EVALUATION OF A QUANTITATIVE PHOSPHORUS TRANSPORT MODEL FOR POTENTIAL IMPROVEMENT OF SOUTHERN PHOSPHORUS INDICES

by

THOMAS ADAM FORSBERG

(Under the Direction of David E. Radcliffe)

ABSTRACT

Management of agricultural nonpoint source phosphorus (P) requires identification of fields susceptible to P loss. P-Indices are the most common tools used to identify critical source areas of P loss. However, the success of the P-Index is impeded by insufficient testing against measured P loss data. Due to a shortage of available P loss data sets, simulated data from a quantitative P transport model may be used to test against a P-Index. The objective of this study was to compare predictions from the Texas Best Management Evaluation Tool (TBET) against measured P loss data to determine whether the model can improve P-Indices in the South. Measured P loss data, representing a range of conditions, were used to test TBET on an event-basis. Our results suggest that TBET can generate satisfactory quantitative predictions of runoff, sediment and P loss with site-specific calibration, but may be better suited for parameterization by physiographic region or state.

INDEX WORDS: TBET, Texas Best Management Evaluation Tool, Phosphorus, Phosphorus Index, Modeling

EVALUATION OF A QUANTITATIVE PHOSPHORUS TRANSPORT MODEL FOR POTENTIAL IMPROVEMENT OF SOUTHERN PHOSPHORUS INDICES

By

THOMAS ADAM FORSBERG

B.S., The University of Georgia, 2010

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2015

EVALUATION OF A QUANTITATIVE PHOSPHORUS TRANSPORT MODEL FOR POTENTIAL IMPROVEMENT OF SOUTHERN PHOSPHORUS INDICES

by

THOMAS ADAM FORSBERG

Major Professor:	David Radcliffe
Committee:	John Dowd
	Adam Milewski

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2015

ACKNOWLEDGEMENTS

I wish to thank the United States Department of Agriculture-Natural Resource Conservation Service for funding this work. I thank David Radcliffe for the opportunity to work on this project, his guidance and friendship throughout. I would also like to thank my committee members John Dowd and Adam Milewski for their support, friendship, and helpful suggestions. I thank Deanna Osmond, Carl Bolster, Dan Storm, and Mike White for their expertise and support. I also thank the departments of Geology and Crop and Soil Sciences as well as the many of the water resources faculty at the University of Georgia for providing an excellent learning experience. Finally, I would like to thank my family and friends who have supported my efforts wholeheartedly.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

The United States Environmental Protection Agency (USEPA) recognizes agricultural nonpoint source pollution as the primary source of stream and lake contamination. Excess nutrients, especially phosphorus (P), in agricultural runoff overstimulate the growth of weeds and algae, a process known as eutrophication (USEPA, 2009). Eutrophication can have profound ecological and societal effects including decreased biodiversity, changes in species composition and dominance, reduction in aesthetics, and impairment to commercial fisheries (Correll, 1998). As a result, the United States Department of Agriculture (USDA) and the USEPA require each state to adopt and implement a management strategy aimed at controlling nonpoint P loading to public waters (Osmond et al., 2006).

The most common strategy for assessing a field's vulnerability to P loss is the P-Index, developed by Lemunyon and Gilbert (1993). A P-index is a qualitative assessment tool used to identify agricultural sites most susceptible to P loss by accounting for and ranking the major source areas and flow pathways controlling P transport (Sharpley, 1995). Environmental P thresholds based on crop requirements or water quality criteria are also commonly used. However, these strategies can be logistically constraining, so most states only use the P-index or a combination of the P-index and P thresholds. Currently, 48 states use the P-index either voluntarily or as mandated by the Natural Resources Conservation Service (NRCS) 590 Nutrient Management Standard (Osmond et al., 2006; Sharpley et al., 2003b).

In spite of widespread adoption, the scientific basis behind the P-index and variations in implementation among states present a threat to the success of the P-index in controlling P runoff (Heathwaite et al., 2005; Sharpley et al., 2013). Due to the inherent unique combination of soils,

land use and hydrology as well as differences in developmental goals chosen by each state, a significant variety of P-index ratings and P application rates exist for similar management conditions. More importantly, few P-indices have been evaluated against measured P loss data (Osmond et al., 2006). The lack of proper development and evaluation of P-Indices is due to state funding constraints and an absence of available field data for testing (Bolster et al., 2012; Heathwaite et al., 2005).

In 2009, the Southern Extension-Research Activity Group 17 (SERA-17), established to review and revise the 590 Nutrient Management Conservation Standard, advised using measured P loss data and predicted P loss from models to evaluate and improve P-Indices (Sharpley et al., 2011). This study evaluated the quantitative performance of the Texas Best Management Practice Evaluation Tool (TBET) against event-based measured P loss data to determine whether the model could be used to improve P-indices for states in the southern region of the United States. TBET is a process-based, daily time step model that applies the Soil and Water Assessment Tool (SWAT) at field scale. TBET was initially developed as a qualitative P loss model for TX and OK to facilitate evaluation of agricultural nonpoint source load reductions from conservation practice implementation (White et al., 2012).

LITERATURE REVIEW

Water quality assessments by federal and state agencies continue to show that nonpoint source pollution is the primary source of impairment to surface waters of the U.S. An impaired waterbody is defined as having chronic or recurring violations of applicable numeric and/or narrative water quality criteria. Impaired waterbodies are listed under Section 303(d) of the 1972 Clean Water Act. Agricultural activities, such as crop production, grazing, and feeding operations impact almost 40% of assessed impaired river and stream miles and about 15% of impaired lakes, ponds and reservoirs. Nonpoint source pollutants associated with agriculture include nutrients, sediment, pesticides and salts. Excess nutrient loading alone is the third leading cause of

impairment to lakes and ponds and the fifth leading cause of impairment to rivers and streams (USEPA, 2009). Nitrogen (N) and phosphorus (P) are the two major nutrients associated with agriculture nonpoint source pollution. These constituents are applied to agricultural land from various sources, however, most commonly through commercial fertilizers and manure. N and P are naturally present in freshwater aquatic environments at mean concentrations of 0.3 to 0.01 mg L^{-1}, respectively. However, excessive loading of these nutrients to waterbodies can radically alter plant productivity and cause severe changes to biodiversity, water quality, and recreational activities. This process is known as eutrophication (USEPA, 2003).

Phosphorus is a key contributor to eutrophication of waterbodies from the input of P in agricultural runoff (Reed-Andersen et al., 2000; Sharpley et al., 2003a). Soil solution P concentrations necessary for adequate plant growth range from 0.2 to 0.3 mg L^{-1}, whereas P concentrations greater than or equal to 0.02 mg L^{-1} can accelerate eutrophication. Thus, soil P levels have increased to levels above crop needs due to ongoing application of fertilizer and manure causing local and regional scale disparities between P inputs and P losses in runoff (Sharpley et al., 2003a).

The main point source of P to a waterbody is municipal and industrial treatment plant effluent. Diffuse, or nonpoint, sources of include commercial fertilizer, animal manure, herbicide, insecticides and naturally occurring P in soils (Sharpley et al., 2003a). More recently, research has highlighted the importance of the fraction of point and nonpoint source P that accumulates in soils, sediments, and biomass. Accumulated P, also known as *legacy P*, is transitionally stored then intermittently remobilized along P transport pathways, thus functioning as a continuous source of P over time (Sharpley et al., 2013). Point sources of P have been effectively reduced in many areas, whereas nonpoint source inputs and legacy P continue to dominate P water quality issues (Jarvie et al., 2013; Sharpley et al., 2003a).

Phosphorus loading from agricultural runoff is a function of interactions among areas on the landscape with high P concentrations with areas of high P transport potential. High-P soils and

locations with concentrated amounts of fertilizer or manure make up P sources. Areas in the landscape prone to surface runoff generation have high P transport potential. A location in which P source areas and transport pathways overlap is known as a critical source area (CSA) (Heathwaite and Johnes, 1996; Pionke et al., 1996). The importance of CSAs on the management of P loss is emphasized by the understanding that greater than 80% of agricultural P lost in runoff originates from less than 20% of the catchment area (Pionke et al., 2000). Furthermore, successful management of P loss from agricultural watersheds requires an understanding of how management of CSAs at the field scale translates P concentrations at the catchment outlet.

The form and concentration of P in a soil are primary factors of P source areas. Soil P is characterized by three main pools: soluble P, reactive P, and stable P. The soluble P pool is primarily composed of orthophosphate ions (ie. $H_2PO_4^-$, HPO_4^{2-}) and provides the most plant available P. Phosphorus in the reactive and stable pools is attached to the soil solid phase and exists concurrently in organic and inorganic forms. The reactive P pool is in dynamic equilibrium with the soluble P pool. As P in the soluble pool is taken up by plants or lost through hydrological processes, replacement of P from the reactive P pool occurs through various soil physiochemical processes. Stable P constitutes the largest portion of soil P and is insoluble, sorbed or occluded, thus not readily available for biological uptake. Reactions with P forms in the stable pool are slow relative to the length of the growing season (Hansen et al., 2002). Therefore, the variable residence and travel times occurring among different P pools is directly translated to P availability (Sharpley et al., 2013).

Phosphorus stored in soils is transported in runoff from the landscape as dissolved P (DP) or particulate P (PP). Particulate P is associated with P bound to soil particles and is made available via erosion—a function of soil type, management history, and the duration and intensity of rainfall events. Physiochemical processes such as sorption, dissolution and diffusion enable movement of DP. The concentration of DP in runoff is a product of the quantity and reactivity of P within 1 to 5 cm of soil depth, and concentrations of DP in runoff from a field are positively correlated with soil

test phosphorus (STP) (Hansen et al., 2002; Pote et al., 1996). PP is not directly related to STP due to its connection to eroded soil particles (Sharpley et al., 1993).

In general, erosion and either saturation excess or infiltration excess overland flow (McDonnell, 2009) are the primary pathways for soil P loss to nearby waterbodies (Sharpley et al., 2002). However, soils lacking substantial subsoil P fixing capacity such as sandy soils, soils with low cation exchange capacity, or soils exhibiting rapid macropore flow also have potential for subsurface drainage of soil P (Hansen et al., 2002; Laubel et al., 1999). In addition, farms containing tile drainage systems are particularly at risk of direct applied P loss through preferential flow (Simard et al., 2000).

Despite the understanding of CSA controls on P loss, questions regarding the effectiveness of source and transport remedial strategies remain. This is especially apparent for P loss at the watershed scale due to field level interactions, rainfall-runoff dynamics, in-stream P processing, and contributions from more indefinite rural and urban sources, such as field-scale runoff and septic systems. Therefore, much of the current research aimed at reducing P loss is focused on management of P at the farm or field scale in hopes of reducing P loss to waterbodies at the watershed scale (Sharpley et al., 2009).

Nutrient management policy in the U.S. is shaped by the USDA—Natural Resource Conservation Service (NRCS) 590 Nutrient Management Conservation Standard (Code 590). This policy originates from a joint strategy between the USEPA and USDA to implement Comprehensive Nutrient Management Plans on Animal Feeding Operations. NRCS Code 590 requires all states to mitigate nonpoint source P losses from fields receiving manure with one of three accepted methods: 1) a soil test P threshold based on crop requirements; 2) a soil test P threshold using water quality criteria; 3) a P-index to identify fields with a high potential of P loss in runoff (Osmond et al., 2006; Sharpley et al., 2008).

Currently, 48 states use the P-index to identify CSAs and target BMPs to reduce P loss (Osmond et al., 2006). Lemunyon and Gilbert (1993) developed the P-index due to economic and

logistic limitations as well as the poor representation of P loss transport mechanisms associated with environmental P thresholds (Sharpley et al., 1993). P-Indices are calculated using input data such as soil test P, soil erosion, distance to stream, and fertilization and application rates. They are intended to identify CSAs in a field and assign a unit-less P loss risk ranking of "low", "medium", "high", and "very high" (Sharpley et al., 1993).

Each state is permitted to design their own Code 590 Standard and P-index due to unique local climate, soil, land use, topographic, and hydrological conditions. Consequently, variations in design and implementation of the P-indices among states has caused a diversity of P-index ratings and permissible P application rates for similar site conditions (Osmond et al., 2006). Moreover, most P-based nonpoint source conservation programs and watershed-scale BMPs have failed to produce water quality improvements (Jarvie et al., 2013). The failure of such measures is twofold. First, the elapsed time between full operation of an installed BMP and measurable detection of water quality recovery, known as lag time, has been slower than predicted (Meals et al., 2010). Second, and more importantly, current conservation practices are inadequately designed to manage sinks and stores of legacy P deposited by past land management practices (Sharpley et al., 2013). Overall P-index success, along with alternative nonpoint mitigation practices, is primarily limited by the issue of legacy P and lag time between treatment and response. This is compounded by the fact that few P-indices have been validated against measured P loss data. Thus, due to widespread adoption of the P-index for compliance with USDA-NCRS Code 590, a comprehensive evaluation of P-indices is critical to the success of nutrient management policy (Osmond et al., 2006).

In response to the flaws of the P-index, USDA-NRCS solicited scientific advisement from the Southern Extension-Research Activity Group 17 (SERA-17) for the purpose of reviewing and revising NRCS Code 590. SERA-17 was specifically tasked with designing a process for improving the P-index. However, a key reason the P-index has not undergone sufficient evaluation is due an absence of readily available P loss data. In addition, a P-Index should be evaluated against measured data under a variety of management scenarios and against runoff data that was collected

at the field edge (Sharpley et al., 2012). Studies comparing edge-of-field P loss data against P-Index ratings are few in number (DeLaune et al., 2004a, b; Eghball and Gilley, 2001; Good et al., 2012; Harmel et al., 2005; Sharpley et al., 2001; Sonmez et al., 2009). Moreover, there are relatively few edge-of-field P loss datasets currently available (Sharpley et al., 2012).

If observed P loss data are unavailable for calibration of a P-Index, a validated process-based model may be used to generate alternative P loss data. It is not necessary for a P-Index to quantify P loss, but its output must be correlated with output from process-based P loss model. Accordingly, SERA-17 advocated using simulated output from P transport models as a substitute for unavailable measured P loss data, also known as the meta-model approach (Schoumans et al., 2002; Sharpley et al., 2012). Under the meta-model approach, quantitative P loss model-generated data would be used to formulate and derive P-index factors. Thus, transforming the P-index into an approximation of the more complex "mother" model. Due to the inherent link between the meta-model and the mother model, the meta-model is only valid for the geographic domain of the mother model. In addition, any uncertainty or incorrect results tied to the mother model are directly transferred to the meta-model (Radcliffe et al., 2009).

The meta-model approach has been successfully applied in multiple studies. Some have calibrated the meta-model with field measurements from runoff studies (DeLaune et al., 2004b; Harmel et al., 2005; Sonmez et al., 2009), while others have used model-generated data (Bolster et al., 2012; Leone et al., 2008; Schoumans et al., 2002; Veith et al., 2005).

Bolster (2011) compared simulated P loss data from APLE against output from the Kentucky P-index. The results of this analysis showed a significant correlation between P loss predicted with the APLE model and the KY P-index. However, the authors noted several limitations of the P-index. First, they noted that the KY P-index lacked soil erosion and P application rates, which fail criteria mandated by NRCS Code 590. More importantly, they found that the weighting factors used in the KY P-index were not empirically based but were the result of best professional judgment.

Similarly, Bolster et al. (2012) applied APLE to inform the Pennsylvania P-index. They demonstrated that APLE simulations could successfully derive more accurate P-index weighting factors, and noted that correlating P-index ratings with quantitative P loss model output can provide valuable estimates of uncertainty in the P-index.

State P-Indices must be assessed and revised to provide consistent recommendations among states, basic guidelines for tool development, and test whether P-Index predictions are correct in direction and magnitude (Sharpley et al., 2012). Given the scarcity of sufficient water quality data, calibrated and validated P loss models are essential for providing P loss data under various climatic and management scenarios. Furthermore, adequate evaluation of a P-Index should include assessment of the P-Index against different model types (Vadas et al., 2013). Several models such as the Annualized Agricultural Nonpoint Source model (AnnAGNPS) (Needham and Young, 1993), the Agricultural Policy/Environmental Extender (APEX) (Williams, 1990), the Soil and Water Assessment Tool (SWAT) (Arnold et al., 1998), the Environmental Policy Integrated Climate model (EPIC) (Williams et al., 1983) and the Annual Phosphorus Loss Estimator (APLE) (Vadas et al., 2009), are capable of simulating P loss at the field to small-catchment scale.

P loss models are capable tools for guiding land management decisions and testing hypotheses with respect to water quality given a properly developed model framework and estimates of measurement and model uncertainty (Bolster and Vadas, 2013; Radcliffe et al., 2009).

Each P model differs in its model development framework. Model development includes the perceptual, conceptual and procedural models, as well as the calibration and validation of the model. The perceptual model is an illustrative interpretation of how a system function, the conceptual model describes the perceptual model through an array of mathematical equations, and the procedural model is how the conceptual model equations are rendered and solved in computer code (Vadas et al., 2013). Calibration is the systematic adjustment of model parameters to find the best fit between simulated and a particular set of measured data, and validation is the application of the calibration parameter set on a separate set of measured data than used in calibration.

Error in a model development framework may occur due to many factors and can subsequently impact the validation of a P model. For example, a P model may be perceptually correct based on the understanding of CSA factors controlling P loss, but may be conceptually incorrect due to the many equations used to define CSA factors (Vadas et al., 2013). Conceptual models used in the development and application of P-Indices are often perceptually correct, but vary among states and therefore produce different predictions of P loss when using the same inputs (Osmond et al., 2006).

In addition to discrepancies between perceptual and conceptual models, a procedural model may be applied according to a different translation of a single conceptual model. For example, the Jones et al. (1984) soil P model is used in AnnAGNPS (Yuan et al., 2005), EPIC (Williams et al., 1983) and SWAT (Arnold et al., 1998) is reliant on the P sorption parameter (PSP). AnnAGNPS computes a daily PSP value, whereas EPIC and SWAT use a constant PSP value for an entire simulation. Due to the daily calculation of PSP in AnnAGNPS and the union of PSP to the labile P pool, the labile P fraction of the AnnAGNPS P model rapidly increases over a simulation period. As a result, differences in the P routines between AnnAGNPS, EPIC and SWAT yield varying predictions of P (Vadas et al., 2013).

Effective P loss models must also provide measurements of uncertainty. Both modelers and decision-makers benefit from uncertainty estimates. Uncertainty facilitates the assessment of confidence in measured and simulated values, which can have a profound impact on analysis, communication and judgment of model performance and its intended use (Harmel et al., 2010; Radcliffe et al., 2009). All P transport models have sources of error including structure error, input error, and parameter error (Bolster and Vadas, 2013). Uncertainty is also inherent in all measured water quality data due to error introduced during streamflow measurement, sample collection, sample preservation, and laboratory analysis. Data management is also a source of error in measured data due to missing data, the assumptions made to approximate missing values, and unit conversion or calculation errors (Harmel et al., 2006). Ignoring uncertainty in P loss modeling

efforts can provide a false sense of accuracy in model results. Thus, reporting model results and the uncertainty associated with those predictions is imperative in communicating the performance of a model to the public, regulators and the scientific community (Bolster and Vadas, 2013).

Successful management of P in the environment requires effective identification of agricultural fields most susceptible to P loss. P-Indices are the most common tools used by states to identify CSAs and select BMPs to reduce P loss, but their success is impeded by several factors. First, states are allowed to develop their own P-indices due to location-specific climate, soil, land use, and hydrological conditions, thus ratings and allowable P applications rates vary widely for similar site conditions. Second, P-index based management is not producing substantial decreases in soil P or runoff P concentrations mainly due to legacy P and the failure to limit P application. Lastly, P-indices have not been properly tested against measured P loss data due to a lack of available resources to calibrate and validate P-indices at the state level. The SERA-17 Working Group, tasked with revising NRCS Code 590, recommends using measured P loss data and, due to the lack of readily available water quality data sets, simulated P loss data from P transport models to evaluate P-Indices. State P-Indices should be tested relative to different model types, such as APEX, APLE, EPIC, SWAT etc., due to differences in model development frameworks among models. P transport model should also include estimates of measurement and model uncertainty associated with measurement and simulation of P loss.

The primary objective of this study was to assess the accuracy of the Texas Best Management Evaluation Tool (TBET), a field-scale version of SWAT, in predicting field-scale P loss, so that it might be used as a transport model for comparison with P-indices. Pre-existing edge-of-field water quality data from three field sites in the southern region of the U.S. representing a range of soil, crop systems, nutrient application rates, and management practices were used to test TBET. The evaluation of TBET included assessments of the baseline, calibration, and validation performance of the model on an event basis. The validation predictions

were accompanied by estimates of low and high uncertainty scenarios associated with typical sources of error present in water quality measurements and P transport models.

REFERENCES

Arnold, J. G., Srinivasan, R., Muttiah, R. S., and Williams, J. R., 1998, Large area hydrologic modeling and assessment - Part 1: Model development: Journal of the American Water Resources Association, v. 34, no. 1, p. 73-89.

Bolster, C., and Vadas, P., 2013, Sensitivity and uncertainty analysis for the Annual Phosphorus Loss Estimator Model: Journal of Environment Quality, v. 42, p. 1109-1118.

Bolster, C. H., Vadas, P. A., Sharpley, A. N., and Lory, J. A., 2012, Using a phosphorus loss model to evaluate and improve Phosphorus Indices: Journal of Environment Quality, v. 41, no. 6, p. 1758-1766.

Correll, D. L., 1998, The role of Phosphorus in the eutrophication of receiving waters: A review: Journal of Environment Quality, v. 27, no. 2, p. 261-266.

DeLaune, P. B., Moore, P. A., Carman, D. K., Sharpley, A. N., Haggard, B. E., and Daniel, T. C., 2004a, Development of a phosphorus index for pastures fertilized with poultry litter - Factors affecting phosphorus runoff: Journal of Environmental Quality, v. 33, no. 6, p. 2183-2191.

-, 2004b, Evaluation of the phosphorus source component in the phosphorus index for pastures: Journal of Environmental Quality, v. 33, no. 6, p. 2192-2200.

Eghball, B., and Gilley, J. E., 2001, Phosphorus risk assessment index evaluation using runoff measurements: Journal of Soil and Water Conservation, v. 56, no. 3, p. 202-206.

Good, L. W., Vadas, P., Panuska, J. C., Bonilla, C. A., and Jokela, W. E., 2012, Testing the Wisconsin Phosphorus Index with year-round, field-scale runoff monitoring: Journal of Environmental Quality, v. 41, no. 6, p. 1730-1740.

Hansen, N. C., Daniel, T. C., Sharpley, A. N., and Lemunyon, J. L., 2002, The fate and transport of phosphorus in agricultural systems: Journal of Soil and Water Conservation, v. 57, no. 6, p. 408-417.

Harmel, R. D., Cooper, R. J., Slade, R. M., Haney, R. L., and Arnold, J. G., 2006, Cumulative uncertainty in measured streamflow and water quality data for small watersheds: Transactions of the ASABE, v. 49, no. 3, p. 689-701.

Harmel, R. D., Smith, P. K., and Migliaccio, K. W., 2010, Modifying Goodness-of-Fit Indicators to Incorporate Both Measurement and Model Uncertainty in Model Calibration and Validation: Transactions of the ASABE, v. 53, no. 1, p. 55-63.

Harmel, R. D., Torbert, H. A., Delaune, P. B., Haggard, B. E., and Haney, R. L., 2005, Field evaluation of three phosphorus indices on new application sites in Texas: Journal of Soil and Water Conservation, v. 60, no. 1, p. 29-42.

Heathwaite, A. L., and Johnes, P. J., 1996, Contribution of nitrogen species and phosphorus fractions to stream water quality in agricultural catchments: Hydrological Processes, v. 10, no. 7, p. 971-983.

Heathwaite, A. L., Quinn, P. F., and Hewett, C. J. M., 2005, Modelling and managing critical source areas of diffuse pollution from agricultural land using flow connectivity simulation: Journal of Hydrology, v. 304, no. 1–4, p. 446-461.

Jarvie, H. P., Sharpley, A. N., Withers, P. J. A., Scott, J. T., Haggard, B. E., and Neal, C., 2013, Phosphorus mitigation to control river eutrophication: murky waters, inconvenient truths, and "postnormal" science: Journal of Environmental Quality, v. 42, no. 2, p. 295-304.

Jones, C. A., Williams, J. R., Sharpley, A. N., and Cole, C. V., 1984, A simplified soil and plant phosphorus model: Journal - Soil Science Society of America, v. 48, no. 4, p. 800-805.

Laubel, A., Jacobsen, O. H., Kronvang, B., Grant, R., and Andersen, H. E., 1999, Subsurface drainage loss of particles and phosphorus from field plot experiments and a tile-drained catchment: Journal of Environmental Quality, v. 28, no. 2, p. 576-584.

Lemunyon, J. L., and Gilbert, R. G., 1993, The concept and need for a phosphorus assessment-tool: Journal of Production Agriculture, v. 6, no. 4, p. 483-486.

Leone, A., Ripa, M. N., Boccia, L., and Lo Porto, A., 2008, Phosphorus export from agricultural land: a simple approach: Biosystems Engineering, v. 101, no. 2, p. 270-280.

McDonnell, J. J., 2009, Hewlett, J.D. and Hibbert, A.R. 1967: Factors affecting the response of small watersheds to precipitation in humid areas. In Sopper, W.E. and Lull, H.W: Progress in Physical Geography, v. 33, no. 2, p. 288-293.

Meals, D. W., Dressing, S. A., and Davenport, T. E., 2010, Lag time in water quality response to best management practices: a review: Journal of Environmental Quality, v. 39, no. 1, p. 85-96.

Needham, S. E., and Young, R. A., 1993, ANN-AGNPS; a continuous simulation watershed model: U. S. Geological Survey : [Reston, VA], United States, 0092332X.

Osmond, D. L., Cabrera, M. L., Feagley, S. E., Hardee, G. E., Mitchell, C. C., Moore, P. A., Mylavarapu, R. S., Oldham, J. L., Stevens, J. C., Thom, W. C., Walker, F., and Zhang, H., 2006, Comparing ratings of the southern phosphorus indices: Journal of Soil and Water Conservation, v. 61, no. 6, p. 325-337.

Pionke, H. B., Gburek, W. J., and Sharpley, A. N., 2000, Critical source area controls on water quality in an agricultural watershed located in the Chesapeake basin: Ecological Engineering, v. 14, no. 4, p. 325-335.

Pionke, H. B., Gburek, W. J., Sharpley, A. N., and Schnabel, R. R., 1996, Flow and nutrient export patterns for an agricultural hill-land watershed: Water Resources Research, v. 32, no. 6, p. 1795-1804.

Pote, D. H., Edwards, D. R., Nichols, D. J., Moore, P. A., Jr., Daniel, T. C., and Sharpley, A. N., 1996, Relating extractable soil phosphorus to phosphorus losses in runoff. [electronic resource]: Soil Science Society of America. Soil Science Society of America journal, v. 60, no. 3, p. 855-859.

Radcliffe, D. E., Freer, J., and Schoumans, O. F., 2009, Diffuse Phosphorus models in the United States and Europe: Their usages, scales, and unceratainties: Journal of Environment Quality, v. 38, p. 1956-1967.

Reed-Andersen, T., Carpenter, S. R., and Lathrop, R. C., 2000, Phosphorus flow in a watershed-lake ecosystem, Springer-Verlag, p. 561.

Schoumans, O. F., Mol-Dijkstra, J., Akkermans, L. M. W., and Roest, C. W. J., 2002, SIMPLE: Assessment of non-point phosphorus pollution from agricultural land to surface waters by means of a new methodology: Water Science and Technology, v. 45, no. 9, p. 177-182.

Sharpley, A., 1995, Identifying sites vulnerable to Phosphorus loss in agricultural runoff: Journal of Environment Quality, v. 24, no. 5, p. 947-951.

Sharpley, A., Beegle, D., Bolster, C., Good, L., Joern, B., Ketterings, Q., Lory, J., Mikkelsen, R., Osmond, D., and Vadas, P., 2012, Phosphorus Indices: why we need to take stock of how we are doing: Journal of Environmental Quality, v. 41, no. 6, p. 1711-1719.

Sharpley, A., Jarvie, H. P., Buda, A., May, L., Spears, B., and Kleinman, P., 2013, Phosphorus legacy: Overcoming the effects of past management practices to mitigate future water quality impairment: Journal of Environment Quality, v. 42, no. 5, p. 1308-1326.

Sharpley, A. N., Daniel, T., Sims, T., Lemunyon, J., Stevens, R., and Parry, R., 2003a, Agricultural phosphorus and eutrophication - 2nd edition: University Park, PA, U.S. Dept. of Agriculture, Agricultural Research Service,.

Sharpley, A. N., Daniel, T. C., and Edwards, D. R., 1993, Phosphorus movement in the landscape: Journal of Production Agriculture, v. 6, no. 4, p. 492-500.

Sharpley, A. N., Kleinman, P. J. A., Heathwaite, A. L., Gburek, W. J., Weld, J. L., and Folmar, G. J., 2008, Integrating contributing areas and indexing phosphorus loss from agricultural watersheds: Journal of Environmental Quality, v. 37, no. 4, p. 1488-1496.

Sharpley, A. N., Kleinman, P. J. A., Jordan, P., Bergstrom, L., and Allen, A. L., 2009, Evaluating the Success of Phosphorus Management from Field to Watershed: Journal of Environmental Quality, v. 38, no. 5, p. 1981-1988.

Sharpley, A. N., Kleinman, P. J. A., McDowell, R. W., Gitau, M., and Bryant, R. B., 2002, Modeling phosphorus transport in agricultural watersheds: Processes and possibilities: Journal of Soil and Water Conservation, v. 57, no. 6, p. 425-439.

Sharpley, A. N., McDowell, R. W., Weld, J. L., and Kleinman, P. J. A., 2001, Assessing site vulnerability to phosphorus loss in an agricultural watershed: Journal of Environmental Quality, v. 30, no. 6, p. 2026-2036.

Sharpley, A. N., Weld, J. L., Beegle, D. B., Kleinman, P. J. A., Gburek, W. J., Moore, P. A., and Mullins, G., 2003b, Development of phosphorus indices for nutrient management planning strategies in the United States: Journal of Soil and Water Conservation, v. 58, no. 3, p. 137-152.

Simard, R. R., Beauchemin, S., and Haygarth, P. M., 2000, Potential for Preferential Pathways of Phosphorus Transport: Journal of Environment Quality, v. 29, no. 1, p. 97-105.

Sonmez, O., Pierzynski, G. M., Frees, L., Davis, B., Leikam, D., Sweeney, D. W., and Janssen, K. A., 2009, A field-based assessment tool for phosphorus losses in runoff in Kansas: Journal of Soil and Water Conservation, v. 64, no. 3, p. 212-222.

USEPA, 2003, National management measures for the control of nonpoint pollution from agriculture, *in* Water, O. o., ed.: Washington, D.C., United States Environmental Protection Agency, Office of Water.

-, 2009, National water quality inventory 2004 report.

Vadas, P., Good, L. W., Moore, P. A., and Widman, N., 2009, Estimating phosphorus loss in runoff from manure and fertilizer for a phosphorus loss quantification tool: Journal of Environment Quality, v. 38, no. 4, p. 1645-1653.

Vadas, P. A., Bolster, C. H., and Good, L. W., 2013, Critical evaluation of models used to study agricultural phosphorus and water quality: Soil Use and Management, v. 29, p. 36-44.

Veith, T. L., Sharpley, A. N., Weld, J. L., and Gburek, W. J., 2005, Comparison of measured and simulated phosphorus losses with indexed site vulnerability: Transactions of the ASAE, v. 48, no. 2, p. 557-565.

White, M. W., Harmel, R. D., and Haney, R. L., 2012, Development and validation of the Texas Best Management Practice Evaluation Tool (TBET): Journal of Soil and Water Conservation, v. 67, no. 6, p. 525-535.

Williams, J. R., 1990, The erosion-productivity impact calculator (EPIC) model: A case history.: Philosophical Transactions of the Royal Society of London: Biological Sciences, v. 329, no. 1255, p. 421-428.

Williams, J. R., Dyke, P. T., and Renard, K. G., 1983, EPIC: a new method for assessing erosion's effect on soil productivity: Journal of Soil and Water Conservation, v. 38, no. 5, p. 381-383.

Yuan, Y., Bingner, R. L., Theurer, F. D., Rebich, R. A., and Moore, P. A., 2005, Phosphorus component in AnnAGNPS: Transactions of the ASAE, v. 48, no. 6, p. 2145-2154.

CHAPTER 2

EVALUATION OF A QUANTITATIVE PHOSPHORUS TRANSPORT MODEL FOR POTENTIAL IMPROVEMENT OF SOUTHERN PHOSPHORUS INDICES[1]

[1] Forsberg, T.A., D.E. Radcliffe, C.H. Bolster, D.E. Storm, and D.L. Osmond. To be submitted to the Journal of Environmental Quality.

ABSTRACT

Management of agricultural nonpoint source phosphorus (P) requires identification of fields susceptible to P loss. P-indices are the most common tools used to identify critical source areas of P loss. However, the success of the P-index is impeded by insufficient testing against measured P loss data. Due to a shortage of available P loss data sets, simulated data from a quantitative P transport model may be used to test against a P-index. The objective of this study was to compare predictions from the Texas Best Management Evaluation Tool (TBET) against measured P loss data to determine whether the model can improve P-Indices in the South. Field-scale measured P loss data from study sites in Washington Co., AR, Putnam Co., GA, and Henderson Co., NC were used to assess quantitative performance of TBET. The Kendall-Theil robust line (KTR) method, Nash-Sutcliffe efficiency (NSE), percent bias (PBIAS), and mean absolute error (MAE) were used to evaluate model performance. In addition, the correction factor method was applied to estimate the effect of measurement and model uncertainty on TBET predictions. Our results suggest that TBET can generate satisfactory event-based predictions (NSE $\geq$ 0.3) of runoff, sediment and P loss with site-specific calibration. Lastly, we compared TBET against the P-index in predictions of average annual P loss. Goodness-of-fit between measured average annual total P loss and the calibrated TBET model (PBIAS = -6%) was stronger than the P-index (PBIAS = -78%). However, the significant variation in best-fit parameters, goodness-of-fit, and estimated uncertainties in TBET predictions among study sites suggested that TBET may not be appropriate for applying a regional set of parameters for the South, but may be more suited for application on a state-by-state or physiographic region basis.

INDEX WORDS: TBET, Texas Best Management Evaluation Tool, Phosphorus, Phosphorus Index, Modeling

INTRODUCTION

Water quality assessments by federal and state agencies continue to show that nonpoint source pollution is the primary source of impairment to surface waters of the U.S. Agricultural activities, such as crop production, grazing, and feeding operations impact almost 40% of assessed impaired river and stream miles and about 15% of impaired lakes, ponds and reservoirs. Excess nutrient loading alone is the third leading cause of impairment to lakes and ponds and the fifth leading cause of impairment to rivers and streams (USEPA, 2009). Phosphorus (P) is a major nutrient associated with agriculture nonpoint source pollution, and despite the understanding of the physical processes controlling P loss, questions regarding the effectiveness of source and transport remedial strategies remain. This is especially apparent for P loss at the watershed scale due to field level interactions, rainfall-runoff dynamics, in-stream P processing, and contributions from more indefinite rural and urban sources, such as field-scale runoff and septic systems. Therefore, much of the current research aimed at reducing P loss is focused on management at the farm or field scale in hopes of reducing P loss to waterbodies at the watershed scale (Sharpley et al., 2009).

The management policy in the U.S. for P and other nutrients is shaped by the U.S. Department of Agriculture (USDA)–Natural Resource Conservation Service (NRCS) 590 Nutrient Management Conservation Standard (Code 590). NRCS Code 590 requires all states to mitigate nonpoint source P losses from fields receiving manure with one of three accepted methods: 1) a soil test P threshold based on crop requirements; 2) a soil test P threshold using water quality criteria; 3) a P-index to identify fields with a high potential of P loss in runoff (Osmond et al., 2006; Sharpley et al., 2008).

Currently, 48 states use the P-index to identify critical source areas (CSAs) and target best management practices (BMPs) to reduce P loss (Osmond et al., 2006). Each state is permitted to design their own Code 590 Standard and P-index due to unique local climate, soil, land use, topographic, and hydrological conditions. Consequently, variations in design and implementation of the P-indices among states has caused a diversity of P-index ratings and permissible P application

rates for similar site conditions (Osmond et al., 2006). Moreover, most P-based nonpoint source conservation programs and watershed-scale BMPs have failed to produce water quality improvements (Jarvie et al., 2013). Overall P-index success, along with alternative nonpoint mitigation practices, is primarily limited by the issue of legacy P and lag time between treatment and response (Meals et al., 2010; Sharpley et al., 2013). This is compounded by the fact that few P-indices have been validated against measured P loss data. Thus, due to widespread adoption of the P-index for compliance with USDA-NCRS Code 590, a comprehensive evaluation of P-indices is critical to the success of nutrient management policy (Osmond et al., 2006).

In response to the flaws of the P-index, USDA-NRCS solicited scientific advisement from the Southern Extension-Research Activity Group 17 (SERA-17) for the purpose of reviewing and revising Code 590. SERA-17 was specifically tasked with designing a process for improving the P-index, but also advised that P-indices must be assessed and revised to provide consistent recommendations among states, basic guidelines for tool development, and test whether P-Index predictions are correct in direction and magnitude (Sharpley et al., 2012). A key reason the P-index has not undergone sufficient evaluation is due an absence of readily available edge-of-field P loss data, and studies comparing edge-of-field P loss data against P-index ratings are few in number (DeLaune et al., 2004a, b; Eghball and Gilley, 2001; Good et al., 2012; Harmel et al., 2005; Sharpley et al., 2001; Sonmez et al., 2009).

If observed P loss data are unavailable for calibration of a P-Index, a process-based P transport model may be used to generate alternative P loss data. Accordingly, SERA-17 advocated using simulated output from P transport models as a substitute for unavailable measured P loss data, also known as the meta-model approach (Schoumans et al., 2002; Sharpley et al., 2012). The meta-model approach has been successfully applied in multiple studies. Some have calibrated the meta-model with field measurements from runoff studies (DeLaune et al., 2004b; Harmel et al., 2005; Sonmez et al., 2009), while others have used model-generated data (Bolster et al., 2012; Leone et al., 2008; Schoumans et al., 2002; Veith et al., 2005). Bolster (2011) compared simulated

P loss data from APLE against output from the Kentucky P-index. The results of this analysis showed a significant correlation between P loss predicted with the APLE model and the KY P-index. However, the KY P-index lacked soil erosion and P application rates, which fail criteria mandated by Code 590. More importantly, he found that the weighting factors used in the KY P-index were not empirically based but were the result of best professional judgment. Similarly, Bolster et al. (2012) applied APLE to inform the Pennsylvania P-index. They demonstrated that APLE simulations could successfully derive more accurate P-index weighting factors, and noted that correlating P-index ratings with quantitative P loss model output can provide valuable estimates of uncertainty in the P-index.

Effective P loss models must also provide measurements of uncertainty. Both modelers and decision-makers benefit from uncertainty estimates. Uncertainty facilitates the assessment of confidence in measured and simulated values, which can have a profound impact on analysis, communication and judgment of model performance and its intended use (Harmel et al., 2010; Radcliffe et al., 2009a). All P transport models have sources of error including structure error, input error, and parameter error (Bolster and Vadas, 2013). Uncertainty is also inherent in all measured water quality data due to error introduced during streamflow measurement, sample collection, sample preservation, and laboratory analysis. Data management is also a source of error in measured data due to missing data, the assumptions made to approximate missing values, and unit conversion or calculation errors (Harmel et al., 2006). Ignoring uncertainty in P loss modeling efforts can provide a false sense of accuracy in model results. Thus, reporting model results and the uncertainty associated with those predictions is imperative in communicating the performance of a model to the public, regulators and the scientific community (Bolster and Vadas, 2013).

The objective of this study was to evaluate the quantitative performance of the Texas Best Management Practice Evaluation Tool (TBET) against event-based measured P loss data to determine whether the model could be used to improve P-indices for states in the southern region of the United States. TBET is a process-based, daily time step model that applies the Soil and Water

Assessment Tool (SWAT) at field scale. TBET was initially developed as a qualitative P loss model for TX and OK to facilitate evaluation of agricultural nonpoint source load reductions from conservation practice implementation (Mittelstet et al., 2012; White et al., 2012b). The model has also been used to compare surface and subsurface P transport in alluvial floodplains (Mittelstet et al., 2011). Pre-existing edge-of-field water quality data from three field sites in the southern region of the U.S. representing a range of soil, crop systems, nutrient application rates, and management practices were used to test TBET. This evaluation of TBET included assessments of the uncalibrated, calibration, and validation performance of the model on an event basis. The validation predictions were accompanied by estimates of low and high uncertainty scenarios associated with typical sources of error present in water quality measurements and P transport models. Lastly, the validated TBET model was used to generate long-term average annual predictions of P loss for comparison with the southern P-indices.

METHODS

Study sites

The study sites were in a humid, temperate climate and average annual rainfall among the sites ranged from 80 cm to 157 cm (Table 2.1). Annual surface runoff from all fields ranged from 1.5 cm to 83 cm. Fields ranged in area from 0.017 to 0.8 hectares and slopes ranged from 2% to 8%. The maximum annual fertilizer/manure rates among all sites were 424 kg ha^{-1} N and 388 kg ha^{-1} P with Mehlich III STP ranging from 14 to 183 mg kg^{-1}.

Each study site consisted of a unique crop systems and soils (Table 2.2). The AR study measured surface runoff, sediment yield, and P load on seven pasture fields from 2009 to 2011, totaling 21 field-years of data, the GA study measured surface runoff and P load on six pasture fields from 1995 to 1998, totaling 24 field-years of data, and the NC study measured surface runoff, sediment yield and P load on five corn with wheat cover fields from 2011 to 2013, totaling 15 field-

years of data. The supplemental information (Supplemental Fig. S1) contains the full site descriptions and sample collection methods.

Washington County, Arkansas

The AR study site (36°04'51.4"N, 94°17'23.0"W) was located on research fields managed by University of Arkansas in Washington County, AR. Seven 0.4-ha permanent pasture fields were monitored between spring 2009 and fall 2011. The only soil series present at the site was Captina (fine-silty, siliceous, active, mesic Typic Fragiudults). The site contained one control field. The other six fields were unique treatments: (1) continuous grazing with broadcast litter; (2) rotational grazing with broadcast litter; (3) hay production with broadcast litter; (4) replicate hay production with broadcast litter; (5) hay production with a single application of injected litter; and (6) hay with two applications of injected litter. Each treatment, except the two control treatments and HI2, received 537 kg ha^{-1} litter in mid-May of each year. The Mehlich III STP for all site years ranged from 73 to 150 mg kg^{-1}.

Putnam County, Georgia

These data were collected at the Central Georgia Branch Station in Putnam County, GA (33°25'05.4"N, 83°29'15.5"W) (Pierson et al., 2001a; Pierson et al., 2001b). Six 0.75-ha fescue-common bermudagrass paddocks were sampled between January 1995 and April 1998 for surface runoff and P loads. Sediment yield was not measured. Soil series at the site were Cecil (fine, kaolinitic, thermic Typic Kanhapludults), Altavista (fine-loamy, mixed, semiactive, thermic Aquic Hapludults), Helena (fine, mixed, semiactive, thermic Aquic Hapludults), and Sedgefield (fine. mixed, active, thermic Aquultic Hapludalfs). The plots had relatively low STP (14 to 142 ppm) and high rates of poultry litter application. Poultry litter was applied in March and September-October 1995 and 1996, while urea-ammonium nitrogen solution was applied in March 1997 and 1998. The

maximum total P added on each paddock was 159 kg ha^{-1}. In addition to fertilizer applications, each paddock received put and take stocking management.

Henderson County, North Carolina

The NC study site was located at the Mountain Horticulture Crop Research Station (35°25'36.4"N, 82°33'51.5"W) (Larsen et al., 2014). The soil series at the site was Delanco (fine-loamy, mixed, semiactive, mesic Aquic Hapludults). Surface runoff, TP, and DP were sampled from twenty 0.017-ha, corn-wheat rotation plots (four replications of five treatments) from January 2011 through December 2013. The five treatments include: (1) conventional tillage with pesticide management denoted by CTP; (2) conservation no-till with pesticide management (NTP); (3) conventional tillage with organic management (CTO); (4) conservation no-till with organic management (NTO); and (5) conventional tillage with no fertilizer or pesticide applications (Control). Mehlich III STP on all treatments ranged from 41 to 121 ppm. Organic treatments received 80 kg ha^{-1} P through commercial fertilizer, while conventional treatments received no additional P applications.

Data acquisition

The climate, soils, management schedules and measured field data from AR, GA, and NC were obtained through publications and from study authors. The entire dataset contained a combined total of 60 field-years of measured runoff, sediment, and P loss data. Each dataset was uniquely formatted due to the specific objectives and monitoring strategy of the respective studies. Measured daily weather data and rainfall depths associated with each sampled runoff event were provided for each site during the period of investigation.

The AR dataset lacked complete weather data for the period of investigation, thus existing weather data were supplemented with observations from three National Oceanic and Atmospheric Administration's National Centers for Environmental Information (NCEI) stations within 15 km of

the study site, as well as data from the National Centers for Environmental Prediction Climate Forecast System Reanalysis. NCEI precipitation and temperature data between the years 1993 and 1995 were used for model warm-up for the GA site. Daily weather data for NC between the years 2006 and 2013 was supplied through the State Climate Office of North Carolina CRONOS database.

Measured data amendments

Due to inherent sampling and laboratory analysis error present in measured water quality data and differences in precipitation data from public access databases and field measured precipitation several amendments were made to make the measured data comparable to the output of TBET. Precipitation data from public access databases were amended to reflect precipitation values measured at each study site. For example, if the field measured precipitation data reported 25 mm of rainfall on a specific date and the precipitation data accessed from a public database only displayed 15 mm for that same date the value was changed from 15 mm to 25 mm to reflect the most accurate precipitation measurement. Field measured runoff for a single event often extended beyond one day due to storm hydrograph lag time. Therefore, the daily runoff value from the TBET output for the entire event (up to three days) was combined for comparison to the measured runoff. Each measured storm event total was then matched with the corresponding daily, or multiple day, simulation values extracted from the model output.

Errors and missing data due to minimum detection limits, equipment malfunction, and laboratory analysis were prevalent in measured sediment and P data. Often, a dissolved P measurement was reported without an associated sediment and/or total P measurement or vice versa. Because TBET output distinguishes separate P pools, total P and dissolved P were compared against their respective output variables. For each event measured total P was compared against a sum of three P pools: P bound to sediment, organic P, and soluble P. Measured dissolved P was evaluated against the soluble P pool output. In addition to missing data, there were events where

the reported dissolved P measurement was greater than the reported total P measurement. In these cases, the total P value was modified to be equal to the dissolved P measurement due to the larger amount of error associated with analysis of total P concentrations.

TBET model description

TBET is an extensively simplified graphical user interface for a modified version of SWAT 2009. The model includes several procedural updates and modifications to account for local climate, soils, topography, and management in the south central U.S. The model is deterministic and semi-distributed using a mix of process based and empirical techniques to predict runoff, sediment yield, nutrient loads and crop yields. It is a continuous time model and runs on a daily time step. Unlike SWAT, which allows a subbasin to contain many unique hydrologic response units (HRUs), TBET only permits simulation of seven predefined HRUs. Thus, TBET does not require a Geographic Information System interface. The minimum inputs required to run TBET include field area, slope, distance to stream, soil type (maximum of three), daily weather (precipitation and temperature), and STP. The standard output produced by TBET includes annual totals of runoff, sediment, total N, total P, and dissolved P.

The predictions of P loss from TBET are governed by runoff volume, sediment load, P fertilizer, P cycling, and plant nutrient uptake. Runoff is predicted using the modified curve number and sediment load is calculated using the Modified Universal Soil Loss Equation (White et al., 2012b). The methodology used by SWAT, as well as TBET, to simulate P processes incorporates six different pools of P. Three pools are devoted to both organic and inorganic P. Soil inorganic P, or mineral P, is separated into soluble, reactive, and stable pools. The solution pool is in rapid equilibrium with the reactive pool, and the stable pool is in slow equilibrium with the reactive pool. Soil organic P is broken down into fresh, reactive, and stable pools. The reactive and stable P pools are associated with soil humic substances, while fresh organic P is associated with crop residue and microbial biomass. In TBET, the initial level of the solution pool is entered by the user as the

Mehlich III STP. The initial solution pool is then distributed to the active and stable P pools via the respective equilibrium ratio (White et al., 2010). Comprehensive descriptions of the subbasin processes and P model used in TBET can be found in White et al. (2010).

TBET includes databases for weather, soils, and crop systems for TX and OK. The soils database is derived from the Soil Survey Geographic (SSURGO) databases (USDA, 1991). The climate database is derived from the National Oceanic and Atmospheric Administration Cooperative Observer network and the Weather-Bureau-Army-Navy station. The TBET crop system database was developed using typical systems in the Texas State Soil and Water Conservation Board (TSSWCB) regions. User-defined options in the crop database include crop type, tillage, irrigation, grazing, stocking rate, and cover crop options. Values for the crop system options, typical operation schedule dates, and default fertilizer types and rates were predetermined to simplify user input requirements. However, fertilizer type and rates are required to be approved by the user and can be modified prior to simulation (White et al., 2012b).

Currently, the soils, climate and crop system databases for TBET mainly include information for the TSSWCB region. Thus, most soils, weather and management input files for this study were prepared manually by accessing and modifying the text input files in the TBET file directory.

Model evaluation process

White et al. (2012b) developed, calibrated and validated TBET for Texas and Oklahoma. The model performance results presented by White et al. (2012b) compared mean annual observed and predicted values. Validation percent bias (PBIAS) values for TBET predictions of mean annual runoff, sediment, and total P loss were -9.9, 3.9, and -14.7%, respectively. These statistics were acceptable, based on recommendations from Moriasi et al. (2007), for its use as a qualitative conservation practice and planning tool intended to replace a P-Index. This evaluation of TBET examined the event-based comparisons of measured and predicted values to determine if TBET

was an appropriate tool for the Southern region to generate simulated P loss data which could act as a substitute for measured P loss data needed to improve the P-Index.

The process of evaluating TBET as a quantitative P transport model included uncalibrated, calibration, and validation evaluation processes (Table 2.3). The validation evaluation also includes estimates of uncertainty. The data sets used in the calibration and validation data sets were chosen based on total annual rainfall. A site-specific manual calibration of the model was used to select the "best-fit" parameter set for each site, and validation was used to test the performance of the model after calibration.

Model simulations were run for a single calendar year and each run incorporated a two-year model warm up to initialize soil moisture and nutrient levels. For example, to run a field in GA for the year 1995 the simulation was started in 1993. The decision to use a two-year warm up period was determined by making ten-year, annual time step model runs and examining the model output to see in which year the output stabilized.

Statistical analyses

Model performance was evaluated by comparing measured runoff, sediment yield, and P load to event-based model predictions. As suggested by Coffey et al. (2004) and Moriasi et al. (2007), both graphical techniques and goodness-of-fit statistics were used to gauge model performance. Measured and predicted values for each field-year of data were combined by site and water quality constituent for statistical analyses in all steps of the evaluation process. For example, any statistic measuring the goodness-of-fit between measured and predicted runoff in AR includes events for all seven fields and three years of measured and predicted data on an event-basis.

Scatter plots and regression supplied an initial visual comparison of simulated and measured data and the 95% confidence interval of the regression slope provided a visual measure of uncertainty. Regression was used in conjunction with statistical measures of goodness-of-fit for proper model performance evaluation.

An analysis was conducted to assess whether the residuals of ordinary least squares regressions (OLS) between uncalibrated observed and predicted data obeyed the assumptions of normality and homoscedasticity. The Shapiro-Wilk test was used to assess whether residuals were normally distributed for a confidence level $\alpha = 0.05$. The null hypothesis of the Shapiro-Wilk test states that the data were normally distributed. Based on the results of the Shapiro-Wilk tests, transformations were performed on the data according to the "ladder of powers" (Helsel and Hirsch, 1992) to attempt to make the data more symmetric. Efforts to coerce the residuals of the OLS regressions to normality and homoscedasticity with transformations were unsuccessful. As a result, nonparametric regression in the form of the Kendall-Theil robust line method (KTR) was used.

The KTR was nearly as efficient as the OLS estimator when the assumptions of normality and homoscedasticity were satisfied, and was stronger when those assumptions were not satisfied. Moreover, the best use of the KTR was for a study where multiple constituents were tested at multiple sites and adequate data transformations differ for site or constituent (Helsel and Hirsch, 1992). The KTR slope estimate was calculated as the median slope of all possible pairwise slopes for each pair of points in the data set. The intercept was calculated as the median of all possible intercepts computed by solving the KTR using the median slope and each data point. The KTR estimates for this study were calculated using R package "mblm" version 0.12 (Komsta, 2013).

Three quantitative goodness-of-fit indicators in addition to the KTR graphical technique were used: Nash-Sutcliffe efficiency (NSE), mean absolute error (MAE), and PBIAS. The calculation of these statistics was performed using R package "hydroGOF" version 0.3-8 (Zambrano-Bigiarini, 2014).

NSE is a normalized metric that calculates the relative magnitude of the residual variance compared to the variance of the observed data. NSE ranges from $-\infty$ to 1.0, with NSE = 1 being a perfect fit. Functionally, NSE is used to indicate how well the observed versus predicted data fit the 1:1 line on a scatter plot. NSE is calculated as follows in equation 1 (Moriasi et al., 2007):

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}(e_i)^2}{\sum_{i=1}^{n}(O_i - O_{mean})^2}\right] \quad (1)$$

where e_i is the *i*th residual error (calculated as the *i*th observed value minus the *i*th predicted value), O_i is the *i*th observation, O_{mean} is the observed mean, and *n* is the observation sample size. NSE was used due to its prevalence in model evaluation studies and thus ability to be compared to other published values.

MAE uses absolute differences between simulated and observed values to prohibit opposite-signed error cancellation and is reported in the units of the constituent of interest. A value of MAE = 0 indicates a perfect fit, and minimizing MAE can aid in model parameter selection. MAE was calculated as shown in equation 2 (Coffey et al., 2004):

$$MAE = \frac{\sum_{i=1}^{n}|e_i|}{n} \quad (2)$$

where *n* is the observation sample size.

PBIAS is a measure of the average tendency of the predicted data to be greater or less than the observed data, expressed as a percentage. A value of PBIAS = 0 is optimal, with positive values indicating underestimation and negative values indicting overestimation. PBIAS is calculated as follows in equation 3 (Moriasi et al., 2007):

$$PBIAS = \left[\frac{\sum_{i=1}^{n}(e_i) * (100)}{\sum_{i=1}^{n}(O_i)}\right] \quad (3)$$

where PBIAS is the average amount by which the predicted data deviates from the observed data.

The statistical criteria used to assess whether model runs were satisfactory or unsatisfactory were based on Moriasi et al. (2007) with modifications for event-based comparisons (Table 2.4). The performance ratings supplied by Moriasi et al. (2007) for NSE and PBIAS were for a monthly time step, and in general, poorer model performance ratings are associated with model simulations of shorter time steps (Engel et al., 2007). Therefore, acceptable performance criteria with respect to event-based evaluations were relaxed.

Uncalibrated performance

Evaluation of TBET on sites in AR, GA and NC was first performed with the parameter set chosen by White et al. (2012) for sites in TX and OK. This analysis will be referred to as the uncalibrated performance. The uncalibrated dataset for AR consisted of 42 observations of runoff and 41 observations of sediment, total P, and dissolved P for years 2009 and 2010. The GA uncalibrated dataset used the years of 1995 and 1998 and consisted of 119 runoff and dissolved P observations, and 116 total P observations. Sediment yield was not measured on the GA site. The NC dataset used years 2011 and 2012 consisting of 139 runoff observations, 127 sediment yield observations, 131 total P observations, and 107 dissolved P observations. The KTR, MAE, NSE, and PBIAS were used in the uncalibrated performance evaluation.

An analysis of the contribution of P versus depth of runoff generated during an event showed that both measured and predicted storms resulting in less than 1 mm of runoff contributed only 0.3 to 1.4 percent of total P from all storms (Figures 2.1 and 2.2). Subsequently, only simulated or observed events with runoff greater than or equal to 1 mm were used in the uncalibrated and calibration data sets. Filtering out events with less than 1 mm of runoff biased our calibration toward the events that contribute most of the P. Events with less than 1 mm of runoff were included in the validation analysis.

Model Calibration

Each site was calibrated separately. A manual calibration using event-based measured and simulated data was performed by individually calibrating each constituent in the following order: runoff, sediment, total P, and dissolved P. Each model run for calibration followed the same procedure as the uncalibrated process with single year runs, two-year warm up and event-based comparisons.

The calibration parameters and ranges used for each constituent are listed in Table 2.5. The use of a manual calibration procedure limited the number of parameters and variation of parameters.

All parameters chosen for calibration, except the peak rate adjustment factor (ADJ_PKR), were cited by Arnold et al. (2012) as commonly applied parameters for calibration of SWAT. In addition, Arnold et al. (2013) recommend using curve number moisture condition II (CN_{II}), slope length (SLSUBBSN), the USLE minimum cropping factor (C_{min}), the USLE erosion factor (USLE K) the P percolation coefficient (PPERCO) and the P soil partitioning coefficient (PHOSKD) for calibration of runoff, sediment and nutrients in SWAT. ADJ_PKR was found to be sensitive for predictions of sediment and thus was included in the calibration parameter.

The parameters were altered according to a unique uniform distribution assigned to each parameter. The calibration of runoff analyzed nine CN_{II} values (± 16 points from the default). Sediment, and as a result the sediment P fraction of total P, were calibrated using seven USLE K (± 60% of the default), seven C_{min} (± 60% of the default), seven ADJ_PKR (ranging from 0.25 to 1.75), and seven SLSBBSN (± 60% of the default) values. Due to the restraints of manual calibration a two-step process was used to calibrate the sediment and sediment P parameters. First, C_{min} factor and SLSBBSN were calibrated in a full factorial 7x7 analysis. The best-fit parameter values for C_{min} and SLSBBSN were chosen and were thereafter used in the 7x7 factorial calibration of USLE K and ADJ_PKR. The calibration of DP analyzed five PPERCO (ranging between 5 and 15) and 10 PHOSKD (ranging between 100 and 200) values.

The manual calibration resulted in a total of 4,752 model simulations for all three sites combined with 132 parameter alterations (nine runoff, 98 sediment/sediment P, and 25 dissolved P),. The AR calibration consisted of 14 field-years multiplied by 132 parameter alterations for a total of 1,848 model runs. The GA and NC calibrations consisted of 12 and 10 field-years for a total of 1,584 and 1,320 model runs, respectively.

The calibration model runs were performed in batches using a series of custom Visual Basic for Application programs designed to modify parameters in the TBET input files and run the TBET.exe in batch mode. Several custom R programming files were then created to extract and process the resulting calibration output.

The three goodness-of-fit metrics presented in the previous section for evaluating hydrologic/water quality model results were used to assess the performance of calibration model runs. NSE was used due to its prevalence in model evaluation studies and its ability to assess model accuracy on the 1:1 line. MAE was used during calibration to evaluate changes in a constituents individual units. Lastly, PBIAS was used to assess over- or underestimation. The best-fit parameters for each constituent were chosen by selecting the local maximum for the event-based NSE value with consideration of MAE and PBIAS values.

Model validation

The AR validation data set used year 2011 with a total of 48 measurements for all constituents. The GA validation data set used years 1996 and 1997 with a total of 187 measurements for runoff and dissolved P and a total of 186 measurements for total P. The NC validation data set used year 2013 with a total of 245, 134, 160, and 152 measurements of runoff, sediment, total P and dissolved P, respectively. Like the assessment of the uncalibrated model, the performance of the validation simulations was assessed using KTR, MAE, NSE, and PBIAS. The goodness-of-fit statistics were also evaluated including estimates of measurement data and model uncertainty.

Uncertainty

The validation evaluation of TBET against measured P loss data accounted for measurement and model uncertainty. Uncertainty in measured P loss data was a function of the cumulative probable uncertainty acquired during sample collection, sample preservation and storage, laboratory analysis, and data management and processing (Harmel et al., 2006). Model uncertainty was a function of model structure error, model input error, and model parameter error (Bolster and Vadas, 2013). Comprehensive estimates of measurement and model uncertainty was rarely available for measured P loss data sets and P transport models. However, typical uncertainty

associated with certain measurement procedures and distributional properties of model predictions can be used in place of actual uncertainty information (Harmel et al., 2010).

The correction factor method was applied to incorporate uncertainty into goodness-of-fit indicators. This requires modification of the residual error term, equation 4, in equations 1, 2, and 3. The residual error (e_i) was modified by a correction factor (CF), equation 6, designed to estimate the degree of overlap (DO), equation 5, between the distributions for each measured and predicted data pair. The DO between observed and predicted probability density functions is an indication of model predictive power. The larger the variance in the respective probability density functions and/or the lower the residual error between the measured and simulated values, the greater the DO (Harmel et al., 2010):

$$e_i = O_i - P_i \tag{4}$$

$$DO_i = \left[Pr\left(O_i < P_{i\,(u)}\right) - Pr\left(O_i < P_{i\,(l)}\right)\right] \cdot \left[Pr\left(P_i < O_{i\,(u)}\right) - Pr\left(P_i < O_{i\,(l)}\right)\right] \tag{5}$$

$$CF(meas + pred)_i = 1 - DO_i \tag{6}$$

$$e(meas + pred)_i = CF(meas + pred)_i \cdot (O_i - P_i) \tag{7}$$

where subscripts *u* and *l* represent the respective upper and lower limits of the probability distributions for each measured (O_i) and predicted (P_i) value, Pr was the cumulative probability density function, and $e(meas + pred)_i$ was the modified residual error.

The uncertainty limits ($O_{i\,(l)}$, $O_{i\,(u)}$, $P_{i\,(l)}$, $P_{i\,(u)}$) in equation 5 were estimated due to the absence of uncertainty information for both the measured P loss data and TBET model uncertainty. Triangular probability distributions were commonly used when actual distributions are unknown (Bolster and Vadas, 2013). Thus, probability distributions were assumed to be symmetric triangular distributions, where the mean and median value of the distribution was represented with the

measured (O_i) or predicted (P_i) value and the limits were set uniformly using equations 9 and 11 (Harmel and Smith, 2007).

Probable error range (E_p) for water quality data sampled from small watersheds reported by Harmel et al. (2006) were used as estimates for the uncertainty boundaries about measured data values. Low and high uncertainty scenarios corresponding to Harmel et al. (2006) minimum and maximum E_p values for "typical" water quality measurements were evaluated for each measured value (Table 2.6).

E_p for each error source and each water quality constituent is calculated according to the root mean square error method (Harmel et al. 2006):

$$E_p = \sqrt{\sum_{i=1}^{n}(E_1^2 + E_2^2 + E_3^2 + \cdots + E_n^2)} \tag{8}$$

$$O_{i(l)} = O_i - \frac{E_p \times O_i}{100} \qquad O_{i(u)} = O_i + \frac{E_p \times O_i}{100} \tag{9}$$

where E_p is the cumulative percent error (± %) for a given water quality constituent, *n* is the total number error sources, and E_n is the error (± %) associated with the *n*th error source. The upper and lower observed data uncertainty limits, $O_{i(u)}$ and $O_{i(l)}$, were then calculated using E_p in equation 9.

Model uncertainty information was not provided for TBET, but Harmel et al. (2010) present typical values for the coefficient of variation (CV) in modeled data at field and watershed scales. Like the estimates of E_p for each measured value, low error and high error CV values (0.026 and 0.256) were used to evaluate the uncertainty about each predicted value (Harmel et al. 2010):

$$Cv = \frac{SD_i}{\bar{x}_i} \tag{10}$$

$$P_{i(l)} = \bar{x}_i - \sqrt{3} \cdot (SD_i) \qquad P_{i(u)} = \bar{x}_i + \sqrt{3} \cdot (SD_i) \tag{11}$$

where SD_i is the standard deviation of the *i*th predicted value, $\bar{x}_i$ is the sample mean (assumed to the predicted (p_i) value for the triangular distribution), and $P_{i(u)}$ and $P_{i(l)}$ are the upper and lower model uncertainty boundaries, respectively.

TBET and the P-index

Lastly, we evaluated TBET against the P-index in predictions of average annual P loss. In its default mode, TBET predicts 25-year average annual loads for P loss. The P-index provides a long-term estimate of risk of P loss in the form of discrete, ratio values ranging from 0 to 100. P-index ratings were provided from study authors for three plots in GA for years 1995 and 1996, and four plots in NC for years 2011 and 2012. To convert the GA and NC P-index ratings from numerical ratings to a load in kg ha^{-1} yr^{-1} the values are divided by 10. To compare measured P loss with the TBET 25-yr average P loss predictions and P-index loads we summed the measured P loss from each field-year in GA and NC and took an annual average. The annual averages reflect a four- and three-year average measured total P loss in GA and NC, respectively. We also provided the TBET predictions using the default parameters from White et al. (2012) to show the improvement in the validated TBET predictions. Goodness-of-fit between measured average annual P loss to TBET and the P-index was evaluated using MAE and PBIAS. NSE was not used in this case due to the small sample size, however we used the KTR method of nonparametric regression to compare the slopes to the 1:1 line.

RESULTS AND DISCUSSION

Uncalibrated model results

The parameter set for TBET chosen by White et al. (2012b) provided mixed performance results among constituents and study sites (Table 2.7). The uncalibrated parameters for AR (CN_{II}: 73.5; C_{min}: 0.001; SLSBBSN: 152 m; USLE_K: 0.43; ADJ_PKR: 1.00; PPERCO: 10; PHOSKD: 175) produced unsatisfactory predictions of runoff (NSE = -7) and sediment (NSE = -59), total P (NSE = -1.5), and dissolved P (NSE = -1.6). The uncalibrated parameters for GA (CN_{II}: 61.5 and 73.5; C_{min}: 0.001; SLSBBSN: 37.3 and 46.1 m; USLE_K: 0.24 and 0.28; ADJ_PKR: 1.00; PPERCO: 10; PHOSKD: 175) produced satisfactory predictions of runoff (NSE = 0.57) and total P (NSE = 0.34), and unsatisfactory results for total P (NSE = 0.19). The uncalibrated parameters for NC (CN_{II}: 74.4; C_{min}: 0.2 and 0.03; SLSBBSN: 46.1 m; USLE_K: 0.17; ADJ_PKR: 1.00; PPERCO: 10; PHOSKD: 175) produced unsatisfactory predictions of runoff (NSE = 0.23), sediment (NSE = -88), total P (NSE = -78), and dissolved P (NSE = -0.08). Figure 2.3 provides the scatter plots and KTR regression statistics for the NC uncalibrated dataset as an example. The supplemental information (Supplemental Figs. S2, S3, S4, and S5) contains the scatter plots and KTR regression statistics for the uncalibrated performance organized by constituent.

Uncalibrated runoff

The uncalibrated runoff predictions at all three sites were overpredicted with PBIAS values of -179%, -48%, and -39% for AR, GA, and NC, respectively. Further visual inspection of the uncalibrated KTR slope and intercept estimates for runoff show overprediction of low to moderate runoff (intercepts greater than zero) and underprediction of high runoff (slope less than 1) in GA These findings are similar to the results of runoff performance observed by White et al. (2012b). Because TBET only predicts surface runoff, and ignores subsurface return flow, there is a tendency for the model to underpredict high runoff (White et al., 2012b).

The overprediction of runoff in NC may be attributed to the presence of no-till treatments in two of the five fields, which are not reflected in the default CN_{II} value. Anand et al. (2007) found that SWAT poorly predicted field-scale runoff on no-till plots with default parameters, and point out that uncalibrated models often provide poor and inconsistent results. Bonta and Shipitalo (2013) and Endale et al. (2011) both found CN values representing long-term, no-till management to be 16 units larger than standard handbook table values.

The poor performance of runoff in AR may be attributed to the lack of site-specific rainfall input data. As previously mentioned, the AR dataset lacked complete weather data for the period of investigation, thus existing weather data were supplemented with observations from three nearby weather stations. Runoff predictions from TBET are primarily influenced by the quality of the precipitation input file. The precipitation file generated for the AR site may have included storms that did not occur at the field site, which may have increased antecedent moisture conditions prior to measured storm events.

Uncalibrated sediment

Uncalibrated sediment predictions for AR and NC were vastly overpredicted with PBIAS values of -256% and -963%, respectively. The sediment regression slope estimate of 5.4 ($p < 0.001$) in NC suggests the extreme overprediction of sediment was due to the overprediction of low sediment loads. In contrast, White et al. (2012) observed a consistent underprediction of sediment and calibrated C_{min} and SLSBBSN resulting in an increase in sediment predictions. Because sediment losses on the GA site are very low sediment loss data were not supplied in the GA dataset. Moreover, the annual sediment prediction among all 24 field-years of data in GA ranged from 0.05 to 0.15 ton ha^{-1} yr^{-1} in contrast to 0.02 to 1.26 ton ha^{-1} yr^{-1} and 2.29 to 105.31 ton ha^{-1} yr^{-1} in AR and NC, respectively.

Sediment loss is a function climate, land use, management activities, hillslope gradient, and rainfall energy. The overprediction of sediment losses could be a result of improper

representation of tillage and grazing activities, underrepresentation of vegetative cover, or improper characterization of hillslope slope and slope length.

Two of the five fields in NC were managed as no-till treatments, and the increased infiltration and reduction of surface runoff commonly associated with no-till may have been insufficiently represented with the default CN_{II}. Also, two of the AR fields were grazed, one continuously and rotationally. Although details of the stocking density (two cows per hectare) and rotational rate (six weeks on and six weeks off) were provided and replicated in the management operations, the effect of grazing may have been overestimated.

The governing equation of sediment losses in TBET is MUSLE, which is highly dependent on subbasin characteristics such as drainage area and flow length. The development of TBET included an analysis of topography in various climate zones within Oklahoma to determine an average subbasin size. The default subbasin area was set to 42 ha, which reflects the typical contributing area to establish a stream on a 1:24,000 topographic map in Oklahoma (White et al., 2012a). TBET predicts delivery to stream and the field data in this study were collected as edge-of-field losses (White et al., 2012b). Therefore, sediment predictions in NC may be highly affected by the use of an average subbasin size developed for Texas and Oklahoma due to climatic and topographical differences in these two regions.

Uncalibrated dissolved phosphorus

Uncalibrated dissolved P was underpredicted at all sites. Although the PBIAS values of -35% in AR and -21% in NC would suggest overprediction of dissolved P for those sites, the KTR slope estimates suggest systematic underprediction (Supplemental Fig. S5). This result is consistent with the findings of White et al. (2012a) in regards to the underprediction of dissolved P with Pasture Phosphorus Management Plus (PPM Plus), the precursor model to TBET. They found that PPM Plus predicted dissolved P more accurately on cultivated land than pasture. This result is reflected in the uncalibrated results for TBET. Although the NSE values for dissolved P prediction

among sites do not vary significantly. The MAE in NC (0.02 kg ha^{-1}) is significantly lower than AR (0.16 kg ha^{-1}) and GA (0.35 kg ha^{-1}). This result is also evident upon visual inspection of the uncalibrated dissolved P scatter plots.

The systematic underestimation of dissolved P may be due to the lack of a specific P pool for manure application in TBET. SWAT assumes that surface applied manure P that is not incorporated immediately becomes a fraction of the soil P pool, however when manure is surface applied and not incorporated, P is readily transported via surface runoff. Thus, SWAT may be expected to underpredict dissolved P losses in systems where manure is not mixed into the soil (Radcliffe et al., 2009a; Vadas et al., 2007).

Uncalibrated total phosphorus

The uncalibrated sediment and dissolved P predictions heavily influenced the uncalibrated total P performance. Uncalibrated total P in NC, like sediment predictions in NC, was extremely overpredicted (PBIAS = -654%). The response of uncalibrated total P to the overprediction of sediment in AR was less severe, but still present (PBIAS = -55%). Predictions of total P in GA were underpredicted (PBIAS = 20%).

Predictions of total P loss are directly dependent on particulate P predictions associated with prediction of sediment loss and dissolved P loss predictions. For example, in NC, the extreme overprediction of sediment overpowered the effect of underpredicted dissolved P, resulting in overprediction of total P. The underprediction bias of total P in GA is greater than the underprediction bias of GA dissolved P. This may be due to the site's relatively low sediment prediction.

Calibration model results

Despite the poor performance using default parameters, goodness-of-fit improved with calibration (Table 2.7), but the best-fit parameters for each site varied significantly (Table 2.8). The

best-fit parameters for AR (CN_{II}: 58; C_{min}: 0.0004; SLSBBSN: 243 m; USLE_K: 0.17; ADJ_PKR: 0.75; PPERCO: 5; PHOSKD: 200 m^3 Mg^{-1}) produced unsatisfactory predictions of runoff (NSE = -2.43), sediment (NSE = -0.57), total P (NSE = 0.01) and dissolved P (NSE = -0.14). Best-fit parameters for GA (CN_{II}: 69.5 and 81.5; C_{min}: 0.001; SLSBBSN: 37.3 and 46.1 m; USLE_K: 0.24 and 0.28; ADJ_PKR: 1.00; PPERCO: 10; PHOSKD: 200 m^3 Mg^{-1}) produced satisfactory predictions of runoff (NSE = 0.58), total P (NSE = 0.51), and dissolved P (NSE = 0.35). Best-fit parameters in NC (CN_{II}: 58; C_{min}: 0.08 and 0.012; SLSBBSN: 18.4 m; USLE_K: 0.07; ADJ_PKR: 0.25; PPERCO: 5; PHOSKD: 200 m^3 Mg^{-1}) produced satisfactory predictions of runoff (NSE = 0.33) and unsatisfactory fit for sediment (NSE = -0.67), total P (NSE = -1.71), and dissolved P (NSE = 0.21).

Runoff calibration

The best-fit CN_{II} values for AR (years 2009 and 2010), and NC (years 2011 and 2012) were 58. The best-fit CN_{II} values for GA (years 1995 and 1998) were 70 (Cecil series) and 82 (Altavista, Helena, Sedgefield series). The AR calibrated CN_{II} value of 58 is more representative of soil hydrologic group A (range 39 to 68) than its classified soil group C (range 74 to 86). Note that the default CN_{II} for AR, 74, also falls within the range of soil group B (range 61 to 79). The GA calibrated CN_{II} value of 70 for the Cecil series remained within the range of CNs for soil hydrologic group B. The GA calibrated CN_{II} value of 82 for soil series Altavista, Helena and Sedgefield also remained within their classified soil hydrologic group of C. The best-fit CN_{II} for AR and NC were both 16 units lower than their default CN_{II}, and the GA CN_{II} values were 8 points greater than the default CN_{II}.

The calibrated CN_{II} value of 58 for NC was in agreement with values reported from Endale et al. (2011) and Bonta and Shipitalo (2013) for fields containing long-term no-till management. Both studies found that CN values for long-term no-till were 16 units less than the average of the range of CN values found in the standard handbook values. The default CN_{II} value of 73.5 in AR

seemed high for fields in continuous pasture as compared to the default CN_{II} value of 61.5 for similar land use and management in GA In addition, Bonta and Shipitalo (2013) found that a simple grass establishment can be as effective as long-term no-till in reductions of runoff. Thus, the calibrated CN_{II} value of 58 in AR seems reasonable. The 8 unit CN_{II} increase in GA remained within the range of values for hydrologic group C soils and may be a reflection of the influence of grazing activities and hay production on these fields.

Calibrated results for daily runoff were satisfactory in GA (NSE = 0.58) and NC (NSE = 0.33). These results were comparable to those presented by Anand et al. (2007). They show calibrated SWAT daily field-scale runoff NSE values between 0.22 and 0.62. The calibrated runoff results in AR were unsatisfactory (NSE = -2), however, they were an improvement from the performance of the data subset prior to calibration (NSE = -7).

Sediment calibration

TBET extremely overpredicted sediment with default parameter values in AR and NC As a result, the calibration of sediment required significant modification default parameters. The best-fit C_{min} in AR and NC were both 60% reductions from the default values. In AR the default value of 0.001 was reduced to 0.0004. In NC the default values of 0.2 (sweet corn) and 0.03 (winter wheat) were reduced to 0.08 and 0.012, respectively. The best-fit SLSBBSN value for AR was 243 m, a 60% increase from the default value of 152 m. In contrast, the best-fit SLSBBSN for NC was achieved at 60% below the uncalibrated value (46.1 m) at 18.4 m. The best-fit USLE K factors for AR and NC were 60% reductions at 0.17 and 0.07, respectively. Lastly, the best-fit ADJ_PKR in AR was 0.75 and in NC the best-fit value was 0.25 with default values of 1.0.

The sediment calibration did not achieve satisfactory goodness-of-fit in AR (NSE = -0.57) or NC (NSE = -0.67). However, all goodness-of-fit metrics showed large improvements from the uncalibrated fit.

Maski et al. (2008) modeled daily sediment yields from three sorghum-soybean field plots (two no-till systems and one conventional-till system) with SWAT and observed mixed results in uncalibrated, calibration and validation data sets. Uncalibrated NSE values ranged from 0.03 to 0.65, calibration values ranged from 0.04 to 0.69, and validation values ranged from -0.81 to 0.49. They also found that SWAT systematically underpredicted sediment in the uncalibrated and calibration data, but overpredicted sediment on one of the no-till treatments in validation.

A significant difference between the Maski et al. (2008) and this study is the amount of measured sediment observed from the field sites. The average annual measured sediment yield for the three fields in Maski et al. (2008) ranged from 0.48 to 2.51 Mg ha^{-1} yr^{-1}, whereas the average annual sediment yield for AR and NC sites ranged from 0.004 to 0.015 Mg ha^{-1} yr^{-1} and 0.013 to 0.148 Mg ha^{-1} yr^{-1}, respectively. Thus, perhaps the greatest contributing factor influencing the overprediction of sediment is the low amount of measured sediment in the field data.

Dissolved Phosphorus calibration

The dissolved P calibration produced a satisfactory result in GA (NSE = 0.35) and unsatisfactory results in AR (NSE = -0.14) and NC (NSE = 0.21). The best-fit parameters in both AR and NC were PHOSKD values of 200 m^3 Mg^{-1} and PPERCO values of 5, while GA achieved a best-fit with a PHOSKD value of 200 m^3 Mg^{-1} and the default PPERCO value of 10. The calibrated PHOSKD value of 200 m^3 Mg^{-1} is closer to the value of 242 m^3 Mg^{-1} derived by Radcliffe et al. (2009b) using the ratio of labile P in the soil to dissolved P in runoff for typical Piedmont soils, but slightly higher than the default value of 175 m^3 Mg^{-1}. Like the uncalibrated dissolved P predictions, the calibrated dissolved P predictions were underpredicted at all three sites, however met the criteria for satisfactory performance with respect to PBIAS.

Total phosphorus calibration

The total P best-fit parameters produced a satisfactory result in GA (NSE = 0.51) and unsatisfactory results in AR (NSE = 0.01) and NC (NSE = -1.71). However, unlike the default parameters, which resulted in large overpredictions of total P in AR and NC, the best-fit parameters substantially improved the PBIAS of total P. For example, in NC PBIAS was improved from -607 to 8.9%. In GA, total P achieved a satisfactory NSE value of 0.5 after increasing CN_{II} by eight units and did not improve with further calibration of sediment related parameters. Total P in GA did improve with calibration of dissolved P using a PHOSKD value of 200 m^3 Mg^{-1}.

Validation results

TBET provided a varied performance in simulating the validation data sets with the best-fit calibration parameters (Table 2.9). Both increases and decreases in performance metrics from the uncalibrated and calibration data sets were observed. The AR dataset improved significantly in all four variables as compared to the calibration dataset, however produced a satisfactory result for runoff (NSE = 0.35) and unsatisfactory results for sediment (NSE = 0.24), total P (NSE = 0.27), and dissolved P (NSE = 0.27). The GA dataset showed satisfactory results for runoff (NSE = 0.57) and dissolved P (NSE = 0.38), while total P performance was unsatisfactory (NSE = -0.02). Of the three sites modeled, NC performed the best with satisfactory daily NSE values for all four output variables with NSE values of 0.65, 0.31, 0.35, and 0.32 for runoff, sediment, total P, and dissolved P, respectively. Figure 2.4 provides the scatter plots and KTR regression statistics for the NC validation performance as an example. The supplemental information contains the scatter plots and KTR regression statistics for the validation performance evaluation organized by constituent (Supplemental Figs. S6, S7, S8, and S9).

The improvement in runoff prediction in AR is encouraging since negative NSE values were observed using in the uncalibrated and calibration data sets. Likewise, the positive NSE of 0.24 for the sediment prediction in AR was greatly improved from -59 and -0.57 for the uncalibrated

and calibration data sets, respectively. The P loss NSE values in the AR validation dataset also improved from the uncalibrated performance and may be more indicative of the ability of TBET to predict P loss from this site since the runoff and sediment predictions were more accurate.

The GA dataset validated well with respect to runoff (NSE = 0.57) and dissolved P (NSE = 0.38). However, total P validated poorly (NSE = -0.02). This result is surprising considering the much improved performance of the calibration dataset with respect to P loss. The annual rainfall in each of the four years of data in GA was below the 50-yr average for the region (Pierson et al., 2001a). Thus, the fact that TBET overestimated runoff using default parameters may be expected. However, the overprediction of runoff in the GA validation dataset (PBIAS = -40%) indicates that the 8-unit increase in the CN_{II} may not be appropriate despite improvements in runoff and P loss predictions in the calibration dataset. Comparing the runoff scatter plots of the uncalibrated and validation data sets shows that the uncalibrated model provided greater confidence in the mean response, indicated by narrower confidence intervals. This is compounded by the fact that the KTR slope estimate for the validation dissolved P model falls outside the 95% confidence interval, indicating a statistically non-significant result.

Regardless of the decrease in performance of the GA validation dataset relative to the calibration dataset, TBET produced satisfactory goodness-of-fit results on a daily basis for runoff (NSE = 0.57) and dissolved P (NSE = 0.38).

The NC dataset validated well regardless of the extreme overprediction observed in the uncalibrated sediment and total P. In fact, the NC validation sediment and total P performed satisfactorily with respect to NSE (0.31 and 0.35, respectively) and PBIAS (12.2% and -16.9%, respectively). The fit of dissolved P in NC was slightly worse than the uncalibrated, however the validation runoff model was improved, which may be a better indicator of the ability for TBET to predict P loss.

Uncertainty results

Performance improvements were minimal when including the correction factor to evaluate the influence of model and measurement uncertainty (Table 2.10). Harmel et al. (2010) found that the correction factor produced insignificant changes in goodness-of-fit results for very good and poor model simulations. They also note that important improvements in fit should occur for datasets in moderate agreement. These statements seem to align with the changes in fit due to the correction factor in this study. The best performing validation dataset, NC runoff, with a NSE value of 0.65, showed no improvement, even in the high uncertainty scenario. Similarly, the poorest performing dataset, GA total P (NSE = -0.02), showed marginal improvements in the high uncertainty scenario despite having the second largest number of overlapping confidence intervals (75% of the data points). The greatest improvements were observed in the dissolved P validation models, which performed moderately relative to the other constituents. For example, the NC dissolved P NSE of 0.32 was improved to 0.4 in the high uncertainty scenario with 60% of the data point probability density functions overlapping.

These results are similar to those presented by Bolster and Vadas (2013) who found no significant increases in goodness-of-fit statistics when applying the correction factor method to P loss predictions from the APLE model. They attributed the negligible changes in fit to underestimation of model uncertainty and showed that a large number of their data point probability density functions either had a low degree of overlap or did not overlap at all. Likewise, we found a low percentage of overlap (0 to 6% among all sites and constituents) in the low uncertainty scenario. Moreover, 17 to 82% probability density functions overlapped in the high scenario, but the degree of overlap was predominantly low with a median degree of overlap among all 724 overlapping data points of 0.937. Note that the closer the degree of overlap is to 1.0 the smaller the effect of the correction factor.

The relatively few overlapping probability density functions in the low uncertainty scenario and the generally low degree of overlap may be a result of underestimated model uncertainties. The

estimates of measurement uncertainty from Harmel et al. (2006) are likely reasonable for this application considering the extensive compilation of journal articles used in the development of their estimates and that the estimates are specifically designated for small catchment and field scale measurements. However, the coefficients of variation used to estimate model uncertainty from Harmel et al. (2010) may have underestimated model uncertainty due to the fact that the values were derived from multiple hydrologic and water quality models and do not reflect the specific parameter uncertainties and model structure errors in TBET.

The estimated uncertainties in TBET predictions based on assumed errors in measured field data and model inputs varied substantially. In AR, the interquartile range (IQR) of the absolute errors in the low uncertainty scenario were -24.7 mm, -0.008 ton ha^{-1}, -0.06 kg ha^{-1}, and -0.04 kg ha^{-1} for runoff, sediment, total P and dissolved P, respectively. The IQR of the absolute errors in the high scenario were -22.1 mm, -0.008 ton ha^{-1}, -0.05 kg ha^{-1}, and -0.04 kg ha^{-1} for runoff, sediment, total P and dissolved P, respectively. In GA, IQR of the absolute errors in the low uncertainty scenario were -7.8 mm, -0.72 kg ha^{-1}, and -0.44 kg ha^{-1} for runoff, total P and dissolved P, respectively. In the high scenario, the IQR for the absolute errors are -7.6 mm, -0.72 kg ha^{-1}, and -0.43 kg ha^{-1} for runoff, sediment, total P and dissolved P, respectively. The IQR of the absolute errors in the NC low uncertainty scenario are -5.9 mm, -0.06 ton ha^{-1}, -0.11 kg ha^{-1}, and -0.01 kg ha^{-1} for runoff, sediment, total P and dissolved P, respectively. The IQR of the absolute errors in the high scenario are -5.8 mm, -0.06 ton ha^{-1}, -0.1 kg ha^{-1}, and -0.01 kg ha^{-1} for runoff, sediment, total P and dissolved P, respectively. The supplemental information contains cumulative frequency distributions of absolute errors for the low and high uncertainty scenarios organized by constituent (Supplemental Figs. S10, S11, S12, and S13).

The absolute errors for TBET total P loss on all three sites ranged from -2.5 to 7.8 kg ha^{-1} for the low uncertainty scenario and -2.5 to 7.7 kg ha^{-1} for the high uncertainty scenario. These ranges are larger than absolute error ranges that Bolster and Vadas (2013) presented for the APLE model. They found that the absolute error for APLE P loss ranged from -3.5×10^{-5} to 2.8 kg ha^{-1}

and -9.4 x 10^{-5} to 7.2 kg ha^{-1} for their low and high uncertainty scenarios, respectively. This result is not surprising since APLE is an annual time step model with fewer inputs and parameters than TBET.

Absolute errors for all constituents increased with increasing storm size. The ability of TBET to precisely predict runoff, sediment, and P loss for larger events is influenced by several factors. As mentioned before, TBET only predicts surface runoff and ignores subsurface flow, and measurements from the field data inevitably include lateral flow. Thus, it is likely that TBET will underestimate runoff, and subsequently dissolved P. In addition, TBET uses a daily precipitation value and inherently cannot precisely simulate the intensity and timing of storms. This may affect the timing of runoff after applications of P as well as the rate of erosion.

TBET and the P-index

The validated TBET model provided slightly better goodness-of-fit against measured average annual total P loss than the P-index and substantially better goodness-of-fit compared to TBET using default parameters from White et al. (2012b) (Tables 2.11 and 2.12). The P-index overpredicted both low and high measured P loss with a PBIAS value of -78% and KTR slope estimate of 1.67 ($p = 0.001$). The validated TBET model tended to overestimate high measured P losses and underestimate low measured P loss with a PBIAS of -6.3% and KTR slope estimate of 0.96 ($p < 0.001$) (Figure 2.5). However, this PBIAS value corresponds to a performance rating of "very good" for P loss predictions according to Moriasi et al. (2007), and is similar to the original TBET validation results for total P (PBIAS = -14%) presented by White et al. (2012b). Moreover, the mean absolute error for the validated TBET model was lower than the P-index at 4.4 and 6.9 kg ha^{-1} yr^{-1}, respectively. The results also showed a significant improvement from the uncalibrated TBET model and the validated TBET model. The poor performance of the uncalibrated TBET model was due to the extreme overprediction of P loss in NC.

CONCLUSIONS

Our results suggested that TBET is more accurate in predictions of event-based runoff and less accurate in predictions of event-based sediment and P loss. Sediment predictions were influenced by the low amounts of measured sediment loss from the study sites, while P loss predictions were most likely affected by the systematic underprediction of dissolved P. TBET does not account for specific interactions between runoff and surface-applied manure P, which can result in underestimation of P loss for events following a surface application of P (Vadas et al., 2007). The combination of TBETs inaccuracy in predicting low sediment losses and underprediction of dissolved P may have been the cause of the poor validation results for total P loss in the pasture system of GA. Thus, if TBET were used as a meta-model for the southern P-indices it may result in less restrictive P application rates than appropriate for system where dissolved P is a dominant or large fraction of total P loss. In addition, the comparison between measured average annual P loss and the 25-year P loss predictions from the validated TBET model illustrates the potential for TBET to result in more restrictive P application rates for fields with high P loss potential due to overprediction of high measured P loss.

The overall improvement in performance of TBET using parameters calibrated for conditions in TX and OK (uncalibrated) compared with best-fit parameters for AR, GA, and NC (validation) suggests that TBET is able to produce satisfactory event-based predictions (NSE $\geq$ 0.3) of runoff, sediment, total P, and dissolved P in the southern region with site-specific calibration. Moreover, the validated TBET model provided "very good" performance (PBIAS = -6.3%) for predictions of long-term P loss according to Moriasi et al. (2007). Still, the relatively broad variation in best-fit parameters, goodness-of-fit, and estimated uncertainties in TBET predictions among study sites suggests TBET may not be appropriate for applying a regional set of parameters for the South. Therefore, it may be more appropriate to develop TBET models by physiographic region (i.e. Coastal Plain, Piedmont, Blue Ridge, etc.) or state-by-state. This would require rigorous model input database construction and calibration of TBET predictions against measured data for

a range conditions within each state or physiographic region, but may reduce variation in TBET best-fit parameters and prediction uncertainties.

REFERENCES

Anand, S., Mankin, K. R., McVay, K. A., Janssen, K. A., Barnes, P. L., and Pierzynski, G. M., 2007, Calibration and validation of ADAPT and SWAT for field-scale runoff prediction: Journal of the American Water Resources Association, v. 43, no. 4, p. 899-910.

Arnold, J. G., Kiniry, J. R., Srinivasan, M. S., Williams, J. R., Haney, R. L., and Neitsch, S. L., 2013, Soil and Water Assessment Tool input/output documentation; version 2012: Texas A&M University, Texas Water Resources Institute, TR-439.

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., Santhi, C., Harmel, R. D., Griensven, A. v., Liew, M. W. v., Kannan, N., and Jha, M. K., 2012, SWAT: model use, calibration, and validation: Transactions of the ASABE, v. 55, no. 4, p. 1491-1508.

Bolster, C., and Vadas, P., 2013, Sensitivity and uncertainty analysis for the Annual Phosphorus Loss Estimator Model: Journal of Environment Quality, v. 42, p. 1109-1118.

Bolster, C. H., 2011, A critical evaluation of the Kentucky phosphorus index: Journal of the Kentucky Academy of Sciences, no. 72, p. 46-58.

Bolster, C. H., Vadas, P. A., Sharpley, A. N., and Lory, J. A., 2012, Using a phosphorus loss model to evaluate and improve Phosphorus Indices: Journal of Environment Quality, v. 41, no. 6, p. 1758-1766.

Bonta, J. V., and Shipitalo, M. J., 2013, Curve numbers for long-term no-till corn and agricultural practices with high watershed infiltration: Journal of Soil and Water Conservation, v. 68, no. 6, p. 487-500.

Coffey, M. E., Workman, S. R., Taraba, J. L., and Fogle, A. W., 2004, Statistical procedures for evaluating daily and monthly hydrologic model predictions: Transactions of the ASAE, v. 47, no. 1, p. 59-68.

DeLaune, P. B., Moore, P. A., Carman, D. K., Sharpley, A. N., Haggard, B. E., and Daniel, T. C., 2004a, Development of a phosphorus index for pastures fertilized with poultry litter - Factors affecting phosphorus runoff: Journal of Environmental Quality, v. 33, no. 6, p. 2183-2191.

-, 2004b, Evaluation of the phosphorus source component in the phosphorus index for pastures: Journal of Environmental Quality, v. 33, no. 6, p. 2192-2200.

Eghball, B., and Gilley, J. E., 2001, Phosphorus risk assessment index evaluation using runoff measurements: Journal of Soil and Water Conservation, v. 56, no. 3, p. 202-206.

Endale, D. M., Schomberg, H. S., Fisher, D. S., and Jenkins, M. B., No-till and curve numbers - A closer look, *in* Proceedings Georgia Water Resources Conference, The University of Georgia, April 11, 2011 2011, The University of Georgia.

Engel, B., Storm, D. E., White, M., Arnold, J., and Arabi, M., 2007, A hydrologic/water quality model application protocol: Journal of the American Water Resources Association, v. 43, no. 5, p. 1223 - 1236.

Good, L. W., Vadas, P., Panuska, J. C., Bonilla, C. A., and Jokela, W. E., 2012, Testing the Wisconsin Phosphorus Index with year-round, field-scale runoff monitoring: Journal of Environmental Quality, v. 41, no. 6, p. 1730-1740.

Harmel, R. D., Cooper, R. J., Slade, R. M., Haney, R. L., and Arnold, J. G., 2006, Cumulative uncertainty in measured streamflow and water quality data for small watersheds: Transactions of the ASABE, v. 49, no. 3, p. 689-701.

Harmel, R. D., and Smith, P. K., 2007, Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling: Journal of Hydrology, v. 337, p. 326-336.

Harmel, R. D., Smith, P. K., and Migliaccio, K. W., 2010, Modifying Goodness-of-Fit Indicators to Incorporate Both Measurement and Model Uncertainty in Model Calibration and Validation: Transactions of the ASABE, v. 53, no. 1, p. 55-63.

Harmel, R. D., Torbert, H. A., Delaune, P. B., Haggard, B. E., and Haney, R. L., 2005, Field evaluation of three phosphorus indices on new application sites in Texas: Journal of Soil and Water Conservation, v. 60, no. 1, p. 29-42.

Helsel, D. R., and Hirsch, R. M., 1992, Statistical methods in water resources, Studies in environmental science, Volume 49: Amsterdam ; New York, Elsevier.

Jarvie, H. P., Sharpley, A. N., Withers, P. J. A., Scott, J. T., Haggard, B. E., and Neal, C., 2013, Phosphorus mitigation to control river eutrophication: murky waters, inconvenient truths, and "postnormal" science: Journal of Environmental Quality, v. 42, no. 2, p. 295-304.

Komsta, L., 2013, mblm: Median-Based Linear Models. R package version 0.12. http://CRAN.R-project.org/package=mblm.

Larsen, E., Grossman, J., Edgell, J., Hoyt, G., Osmond, D., and Hu, S. J., 2014, Soil biological properties, soil losses and corn yield in long-term organic and conventional farming systems: Soil & Tillage Research, v. 139, p. 37-45.

Leone, A., Ripa, M. N., Boccia, L., and Lo Porto, A., 2008, Phosphorus export from agricultural land: a simple approach: Biosystems Engineering, v. 101, no. 2, p. 270-280.

Maski, D., Mankin, K. R., Janssen, K. A., Tuppad, P., and Pierzynski, G. M., 2008, Modeling runoff and sediment yields from combined in-field crop practices using the Soil and Water Assessment Tool, Journal of Soil and Water Conservation (Ankeny), Volume 63: Ankeny; USA, Soil and Water Conservation Society, p. 193-203.

Meals, D. W., Dressing, S. A., and Davenport, T. E., 2010, Lag time in water quality response to best management practices: a review: Journal of Environmental Quality, v. 39, no. 1, p. 85-96.

Mittelstet, A. R., Daly, E. R., Storm, D. E., White, M. J., and Kloxin, G. A., 2012, Field scale modeling to estimate phosphorus and sediment load reductions using a newly developed graphical user interface

for the Soil and Water Assessment Tool.: American Journal of Environmental Science, no. 8, p. 605-614.

Mittelstet, A. R., Heeren, D. M., Fox, G. A., Storm, D. E., White, M. J., and Miller, R. B., 2011, Comparison of subsurface runoff phosphorus transport rates in alluvial floodplains.: Agriculture, Ecosystems and Environment, no. 141, p. 417-425.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L., 2007, Model evaluation guidelines for systematic quantification of accuracy in watershed simulations: Transactions of the ASABE, v. 50, no. 3, p. 885-900.

Osmond, D. L., Cabrera, M. L., Feagley, S. E., Hardee, G. E., Mitchell, C. C., Moore, P. A., Mylavarapu, R. S., Oldham, J. L., Stevens, J. C., Thom, W. C., Walker, F., and Zhang, H., 2006, Comparing ratings of the southern phosphorus indices: Journal of Soil and Water Conservation, v. 61, no. 6, p. 325-337.

Pierson, S. T., Cabrera, M. L., Evanylo, G. K., Kuykendall, H. A., Hoveland, C. S., McCann, M. A., and West, L. T., 2001a, Phosphorus and Ammonium concentrations in surface runoff from grasslands fertilized with broiler litter: Journal of Environment Quality, v. 30, no. 5, p. 1784-1789.

Pierson, S. T., Cabrera, M. L., Evanylo, G. K., Schroeder, P. D., Radcliffe, D. E., Kuykendall, H. A., Benson, V. W., Williams, J. R., Hoveland, C. S., and McCann, M. A., 2001b, Phosphorus Losses from Grasslands Fertilized with Broiler Litter: Journal of Environment Quality, v. 30, no. 5, p. 1790-1795.

Radcliffe, D. E., Freer, J., and Schoumans, O. F., 2009a, Diffuse Phosphorus models in the United States and Europe: Their usages, scales, and unceratainties: Journal of Environment Quality, v. 38, p. 1956-1967.

Radcliffe, D. E., Lin, Z., Risse, L. M., Romeis, J. J., and Jackson, C. R., 2009b, Modeling Phosphorus in the Lake Allatoona Watershed Using SWAT: I. Developing Phosphorus Parameter Values: Journal of Environmental Quality, v. 38, no. 1, p. 111-120.

Schoumans, O. F., Mol-Dijkstra, J., Akkermans, L. M. W., and Roest, C. W. J., 2002, SIMPLE: Assessment of non-point phosphorus pollution from agricultural land to surface waters by means of a new methodology: Water Science and Technology, v. 45, no. 9, p. 177-182.

Sharpley, A., Beegle, D., Bolster, C., Good, L., Joern, B., Ketterings, Q., Lory, J., Mikkelsen, R., Osmond, D., and Vadas, P., 2012, Phosphorus Indices: why we need to take stock of how we are doing: Journal of Environmental Quality, v. 41, no. 6, p. 1711-1719.

Sharpley, A., Jarvie, H. P., Buda, A., May, L., Spears, B., and Kleinman, P., 2013, Phosphorus legacy: Overcoming the effects of past management practices to mitigate future water quality impairment: Journal of Environment Quality, v. 42, no. 5, p. 1308-1326.

Sharpley, A. N., Kleinman, P. J. A., Heathwaite, A. L., Gburek, W. J., Weld, J. L., and Folmar, G. J., 2008, Integrating contributing areas and indexing phosphorus loss from agricultural watersheds: Journal of Environmental Quality, v. 37, no. 4, p. 1488-1496.

Sharpley, A. N., Kleinman, P. J. A., Jordan, P., Bergstrom, L., and Allen, A. L., 2009, Evaluating the Success of Phosphorus Management from Field to Watershed: Journal of Environmental Quality, v. 38, no. 5, p. 1981-1988.

Sharpley, A. N., McDowell, R. W., Weld, J. L., and Kleinman, P. J. A., 2001, Assessing site vulnerability to phosphorus loss in an agricultural watershed: Journal of Environmental Quality, v. 30, no. 6, p. 2026-2036.

Sonmez, O., Pierzynski, G. M., Frees, L., Davis, B., Leikam, D., Sweeney, D. W., and Janssen, K. A., 2009, A field-based assessment tool for phosphorus losses in runoff in Kansas: Journal of Soil and Water Conservation, v. 64, no. 3, p. 212-222.

USDA, 1991, Soil Survey Geographic (SSURGO) Database: Data Use Information: Washington, DC.

Vadas, P. A., Gburek, W. J., Sharpley, A. N., Kleinman, P. A., Moore, P. A., Cabrera, M. L., and Harmel, R. D., 2007, A model for phosphorus transformation and runoff loss for surface-applied manures: Journal of Environment Quality, v. 36, p. 324-332.

Veith, T. L., Sharpley, A. N., Weld, J. L., and Gburek, W. J., 2005, Comparison of measured and simulated phosphorus losses with indexed site vulnerability: Transactions of the ASAE, v. 48, no. 2, p. 557-565.

White, M. J., Storm, D. E., Busteed, P. R., Smolen, M. D., Zhang, H., and Fox, G. A., 2010, A quantitative phosphorus loss assessment tool for agricultural fields: Environmental Modelling & Software, v. 25, no. 10, p. 1121-1129.

White, M. J., Storm, D. E., Smolen, M. D., Busteed, P. R., Zhang, H., and Fox, G. A., 2012a, Validation of a Quantitative Phosphorus Loss Assessment Tool: Journal of Environment Quality, v. 0, no. 0, p. 0.

White, M. W., Harmel, R. D., and Haney, R. L., 2012b, Development and validation of the Texas Best Management Practice Evaluation Tool (TBET): Journal of Soil and Water Conservation, v. 67, no. 6, p. 525-535.

Zambrano-Bigiarini, M., 2014, hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package version 0.3-8. http://CRAN.R-project.org/package=hydroGOF.

Table 2.1. Summary of study area physical characteristics.

Site	Field size	Duration	Annual rainfall	Annual runoff	Annual erosion	Fertilizer P_2O_5	Manure P_2O_5	Soil test P
	ha	yrs	cm	cm	tons ac^{-1}	lbs ac^{-1}	lbs ac^{-1}	mg kg^{-1}
Washington Co., AR	0.4	3	98 - 137	1.5 - 6.7	0.003 - 0.08	0	36 - 72	81 – 183
Putnam Co., GA	0.72 - 0.79	4	101 - 118	12 - 23	< 1	0	0 - 322	14 - 142
Henderson Co., NC	0.04	3	80 - 157	3.4 - 83	0 - 3.5	0 - 22.5	0 - 231	41 - 121

Table 2.2. Summary of study area crop systems and soils.

Site	Crop	Soil series (hydro group)			
		1	*2*	*3*	*4*
Washington Co., AR	Pasture	Captina (C)	Altavista (C)	Sedgefield (C)	Helena (C)
Putnam Co., GA	Pasture	Cecil (B)			
Henderson Co., NC	Corn with wheat cover	Delanco (C)			

Table 2.3. Site-years and number of observations for the uncalibrated, calibration, and validation data sets.

Site	# Fields	Years	Site-years	Number of measurements			
				Runoff	Sediment	Total P	Dissolved P
Uncalibrated, Calibration							
Washington Co., AR	7	2009, 2010	14	42	41	41	41
Putnam Co., GA	6	1995, 1998	12	119	0	116	119
Henderson Co., NC	5	2011, 2012	10	139	127	131	107
Validation							
Washington Co., AR	7	2011	7	48	48	48	48
Putnam Co., GA	6	1996, 1997	12	187	0	165	187
Henderson Co., NC	5	2013	5	245	134	160	152

Table 2.4. Statistical criteria for event-based observed versus predicted comparisons.

Constituent	NSE	PBIAS
Runoff	≥ 0.3	± 35
Erosion	≥ 0.3	± 60
P loss	≥ 0.3	± 60

Table 2.5. Parameters and ranges of values used for the Texas Best Management Evaluation Tool calibration.

Variable	Parameter	Description	Adjustment type	Calibration range
Runoff	CN	Curve number	absolute	(-16, -8, -4, -2, 0, +2, +4, +8, +16)
Erosion; sediment bound P	USLE_K	USLE soil erosion factor	relative	(-60%, -40%, -20%, 0, +20%, +40%, +60%)
	C_{min}	USLE crop factor	relative	(-60%, -40%, -20%, 0, +20%, +40%, +60%)
	ADJ_PKR	Peak rate adjustment factor	absolute	(0.25, 0.5, 0.75, 1, 1.25, 0.5, 1.75)
	SLSBBSN	Subbasin slope length	relative	(-60%, -40%, -20%, 0, +20%, +40%, +60%)
Dissolved P	PPERCO	P percolation coefficient	absolute	(5, 7.5, 10, 12.5, 15)
	PHOSKD	Soil P partitioning coefficient	absolute	(100, 125, 150, 175, 200)

Table 2.6. Probable error ranges as reported by Harmel et al. (2006) for "typical" water quality measurements.

Constituent	Probable error range (%)	
	Low	High
Runoff	6	19
Sediment	7	53
Total P	8	110
Dissolved P	8	104

Table 2.7. Uncalibrated and calibration goodness-of-fit statistics for TBET model predictions on field sites in Washington Co., AR, Putnam Co., GA, and Henderson Co., NC for years 2009 and 2010, 1995 and 1998, and 2011 and 2012, respectively.

Statistic*	Uncalibrated			Calibration		
	AR	GA	NC	AR	GA	NC
Runoff						
MAE, mm	20.54	9.74	6.72	13.82	9.33	6.8
NSE	-7	0.57	0.23	-2	0.58	0.33
PBIAS, %	-179	-48	-39	-112	-48	-6
Sediment						
MAE, ton ha^{-1}	0.02	-	0.53	0.01	-	0.07
NSE	-59	-	-88	-0.57	-	-0.67
PBIAS, %	-256	-	-963	53	-	16
Total P						
MAE, kg P ha^{-1}	0.17	0.35	0.89	0.12	0.3	0.17
NSE	-1.5	0.34	-78	0.01	0.51	-1.7
PBIAS, %	-55	20	-654	34	10	9
Dissolved P						
MAE, kg P ha^{-1}	0.13	0.32	0.01	0.1	0.29	0.01
NSE	-1.6	0.19	-0.08	-0.14	0.35	0.21
PBIAS, %	-35	11	-21	24	2	15

*MAE , mean absolute error; NSE, Nash-Sutcliffe efficiency; PBIAS, percent bias

Table 2.8. Calibration best-fit parameters for TBET predictions of runoff, sediment, total phosphorus, and dissolved phosphorus on field sites in Washington Co., AR, Putnam Co., GA, and Henderson Co., NC.

Parameter*	AR	GA	NC
CN_{II}	-16	+8	-16
C_{min}	-60%	default	-60%
SLSBBSN	+60%	default	-60%
ADJ_PKR	0.75	default	0.25
USLE_K	-60%	default	-60%
PPERCO	5	default	5
PHOSKD	200	200	200

*CN_{II}, Curve number (moisture condition II); $C_{min,}$ USLE minimum crop factor; SLSBBSN, subbasin slope length; ADJ_PKR, Peak rate adjustment factor; USLE_K, USLE erosion factor; PPERCO, phosphorus percolation coefficient; PHOSKD, phosphorus soil partitioning coefficient.

Table 2.9. Validation goodness-of-fit statistics without uncertainty estimates for TBET model predictions on field sites in Washington Co., AR, Putnam Co., GA, and Henderson Co., NC for years 2011, 1996 and 1997, and 2013, respectively.

Statistic*	Validation, no uncertainty		
	AR	GA	NC
Runoff			
MAE, mm	18.66	6.05	6.89
NSE	0.35	0.57	0.65
PBIAS, %	-81.1	-40.1	-41.8
Sediment			
MAE, ton ha^{-1}	0.01	-	0.13
NSE	0.24	-	0.31
PBIAS, %	39.4	-	12.2
Total P			
MAE, kg P ha^{-1}	0.25	0.59	0.22
NSE	0.27	-0.02	0.35
PBIAS, %	56.8	69.5	-16.9
Dissolved P			
MAE, kg P ha^{-1}	0.24	0.42	0.04
NSE	0.27	0.38	0.32
PBIAS, %	59.9	57.6	46.1

*MAE, mean absolute error; NSE, Nash-Sutcliffe efficiency; PBIAS, percent bias

Table 2.10. Validation goodness-of-fit statistics for TBET model predictions with no uncertainty, low uncertainty, and high uncertainty in measured data and model predictions on field sites in Washington Co., AR, Putnam Co., GA, and Henderson Co., NC for years 2011, 1996 and 1997, and 2013, respectively.

Statistic*	AR			GA			NC		
	None	Low	High	None	Low	High	None	Low	High
Runoff									
MAE, mm	18.66	18.66	18.32	6.05	6.04	5.87	6.89	6.89	6.69
NSE	0.35	0.35	0.36	0.57	0.57	0.58	0.65	0.65	0.65
PBIAS, %	-81.1	-81.1	-79.8	-40.1	-40.1	-39.7	-41.8	-41.8	-41.7
Overlapping PDFs (% of dataset)		1 (2%)	8 (17%)		11 (6%)	67 (36%)		8 (3%)	73 (30%)
Sediment									
MAE, ton ha^{-1}	0.01	0.01	0.01	-	-	-	0.13	0.13	0.13
NSE	0.24	0.24	0.25	-	-	-	0.31	0.31	0.32
PBIAS, %	39.4	39.4	40.7	-	-	-	12.2	12.2	11.8
Overlapping PDFs (% of dataset)		0 (0%)	8 (17%)	-	-	-		1 (<1%)	33 (25%)
Total P									
MAE, kg P ha^{-1}	0.25	0.25	0.25	0.59	0.59	0.58	0.22	0.22	0.21
NSE	0.27	0.27	0.27	-0.02	-0.02	-0.01	0.35	0.35	0.4
PBIAS, %	56.8	56.8	57.1	69.5	69.5	69.4	-16.9	-16.9	-17.3
Overlapping PDFs (% of dataset)		1 (2%)	11 (23%)		4 (2%)	139 (75%)		6 (4%)	75 (47%)
Dissolved P									
MAE, kg P ha^{-1}	0.24	0.24	0.23	0.42	0.42	0.41	0.04	0.04	0.03
NSE	0.27	0.27	0.31	0.38	0.38	0.44	0.32	0.32	0.4
PBIAS, %	59.9	59.9	57.6	57.6	57.6	54.9	46.1	46.1	43.4
Overlapping PDFs (% of dataset)		1 (2%)	21 (44%)		7 (4%)	154 (82%)		7 (5%)	89 (59%)

*PDF, probability density function; MAE, mean absolute error; NSE, Nash-Sutcliffe efficiency; PBIAS, percent bias

Table 2.11. Kendall-Theil robust line regressions and goodness-of-fit statistics for measured average annual P loss versus P-index, uncalibrated, and validated Texas Best Management Practice Evaluation Tool 25-year average annual total phosphorus loss for Putnam Co., GA and Henderson Co., NC.

Statistics*		P-index	Uncalibrated TBET	Validated TBET
KTR	slope	1.67	-2.2	0.96
	intercept	1.65	35.63	-1.26
MAE, kg ha^{-1}		6.9	20.6	4.4
PBIAS, %		-78	-320	-6.3

*KTR, Kendall-Theil robust line; MAE, mean absolute error; PBIAS, percent bias

Table 2.12. P-index loads, values, and ratings and 25-year average annual P loss predictions for TBET using default parameters and best-fit parameters for selected field-years in Putnam Co., GA and Henderson Co., NC.

Site	Field	Year	Average annual measured P loss	Default TBET	Validated TBET	P-index load	P-index value	P-index rating*
			(kg ha^{-1})	(kg ha^{-1})	(kg ha^{-1})	(kg ha^{-1})		
GA	P2	1995	6.02	12.51	13.48	19.6	196	VH
	P2	1996	6.02	16.22	17.56	21.0	210	VH
	P4	1995	8.81	10.95	11.82	10.2	102	VH
	P4	1996	8.81	14.43	15.62	14.1	141	VH
	P6	1995	13.41	12.79	13.78	22.0	220	VH
	P6	1996	13.41	16.64	18.05	29.7	297	VH
NC	CTC	2011	6.85	31.31	0.21	0.7	7	L
	CTC	2012	6.85	31.43	0.25	0.8	8	L
	NTC	2011	1.30	30.68	0.48	0.9	9	L
	NTC	2012	1.30	24.59	0.41	0.8	8	L
	CTO	2011	5.41	51.38	0.59	12.5	125	VH
	CTO	2012	5.41	44.26	0.47	9.0	90	H
	NTO	2011	3.06	48.84	1.83	11.8	118	VH
	NTO	2012	3.06	31.09	0.86	6.8	68	M

*L, low; M, medium; H, high; VH, very high

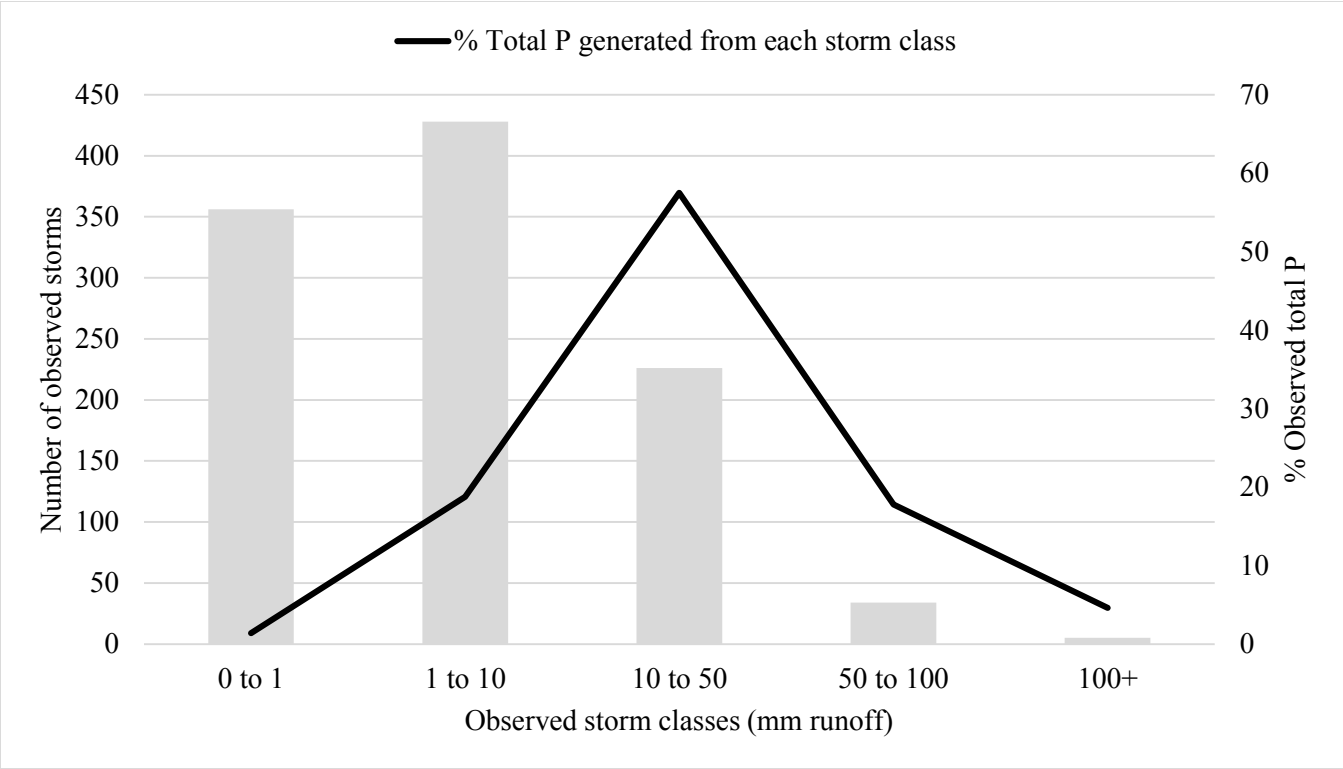


Figure 2.1. Percent observed total P generated from each observed storm class for all sites combined.

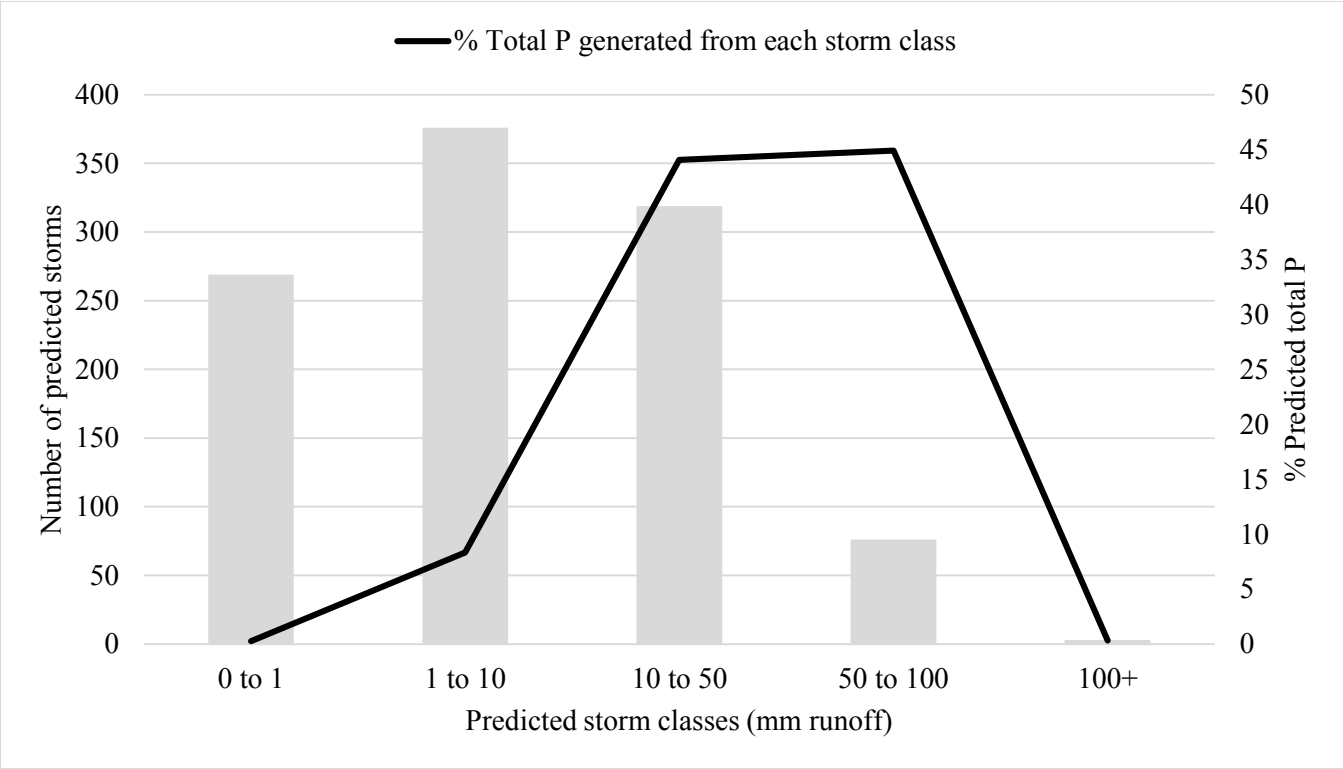
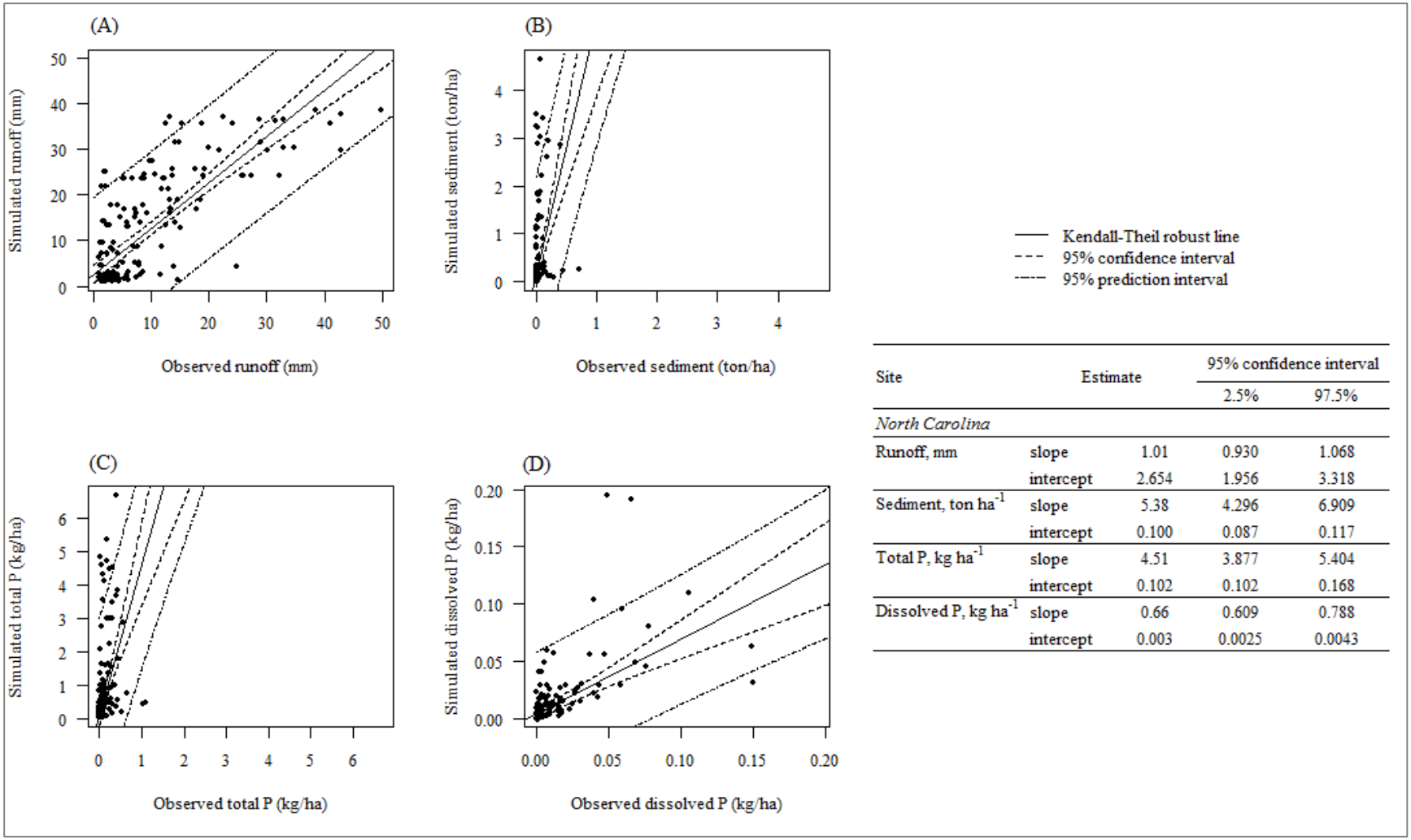
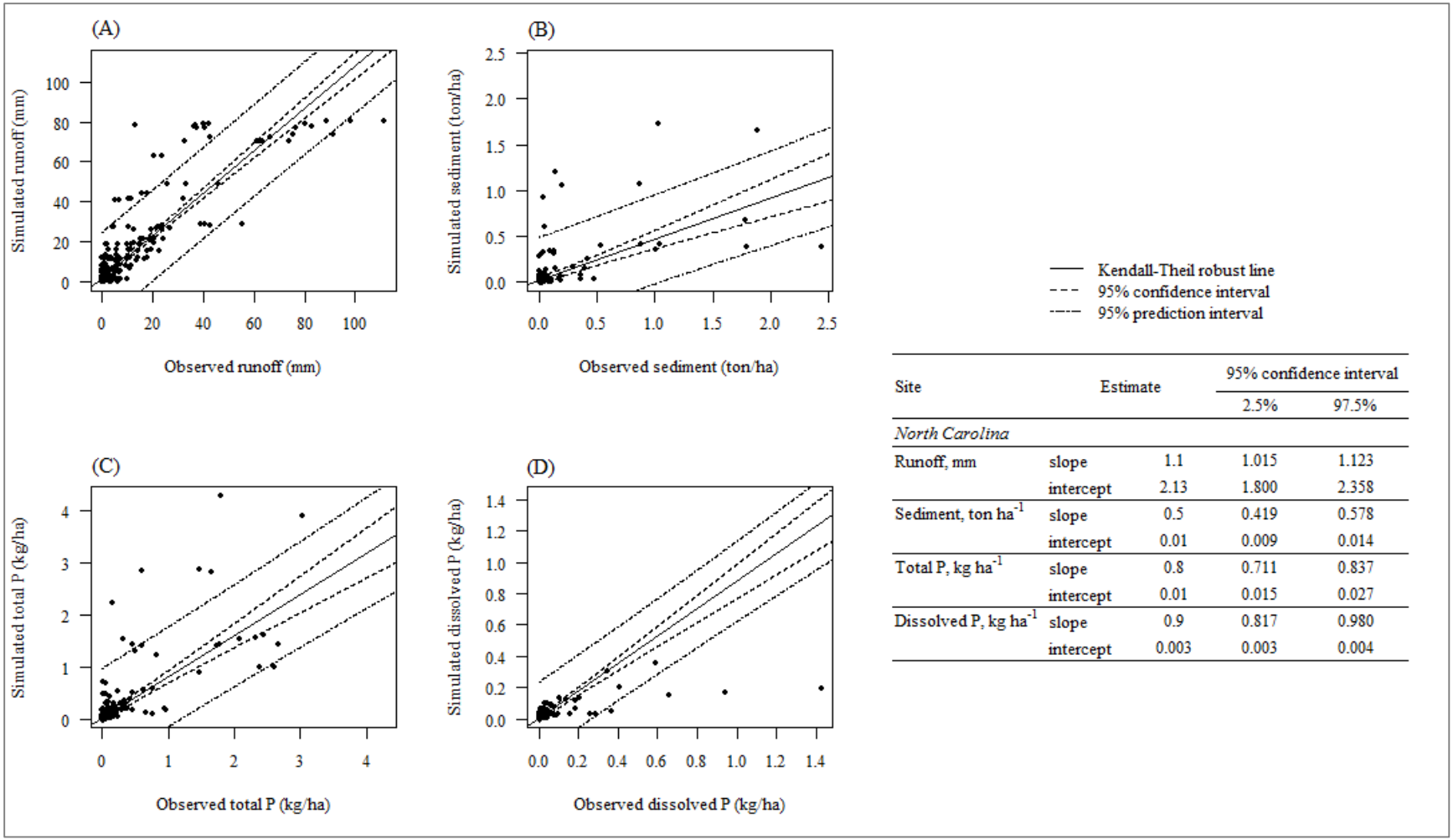


Figure 2.2. Percent of predicted total P generated from each predicted storm class for all sites combined.



Site		Estimate	95% confidence interval	
			2.5%	97.5%
North Carolina				
Runoff, mm	slope	1.01	0.930	1.068
	intercept	2.654	1.956	3.318
Sediment, ton ha^{-1}	slope	5.38	4.296	6.909
	intercept	0.100	0.087	0.117
Total P, kg ha^{-1}	slope	4.51	3.877	5.404
	intercept	0.102	0.102	0.168
Dissolved P, kg ha^{-1}	slope	0.66	0.609	0.788
	intercept	0.003	0.0025	0.0043

Figure 2.3. Measured event-based and Texas Best Management Practice Evaluation Tool predicted runoff (A), sediment (B), total phosphorus (C), and dissolved phosphorus (D) for the Henderson Co., NC uncalibrated dataset. The statistics of the Kendall-Theil robust line regressions are listed in the table.



Site		Estimate	95% confidence interval	
			2.5%	97.5%
North Carolina				
Runoff, mm	slope	1.1	1.015	1.123
	intercept	2.13	1.800	2.358
Sediment, ton ha^{-1}	slope	0.5	0.419	0.578
	intercept	0.01	0.009	0.014
Total P, kg ha^{-1}	slope	0.8	0.711	0.837
	intercept	0.01	0.015	0.027
Dissolved P, kg ha^{-1}	slope	0.9	0.817	0.980
	intercept	0.003	0.003	0.004

Figure 2.4. Measured event-based and Texas Best Management Practice Evaluation Tool predicted runoff (A), sediment (B), total phosphorus (C), and dissolved phosphorus (D) for the Henderson Co., NC validation dataset. The statistics of the Kendall-Theil robust line regressions are listed in the table.

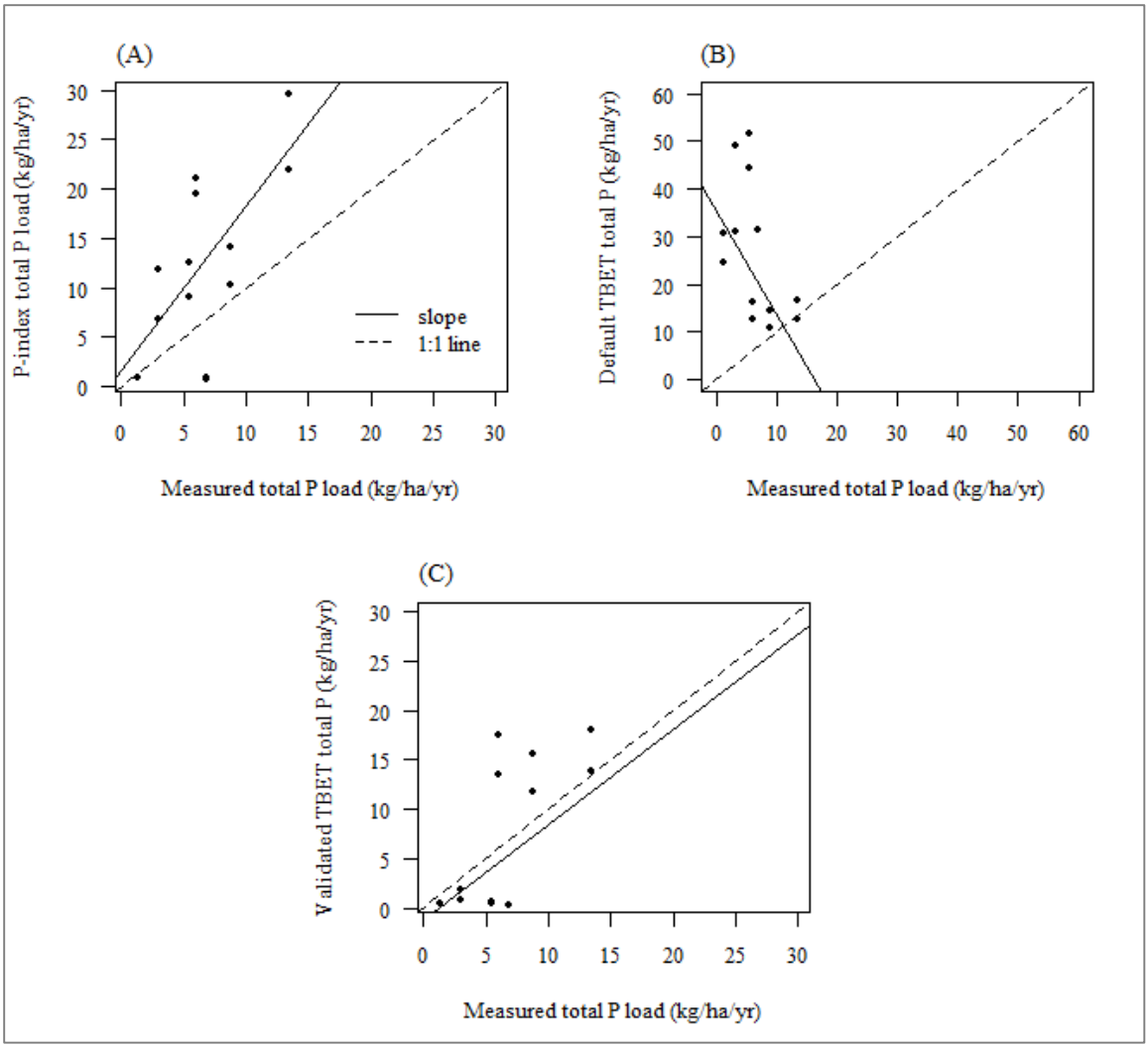


Figure 2.5. Measured average annual P loss versus P-index (A), default (B) and validated (C) Texas Best Management Practice Evaluation Tool 25-year average annual total phosphorus loss for Putnam Co., GA and Henderson Co., NC

APPENDIX A

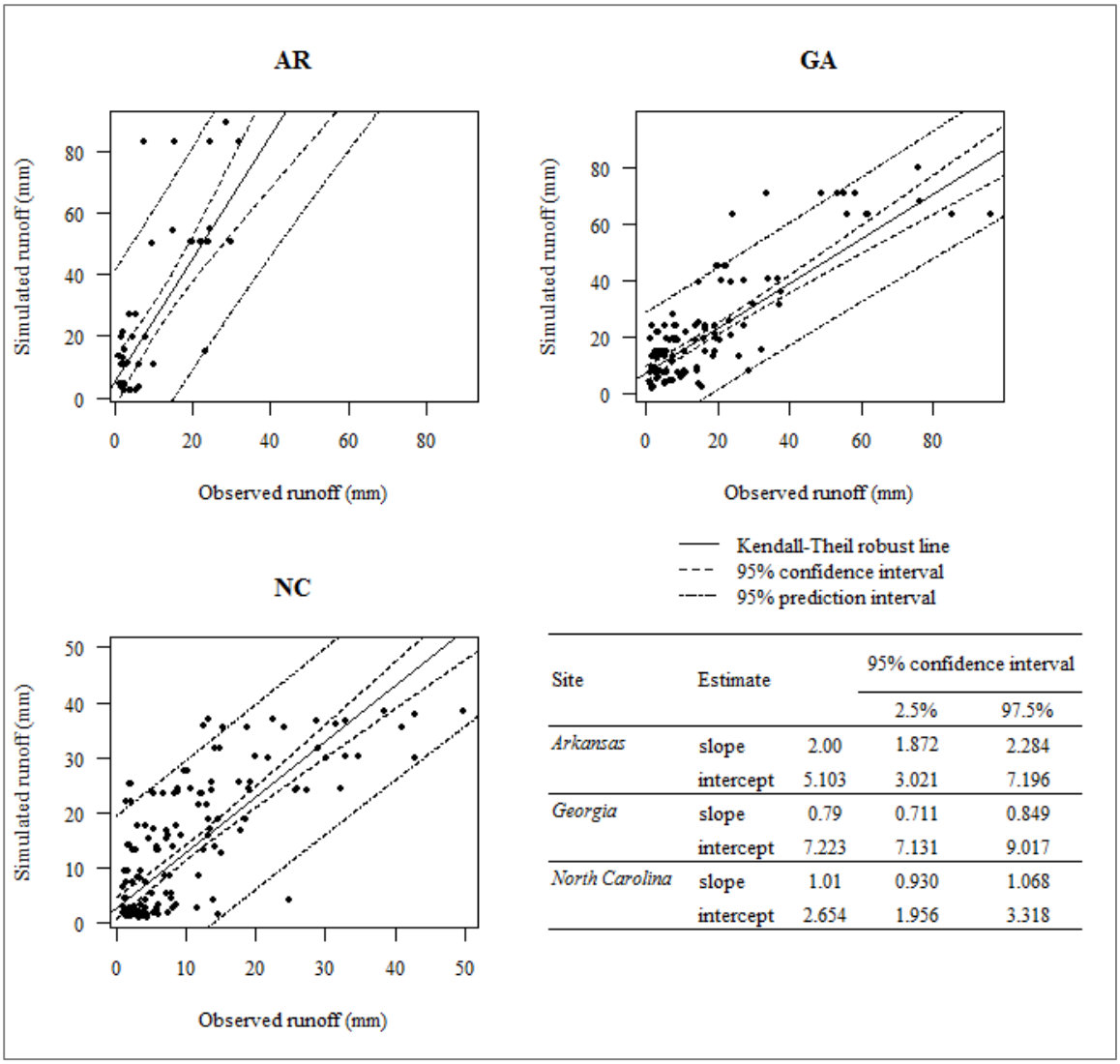
SITE DESCRIPTIONS AND SAMPLE COLLECTION METHODS

Table A.1 Site descriptions and sample collection, preservation, and analysis information.

Site	Coordinates	Site description	Sample collection/preservation	Reference
Washinton Co., AR	36°04'51.4"N, 94°17'23.0"W	Seven 0.4-ha permanent pasture fields were monitored between spring 2009 and fall 2011. The only soil series present at the site was Captina (fine-silty, siliceous, active, mesic Typic Fragiudults). The site contained one control field. The other six fields were unique treatments: (1) continuous grazing with broadcast litter; (2) rotational grazing with broadcast litter; (3) hay production with broadcast litter; (4) replicate hay production with broadcast litter; (5) hay production with a single application of injected litter; and (6) hay with two applications of injected litter. Each treatment, except the two control treatments and HI2, received 537 kg ha^{-1} litter in mid-May of each year. The Mehlich III STP for all site years ranged from 73 to 150 mg kg^{-1}.	A 0.45-m H-flume equipped with an ISCO and 720 flow transducer was used to collect flow-weighted samples and measure flow automatically. A 100 mL water sample was collected every 500 gallons of runoff. Water samples were refrigerated (4°C) within eight hours of each runoff event. The samples were then brought to laboratory and were filtered within 24 hours of sample collection. DRP Samples were syringe-filtered through 0.45 micron millipore membrane and were analyzed by the ascorbic acid-molybdate blue method on a Lachet autoanalyzer. Analysis of TSS and TP occurred within 14 days of sample collection. Total P and TSS were analyzed by nitric acid microwave digestion and gravimetric analysis, respectively.	Andrew Sharpley, sharpley@uark.edu
Putnam Co., GA	33°25'05.4"N, 83°29'15.5"W	Six 0.75-ha fescue-common bermudagrass paddocks were sampled between January 1995 and April 1998 for surface runoff and P loads. Sediment yield was not measured. Soil series at the site were Cecil (fine, kaolinitic, thermic Typic Kanhapludults), Altavista (fine-loamy, mixed, semiactive, thermic Aquic Hapludults), Helena (fine, mixed, semiactive, thermic Aquic Hapludults), and Sedgefield (fine. mixed, active, thermic Aqualtic Hapludalfs). The plots had relatively low STP (14 to 142 ppm) and high rates of poultry litter application. Poultry litter was applied in March and September-October 1995 and 1996, while urea-ammonium nitrogen solution was applied in March 1997 and 1998. The maximum total P added on each paddock was 159 kg ha^{-1}. In addition to fertilizer applications, each paddock received put and take stocking management.	A 0.45-m H-flume equipped with a SENIX ultrasonic sensor was used to measure depth of surface runoff from each the six plot separately. A 0.6-m Coshocton wheel subsampled the surface runoff which were subsequently stored in an ISCO 3700FR refrigerated sampler. Precipitation and runoff volume were recorded by CR10 dataloggers (Campbell Scientific, Logan, UT). All samples were refrigerated (4°C) for up to 24hrs, brought to laboratory, filtered and either analyzed immediately or frozen for later analysis. Runoff samples were vacuum-filtered through Whatman 0.45-µm cellulose nitrate membranes. Filtered samples were analyzed for DRP by the acid-molybdate blue method.	Pierson et al., 2001a; Pierson et al., 2001b
Henderson Co., NC	35°25'36.4"N, 82°33'51.5"W	The soil series at the site was Delanco (fine-loamy, mixed, semiactive, mesic Aquic Hapludults). Surface runoff, TP, and DP were sampled from twenty 0.017-ha, corn-wheat rotation plots (four replications of five treatments) from January 2011 through December 2013. The five treatments include: (1) conventional tillage with pesticide management denoted by CTP; (2) conservation no-till with pesticide management (NTP); (3) conventional tillage with organic management (CTO); (4) conservation no-till with organic management (NTO); and (5) conventional tillage with no fertilizer or pesticide applications (Control). Mehlich III STP on all treatments ranged from 41 to 121 ppm. Organic treatments received 80 kg ha^{-1} P through commercial fertilizer, while conventional treatments received no additional P applications.	A 0.45-m H-flume or v-notch weir equipped with an ISCO 6700 flow meter was used for measurement of flow and collecting flow-weighted samples. Runoff volume for storms with equipment malfunction was estimated using a 50-storm rainfall-runoff regression relationship established from storms where the data was collected. Runoff samples were retrieved within 48 hours of a storm. TP samples preserved with sulfuric acid. DRP samples were filtered within 48 hours and frozen for preservation. TSS and TP samples were refrigerated (4°C). Analysis of TSS, TP, and DRP occurred within 14 days of sample collection. TP was digested with a Kjeldatherm block digestion system and analyzed with a Quick Chem 8000 Automated Ion Analyzer. DRP was filtered with 0.45-µm membrane and analyzed with Quick Chem 8000 Automated Ion Analyzer. TSS concentrations were measured from non-acidified samples using the 2540D method.	Larsen et al., 2014

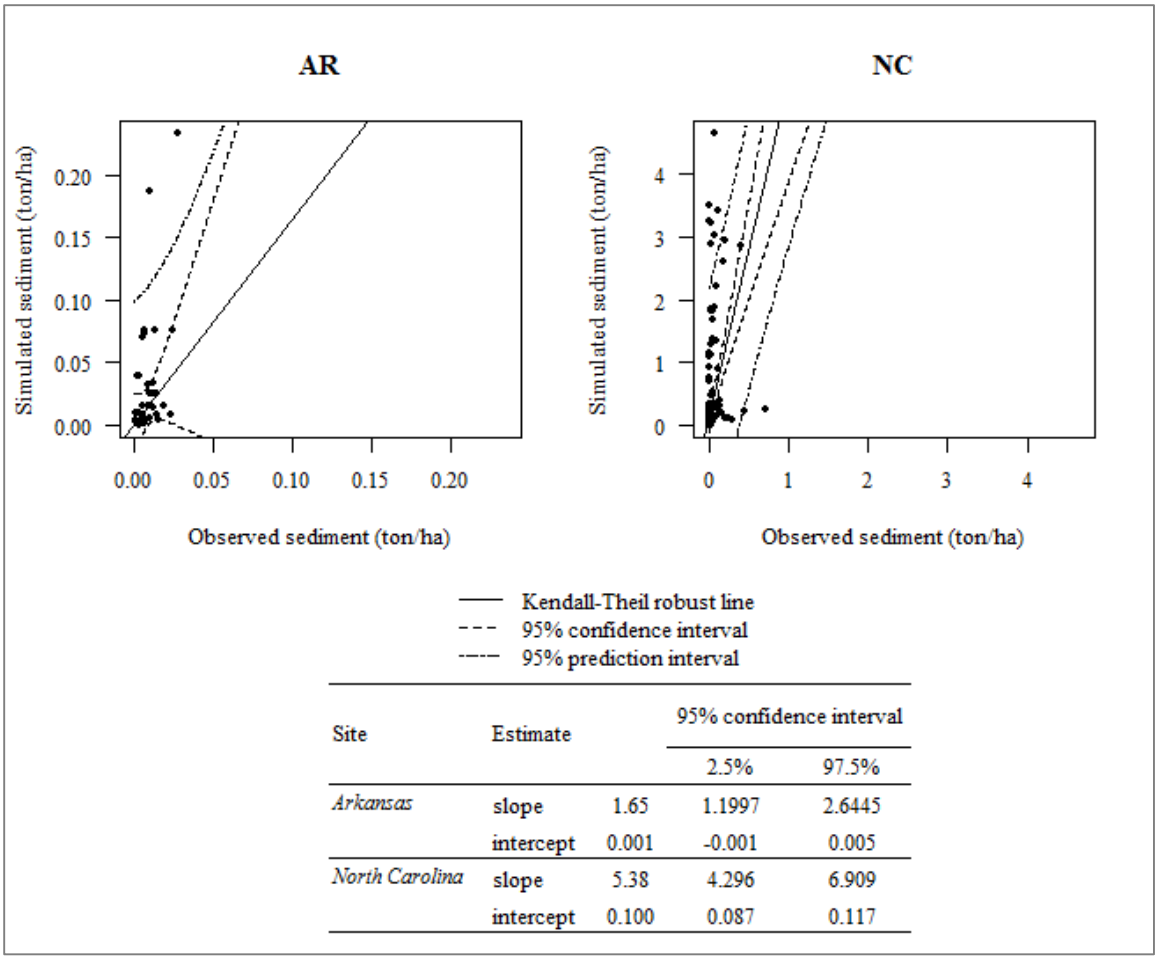
APPENDIX B

BASELINE PERFORMANCE SCATTER PLOTS AND KENDALL-THEIL ROBUST LINE STATISTICS



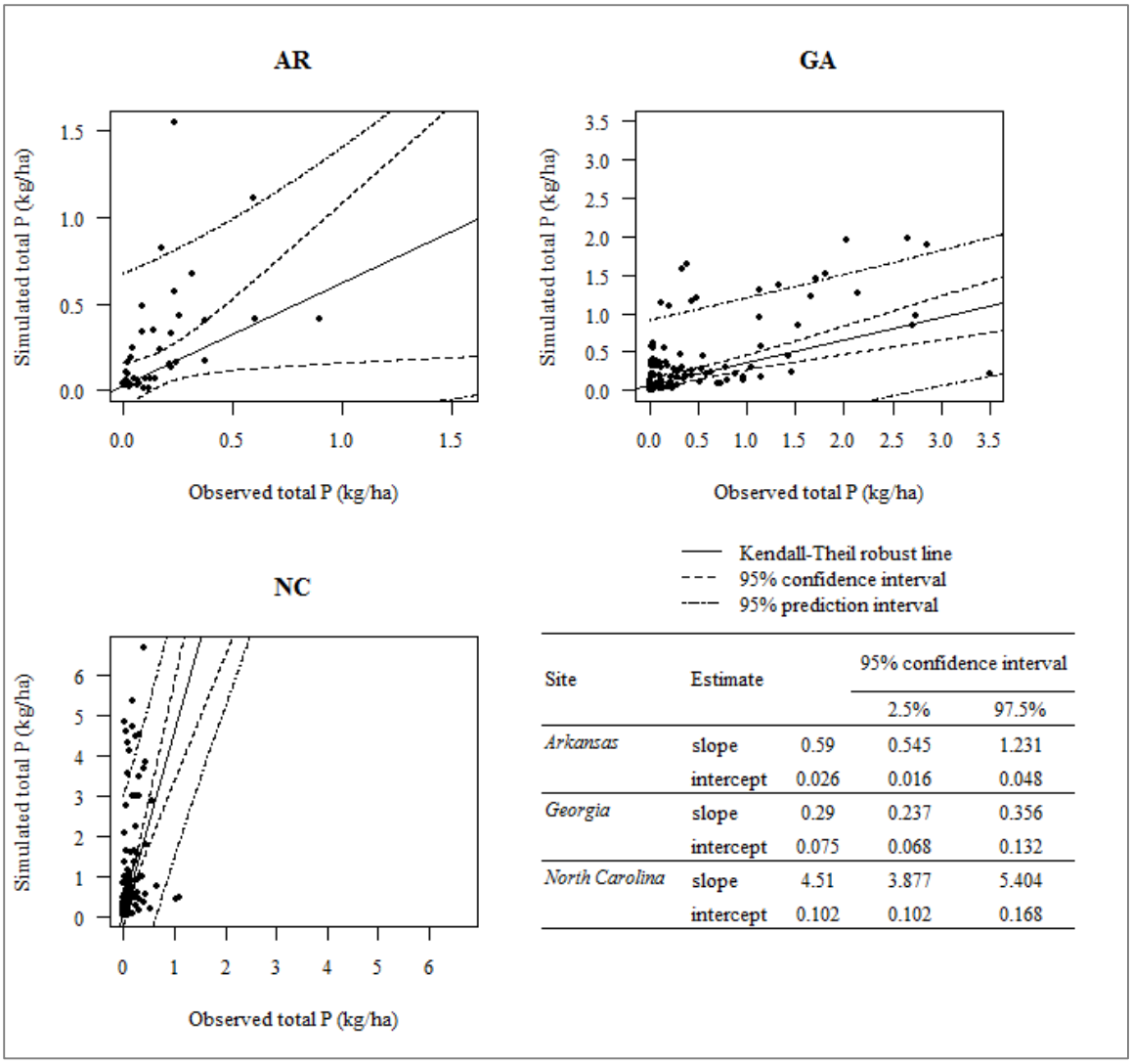
Site	Estimate		95% confidence interval	
			2.5%	97.5%
Arkansas	slope	2.00	1.872	2.284
	intercept	5.103	3.021	7.196
Georgia	slope	0.79	0.711	0.849
	intercept	7.223	7.131	9.017
North Carolina	slope	1.01	0.930	1.068
	intercept	2.654	1.956	3.318

Figure B.1. Kendall-Theil robust line regression statistics and scatter plots for event-based measured and Texas Best Management Practice Evaluation Tool predicted runoff using the baseline data sets and default parameters.



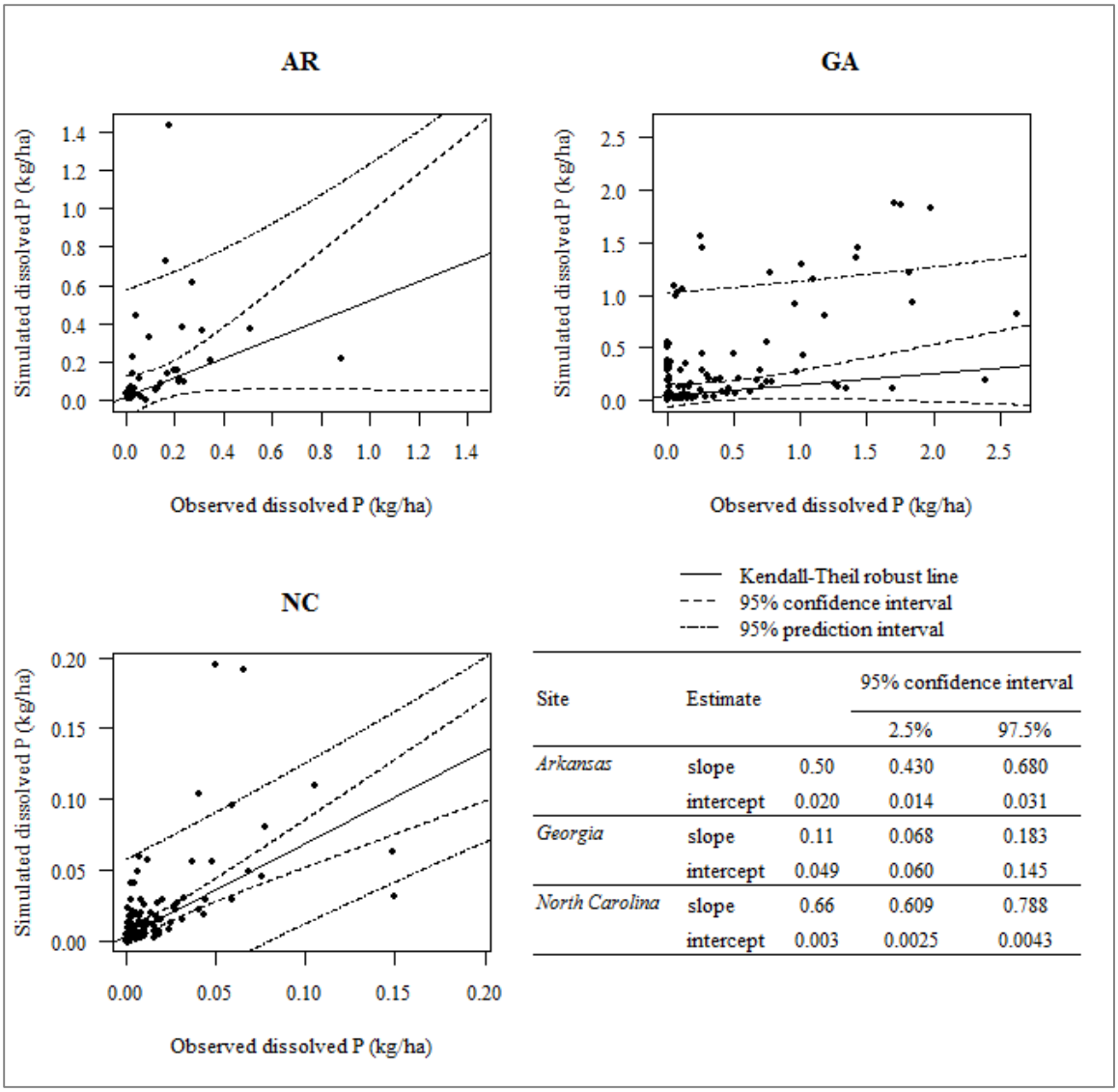
Site	Estimate		95% confidence interval	
			2.5%	97.5%
Arkansas	slope	1.65	1.1997	2.6445
	intercept	0.001	-0.001	0.005
North Carolina	slope	5.38	4.296	6.909
	intercept	0.100	0.087	0.117

Figure B.2. Kendall-Theil robust line regression statistics and scatter plots for event-based measured and Texas Best Management Practice Evaluation Tool predicted sediment using the baseline data sets and default parameters.



Site	Estimate		95% confidence interval	
			2.5%	97.5%
Arkansas	slope	0.59	0.545	1.231
	intercept	0.026	0.016	0.048
Georgia	slope	0.29	0.237	0.356
	intercept	0.075	0.068	0.132
North Carolina	slope	4.51	3.877	5.404
	intercept	0.102	0.102	0.168

Figure B.3. Kendall-Theil robust line regression statistics and scatter plots for event-based measured and Texas Best Management Practice Evaluation Tool predicted total phosphorus using the baseline data sets and default parameters.

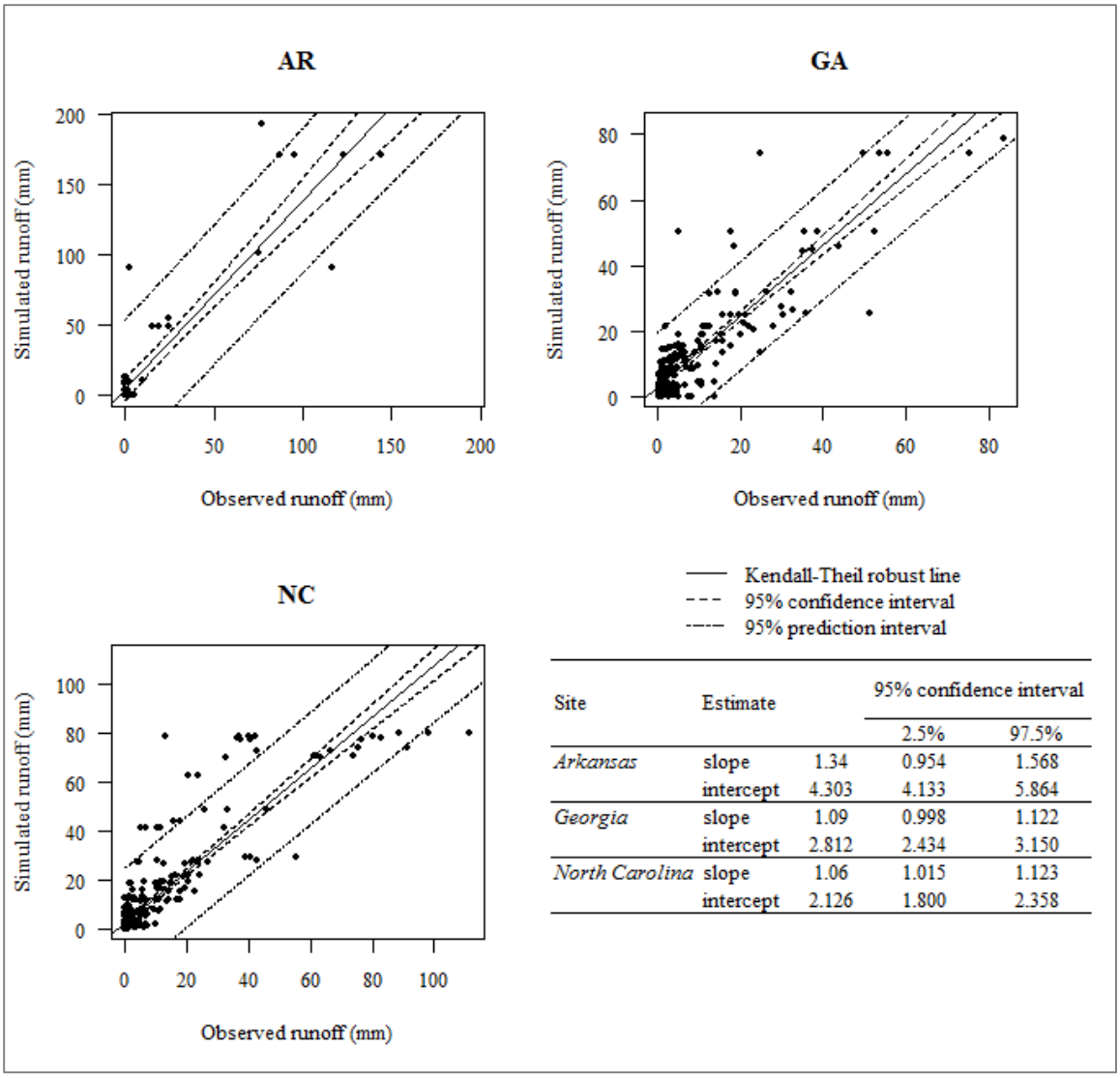


Site	Estimate		95% confidence interval	
			2.5%	97.5%
Arkansas	slope	0.50	0.430	0.680
	intercept	0.020	0.014	0.031
Georgia	slope	0.11	0.068	0.183
	intercept	0.049	0.060	0.145
North Carolina	slope	0.66	0.609	0.788
	intercept	0.003	0.0025	0.0043

Figure B.4. Kendall-Theil robust line regression statistics and scatter plots for event-based measured and Texas Best Management Practice Evaluation Tool predicted total phosphorus using the baseline data sets and default parameters.

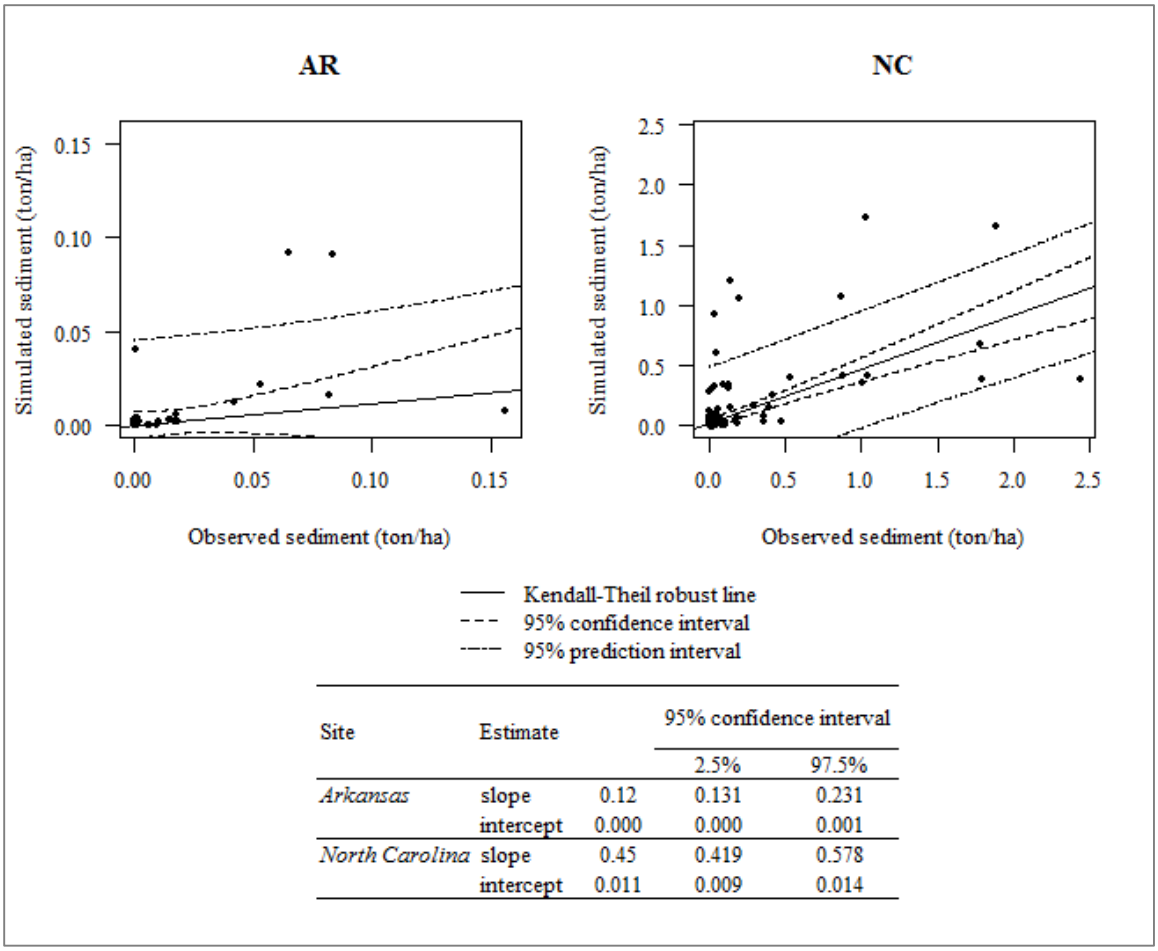
APPENDIX C

VALIDATION PERFORMANCE SCATTER PLOTS AND KENDALL-THEIL ROBUST LINE STATISTICS



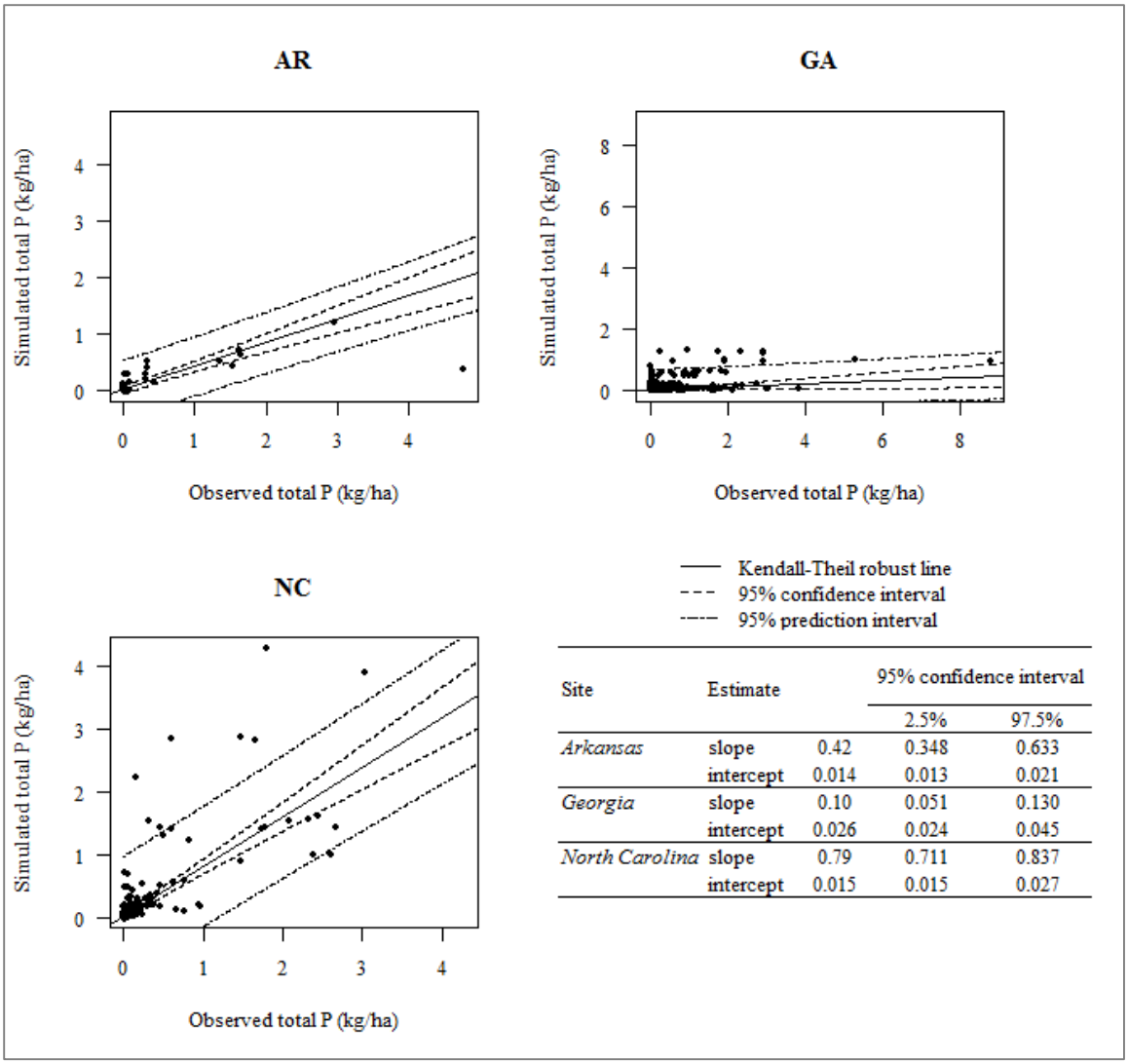
Site	Estimate		95% confidence interval	
			2.5%	97.5%
Arkansas	slope	1.34	0.954	1.568
	intercept	4.303	4.133	5.864
Georgia	slope	1.09	0.998	1.122
	intercept	2.812	2.434	3.150
North Carolina	slope	1.06	1.015	1.123
	intercept	2.126	1.800	2.358

Figure C.1. Kendall-Theil robust line regression statistics and scatter plots for event-based measured and Texas Best Management Practice Evaluation Tool predicted runoff using the validation data sets and calibration best-fit parameters.



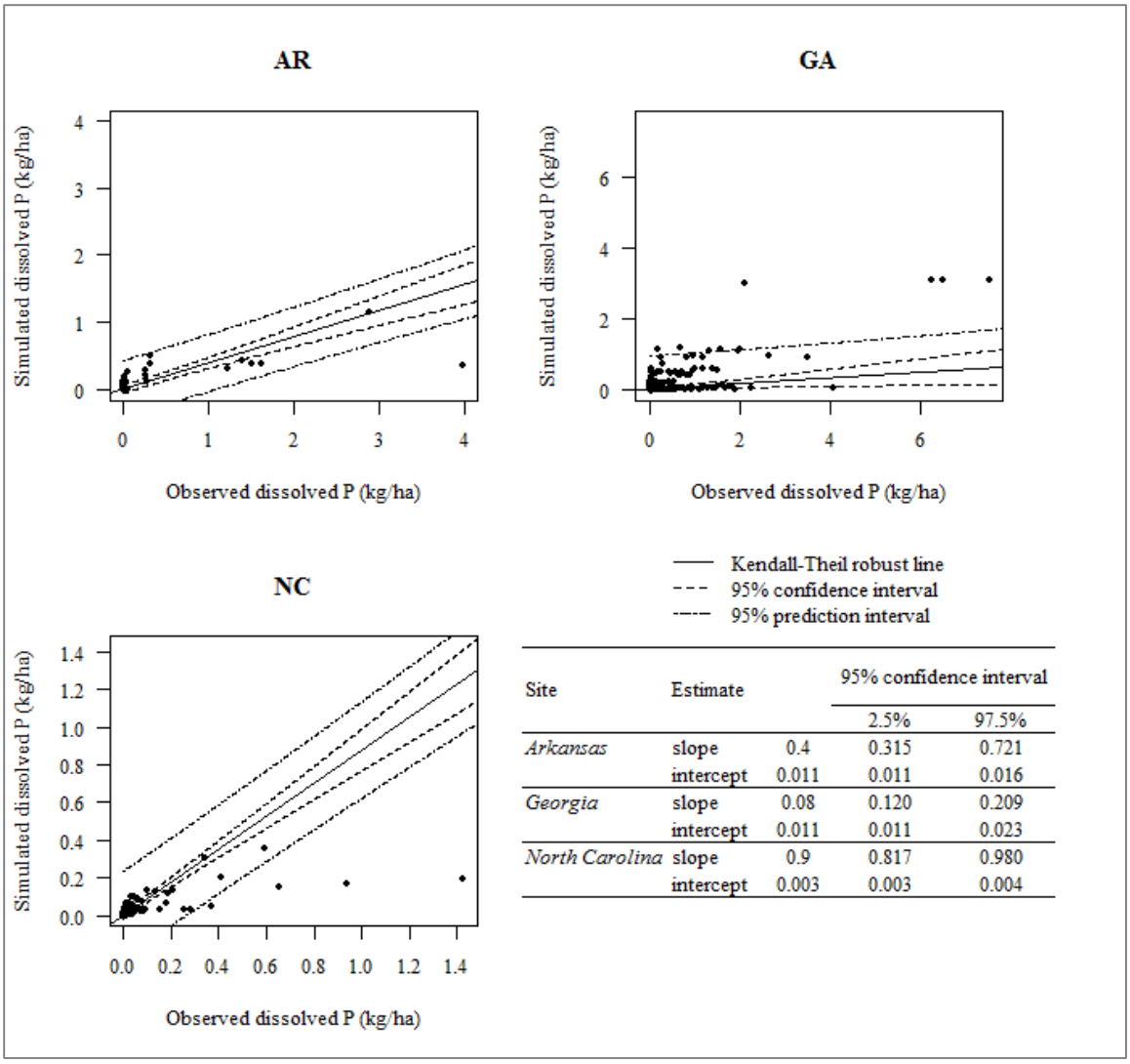
Site	Estimate		95% confidence interval	
			2.5%	97.5%
Arkansas	slope	0.12	0.131	0.231
	intercept	0.000	0.000	0.001
North Carolina	slope	0.45	0.419	0.578
	intercept	0.011	0.009	0.014

Figure C.2. Kendall-Theil robust line regression statistics and scatter plots for event-based measured and Texas Best Management Practice Evaluation Tool predicted sediment using the validation data sets and calibration best-fit parameters.



Site	Estimate		95% confidence interval	
			2.5%	97.5%
Arkansas	slope	0.42	0.348	0.633
	intercept	0.014	0.013	0.021
Georgia	slope	0.10	0.051	0.130
	intercept	0.026	0.024	0.045
North Carolina	slope	0.79	0.711	0.837
	intercept	0.015	0.015	0.027

Figure C.3. Kendall-Theil robust line regression statistics and scatter plots for event-based measured and Texas Best Management Practice Evaluation Tool predicted total phosphorus using the validation data sets and calibration best-fit parameters.

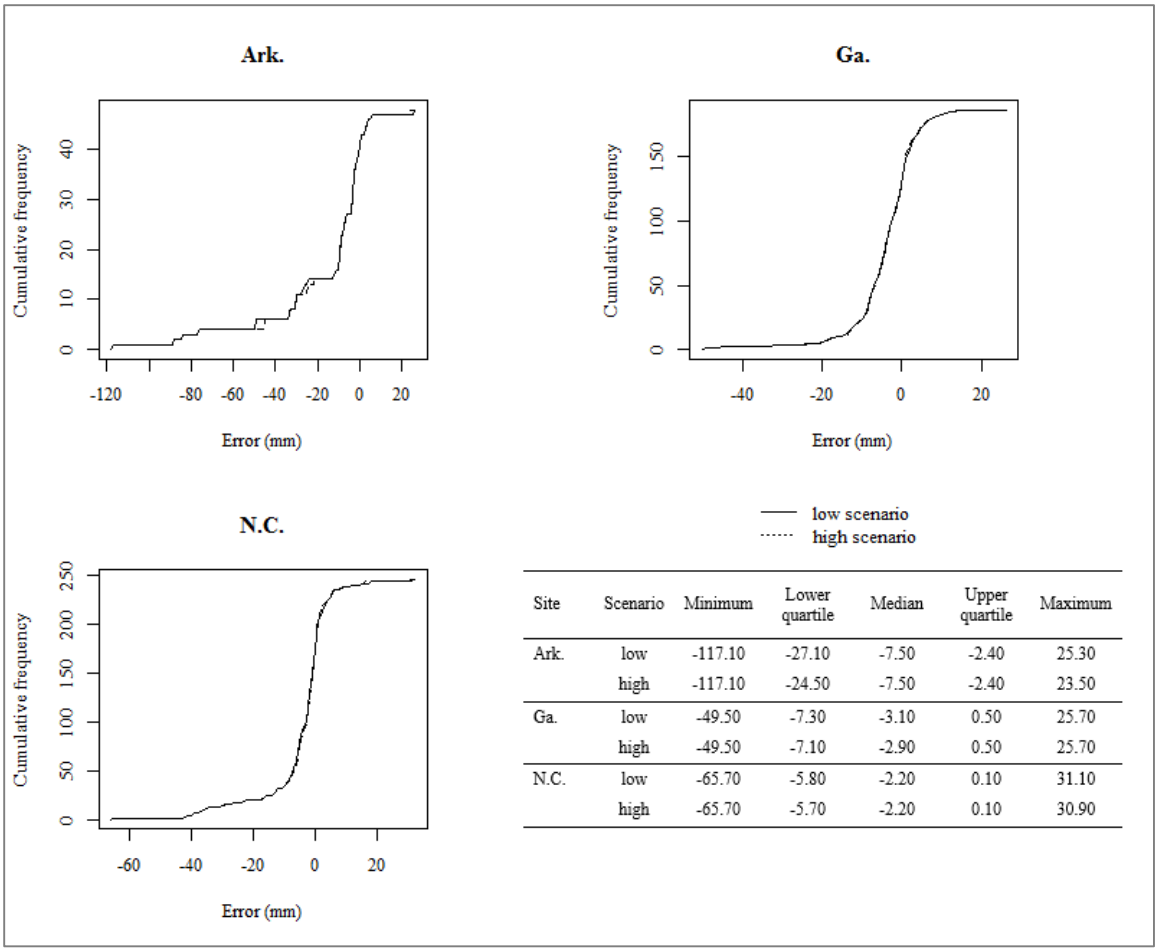


Site	Estimate		95% confidence interval	
			2.5%	97.5%
Arkansas	slope	0.4	0.315	0.721
	intercept	0.011	0.011	0.016
Georgia	slope	0.08	0.120	0.209
	intercept	0.011	0.011	0.023
North Carolina	slope	0.9	0.817	0.980
	intercept	0.003	0.003	0.004

Figure C.4. Kendall-Theil robust line regression statistics and scatter plots for event-based measured and Texas Best Management Practice Evaluation Tool predicted dissolved phosphorus using the validation data sets and calibration best-fit parameters.

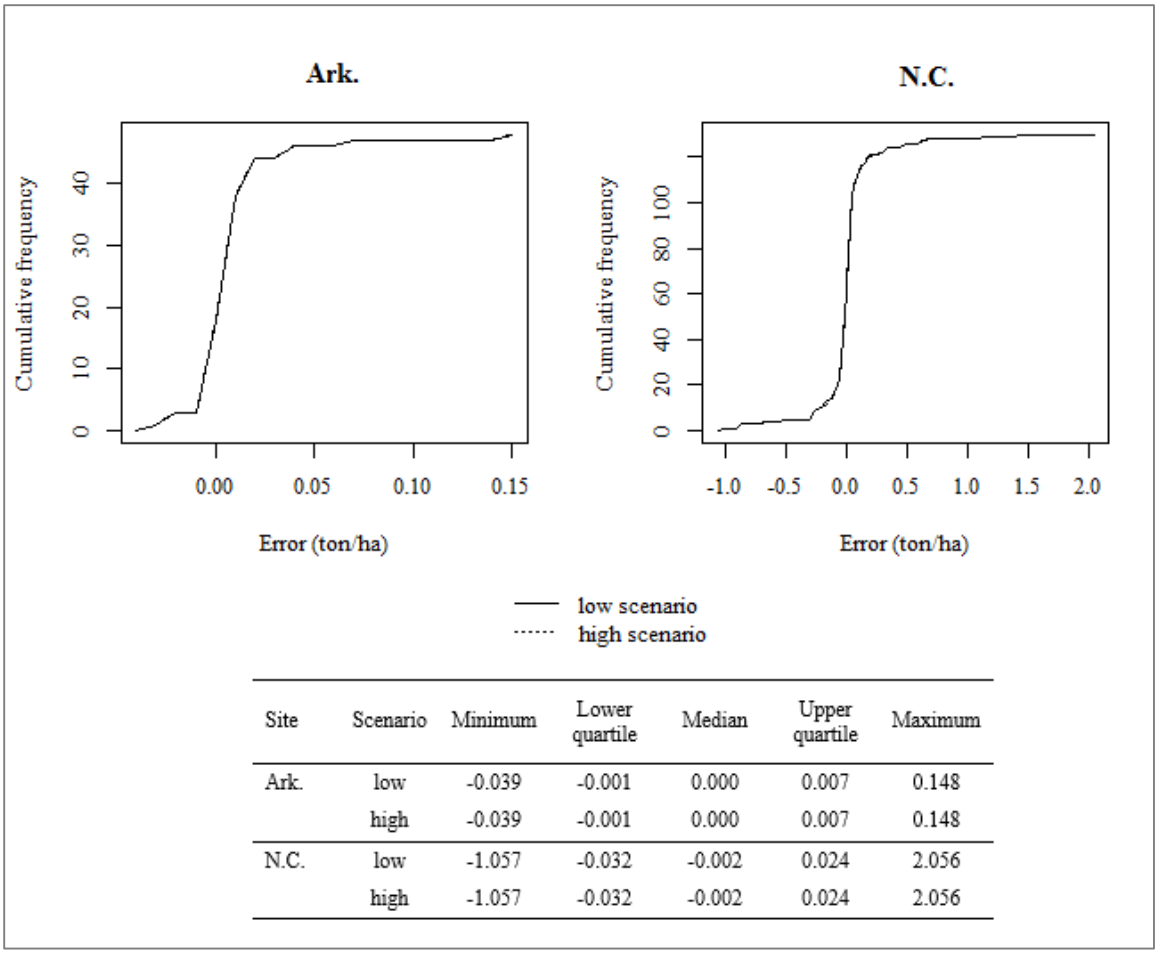
APPENDIX D

CUMULATIVE FREQUENCY DISTRIBUTIONS OF ABSOLUTE ERRORS FOR VALIDATION DATA SETS WITH LOW AND HIGH UNCERTAINTY SCENARIOS



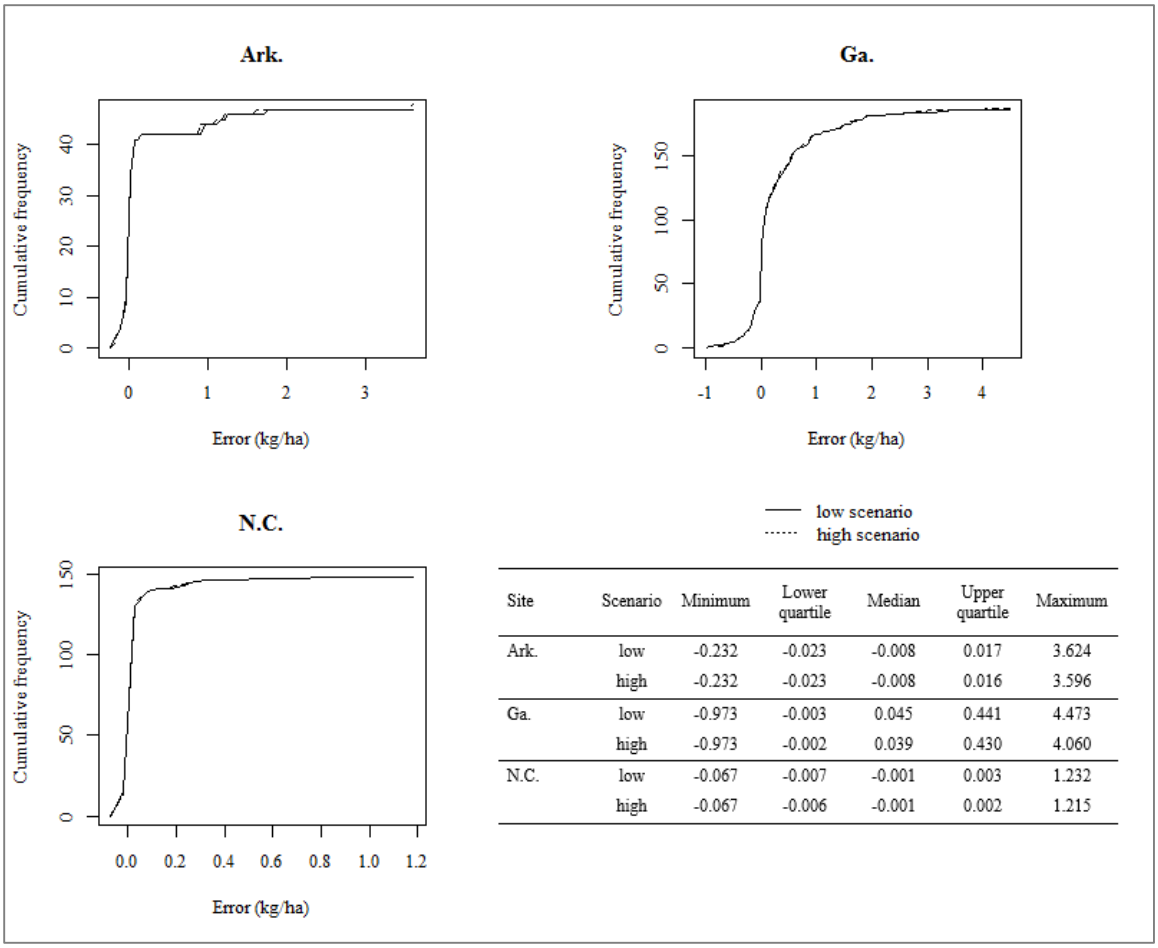
Site	Scenario	Minimum	Lower quartile	Median	Upper quartile	Maximum
Ark.	low	-117.10	-27.10	-7.50	-2.40	25.30
	high	-117.10	-24.50	-7.50	-2.40	23.50
Ga.	low	-49.50	-7.30	-3.10	0.50	25.70
	high	-49.50	-7.10	-2.90	0.50	25.70
N.C.	low	-65.70	-5.80	-2.20	0.10	31.10
	high	-65.70	-5.70	-2.20	0.10	30.90

Figure D.1. Cumulative frequency and distribution of absolute errors for the low and high uncertainty scenarios of event-based runoff using the validation data sets and calibration best-fit parameters.



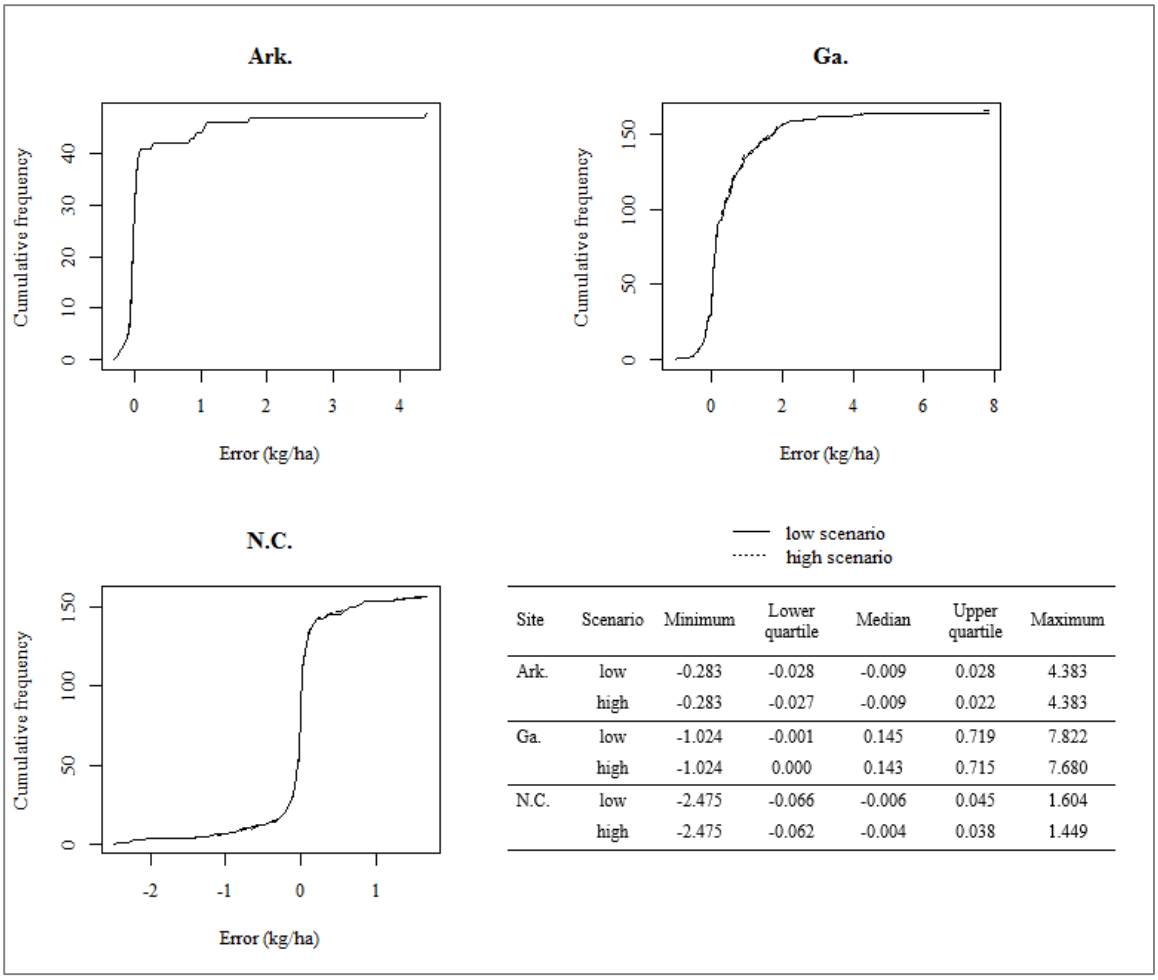
Site	Scenario	Minimum	Lower quartile	Median	Upper quartile	Maximum
Ark.	low	-0.039	-0.001	0.000	0.007	0.148
	high	-0.039	-0.001	0.000	0.007	0.148
N.C.	low	-1.057	-0.032	-0.002	0.024	2.056
	high	-1.057	-0.032	-0.002	0.024	2.056

Figure D.2. Cumulative frequency and distribution of absolute errors for the low and high uncertainty scenarios of event-based sediment using the validation data sets and calibration best-fit parameters.



Site	Scenario	Minimum	Lower quartile	Median	Upper quartile	Maximum
Ark.	low	-0.232	-0.023	-0.008	0.017	3.624
	high	-0.232	-0.023	-0.008	0.016	3.596
Ga.	low	-0.973	-0.003	0.045	0.441	4.473
	high	-0.973	-0.002	0.039	0.430	4.060
N.C.	low	-0.067	-0.007	-0.001	0.003	1.232
	high	-0.067	-0.006	-0.001	0.002	1.215

Figure D.3. Cumulative frequency and distribution of absolute errors for the low and high uncertainty scenarios of event-based total phosphorus using the validation data sets and calibration best-fit parameters.



Site	Scenario	Minimum	Lower quartile	Median	Upper quartile	Maximum
Ark.	low	-0.283	-0.028	-0.009	0.028	4.383
	high	-0.283	-0.027	-0.009	0.022	4.383
Ga.	low	-1.024	-0.001	0.145	0.719	7.822
	high	-1.024	0.000	0.143	0.715	7.680
N.C.	low	-2.475	-0.066	-0.006	0.045	1.604
	high	-2.475	-0.062	-0.004	0.038	1.449

Figure D.4. Cumulative frequency and distribution of absolute errors for the low and high uncertainty scenarios of event-based dissolved phosphorus using the validation data sets and calibration best-fit parameters.