information (S2).

Statistical analyses

Initial model selection for both the national and local data sets (from here on referred to as the 'nonspatial' models) was conducted using best-subsets multiple linear regression on log-transformed covariates and the response variable. Best-subsets regression (BSR) (Furnival and Wilson, 1974) was used because severe multicollinearity of the set of available covariates made other variable selection procedures untenable. The 'best' set of covariates was selected for each data set using Akaike's Information Criterion (AIC), with a maximum of six covariates being allowed in each model to avoid overfitting. Although AIC can overfit compared to some other penalized goodness of fit measures, only AIC was readily available for the SSN models. Therefore, for consistency, AIC was used for model selection and comparison throughout.

Where categorical variables occurred in the six best covariates, interaction terms were included and BSR was repeated to determine the best set of covariates including interactions, again based on AIC. The inclusion of interaction terms required that the first order terms in the interaction also be included in the model, whether or not they were in the initial BSR selection. In addition, because of the high multicollinearity in the covariate set and the accompanying high correlation of the coefficient estimates, the regression coefficient table p-values (Wald test) sometimes indicated that the covariates selected by BSR were not significant. Standard diagnostic procedures were conducted to validate the final nonspatial models including various residual plots, influence plots, and added-variable plots (Fox and Weisberg, 2011) as shown in the supporting information (S2).

Generalized variance inflation factors (GVIF; Fox and Weisberg, 2011) for the variables (and interactions) indicated multicollinearity in the predictors; however, the variance inflation observed in these models would not affect the ability of the models to predict (Shmueli, 2010), which was our primary objective. Other dimension reduction techniques, principal components regression (PCR) and canonical correlations regression (CC), were considered. These methods eliminate the multicollinearity problem by creating predictors that are uncorrelated (only in a pure statistical sense, not in a spatial sense) linear combinations of the full set of available explanatory variables. However, these do not identify the variables that are important to the processes being studied and eliminate those that are not. Since this is a scientific study, prediction models must be validated based in part on whether the selected variables make sense scientifically.

Spatial stream network modeling was then conducted for the national and local data sets (from here on referred to as the 'spatial' models) using the final set of covariates and interactions from the nonspatial models. SSN modeling includes any combination of upstream and downstream spatial autocovariance models and parameters among sites along the stream network, as well as autocovariance parameters in Euclidean (landscape) space. Spatial autocovariance is quantified by the selected model type, range, and partial sill of the semivariogram of the response variable, which is estimated using a moving average approach in SSN modeling (Peterson and Ver Hoef, 2010). The 'best' set of autocovariance parameters to include in the spatial models, along with the nonspatial model covariates and interactions, was determined based on AIC. Standard diagnostic procedures were then conducted to validate the final spatial models for the national and local data sets (S2).

Because the models were fitted using least squares and generalized least squares (spatial models), the normal distribution for errors was not automatically assumed. The leave-one-out cross-validation (LOOCV) studentized residuals from the 85 modeling sites and the standardized residuals at the 20 validation sites were fit very well by the normal distribution. However, the predictive distribution at the 20 validation sites in the original scale is of primary importance. Accordingly, the ratios in the original scale of the observed to the predicted (exponentiated log-scale predictions) were examined. They were fitted by lognormal, Gamma, Weibull, and normal distributions. In each case the appropriate distribution was selected using AIC. The median predicted value, prediction standard error, and 90% prediction interval were subsequently back-transformed to the original scale based on the appropriate distribution.

Multiple criteria for model evaluation were used to compare the performance of the four final models: the national nonspatial and spatial, and the local nonspatial and spatial. Goodness of fit comparisons were based on AIC and decomposition of model variance components for the national and local models. Prediction accuracy was compared using the Root Mean Square of the Percent Prediction Error (RMSPPE) and the width of the 90% prediction interval as a percentage of the median prediction averaged among the 20 validation sites. The latter is conceptually similar to a coefficient of variation. The signed prediction error (prediction minus observation) expressed as a percentage of the median prediction was also calculated for each of the 20 validation sites.

Results

Adding the local covariates representing potentially important sources of phosphorus from septic systems and WWTPs to the covariates in the national data set produced different predictive models for TP concentration in the East Fork watershed. When the local covariates were included the multiple linear regression (MLR) model explained more of the variance in median TP concentration with an adjusted R^2 = 0.552 compared to an adjusted R^2 = 0.483 with the national covariates. A set of four covariates and two interactions produced the 'best' model from the national data set, while a set of five covariates and three interactions emerged when local covariates were included (Table 2). Both cumulative septic area and WWTP TP loads were significant covariates in the local model, whereas septic density did not emerge as a significant covariate. Furthermore, the inclusion of septic

EPA Author Manuscript

EPA Author Manuscript

systems and WWTP TP loads in the data set resulted in a different set of covariates that were significant for predicting TP concentration. In particular, watershed area, low (0–2%) slope area, and Clermont soils were not significant after the septic and WWTP covariates were included, whereas agriculture and Rossmoyne soils were (Table 2).

The coefficients of the significant covariates in the model fit from the national data set suggested that with an increase in watershed area and in the presence of till plain soils there is a decrease in median TP concentration (Table 2). The interaction between Clermont and till plain soils further decreased median TP concentration, although the effect was smaller than that of till plain soils alone, suggesting that as the area of Clermont soils increases, the effect of till plain soils on lowering median TP concentration is reduced. Clermont soils alone did not have a significant effect on median TP concentration. The interaction of low slopes with till plain soils significantly increased median TP concentration in the models fit with the national data set (Table 2). In the models fit with the local covariates included, cumulative septic area significantly increased median TP concentration, as did WWTP TP load in the presence of till plain soils (Table 2). Agricultural area also significantly increased median TP concentration in the local models; however, the interaction between agricultural area and the presence of till plain soils reduced median TP concentration (Table 2). Increasing area of Rossmoyne soils significantly reduced median TP concentration in the model fit with the local covariates included, whereas this covariate was not significant in the model fit from the national data set only. Coefficient estimates varied only modestly and consistently between the nonspatial and spatial models for the national and local covariates.

Spatial stream network modeling revealed that spatial autocovariance existed among samples of median TP concentration in the East Fork watershed. The autocovariance structure was best explained by a linear-with-sill tail-up autocovariance model (see Ver Hoef *et al.*, 2006 for details). This type of autocovariance model suggests that median TP concentration at a site is related to median TP concentration at sites upstream, and that this relationship weakens with distance upstream in a linear manner. The range over which spatial autocovariance existed in median TP concentration (i.e., the distance at which the 'sill' or maximum variance among pairs of sites was reached) was approximately 25 river km, with the maximum downstream travel distance in the watershed being around 135 river km.

In terms of goodness of fit, the spatial model outperformed the nonspatial model, which assumes median TP concentration samples are not spatially autocorrelated, using the national data set and when the local covariates were included. Based on AIC, the 'best' model was the local spatial followed by the local nonspatial, national spatial, and national nonspatial. For the model fitted from the national covariates AIC was reduced by 9.6 points and the nugget, or unexplained variance, declined from 0.480 to 0.139, representing a 71% reduction, by including spatial autocovariance parameters in the model (Table 2). In fact, more of the variance in median TP concentration was explained by the autocovariance structure (in-stream spatial covariance) than by the covariates in the national spatial model. Similarly, in the model fit from the data set that included the local covariates AIC dropped by 1.2 points and the nugget was reduced by 47%, from 0.405 to 0.214, by including spatial parameters. The amount of variance in median TP concentration explained by the covariates (equivalent to the generalized R^2 in MLR) also decreased by including autocovariance parameters for both the national and the local models (Table 2). This suggests that nonspatial models may be artificially inflating the amount of variance in median TP concentration that is being explained by the covariates because spatial autocorrelation inherent in the stream network is not being accounted for; that is, the assumption of independence among samples is being violated.

In terms of prediction accuracy, the nonspatial and spatial models fit from the data set that included the local covariates predicted median TP concentration more accurately than the national models. RMSPPE among the 20 validation sites was lowest for the local spatial model followed by the local nonspatial, national spatial, and national nonspatial (Table 3). However, all RMSPPE values were high, ranging from 89 to 106% (Table 3). These high RMSPPEs appeared to be related to two outliers, sites CWL and 200497, which had observed median TP concentration much higher than predicted in all four models (Fig 2). Removal of these two outliers reduced the RMSPPE from 106 to 56% for the national nonspatial, from 94 to 50% for the national spatial, from 93 to 48% for the local nonspatial, and from 89 to 48% for the local spatial model. Site M04S16 also had a much higher predicted median TP concentration than observed in the national models, although it did not fall outside of the 90% prediction interval (Fig 2a and b).

Predicted median TP concentration values were generally close to those observed up to around 0.2 mg/L in all four models, with the 90% prediction intervals being relatively tight in this prediction range (Fig 2). However, there was a substantial increase in the width of the 90% prediction intervals as median predictions increased, particularly in the models fitted from the national data set (Fig 2a and b). This increase in prediction intervals was associated with fewer extreme values in the modeling data set; in particular, site CWL had the highest observed median TP concentration of any of the 105 monitoring sites but was withheld from the modeling, by chance, in selection of the validation data set. However, by including specific local covariates in the models the prediction intervals at high prediction values were greatly improved and no sites were extremely over-predicted (Fig 2c and d). Inclusion of spatial autocovariance parameters also improved prediction accuracy at sites with high observed median TP concentration. In particular, median TP concentration at site M04S29, located near the mouth of the river, was accurately predicted in both spatial models but under-predicted in both nonspatial models (Fig 2). The averages among the 20 validation sites of the width of the 90% prediction interval as a percentage of the predicted value were smaller for the local models than for the national models, as well as being 11% smaller in the spatial model than the nonspatial for the national data set (Table 3).

Discussion

Model evaluation and prediction using national and local covariates

Nationally-available spatial databases are increasingly being utilized in watershed science and management in the United States and other countries. Our results suggest that while covariates from these databases (e.g., the NLCD) can be used to produce reasonable statistical models of nutrients in stream networks (TP concentrations in this study), the inclusion of additional covariates that are not currently available nationwide, specifically the locations of septic systems and the TP load released from WWTPs, improves both model fit and prediction accuracy. Sferratore *et al.* (2005) found that global land use and lithology data could be used to correctly predict watershed nutrient fluxes from diffuse sources, but that prediction accuracy was sensitive to knowledge of the distribution of point sources of nutrients, particularly for phosphorus. Our results also indicate that knowledge of local point sources improves models of TP concentration. Although our analysis covered only one watershed, septic systems and WWTPs represent explicit sources of phosphorus. Therefore, it is reasonable to suggest that were these covariates available nationally, the effectiveness of these databases for stream network modeling of nutrients in the United States would be greatly improved. Given the relevance of predicting nutrient concentrations to water quality management everywhere, future iterations of these databases should aim to include such covariates. While databases exist nationally on the location of discharge permits and annual discharge summaries of some facilities (e.g., the U.S. EPA's Discharge Monitoring Report Pollutant Loading Tool), these databases could be built upon to become more complete and more frequently updated. Such updates are occurring with the DMR Pollutant Loading Tool as electronic submission of discharge monitoring reports can now be done. However, facilities may be missing from the DMR Pollutant Loading Tool, as noted on the tool website. It should be noted that WWTP nutrient loads are often not static over time. We found it important to have knowledge of the actual loadings from these point sources, as opposed to just their location in our study. If loads are changing over time due to new permit requirements, increased capacity, or upgraded nutrient removal technologies, then it would be necessary to provide adequate metadata and make routine updates to the data contained on point source loadings in the national databases. This could add considerable costs to managing these data.

Our results suggest that an SSN modeling approach also improves goodness of fit and prediction accuracy of in-stream TP concentration models compared to traditional MLR. The importance of including spatial autocovariance parameters was particularly evident for the model based on the nationally-available covariates (Table 2), indicating that an SSN modeling approach should be adopted when using national covariates alone to model TP concentration in stream networks. Accurate prediction at sites with high TP concentrations depended upon the inclusion of spatial autocovariance parameters in the national and local models, particularly in areas around other sites of high TP concentrations. For example, site M04S29 at the mouth of the river (Fig 3) was under-predicted in both nonspatial models (Fig 2a and c), which only have access to the covariate values at that site in order to make a prediction. However, both spatial models accurately predicted median TP concentration at this site (Fig 2b and d) because SSN models are able to draw upon known observed values at nearby, correlated locations in order to improve predictions (Ver Hoef and Peterson, 2010). In this case, there were two sites with high TP concentrations near the mouth of the river that were used by the SSN models to predict high median TP concentration at the validation site M04S29 (Fig 3). These results support the increasing evidence that SSN modeling improves on traditional MLR in stream networks (Frieden *et al.*, 2014; Isaak *et al.*, 2014).

The relatively poor prediction accuracy of models in this study was largely attributable to two outliers in the validation data set (sites CWL and 200497; Fig 2). Site CWL is located on a headwater stream surrounded by intensive agriculture and experienced stagnant water conditions in 2012, resulting in an extremely high TP concentration that was not predicted by the models in this study. Site 200497 is located on a small tributary approximately 2 km downstream of a WWTP release (Fig 1). Although other sites directly downstream of WWTPs were included in the modeling, 200497 is the only example of a small tributary site receiving discharges from a major point source. Dilution in the small tributary is minimal compared to the main channel. Removal of these two outliers resulted in substantial improvements in prediction accuracy among the remaining 18 validation sites.

Differences among the four models inevitably produce different spatial predictions of median TP concentration throughout the East Fork watershed using the NSI prediction sites (Fig 3). These differences are most obvious in the upper part of the watershed (i.e., in the till plain) and around the mouth of the river (Fig 3). In particular, both national models predict very high median TP concentration along the main stem in the upper part of the watershed, while tributaries are predicted to have very low median TP concentration. Conversely, the local models predict lower median TP concentration along the main stem and higher values in the tributaries of the till plain (Fig 3). In the till plain, low slopes only occur in a relatively narrow band along the main valley floor, coincident with the location of high predicted values in the national models. The interaction between low slopes and till plain soils in the national models resulted in these sites having high predicted median TP concentration. However, there are also septic systems in the till plain region of the watershed, as well as three waste water treatment plants (Fig 1). The local model coefficients suggest that septic systems and WWTP TP loads, in the presence of till plain soils, cause the moderately high median TP concentration values observed at the three uppermost modeling sites. Without access to these local covariates, the national model has misattributed these high values to the interaction between low slopes and till plain soils, and over-predicted sites in this area; for example, site M04S16 (Fig 2a and b). Discrepancies in predictions are also evident between the nonspatial and spatial models along the lower main stem near the mouth of the river (Fig 3). In particular, median TP concentration predictions in this area are much lower in both nonspatial models compared to their respective spatial counterpart. The lower predicted values in the nonspatial models along the lower main stem can be attributed to large watershed areas in the national model and large areas of Rossmoyne soils in the local model. In contrast, the spatial models are able to draw on the nearby modeling sites, which have known high values, in order to make better predictions along the lower main stem (Fig 3). This is also the reason why validation site M04S29 was more accurately predicted in both spatial models than in the nonspatial models (Fig 2).

The inferences gleaned from the four TP concentration models in this study were influenced by working retrospectively using existing monitoring data and by the design of our model validation approach. Our objective was to determine how local watershed modeling and prediction based on national spatial databases could be improved by including additional local covariates and adopting the SSN modeling approach. In order to meaningfully evaluate prediction accuracy, it was necessary to withhold known observation data from the modeling process. We chose a spatially-balanced random survey design to select the validation sites because it guarantees that samples represent the entire extent of the 105 monitoring sites within the study area as best as possible (Olsen *et al.*, 2012). Obtaining a good spatial representation throughout the study area is essential because the location of observations is known to influence spatial patterns observed in watersheds (e.g., Scown *et al.*, 2016). However, the statistical distribution of the response variable among the modeling and validation sites was overlooked by adopting a spatially-balanced random survey of monitoring sites, as were the spatial distributions of the covariates. In particular, the sites with the highest and lowest observed median TP concentration values of all 105 sites used in this study (sites CWL and M04P12, respectively) were withheld from the modeling data set and included in the validation data set, by chance. Also the till plain soils covariate was only relevant to ten modeling sites, with the interaction between till plain soils and other covariates represented by even fewer sites. Thus, coefficient estimations and spatial predictions were likely affected by the spatially-balanced random sampling of validation sites, and model results may have been different had the validation sites selected been different. The monitoring program designs adopted by the U.S. and Ohio EPAs, whose data were used in this study, may have further affected the model inferences. The Ohio EPA monitoring program in the East Fork watershed targets stream biotic assessments at subwatersheds around 100 km^2 in area or greater and around the potential effects of permitted point sources. This places sites primarily along the main stem of the East Fork and larger confluent tributaries. The U.S. EPA monitoring program spreads sites more evenly among small tributaries and main stems, but focuses effort in the lower portions of the watershed so that all sites could be visited on the same day and more frequently. Thus, smaller tributaries, particularly in the eastern part of the watershed, were underrepresented in the sample of monitoring sites (Fig 1) and median TP concentration predictions are likely affected by the absence of monitoring data in these areas of the East Fork watershed.

We can make some broad comparisons to other modeling approaches, most notably SPAtially Referenced Regression On Watershed attributes (SPARROW), that have been used to predict TP concentration in the East Fork of the Little Miami River. Although a full comparison of SPARROW and SSN modeling is beyond the scope of this paper, highlighting several important differences between the two is necessary. SPARROW models are developed for national or regional applications (Alexander *et al.*, 2004; Alexander *et al.*, 2007; Robertson and Saad, 2011), and, consequently, use stream network GIS data at a coarser spatial resolution (1:500,000 scale stream lines and 1-km or 100-m DEM) compared to those used in our SSN modeling (1:100,000 scale stream lines and 10-m DEM). The data inputs for the response variable also differ between SPARROW and our SSN analyses. Concentrations of TP used in SPARROW are derived from gages on large rivers with long periods of record (Alexander *et al.*, 2004; Alexander *et al.*, 2007; Robertson and Saad, 2011), of which only one occurs in the East Fork watershed at the outlet. In contrast, we used grab samples taken at 105 locations throughout the watershed, including on smaller streams, for our modeling and validation. These differences in GIS and water quality inputs create substantial differences in the spatial extent and resolution of TP concentration predictions. While both models are predicting over the same areal extent of the watershed, ~ 1300 km^2, they differ substantially in the length of stream network along which those predictions are made. In the headwaters of the East Fork, which are encompassed by the 12-digit hydrologic unit 050902021006, SPARROW predicts a TP concentration of 0.23 mg/L only for the 23 km of stream reach of that main stem using the MRB3 model (USGS, 2015). Our SSN predictions in that same HUC12 are at the midpoints of 49 main stem and tributary reaches, totaling 86 stream km, with TP concentration ranging from 0.08 mg/L to 0.46 mg/L with a median of 0.21 mg/L. As intended, SPARROW models provide predictions of nutrient concentrations at national and regional extents in broad brushstrokes; whereas SSN can provide nutrient concentration predictions at a much finer spatial resolution that can potentially lead to more specific management action. Both SPARROW (Alexander *et al.*, 2004) and SSN (Hagy, 2015) analyses have emphasized the importance of gaining accurate estimates of nutrient loads from point sources to improve those analyses.

Interpretation of national and local covariates

The best-subsets regression using the national spatial data coverages produced a relatively generic set of covariates whose relevance for understanding and managing TP concentrations in the East Fork watershed is limited. In fact, the autocovariance parameters in the national spatial model explained more of the variance in median TP concentration than the covariates (Table 2), indicating that knowledge of TP concentration at nearby locations is more informative than the set of specific national covariates used at a particular site. The decline in median TP concentration with increasing watershed area in the national model reflects a dilution effect and has little consequence for interpretation and management. However, the significantly lower TP concentration in the presence of till plain soils compared to drift plain soils in the national model is consistent with the findings of Daniel *et al.* (2010) who observed lower TP concentrations in catchments in the till plain versus the drift plain of the Little Miami River watershed. The presence or absence of till plain soils also had significant interactions with cumulative Clermont soil area and cumulative area of land having a slope of 0–2%. Clermont soil, which is widespread throughout the drift plain, appears to reduce the effect that till plain soils have on lowering in-stream TP concentration; however, the cumulative area of Clermont soil did not have a significant effect itself. The mechanisms behind the observed effects of areas of soil types are worthy of future investigation; for example, by incorporating into the model soil attributes such as erodibility and permeability, which are contained in the national database. The national model also appears to have misattributed high median TP concentration in the upper watershed to the interaction between low slopes and till plain soils, as discussed in the preceding section. Despite their generality, the national covariates are readily available, require much less geoprocessing than the local covariates, and produce a reasonable model of median TP concentration in the East Fork when combined with spatial autocovariance parameters.

The covariates that emerged from the local best-subsets regression model are more relevant for interpretation and management of TP concentrations in the East Fork watershed. We observed a positive relationship between TP concentration and cumulative agricultural area in the local model. Agricultural land cover is a nationally-available covariate; however, it did not emerge from the national best-subsets regression, suggesting that local covariates are required to decompose the multiple interacting influences on TP concentrations in the East Fork watershed. The positive effect of agricultural land cover on TP concentrations observed in this study is consistent with previous research in the Little Miami River (Daniel *et al.*, 2010) and other watersheds (Carpenter *et al.*, 1998). The interaction between agricultural area and till plain soils had a negative coefficient in the local model, suggesting that till plain soils may buffer the effect of agriculture in the East Fork watershed. This is consistent with the findings of Daniel *et al.* (2010) who found that in the till plain region of the Little Miami River watershed, TP concentration was not significantly related to the percentage of row crop land cover, whereas in the drift plain there was a positive relationship. The effect till plain soils have on lowering stream nutrient loads is likely related to these soils being more permeable and less erodible than drift plain soils (Daniel *et al.*, 2010), which may enable them to retain nutrients within the soil profile rather than losing them to the stream via runoff and erosion. We also observed an interaction between the presence of till plain soils and WWTP TP loads on stream TP concentrations. Permitted point source discharge locations and WWTP outfalls were excluded from the study design of Daniel *et al.* (2010). In other watersheds, however, WWTP densities are associated with higher stream TP concentrations (Rothenberger *et al.*, 2009), as are greater sewage flows from treatment plants (Zampella, 1994). Septic tanks have been hypothesized to be a low-level, but chronic input of phosphorus into streams and rivers (Arnscheidt *et al.*, 2007), and we observed a significant positive relationship between septic area and median TP concentration in the local model. Although an individual septic system can be considered a potential point source of stream TP, septic systems are so widespread throughout the East Fork watershed that our results suggest they have cumulative effects on TP concentrations and operate more like a diffuse source at the watershed scale.

Because of the prominence of fine clay and poorly infiltrating soil types in the East Fork watershed (i.e., Clermont, Avonburg, and Rossmoyne soils) traditional septic systems that rely on buried leach fields for wastewater treatment are prone to failure. Aerobic septic systems, used frequently in place of the traditional systems in the watershed, require more homeowner attention to remain effective and are often designed with direct discharges to receiving streams. Therefore, these conditions of onsite wastewater management that are somewhat specific to the study watershed likely help promote the significance of septic systems found for predictive modeling of phosphorus in East Fork streams. Other studies conducted in East Fork streams have found septic densities to correlate well with molecular markers of human fecal bacteria (Peed *et al.*, 2011) and other contaminants of emerging concern (Schenck *et al.*, 2015). Although cumulative septic area rather than density emerged as a significant covariate in this study, we note this aspect of the study system as a caveat to the relative importance of having data on septic systems to improve predictions for nutrients in watersheds.

Conclusions

Our objective was to determine whether a predictive model of in-stream phosphorus concentrations based on nationally-available spatial covariates could be improved by including additional locally-derived covariates and adopting an SSN modeling approach. Nationally-available spatial data can be used for spatial predictions of nutrients throughout stream networks; however, additional local covariates provided a more mechanistic interpretation of influences on TP concentrations in the East Fork watershed, as well as increasing model goodness of fit and prediction accuracy. While the national covariates were effective in building generic models of median TP concentration, these models were highly susceptible to prediction errors because of misattribution of mechanisms. Adopting an SSN modeling approach was essential to improve the prediction accuracies of the national model, and inclusion of WWTP TP loads and septic areas in the local models resulted in further improvements. The advantage that the national covariates have is that they are becoming readily available in the format required to conduct SSN modeling (Hill *et al.*, 2015; Nagel *et al.*, 2015), thus dramatically reducing geoprocessing costs. For models built using national covariates, as well as models with additional local covariates, SSN prediction provides researchers with 1) an expected value of a response variable that can then be field tested, and 2) a spatial distribution of prediction errors. These outcomes can be used to inform future monitoring programs or to designate additional monitoring sites around existing programs (Peterson and Ver Hoef, 2010; Isaak *et al.*, 2014). Investigation of the prediction errors associated with the models presented in this study would likely yield valuable information for managers in the East Fork watershed; however, that is beyond the scope of this study.

The effects of monitoring and analysis design on inferences in this study help to inform the design of future stream network modeling studies, which will depend upon the study objective. If the objective is primarily to fit an SSN model to a response variable, perhaps for purely explanatory purposes, all observational data can be retained in the modeling process and validation can be conducted using various cross-validation techniques (e.g., LOOCV). In addition, the initial design of monitoring locations should cover a broad distribution of paired distances among sites, as well as multiple samples around confluences, to enable longitudinal network relationships to be established and autocovariance functions to be quantified (Frieden *et al.*, 2014; McDonnell *et al.*, 2015). If the aim of the study is to evaluate prediction accuracy, withholding of validation sites is necessary; however, the statistical and spatial distributions of the response variable and covariates must be considered when selecting these sites. Conducting a spatially-balanced random sample stratified by certain patchily-distributed covariates (e.g., till plain soils in this study) may be a solution. Imposing further constraints based on the statistical distribution of the response variable may also be necessary to accurately model and predict extreme values. Careful consideration of such stratified sampling approaches is essential (Maas-Hebner *et al.*, 2015). Iteratively conducting modeling and validation with multiple samples could be incorporated into the process within the SSN package in R (Jay Ver Hoef, November 5, 2015, personal communication); however, it is clear from our results that conducting modeling and prediction on a single set of modeling and validation sites can greatly affect the study inferences. Regardless of the approach adopted, thorough initial exploratory data analysis is imperative in SSN modeling studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to acknowledge Balaji Ramakrishnan for preparing the septic system data set, Amr Safwat for providing general GIS support to the SSN modeling, Matthew Heberling for work assignment management and statistical oversight, Sri Panguluri for project management, and Heather Golden and Michael Griffith for their comments on an earlier version of the manuscript. We also thank three anonymous reviewers whose input greatly improved the manuscript, particularly in relation to SPARROW comparisons and the DMR Pollutant Loading Tool. This research was supported in part by an appointment to the ORISE participant research program supported by an interagency agreement between EPA and DOE. The views and interpretations expressed herein are those of the authors and do not reflect those of the U.S. EPA. The mention of equipment or tradenames do not connote official endorsement by the U.S. EPA.

Literature Cited

Alexander R, Smith R, Schwarz G. Estimates of Diffuse Phosphorus Sources in Surface Waters of the United States Using a Spatially Referenced Watershed Model. Water Science and Technology. 2004; 49:1–10.

Alexander RB, Smith RA, Schwarz GE, Boyer EW, Nolan JV, Brakebill JW. Differences in Phosphorus and Nitrogen Delivery to the Gulf of Mexico from the Mississippi River Basin. Environmental Science & Technology. 2007; 42:822–830. DOI: 10.1021/es0716103

Arnscheidt J, Jordan P, Li S, McCormick S, McFaul R, McGrogan HJ, Neal M, Sims JT. Defining the Sources of Low-Flow Phosphorus Transfers in Complex Catchments. Science of the Total Environment. 2007; 382:1–13. DOI: 10.1016/j.scitotenv.2007.03.036 [PubMed: 17512972]

Bureau of Metereology (BOM)Australian Hydrological Geospatial Fabric Version 2 Australian Government; Canberra: 2015

Carpenter SR, Caraco NF, Correll DL, Howarth RW, Sharpley AN, Smith VH. Nonpoint Pollution of Surface Waters with Phosphorus and Nitrogen. Ecological Applications. 1998; 8:559–568. DOI: 10.1890/1051-0761(1998)008[0559:NPOSWW]2.0.CO;2

Daniel FB, Griffith MB, Troyer ME. Influences of Spatial Scale and Soil Permeability on Relationships between Land Cover and Baseflow Stream Nutrient Concentrations. Environmental Management. 2010; 45:336–350. DOI: 10.1007/s00267-009-9401-x [PubMed: 19956950]

Dent CL, Grimm NB. Spatial Heterogeneity of Stream Water Nutrient Concentrations over Successional Time. Ecology. 1999; 80:2283–2298. DOI: 10.2307/176910

Diebel MW, Maxted JT, Robertson DM, Han S, Vander Zanden MJ. Landscape Planning for Agricultural Nonpoint Source Pollution Reduction III: Assessing Phosphorus and Sediment Reduction Potential. Environmental Management. 2009; 43:69–83. DOI: 10.1007/s00267-008-9139-x [PubMed: 18521658]

ESRIArcgis Desktop: Release 10.2.2 Redlands, CA: Environmental Systems Research Institute; 2014

Flotemersch JE, Leibowitz SG, Hill RA, Stoddard JL, Thoms MC, Tharme RE. A Watershed Integrity Definition and Assessment Approach to Support Strategic Management of Watersheds. River Research and Applications. 2015; 32:1654–1671. DOI: 10.1002/rra.2978

Fox J, , Weisberg S. An R Companion to Applied Regression 2. Los Angeles, CA: SAGE Publications, Inc; 2011

Frieden JC, Peterson EE, Angus Webb J, Negus PM. Improving the Predictive Power of Spatial Statistical Models of Stream Macroinvertebrates Using Weighted Autocovariance Functions. Environmental Modelling & Software. 2014; 60:320–330. DOI: 10.1016/j.envsoft.2014.06.019

Furnival GM, Wilson RW. Regressions by Leaps and Bounds. Technometrics. 1974; 16:499–511. DOI: 10.2307/1267601

Garreta V, Monestiez P, Ver Hoef JM. Spatial Modelling and Prediction on River Networks: Up Model, Down Model or Hybrid? Environmetrics. 2010; 21:439–456. DOI: 10.1002/env.995

Hagy JD, III. U.S. Environmental Protection Agency Report EPA/600/R-15/262 National Health and Enironmental Effects Research Laboratory, Gulf Ecology Division, U.S. Environmental Protection Agency; Gulf Breeze, FL: 2015 Science Supporting Numeric Nutrient Criteria for Lakes and Their Watersheds: A Synopsis of Research Completed for the U.S. Environmental Protection Agency.

Hill RA, Weber MH, Liebowitz SG, Olsen AR, Thornbrugh DJ. The Stream-Catchment (Streamcat) Dataset: A Database of Watershed Metrics for the Conterminous United States. Journal of the American Water Resources Association. 2015; 52:1–9. DOI: 10.1111/1752-1688.12372

Hollister JW, Milstead WB, Kreakie BJ. Modeling Lake Trophic State: A Random Forest Approach. Ecosphere. 2016; 7:e01321.doi: 10.1002/ecs2.1321

Homer C, , Fry J. The National Land Cover Database U.S. Geological Survey, U.S. Department of the Interior; Washington, DC: 2012

EPA Author Manuscript

EPA Author Manuscript

Isaak DJ, Luce CH, Rieman BE, Nagel DE, Peterson EE, Horan DL, Parkes S, Chandler GL. Effects of Climate Change and Wildfire on Stream Temperatures and Salmonid Thermal Habitat in a Mountain River Network. Ecological Applications. 2010; 20:1350–1371. DOI: 10.1890/09-0822.1 [PubMed: 20666254]

Isaak DJ, Peterson EE, Ver Hoef JM, Wenger SJ, Falke JA, Torgersen CE, Sowder C, Steel EA, Fortin MJ, Jordan CE. Applications of Spatial Statistical Network Models to Stream Data. Wiley Interdisciplinary Reviews: Water. 2014; 1:277–294. DOI: 10.1002/wat2.1023

Johnson LB, Richards C, Host GE, Arthur J. Landscape Influences on Water Chemistry in Mid-Western Stream Ecosystems. Freshwater Biology. 1997; 37:193–208. DOI: 10.1046/j.1365-2427.1997.d01-539.x

Karcher SC, , Van Briesen J, , Nietch CT, , Heberling MT, , Ramakrishnan B. Examining the Feasibility of Water Quality Trading in the East Fork Watershed Case Study U.S. Environmental Protection Agency; Cincinnati, OH: 2012

Karcher SC, VanBriesen JM, Nietch CT. Alternative Land-Use Method for Spatially Informed Watershed Management Decision Making Using Swat. Journal of Environmental Engineering. 2013; 139:1413–1423.

Maas-Hebner KG, Harte M, Molina N, Hughes RM, Schreck C, Yeakley JA. Combining and Aggregating Environmental Data for Status and Trend Assessments: Challenges and Approaches. Environmental Monitoring and Assessment. 2015; 187:1–16. DOI: 10.1007/s10661-015-4504-8 [PubMed: 25600401]

McDonnell TC, Sloat MR, Sullivan TJ, Dolloff CA, Hessburg PF, Povak NA, Jackson WA, Sams C. Downstream Warming and Headwater Acidity May Diminish Coldwater Habitat in Southern Appalachian Mountain Streams. Plos One. 2015; 10:e0134757.doi: 10.1371/journal.pone.0134757 [PubMed: 26247361]

McGuire KJ, Torgersen CE, Likens GE, Buso DC, Lowe WH, Bailey SW. Network Analysis Reveals Multiscale Controls on Streamwater Chemistry. Proceedings of the National Academy of Sciences. 2014; 111:7030–7035. DOI: 10.1073/pnas.1404820111

McKay L, , Bondelid T, , Dewald T, , Johnston J, , Moore R, , Rea A. NHDPlus Version 2: User Guide Horizon Systems; Herdon, VA: 2012

Mercurio G, , Chaillou JC, , Roth NE. Guide to Using 1995–1997 Maryland Biological Stream Survey Data Maryland Department of Natural Resources; Annapolis, MD: 1999

Nagel DE, , Peterson E, , Isaak DJ, , Ver Hoef JM, , Horan DL. National Stream Internet Protocol and User Guide U.S. Forest Service, Rocky Mountain Research Station Air, Water, and Aquatic Environments Program; Boise, ID: 2015

Nietch CT. Quality Assurance Project Plan Id: 634-Q-2-0 U.S. Environmental Protection Agency, Office of Research and Development; Cincinnati, OH: 2006 Experimental Stream Facility and East Fork Watershed Study: Research Linking Land Use Management Practices to Ecological Structure and Function in Small Stream Ecosystems.

Ohio Environmental Protection Agency (Ohio EPA)Ohio EPA Manual of Surveillance Methods and Quality Assurance Practices, Updated Edition Ohio EPA, Division of Environmental Services; Columbus, OH: 2009

Olsen AR, , Kincaid TM, , Payton Q. Spatially Balanced Survey Designs for Natural Resources. In: Gitzen RA, Millspaugh JJ, Cooper AB, , Licht DS, editorsDesign and Analysis of Long-Term Ecological Monitoring Studies Cambridge University Press; Cambridge, UK: 2012 126150

Peed LA, Nietch CT, Kelty CA, Meckes M, Mooney T, Sivaganesan M, Shanks OC. Combining Land Use Information and Small Stream Sampling with PCR-Based Methods for Better Characterization of Diffuse Sources of Human Fecal Pollution. Environmental Science & Technology. 2011; 45:5652–5659. DOI: 10.1021/es2003167 [PubMed: 21662992]

Peterson EE. Stars: Spatial Tools for the Analysis of River Systems Version 2.0.0 - a Tutorial CSIRO; Canberra: 2014

Peterson EE, Merton AA, Theobald DM, Urquhart NS. Patterns of Spatial Autocorrelation in Stream Water Chemistry. Environmental Monitoring and Assessment. 2006; 121:571–596. DOI: 10.1007/s10661-005-9156-7 [PubMed: 16897525]

EPA Author Manuscript
EPA Author Manuscript

Peterson EE, Ver Hoef JM. A Mixed-Model Moving-Average Approach to Geostatistical Modeling in Stream Networks. Ecology. 2010; 91:644–651. DOI: 10.1890/08-1668.1 [PubMed: 20426324]

Peterson EE, Ver Hoef JM. Stars: An Arcgis Toolset Used to Calculate the Spatial Information Needed to Fit Spatial Statistical Models to Stream Network Data. J Stat Softw. 2014; 56:1–17. DOI: 10.18637/jss.v056.i02

R Core TeamR Foundation for Statistical Computing Vienna, Austria: 2015 R: A Language and Environment for Statistical Computing.

Read EK, Patil VP, Oliver SK, Hetherington AL, Brentrup JA, Zwart JA, Winters KM, Corman JR, Nodine ER, Woolway RI, Dugan HA, Jaimes A, Santoso AB, Hong GS, Winslow LA, Hanson PC, Weathers KC. The Importance of Lake-Specific Characteristics for Water Quality across the Continental United States. Ecological Applications. 2015; 25:943–955. DOI: 10.1890/14-0935.1 [PubMed: 26465035]

Robertson DM, Saad DA. Nutrient Inputs to the Laurentian Great Lakes by Source and Watershed Estimated Using Sparrow Watershed Models. Journal of the American Water Resources Association. 2011; 47:1011–1033. DOI: 10.1111/j.1752-1688.2011.00574.x [PubMed: 22457580]

Rothenberger MB, Burkholder JM, Brownie C. Long-Term Effects of Changing Land Use Practices on Surface Water Quality in a Coastal River and Lagoonal Estuary. Environmental Management. 2009; 44:505–523. DOI: 10.1007/s00267-009-9330-8 [PubMed: 19597872]

Schenck K, Rosenblum L, Ramakrishnan B, Carson J, Macke D, Nietch C. Correlation of Trace Contaminants to Wastewater Management Practices in Small Watersheds. Environmental Science: Processes & Impacts. 2015; 17:956–964. DOI: 10.1039/C4EM00583J [PubMed: 25881834]

Scown MW, , Thoms MC, , De Jager NR. Measuring Spatial Pattern in Floodplains: A Step Towards Understanding the Complexity of Floodplain Ecosystems. In: Gilvear DJ, Greenwood M, Thoms MC, , Wood P, editorsRiver Science: Research and Management for the 21st Century John Wiley and Sons; Chichester, UK: 2016 103131

Sferratore A, Billen G, Garnier J, Thery S. Modeling Nutrient (N, P, Si) Budget in the Seine Watershed: Application of the Riverstrahler Model Using Data from Local to Global Scale Resolution. Global Biogeochemical Cycles. 2005; 19:GB4S07.doi: 10.1029/2005GB002496

Shmueli G. To Explain or to Predict? Statistical Science. 2010; 25:289–310. DOI: 10.1214/10-STS330

Som NA, Monestiez P, Ver Hoef JM, Zimmerman DL, Peterson EE. Spatial Sampling on Streams: Principles for Inference on Aquatic Networks. Environmetrics. 2014; 25:306–323. DOI: 10.1002/env.2284

Strayer DL, Beighley RE, Thompson LC, Brooks S, Nilsson C, Pinay G, Naiman RJ. Effects of Land Cover on Stream Ecosystems: Roles of Empirical Models and Scaling Issues. Ecosystems. 2003; 6:407–423. DOI: 10.1007/PL00021506

United States Environmental Protection Agency (U.S. EPA)Office of Science and Technology U.S. Environmental Protection Agency; 2016a Discharge Monitoring Report Pollutant Loading Tool. https://cfpub.epa.gov/dmr/. Effective late November 2017 new tool location: https://echo.epa.gov/resources/general-info/loading-tool-modernization

United States Environmental Protection Agency (U.S. EPA). Level III and IV Ecoregions 2016b http://catalog.data.gov/dataset/u-s-level-iii-and-iv-ecoregions-u-s-epahttp://catalog.data.gov/dataset/u-s-level-iii-and-iv-ecoregions-u-s-epa

United States Geological Survey (USGS)U.S. Geological Survery U.S. Department of the Interior; Washington, DC: 2015 Sparrow Decision Support System - Upper Midwest Total Phosphorus in Water Model.

Ver Hoef JM, Peterson E, Theobald D. Spatial Statistical Models That Use Flow and Stream Distance. Environmental and Ecological Statistics. 2006; 13:449–464. DOI: 10.1007/s10651-006-0022-8

Ver Hoef JM, Peterson EE. A Moving Average Approach for Spatial Statistical Models of Stream Networks. Journal of the American Statistical Association. 2010; 105:6–18. DOI: 10.1198/jasa.2009.ap08248

Ver Hoef JM, Peterson EE, Clifford D, Shah R. SSN: An R Package for Spatial Statistical Modeling on Stream Networks. Journal of Statistical Software. 2014; 56:1–43. DOI: 10.18637/jss.v056.i03

Williams BS, D’Amico E, Kastens JH, Thorp JH, Flotemersch JE, Thoms MC. Automated Riverine Landscape Characterization: Gis-Based Tools for Watershed-Scale Research, Assessment, and

Management. Environmental Monitoring and Assessment. 2013; 185:7485–7499. DOI: 10.1007/s10661-013-3114-6 [PubMed: 23435849]

Withers P, Jarvie H. Delivery and Cycling of Phosphorus in Rivers: A Review. Science of the Total Environment. 2008; 400:379–395. DOI: 10.1016/j.scitotenv.2008.08.002 [PubMed: 18804845]

Zampella RA. Characterization of Surface-Water Quality Along a Watershed Disturbance Gradient. Water Resources Bulletin. 1994; 30:605–611. DOI: 10.1111/j.1752-1688.1994.tb03315.x

Zampella RA, Procopio NA, Lathrop RG, Dow CL. Relationship of Land-Use/Land-Cover Patterns and Surface-Water Quality in the Mullica River Basin. Journal of the American Water Resources Association. 2007; 43:594–604. DOI: 10.1111/j.1752-1688.2007.00045.x

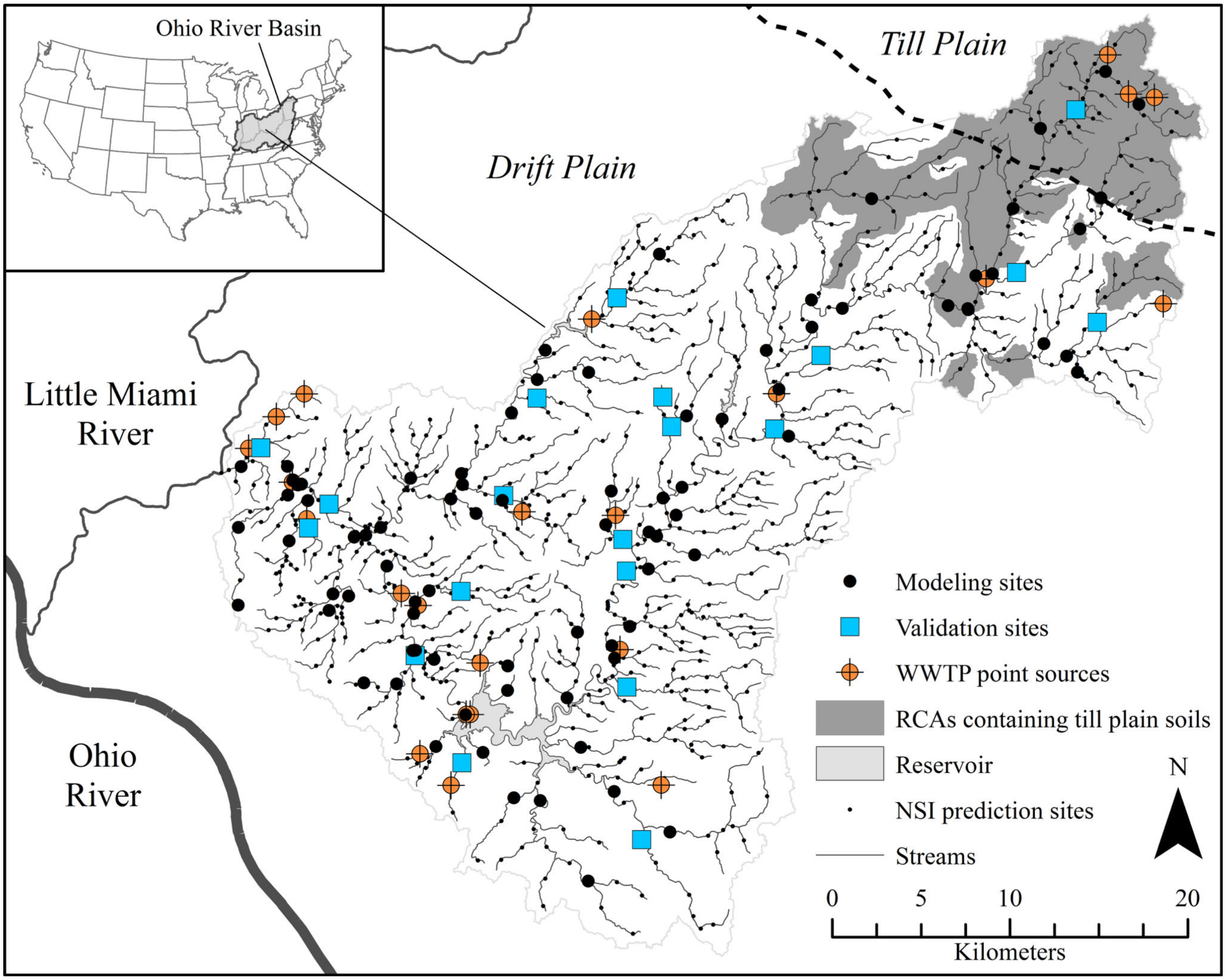

FIGURE 1.

Map of the East Fork Watershed. Regional location and detail is shown of: the stream network; location of modeling, validation, and prediction sites; the location of wastewater treatment plants (WWTP) with known phosphorus release loads; the two Level IV Ecoregions in the watershed—the till plain and drift plain (boundary indicated by dashed line); and the extent of Reach Contributing Areas (RCAs) that contain till plain soils in the upper part of the watershed. Note till plain soils occur in the till plain as well as in some valley bottoms in the upper drift plain.

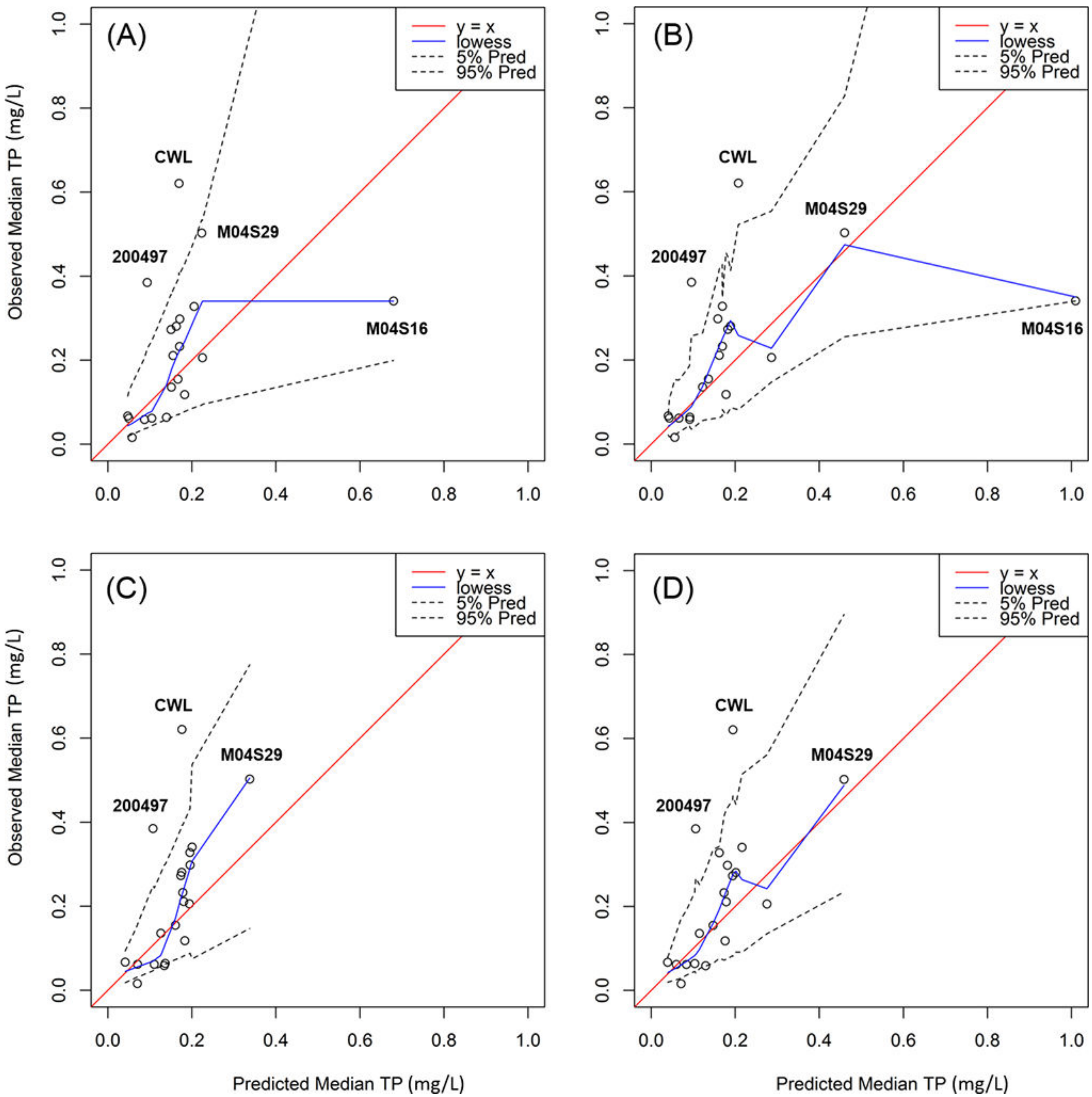

FIGURE 2.

Observed versus predicted median TP concentration at the 20 validations sites. (A) National nonspatial, (B) national spatial, (C) local nonspatial, and (D) local spatial models. Locally weighted smoothing (lowess) lines are shown, as well as the 5 and 95% prediction limits. Sites with extreme observed or predicted values are labeled for discussion.

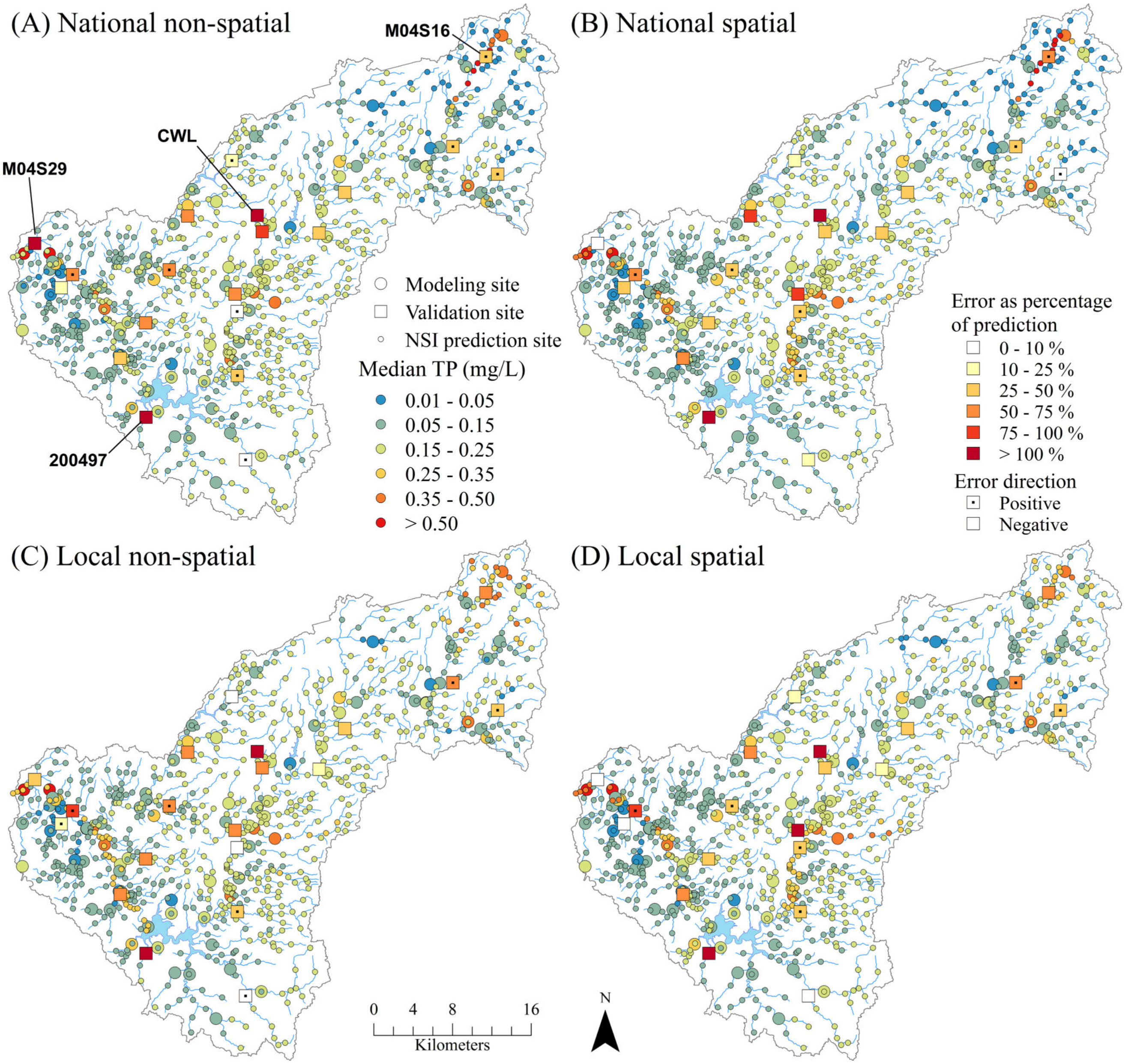

FIGURE 3.

Spatial predictions of median TP concentration. Predictions are shown at 779 National Stream Internet (NSI) sites throughout the East Fork watershed based on the (A) national nonspatial model, (B) national spatial model, (C) local nonspatial model, and (D) local spatial model. Modeling sites are also shown with their observed median TP concentration values and validation sites are shown with their signed error expressed as a percentage of the prediction. High resolution figure supplied in S3.

TABLE 1

List of Covariates, Their Spatial Treatment, and Data Source. Areal covariates were considered either throughout the entire watershed upstream of a point or as a proportion of the local Reach Contributing Area (RCA) around that point (see text for details).

Variable type	Included in model	Covariate	Spatial Treatment	Data Source
Landscape	National and local	Watershed area (km^2)	Cumulative watershed	NED 10 m DEM[1]
		Tributary stream category (WS ≤ 100 km^2)	Cumulative watershed	NED 10 m DEM[1]
		Main stem stream category (WS > 100 km^2)	Cumulative watershed	NED 10 m DEM[1]
		Slope 0–2% area (km^2)	Cumulative watershed	NED 10 m DEM[1]
		Slope > 5% area (km^2)	Proportion of RCA	NED 10 m DEM[1]
		Avonburg soil area (km^2)	Cumulative watershed	SSURGO[1]
		Clermont soil area (km^2)	Cumulative watershed	SSURGO[1]
		Rossmoyne soil area (km^2)	Cumulative watershed	SSURGO[1]
		Till plain soils (Miamian, Russell, Xenia)	Presence/absence in RCA	SSURGO[1]
		Valley soils (Cincinnati, Edenton)	Presence/absence in RCA	SSURGO[1]
Land use	National and local	Agriculture (km^2)	Cumulative watershed	NLCD[1]
		Urban/developed land (km^2)	Proportion of RCA	NLCD[1]
		Deciduous forest (km^2)	Proportion of RCA	NLCD[1]
		Pasture (km^2)	Proportion of RCA	NLCD[1]
	Local	Area of septic systems (km^2)	Cumulative watershed	*See text*
		Density of septic system areas (km^2/km^2)	Density in watershed	*See text*
Point sources	National	NPDES permit address	Presence/absence in watershed	NPDES[2]
	Local	WWTP total P load released in 2012 (kg)	Accumulated downstream	*See text*
		WWTP average TP concentration (mg/L)	Accumulated downstream	*See text*

Abbreviations used in table: NED—National Elevation Dataset; DEM—Digital Elevation Model; WS—Watershed; SSURGO—Soil Survey Geographic Database; RCA—Reach Contributing Area; NLCD—National Land Cover Database; NPDES—National Pollutant Discharge Elimination System; WWTP—Wastewater Treatment Plant.

[1]Data obtained from https://gdg.sc.egov.usda.gov/.

[2]Data obtained from http://www.epa.gov/enviro/geospatial-data-download-service/.

TABLE 2

Modelling Results. Covariates, coefficients, and goodness of fit criteria for the multiple linear regression (nonspatial) and SSN (spatial) models for the national and local covariates.

	Covariate/coefficient/criteria	Nonspatial	Spatial
National coefficients	Watershed area	−1.096**	−1.255**
	Slope 0–2% area	0.979[1]	1.212[1]
	Clermont soil area	0.278[1]	0.173[1]
	Presence of till plain soils	−2.729**	−3.453***
	Slope 0–2% area:Till plain soils	1.936***	2.284***
	Clermont soil:Till plain soils	−1.802***	−1.985***
	Intercept	−1.783***	−1.707***
National AIC	AIC	144.61	135.88
National variance components	Covariates	0.520	0.429
	Autocovariance	n/a	0.432
	Nugget	0.480	0.139
Local coefficients	Agricultural area	0.387***	0.374***
	Rossmoyne soil area	−0.591***	−0.583***
	Presence of till plain soils	1.226[1]	1.226[1]
	Cumulative septic area	1.376**	1.349*
	WWTP TP load	−0.007[1]	0.003[1]
	Agriculture:Till plain soils	−1.187*	−0.990*
	Septic:Till plain soils	9.667[2]	8.667[2]
	WWTP TP load:Till plain soils	0.273*	0.283*
	Intercept	−1.926***	−1.932***
Local AIC	AIC	134.98	133.76
Local variance components	Covariates	0.595	0.491
	Autocovariance	n/a	0.295
	Nugget	0.405	0.214

*Coefficient p-value < 0.05.

**Coefficient p-value < 0.01.

***Coefficient p-value < 0.001.

[1]Added because of significant interaction.

[2]Selected by BSR but not by Wald test.

EPA Author Manuscript

EPA Author Manuscript

TABLE 3

Root mean square percent prediction error (RMSPPE) and the width of the 90% prediction interval as a percentage of the prediction averaged among the 20 validation sites for the four models.

Model	RMSPPE	Average (90% P.I. / prediction) × 100%
National nonspatial	106%	203%
National spatial	94%	192%
Local nonspatial	93%	180%
Local spatial	89%	184%