Comparing an Annual and a Daily Time-Step Model for Predicting Field-Scale Phosphorus Loss

Carl H. Bolster,* Adam Forsberg, Aaron Mittelstet, David E. Radcliffe, Daniel Storm, John Ramirez-Avila, Andrew N. Sharpley, and Deanna Osmond

Abstract

A wide range of mathematical models are available for predicting phosphorus (P) losses from agricultural fields, ranging from simple, empirically based annual time-step models to more complex, process-based daily time-step models. In this study, we compare field-scale P-loss predictions between the Annual P Loss Estimator (APLE), an empirically based annual time-step model, and the Texas Best Management Practice Evaluation Tool (TBET), a process-based daily time-step model based on the Soil and Water Assessment Tool. We first compared predictions of field-scale P loss from both models using field and land management data collected from 11 research sites throughout the southern United States. We then compared predictions of P loss from both models with measured P-loss data from these sites. We observed a strong and statistically significant ($p < 0.001$) correlation in both dissolved ($\rho = 0.92$) and particulate ($\rho = 0.87$) P loss between the two models; however, APLE predicted, on average, 44% greater dissolved P loss, whereas TBET predicted, on average, 105% greater particulate P loss for the conditions simulated in our study. When we compared model predictions with measured P-loss data, neither model consistently outperformed the other, indicating that more complex models do not necessarily produce better predictions of field-scale P loss. Our results also highlight limitations with both models and the need for continued efforts to improve their accuracy.

Core Ideas

- We compared predictions of P loss between an empirically-based and process-based model.
- Predictions from both models were well correlated with each other.
- The process-based model did not result in noticeably better predictions of P loss.
- APLE predicted greater DP loss and TBET predicted greater PP loss.
- Results indicate the need for improving accuracy of both models.

APPLICATION of phosphorus (P) to agricultural lands can lead to increased offsite transport of P via surface runoff, erosion, and/or subsurface leaching to groundwater. Delivery of this P to P-sensitive water bodies can lead to water quality deterioration, primarily by accelerating the natural eutrophication process. Notable examples where excess P loading is contributing to water quality degradation include the Baltic Sea, Chesapeake Bay, the Florida Everglades, the Gulf of Mexico, and Lake Erie (Richardson et al., 2007; Chesapeake Bay Program, 2009; Dale et al., 2010; Andersson et al., 2014; Schoumans et al., 2014). In response to concerns over P losses from agricultural fields, research has focused on improving our understanding of the processes controlling P movement through the landscape (Radcliffe and Cabrera, 2007). This in turn has led to the development, improvement, and testing of models for predicting P fate and transport in the environment. When properly developed and used, these models can be useful tools for evaluating different management strategies for reducing P loss from agricultural fields (Sharpley et al., 2003; Radcliffe et al., 2009).

Models for describing P movement through the landscape range in complexity depending on the theoretical rigor of the governing equations, the number of processes included in the model, and the temporal and spatial scales of the model (Radcliffe and Cabrera, 2007; Radcliffe et al., 2009; Vadas et al., 2013). Model complexity will determine the amount of data required, the number of model parameters needed to be estimated, and the level of expertise required to properly run the model and interpret its results. While strong opinions often exist on the best model to use, no single model or modeling approach is appropriate for all settings. Rather, the most appropriate model in any given context will depend on several factors, including the amount, availability, and accuracy of the data; the level of accuracy and detail required in the model predictions;

C.H. Bolster, USDA–ARS, Food Animal Environmental Systems Research Unit, 2413 Nashville Rd.– B5, Bowling Green, KY 42101; A. Forsberg and D.E. Radcliffe, Crop and Soil Sciences Dep., Univ. of Georgia, Athens, GA 30602; A. Forsberg, current address, 70 Spruce St. NE, Atlanta, GA 30307; A. Mittelstet and D. Storm, Biosystems and Agricultural Engineering Dep., Oklahoma State Univ., Stillwater, OK 74078-6016; A. Mittelstet, current address, Biological Systems Engineering Dep., Univ. of Nebraska, Lincoln, NE 68583; J. Ramirez-Avila, Civil and Environmental Engineering Dep., Mississippi State Univ., Mississippi State, MS 39762; A.N. Sharpley, Crop, Soil, and Environmental Science, Univ. of Arkansas, Fayetteville, AR 72701; D. Osmond, Soil Science Dep., North Carolina State Univ., Raleigh, NC 27695-7619. Assigned to Associate Editor Peter Kleinman.

the expertise of the model users and target audience; and the overall goal of the modeling effort. Tradeoffs will exist regardless of the complexity of the model. For instance, empirically based, annual time-step models generally have fewer data requirements and require less training to implement than more process-based models; however, because empirically based models are derived from measured data and not theoretical considerations, their application should be restricted to conditions similar to those in which the model was developed. Process-based models, on the other hand, are more theoretically rigorous and should, in principle, be more transferable in space and time, though all process-based models have some degree of empiricism in their governing equations. Process-based models, however, generally require significantly more data and technical expertise to run. While better accuracy is often assumed with more complex, process-based models, comparisons of empirically based models with process-based models for predicting field-scale P loss are few.

Two models for predicting P loss that differ notably in their complexity are the Annual P Loss Estimator (APLE) and the Texas Best Management Practice Evaluation Tool (TBET). The APLE model is an empirically based, annual time-step model (Vadas et al., 2009, 2012) designed to predict annual field-scale P loss without requiring a considerable amount of data or technical expertise to run. The Texas Best Management Practice Evaluation Tool (TBET) is a daily time-step model that uses the Soil and Water Assessment Tool (SWAT) at the field scale (White et al., 2012). It was designed to be used by land managers and agency planners to evaluate the effects of best management practices for pasture and croplands and to be simple to use with readily available data, though the data requirements and level of expertise needed to run the model are considerably greater than those for APLE.

Two important differences exist between these models: the time step used and how incidental P losses from surface-applied P sources are simulated. Because APLE is an annual time-step model, it cannot directly account for time-dependent factors affecting P loss, such as the time interval between P application and the first runoff event. The TBET model, on the other hand, uses a daily time step and can therefore account for time-dependent factors when predicting P loss. The current version of TBET, however, does not simulate direct interactions between runoff and surface-applied P; rather, all surface-applied P is assumed to be incorporated into the top 10 mm of soil, leading to an increase in soil P levels. In contrast, APLE uses a set of empirical equations to specifically describe P transformation and loss from a distinct surface layer of P.

The primary objective of this study was to evaluate whether a simple annual P-loss model such as APLE can provide similar estimates of field-scale P loss as the more complex TBET model. A secondary objective was to assess the accuracy of both models for predicting field-scale P loss under a range of conditions representative of the southern United States. For the APLE model, we used runoff and erosion data calculated from the TBET model simulations summed over the year. This allowed us to focus on the differences in how the two models describe runoff P loss from surface-applied P sources and differences in their time steps. We first compared predictions of field-scale P loss using field and land management data collected from multiple locations throughout the southern United States (Arkansas, Georgia, Mississippi, North Carolina, Oklahoma, and Texas) as model inputs. We then compared model predictions with measured P-loss data from these sites.

Materials and Methods

Annual P Loss Estimator (APLE)

The APLE model is an empirically based spreadsheet model developed to describe annual, field-scale P loss when surface runoff is the dominant P-loss pathway (Vadas et al., 2009). The model calculates annual particulate P (PP) loss from eroded soil and annual dissolved P (DP) loss from soil and applied manure and fertilizer. The model calculates P loss from these four pathways based on five empirically based equations developed from P-loss data collected from multiple studies ranging in scale, soil type, physiographic regions, and P application rates (Vadas et al., 2004, 2005, 2007, 2008). A primary difference in this model compared with other models is how it calculates DP loss from surface-applied P sources. The APLE model specifically accounts for a distinct surface layer of P and calculates annual DP loss from this layer using runoff-to-precipitation ratio (Vadas et al., 2004, 2008). Because APLE does not predict runoff or erosion, we used runoff and erosion values calculated from the TBET simulations. See the Supplemental Materials for more details.

Texas Best Management Practice Evaluation Tool (TBET)

The TBET model was developed to use a modified version of SWAT 2009. 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 soil test P. User-defined options in the crop database include crop type, tillage, irrigation, grazing, stocking rate, and cover crop options. The TBET model was run on a single-year annual basis with a 2-yr warmup. All model predictions were conducted using uncalibrated (i.e., default) parameter values, with the exception of the simulations for the North Carolina sites in 2013, due to the unrealistically high erosion rates predicted by TBET (Forsberg et al., 2017). The default parameters initially chosen for Texas and Oklahoma by White et al. (2012) were used in the uncalibrated simulations. For the 2013 North Carolina sites, the manual calibration of TBET was performed by individually calibrating the following parameters: curve number (CN), the Universal Soil Loss Equation soil erosion (K) and crop factors (C_{min}), peak adjustment factor (ADJ_PKR), subbasin slope length (SLSBBSN), the P percolation coefficient (PPERCO), and the soil P partitioning coefficient (PHOSKD) (Forsberg et al., 2017).

Study Sites

Data from 28 field sites were collected from several published and unpublished studies (Supplemental Fig. S1). The field sites were located in Arkansas (Sharpley, unpublished data, 2015), Georgia (Pierson et al., 2001a, 2001b), Mississippi, North Carolina (Larsen et al., 2014; Edgell et al., 2015), Oklahoma (Olness et al., 1975; Sharpley et al., 1985; Smith et al., 1991), and Texas (McFarland et al., 2000; Harmel et al., 2008). These sites represent a wide range in climate, soil type, land management, and measured P losses (Supplemental Table S1). See the Supplemental Material for more details.

Annual measurements of DP and total P (TP) loss were obtained by aggregating measured P losses from individual events. For several sites, not all events were measured due to equipment failure or logistical problems. In these situations, the model predictions were modified as follows. For TBET, model predictions

for DP or TP for events in which one of these constituents was not measured were removed from the total annual predictions of P loss. For APLE, predicted annual P losses were reduced by the same percentage as the reductions in TBET predictions.

Evaluation of Model Predictions

Correlations in predictions of P loss between the two models and between model-predicted and observed values were evaluated using Spearman's rho (ρ). This nonparametric measure was chosen over the more commonly used Pearson's correlation coefficient because it is more resistant to outliers and does not assume a linear correlation or normally distributed data (Helsel and Hirsch, 2002). Model predictions were also evaluated using the Nash–Sutcliffe model efficiency (NSE), the RMSE, median absolute percent error (MAPE), and percent bias (PBIAS) (see the Supplemental Material for equations). To prevent division by zero when calculating MAPE values, we assumed a value of 0.005 kg ha^{-1} when measured ($n = 4$) or predicted ($n = 6$) P losses were zero.

A considerable amount of error can be expected with both measured (Harmel et al., 2006, 2010) and modeled P-loss data (Bolster and Vadas, 2013; Bolster et al., 2016). Therefore, we also calculated goodness-of-fit statistics while accounting for uncertainties in both predicted and measured P loss using a modified residual term, δm_i, based on the degree of overlap (DO) between the distributions for each paired measured and predicted P-loss values (Haan et al., 1995; Harmel et al., 2010; Bolster and Vadas, 2013):

$$\delta m_i = \delta_i \left(1 - DO_i\right) \qquad [1]$$

$$DO_i = \left[Pr\left(o_i < y_{u(i)}\right) - Pr\left(o_i < y_{l(i)}\right)\right] \\ \times \left[Pr\left(y_i < o_{u(i)}\right) - Pr\left(y_i < o_{l(i)}\right)\right] \qquad [2]$$

where o and y are the observed and predicted data, respectively, the subscripts l and u represent the lower and upper values of the 95% confidence intervals, respectively, and Pr is the cumulative probability density function. Based on the work of Harmel (Harmel et al., 2006, 2010; Harmel and Smith, 2007), we assumed that the probable error range (PER) corresponding to ±3.9 standard deviations for the measured data was ±30% of the measured value, resulting in a CV of 7.7%. For the model-predicted values, we assumed three levels of uncertainty with CVs of ±6.4 (PER = 25%), 12.8 (PER = 50%), and 25.6 (PER = 100%) of the model-predicted values (Harmel et al., 2010; Bolster and Vadas, 2013; Bolster et al., 2016). Normally distributed errors were assumed for the predicted and observed data.

The two models were also compared by regressing model predictions on measured P-loss data. We calculated the slope of the best-fit line with the Kendall–Theil Robust (KTR) method using the USGS software program KTRLine version 1.0 (Granato, 2006). This nonparametric method calculates the slope by taking the median value of all slopes that can be calculated between any two data points (Helsel and Hirsch, 2002). Unlike traditional linear regression, the KTR method is not based on the assumption of normally distributed residuals and thus is less sensitive to outliers than traditional linear regression.

In addition to evaluating the correlation between measured and model-predicted P-loss values, we investigated how similarly the two models assigned risk of P loss to each field by assigning P-loss risk values of low, moderate, and high to both observed and model-predicted values of DP and TP loss. Values of P loss associated with each risk category were based on the example values provided by NRCS in their Title 190, National Instruction, Part 302 of the revised 590 Nutrient Management Standard (USDA–NRCS, 2012). The threshold values were <2.2 kg ha^{-1} yr^{-1} (low), 2.2 to 5.5 kg ha^{-1} yr^{-1} (moderate), and >5.5 kg ha^{-1} yr^{-1} (high). To test whether there was a significant correlation between how each model categorized P-loss risk and the actual risk associated with each field based on our assigned thresholds, Kendall's modified Tau (τ_b) for ordinal data (Helsel and Hirsch, 2002) was calculated. Correlations were considered significant at $\alpha < 0.05$.

The accuracy of each model in assigning the correct risk category for each field was calculated using the Heidke skill score (HSS), a metric commonly used for evaluating accuracy of weather forecasts. The HSS is calculated as (Wilks, 2011):

$$HSS = \frac{\sum_{i=1}^{n} p(y_i, o_i) - \sum_{i=1}^{n} p(y_i) p(o_i)}{1 - \sum p(y_i) p(o_i)} \qquad [3]$$

where o and y now refer to the risk associated with each field based on the observed or predicted data, respectively. The first term in the numerator is the joint distribution of observed and predicted risk ratings, and the second term is the marginal distributions for each. A perfect forecast yields a score of one, whereas a value of zero indicates that all correct forecasts (i.e., categorizations) are due to random chance.

Results and Discussion

Comparing Predictions of Phosphorus Loss between Models

Correlations between predictions of DP loss from the APLE and TBET models ranged from 0.62 for the Mississippi sites to 0.98 for the Texas sites (Table 1). With the exception of the Mississippi sites ($\rho = 0.62$, $p = 0.10$), correlations were all statistically significant ($p < 0.001$). When data from all sites were combined, a strong correlation ($\rho = 0.92$; $p < 0.001$) between the two models was observed (Table 1). Inspection of the residual plot shows that the differences between predicted DP loss for the two models generally increased

Table 1. Spearman's correlation coefficient (ρ) and percent bias (PBIAS) comparing predictions of dissolved (DP) and particulate (PP) P loss between the APLE and TBET models for the studied locations. Positive PBIAS values reflect greater predictions by the Annual Phosphorus Loss Estimator (APLE) than the Texas Best Management Practice Evaluation Tool (TBET).

Location	ρ	PBIAS	ρ	PBIAS
	DP		PP	
All sites	0.92***	44	0.87***	−105
AR	0.88***	50	0.98***	−103
GA	0.66***	33	−0.47*	−218
MS	0.62	89	0.71*	−15
NC	0.73***	71	0.89***	−110
OK	0.73***	28	0.96***	−137
TX	0.98***	76	0.89***	−71

* Significant at the 0.05 probability level.

*** Significant at the 0.001 probability level.

with increasing APLE predictions, and this general trend was observed for all sites (Supplemental Fig. S2A). Residuals (TBET–APLE) ranged from −6.9 to 0.09 kg ha⁻¹, with 80% of the absolute values of the residuals being <2.2 kg ha⁻¹ and 60% <0.81 kg ha⁻¹. There were no consistent trends between the residuals and P application rate (total applied P or surface water-extractable P) or annual runoff (Supplemental Fig. S3).

While predictions from both models were strongly correlated with each other, important differences were observed. For the majority of sites, APLE predicted greater DP losses than TBET (Fig. 1A), as reflected in PBIAS values ranging from 28 to 89% for the individual sites and 44% for all sites combined (Table 1). Both models describe DP loss from soil as a function of labile (or solution) P, runoff rate, and an extraction (or partition) coefficient for estimating the amount of dissolved P in runoff from the concentration of labile P. In the treatments in which no P was applied (a total of 18 field years in Arkansas, Georgia, and Mississippi), APLE predicted, on average, 50% greater DP loss than TBET; differences between the P extraction coefficient of APLE (0.005) and the inverse of the P partition coefficient of TBET (175) cannot account for the greater predictions of P loss by APLE. The largest differences between the models was observed for the Mississippi sites, in which APLE predictions of DP loss ranged from 0.23 to 1.1 kg ha⁻¹ (65–93%) more than TBET. At the Mississippi sites, two crops were grown concurrently. To account for this, TBET includes a distance to stream factor to predict DP loss for the portion of the field that does not drain directly to the sampling location. As a result, predicted DP losses are expected to be lower for this scenario, which explains, at least in part, the substantially lower predictions by TBET for this location.

One of the most important differences between the models is in how DP loss is calculated from surface-applied P sources. While studies have shown that direct runoff from surface-applied P can be a significant P-loss pathway (Kleinman et al., 2002; DeLaune et al., 2004; Schroeder et al., 2004; Sistani et al., 2009, 2010), the current version of TBET does not simulate direct interactions between runoff and surface-applied P sources. Rather, all surface-applied P is assumed to be incorporated uniformly into the top 10 mm of soil and partitioned between the different P pools using a modified version of the Environmental Policy Integrated Climate P-cycling model (White et al., 2010). This, in effect, results in an extremely shallow incorporation of all surface-applied P; as a result, the amount of P available for surface runoff loss may be underestimated for recent surface P applications. A workaround to this limitation is to increase soil P values in excess of actual soil P values within the model inputs. This approach, however, may result in overpredictions of P loss in the future, as well as limit the model's ability to accurately reflect changing manure management strategies on future P losses.

In contrast, APLE simulates P transformation and dissolution from a distinct surface layer of P using a set of empirically based equations. As a result, the amount of P vulnerable to runoff loss will generally be greater for APLE than TBET, potentially leading to greater predictions of P loss following manure application (Sen et al., 2012; Collick et al., 2016). Incorporating daily versions of the APLE manure P-loss routines into SWAT (revision 586), Collick et al. (2016) found that the new P routines better represented effects of different manure management practices on P losses at the small watershed scale (it is unclear when these modifications will be incorporated into TBET). Differences in how APLE and TBET simulate incidental P losses from surface-applied manures, however, cannot entirely account for the differences we observed in predictions of DP loss between the two models, as we did not find a strong correlation between residuals in predicted DP loss and surface-applied P rates (Supplemental Fig. S3).

Another important difference between the models is the time step used. Because APLE is an annual time-step model, it cannot directly account for time-dependent processes controlling P loss. For instance, several studies have shown that P loss following manure application tends to decrease with increasing time interval between manure application and first runoff event (Sharpley, 1997; Schroeder et al., 2004; Sistani et al., 2009). This has often been attributed to increased sorption of manure P to soil over time. Vadas et al. (2011), however, reanalyzed several of these studies and found that storm hydrology, specifically the ratio of runoff to precipitation, was the primary factor affecting P loss, rather than P sorption to soil. They concluded that greater runoff-to-rainfall ratios coincided with greater runoff at the beginning of a rainfall event when greater concentrations of DP are expected. Because their analysis applied a daily model to individual runoff events, it is unclear how well this
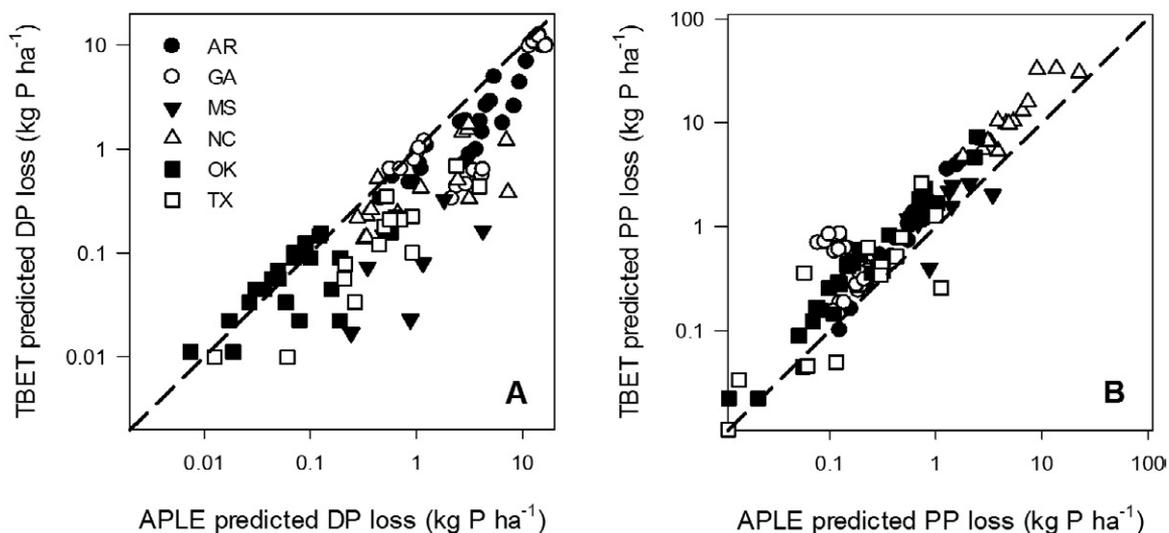

Fig. 1. Scatter plot of Annual Phosphorus Loss Estimator (APLE) and Texas Best Management Practice Evaluation Tool (TBET) predictions of (A) dissolved P (DP) and (B) particulate P (PP) loss. Also included is 1:1 line.

translates to annual P losses, but the use of annual runoff ratio by APLE may help overcome some of the potential limitations of using an annual model to predict P loss from surface-applied manures. Moreover, results from Collick et al. (2016) suggest that the current P routines in TBET are insensitive to manure application timing. Thus, it is unclear whether differences in time steps between the two models is a significant contributor to the observed differences in predicted DP loss from fields receiving manure.

Similar to predictions of DP loss, a strong correlation was observed between predictions of PP loss for the combined dataset ($\rho = 0.87$, $p < 0.001$, Fig. 1B, Table 1). For the individual locations, ρ ranged from −0.47 (indicating an inverse correlation) for the Georgia sites to 0.98 for the Arkansas sites (Table 1). The APLE model consistently predicted lower PP losses than TBET, as reflected in PBIAS values ranging from −15% for Mississippi to −218% for Georgia and −105% for all sites combined (Table 1). Residual values for the PP predictions generally increased with increasing APLE predictions of PP loss (Supplemental Fig. S2B). Residuals ranged from −1.4 to 24 kg ha^{-1}, with 80% of the absolute values being <1.4 kg ha^{-1} and 60% <0.45 kg ha^{-1}. The residuals between the two model predictions generally increased with increasing predictions of erosion rate (Supplemental Fig. S4A). We did not observe any significant relationships between the residuals and soil test P (STP) (Supplemental Fig. S4B).

Both models describe PP loss as a function of total soil P, erosion rate, and a P enrichment ratio defined as the ratio of the P concentration in eroded soil to that in the underlying bulk soil. In calculating total soil P, APLE includes labile P, whereas TBET does not (White et al., 2012), though differences will be relatively minor, as labile P is only a small fraction of total soil P. Another difference in the models is in how the P enrichment ratio is calculated. The APLE model calculates the enrichment ratio from annual sediment loading rates (kg ha^{-1}) based on equations derived from storm-event data by Menzel (1980) and Sharpley (1980). The TBET model, on the other hand, calculates a P enrichment ratio for each individual erosion event using sediment concentration. For the erosion and runoff rates used in our study, P enrichment ratios calculated by TBET for individual storm events were generally greater than enrichment ratios calculated by APLE. While some of these differences can be attributed to the use of different equations, the more important difference between the two models is the time step used. Because P enrichment ratios decrease log-linearly with increasing erosion rate, the use of annual erosion rates by APLE results in lower enrichment ratios than if the same equation were used with storm-event erosion rates. Depending on erosion rates and TP values, this generally resulted in greater predictions of PP loss with TBET compared with APLE. Another important difference between the models is that APLE calculates the P sorption parameter needed to partition added P between the labile and active P pools based on user inputs of soil properties. In contrast, TBET assumes a value of 0.4 unless otherwise specified by the user. Depending on soil texture and STP values, differences in the P sorption parameter between the models can be significant, potentially leading to large differences in predicted PP loss between the two models.

Comparing Model Predictions with Measured Phosphorus-Loss Data

Predictions of DP loss from both models were generally well correlated with measured P loss, with ρ values for the combined sites of 0.68 ($p < 0.001$) and 0.70 ($p < 0.001$) for APLE and TBET, respectively (Table 2). Model efficiency for the combined dataset was slightly higher for APLE (0.52) than TBET (0.41). Surprisingly, when calculated for each separate location, NSE values for APLE were negative for five locations and for four locations for TBET, likely a result of the relatively small number of data points at each location. A significant amount of the observed variability in measured DP loss was not captured by either model (Fig. 2A). For instance, the MAPE between model-predicted and observed DP loss was 81% for APLE and 71% for TBET, with RMSE values of 2.4 and 2.6 kg ha^{-1}, respectively. With the exception of the Georgia and Oklahoma datasets, APLE systematically overpredicted DP loss with an overall PBIAS of −9.6%, whereas TBET consistently underpredicted DP loss with an overall PBIAS of 40%. Including our estimated uncertainties in both the model-predicted and measured data did not result in any meaningful improvements in our goodness-of-fit statistics, even with a model PER of 100% (Supplemental Table S2). Kendall–Theil Robust slopes between observed and predicted P loss were 0.74 (95% confidence intervals of 0.47–0.99) and 0.35 (0.21–0.53) for APLE and TBET, respectively. These values are similar to slopes obtained using traditional linear regression (0.75 and 0.46, respectively).

Residuals for predicted DP ranged from −5.2 to 12.4 kg ha^{-1} for APLE and from −3.6 to 15.4 kg ha^{-1} for TBET (Supplemental Fig. S5). There were no consistent trends between the residuals and P application rate (total applied P or surface water-extractable P) or annual runoff (Supplemental Fig. S6) for either model.

Consistent with our results for the DP predictions, predicted and observed TP losses were significantly correlated, with ρ values of 0.57 and 0.52 ($p < 0.001$, Table 3) for APLE and TBET, respectively. Model efficiencies, however, were negative for both models for the combined dataset, as well as for the majority of locations (Table 3). Similar to the DP predictions, a large amount of the observed variability in TP was not captured by either model (Fig. 3), as reflected in MAPE values of 77 and 74% for APLE and TBET, respectively. Consistent with DP model predictions, goodness-of-fit statistics were similar for the two models, though RMSE values were 25% less for APLE (4.6 kg ha^{-1}) than TBET (6.0 kg ha^{-1}). The PBIAS values calculated on the combined dataset indicate that both models generally underpredicted TP losses, though for TBET, PBIAS was close to zero (1.5%). As discussed previously, the underpredictions of TP by APLE are due in part to the use of annual rather than storm-event erosion rates in calculating the P enrichment ratio. Inclusion of model and measurement uncertainties did not result in any noticeable improvements in any of the goodness-of-fit statistics for TP (Supplemental Table S3). The KTR slopes between observed and predicted TP losses were 0.45 (0.25–0.65) and 0.43 (0.26–0.59) for APLE and TBET, respectively. Slopes from traditional regression analysis were 0.30 and 0.31, respectively. Residuals ranged from −16 to 15 kg ha^{-1} for APLE and −29 to 18 kg ha^{-1} for TBET, with both models generally overpredicting TP loss from the North Carolina sites due to the high erosion rates predicted by TBET (Supplemental Fig. S7). No trends were observed between model residuals for TP and predicted runoff and erosion, surface-applied water-extractable P, and STP (Supplemental Fig. S8).

Based on the model performance classification scheme proposed by Moriasi et al. (2007), model efficiency for DP predictions

Table 2. Spearman's correlation coefficient (ρ), model efficiency (NSE), median absolute percent error (MAPE), percent bias (PBIAS), and RMSE evaluating predictions of dissolved P (DP) losses for the Annual Phosphorus Loss Estimator (APLE) and Texas Best Management Practice Evaluation Tool (TBET) models against the measured DP loss data from each location.

Location	ρ		NSE		MAPE		PBIAS		RMSE	
	APLE	TBET	APLE	TBET	APLE	TBET	APLE	TBET	APLE	TBET
					———— % ————				——— kg ha⁻¹ ———	
All	0.68***	0.70***	0.52	0.41	81	71	−9.6	40	2.4	2.6
AR	0.86***	0.80***	−0.52	0.41	185	43	−89	3.6	1.4	0.84
GA	0.55*	0.33	0.18	−0.23	62	65	8.4	41	4.1	5.0
MS	0.69	0.98***	−18	−0.73	60	75	−158	72	1.3	0.40
NC	0.63**	0.63**	−1.3	0.40	163	60	−107	34	1.7	0.86
OK	0.99***	0.09	−0.44	−1.3	82	95	76	93	0.72	0.90
TX	0.35	0.60**	−1.8	−0.36	99	81	−2.4	74	0.95	0.66

* Significant at the 0.05 probability level.

** Significant at the 0.01 probability level.

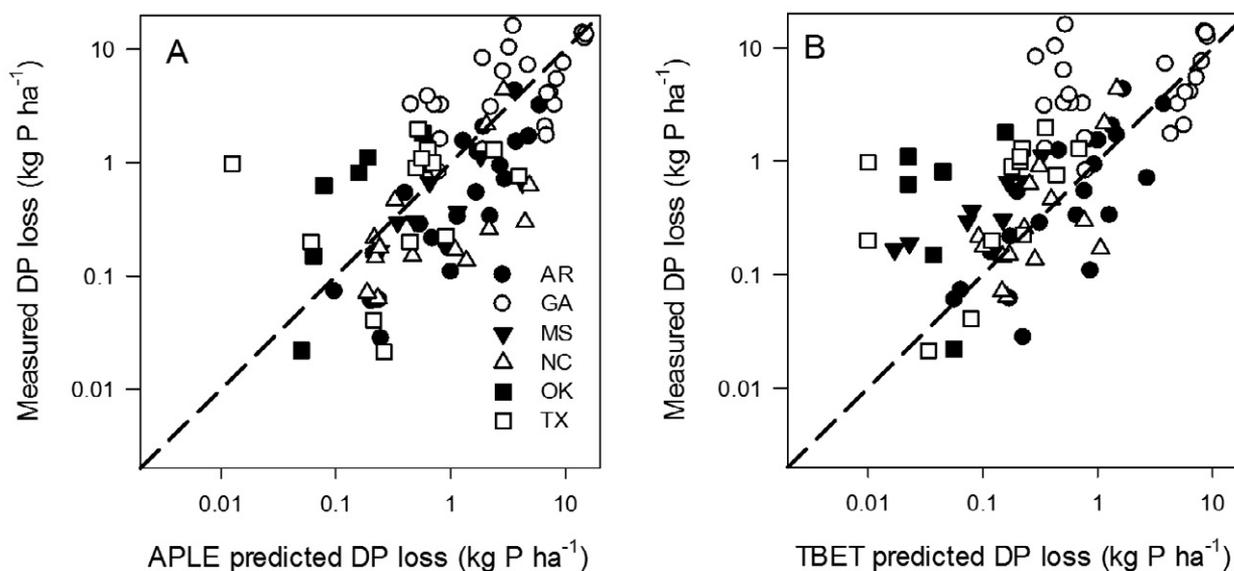*** Significant at the 0.001 probability level.

Fig. 2. Scatter plots of measured and (A) Annual Phosphorus Loss Estimator (APLE)- and (B) Texas Best Management Practice Evaluation Tool (TBET)-predicted dissolved P (DP) loss. Also included is 1:1 line.

Table 3. Spearman's correlation coefficient (ρ), model efficiency (NSE), median absolute percent error (MAPE), percent bias (PBIAS), and RMSE evaluating predictions of total P (TP) losses for the Annual Phosphorus Loss Estimator (APLE) and Texas Best Management Practice Evaluation Tool (TBET) models against the measured TP loss data from each location.

Location	ρ		NSE		MAPE		PBIAS		RMSE	
	APLE	TBET	APLE	TBET	APLE	TBET	APLE	TBET	APLE	TBET
					———— % ————				——— kg ha⁻¹ ———	
All	0.57***	0.52***	−0.11	−1.0	77	74	7.7	1.5	4.5	6.0
AR	0.91***	0.86***	−0.76	0.01	173	90	−93	−44	1.6	1.2
GA	0.55**	0.12	−0.20	−0.78	53	72	35	54	5.4	6.5
MS	0.76*	0.76*	0.32	0.36	30	26	−24	16	1.1	1.1
NC	−0.21	−0.19	−2.1	−6.9	259	539	−65	−175	7.6	12
OK	0.89**	0.89**	−0.17	0.53	77	43	78	49	4.7	3.0
TX	0.58**	0.61**	−0.10	−0.06	91	91	40	67	2.5	2.5

* Significant at the 0.05 probability level.

** Significant at the 0.01 probability level.

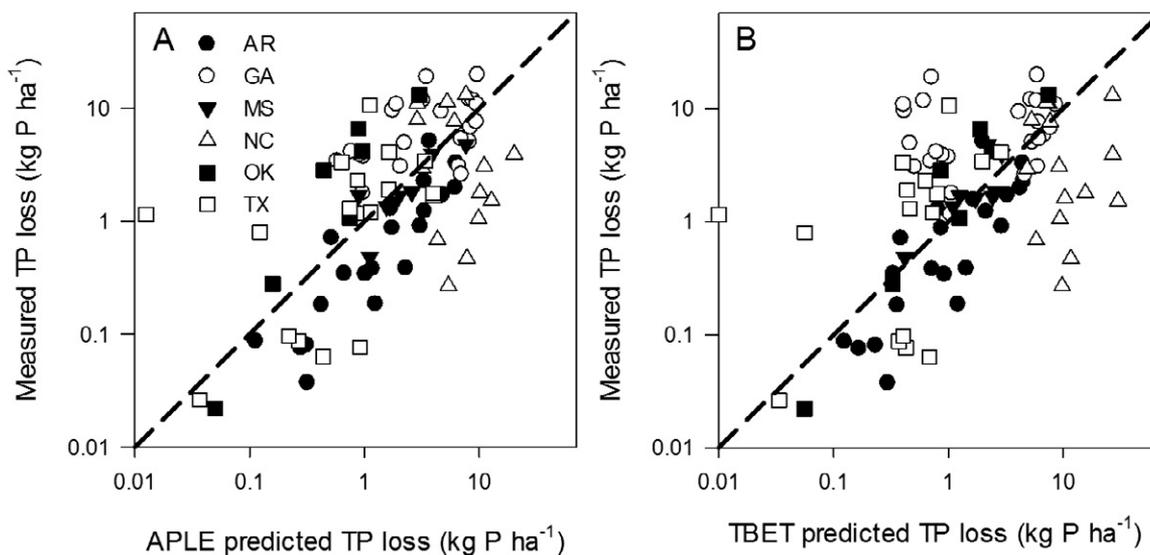*** Significant at the 0.001 probability level.

Fig. 3. Scatter plots of measured and (A) Annual Phosphorus Loss Estimator (APLE)- and (B) Texas Best Management Practice Evaluation Tool (TBET)-predicted total P (TP) loss. Also included is 1:1 line.

by TBET (NSE = 0.41) was below the threshold of 0.5 considered to represent satisfactory model performance, whereas DP predictions by APLE (NSE = 0.52) barely exceeded this threshold; this threshold, however, is based on monthly data—for annual data, Moriasi et al. (2007) suggest that their proposed thresholds be increased. For predictions of TP loss, both models clearly provided unsatisfactory model predictions based on NSE values. The PBIAS values, on the other hand, indicated very good (PBIAS < ±25%) model performance for both DP and TP for APLE, and good to satisfactory performance for DP and very good performance for TP for TBET. These results indicate that, while the model predictions and observations fell far from a 1:1 line, the models provided predictions with relatively low overall bias.

The relatively poor performance of both of these models was unexpected given the relative success at predicting P loss reported for these models in other studies (Gassman et al., 2007; Vadas et al., 2009; Bolster et al., 2012). One potential explanation for this is that, by using an uncalibrated model, our predictions of both runoff and erosion were less accurate than the aforementioned studies in which calibrated (SWAT) or measured (APLE) runoff and erosion rates were used. Model efficiencies indicate borderline unsatisfactory predictions of runoff (NSE = 0.47) and extremely poor predictions of erosion (NSE = −2.4) for the uncalibrated TBET model used in our study (Supplemental Fig. S9). Forsberg et al. (2017) applied a calibrated TBET model to the Arkansas, Georgia, and North Carolina locations, and while they observed some significant improvements in predictions of event-based runoff and erosion rates from these fields, they only reported slight improvements in predictions of P loss for the Georgia sites, with no discernible improvements in predicting DP and TP losses for the Arkansas and North Carolina sites. To test whether poor predictions of the transport factors were the primary source of our relatively poor model performances for APLE, we reran APLE for each of these sites using measured runoff and erosion data. For the entire data set, NSE increased from 0.52 to 0.62 for DP and from −0.13 to 0.43 for TP (Supplemental Table S4). The greatest improvement in NSE values for the TP predictions was observed for the Mississippi and Texas sites, where NSE

values increased from 0.32 to 0.91 and from −0.10 to 0.78, respectively. The greatest improvement with predictions of DP was for the Arkansas dataset, with NSE increasing from −0.52 to 0.35. With the exception of PBIAS values, which increased from −9.8 to 32% for DP and 7.6 to 45% for TP, all other goodness-of-fit statistics improved (Supplemental Table S4). Improvements in DP predictions were generally not as substantial as improvements in TP owing to the fact that our runoff predictions from the uncalibrated model were much more accurate than our erosion predictions. Notwithstanding these improvements in P-loss predictions using measured runoff and erosion values, there still exists a large amount of variability in measured P loss that APLE does not capture (MAPE values of 64 and 52% for DP and TP, respectively), indicating the need for continued refinement of the model.

Another factor that may have adversely affected the performance of the models is the accuracy of the measured data. All measured data are susceptible to measurement and sampling errors (Harmel et al., 2006; Harmel and Smith, 2007). When we included our estimated uncertainties in both the measured and predicted data, we did not see any noticeable improvement in our goodness-of-fit statistics, even with a model PER of 100% (Supplemental Tables S2 and S3). Incorporation of measurement and model uncertainty has been shown to improve goodness-of-fit statistics when model predictions are relatively good; however, in cases where model performance is poor, incorporation of measurement and prediction uncertainties will have minimal impact on these calculations due to minimal overlap in the error distributions between the measured and model-predicted data (Harmel et al., 2010; Bolster and Vadas, 2013; Forsberg et al., 2017). While we took a very simplified approach to estimating our uncertainties in both the measured and predicted P-loss data, our uncertainty estimates are based on values reported in the literature and thus should be fairly representative of the actual errors in our data and model predictions.

Missed runoff events will also adversely affect the accuracy of the measured data and thus perceived model performance. Several runoff events were missed at the Arkansas, Georgia, and North Carolina locations, resulting in measured P loads being

less than actual P loads. To account for these missed events, we modified our model predictions based on certain assumptions. For TBET, we removed the predicted P loss from these events from the annual P-loss total. For APLE, we reduced the predicted P loss by the fraction of P loss assumed to be missed with TBET. Certainly, our simplified approach of accounting for these missed events in the model predictions contributed to some of the modeling errors for these sites.

In addition to assessing the accuracy of the models in predicting actual field-scale P losses, we evaluated how well each model assigned risk of P loss to each field. Correlations between risk ratings assigned to each field based on predicted and measured DP loss were similar for both models (τ_b = 0.41 and 0.38 for APLE and TBET, respectively, $p < 0.001$), with APLE and TBET correctly assigning risks to 76 and 81% of the fields, respectively, resulting in HSS of 0.37 and 0.42. Applying the same risk thresholds to our model predictions of TP loss, APLE correctly assigned risk to 54% of the fields, whereas TBET correctly assigned risk to 62% of the fields, yielding HSS values of 0.25 and 0.35, respectively. Correlations between model and measured risk ratings of TP loss were also similar for the two models (τ_b = 0.28 and 0.27 for APLE and TBET, respectively; $p < 0.001$). Somewhat surprisingly, Osmond et al. (2017) found that P Indices from several southern states were as accurate as these two models in assigning risk ratings to the same fields used in our study.

Differences in how these two models describe field-scale P loss may have important implications in the context of P management strategies. When applied to the same field, APLE will likely predict greater DP losses than TBET and thus may lead to more restrictive P application rates when DP is the primary source of P loss. For instance, compared with TBET, APLE categorized 14 more fields as moderate or high risk, ratings that may result in more restrictive P application to these fields. Conversely, because TBET generally predicted greater PP losses than APLE, under conditions where PP is the dominant form of P loss, TBET may result in more restrictive P application rates than APLE. The reason we did not see more of our fields rated in the higher risk categories with TBET than APLE based on TP losses is that, for the majority of our sites, predicted erosion, and therefore PP loss, was low; in those fields where predicted PP loss was high, these values were extremely high for both models, and thus both models predicted high risk for these fields.

Conclusion

Similar goodness-of-fit statistics for the two models indicate that when using the same runoff and erosion data, APLE, which is an annual, empirically based model, produced similarly accurate predictions of P loss for these sites as TBET, a more process-based daily time-step model. These results demonstrate that increasing model complexity does not necessarily lead to improved model predictions of field-scale P loss. Our findings thus indicate that, depending on the needs of the user and the availability and accuracy of the data, simple models that do not require a significant amount of technical expertise or data to execute may provide valuable insights into the relative risks of field-scale P loss and can serve as a useful tool to guide land management decisions. Additional testing, however, is required to determine if these findings are relevant to other locations throughout the United States. Furthermore, our results highlight important limitations

with both models and the need to continually evaluate and update these models to improve their accuracy.

References

Andersson, I., J. Jarsö, and M. Petersson. 2014. Saving the Baltic Sea, the inlands waters of its drainage basin, or both? Spatial perspectives on reducing P-loads in eastern Sweden. Ambio 43:914–925. doi:10.1007/s13280-014-0523-x

Bolster, C.H., and P.A. Vadas. 2013. Sensitivity and uncertainty analysis for the Annual Phosphorus Loss Estimator model. J. Environ. Qual. 42:1109–1118. doi:10.2134/jeq2012.0418

Bolster, C.H., P.A. Vadas, and D. Boykin. 2016. Parameter uncertainty analysis for the annual phosphorus loss estimator (APLE) model. J. Hydrol. 539:27–37. doi:10.1016/j.jhydrol.2016.05.009

Bolster, C.H., P.A. Vadas, A.N. Sharpley, and J.A. Lory. 2012. Using a phosphorus loss model to evaluate and improve phosphorus indices. J. Environ. Qual. 41:1758–1766. doi:10.2134/jeq2011.0457

Chesapeake Bay Program. 2009. Executive order 13508: Draft strategy for restoring and protecting the Chesapeake Bay. Federal Leadership Committee for the Chesapeake Bay, Annapolis, MD. http://executiveorder.chesapeakebay.net/file.axd?file = 2010%2f9%2fChesapeake+EO+Action+Plan+FY2011.pdf (accessed 24 Oct. 2016).

Collick, A.S., T.L. Veith, D.A. Fuka, P.J.A. Kleinman, A.R. Buda, J.L. Weld et al. 2016. Improved simulation of edaphic and manure phosphorus loss in SWAT. J. Environ. Qual. 45:1215–1225. doi:10.2134/jeq2015.03.0135

Dale, V.H., C.L. Kling, J.L. Meyer, J. Sanders, H. Stallworth, T. Armitage et al. 2010. Hypoxia in the Northern Gulf of Mexico. Springer Science, New York. doi:10.1007/978-0-387-89686-1

DeLaune, P.B., P.A. Moore, Jr., D.K. Carmen, A.N. Sharpley, B.E. Haggard, and T.C. Daniel. 2004. Evaluation of the phosphorus source component in the phosphorus index for pastures. J. Environ. Qual. 33:2192–2200. doi:10.2134/jeq2004.2192

Edgell, J., D.L. Osmond, D.E. Line, G. Hoyt, J.M. Grossman, and E.M. Larsen. 2015. Comparison of surface water quality and yields from organically and conventionally produced sweet corn plots with conservation and conventional tillage. J. Environ. Qual. 44:1861–1870. doi:10.2134/jeq2015.02.0074

Forsberg, T.A., D.E. Radcliffe, C.H. Bolster, A. Mittelstet, D.E. Storm, and D.L. Osmond. 2017. Evaluation of the TBET model for potential improvement of Southern P Indices. J. Environ. Qual. doi:10.2134/jeq2016.06.0210 (in press).

Gassman, P.W., M.R. Reyes, C.H. Green, and J.G. Arnold. 2007. The soil and water assessment tool: Historical development, applications, and future research directions. Trans. ASABE 50:1211–1250. doi:10.13031/2013.23637

Granato, G.E. 2006. Kendall-Theil Robust Line (KTRLine-version 1.0—A visual basic program for calculating and graphing robust nonparametric estimates of linear-regression coefficients between two continuous variables. Techniques and methods of the U.S. Geological Survey 4A7. USGS, Reston, VA.

Haan, C.T., B. Allred, D.E. Storm, G.J. Sabbagh, and S. Prabhu. 1995. A statistical procedure for evaluating hydrologic/water quality models. Trans. ASAE 38:725–733. doi:10.13031/2013.27886

Harmel, R.D., R.J. Cooper, R.M. Slade, R.L. Haney, and J.G. Arnold. 2006. Cumulative uncertainty in measured streamflow and water quality data for small watersheds. Trans. ASAE 49:698–701.

Harmel, R.D., C.G. Rossi, T. Dybala, J.G. Arnold, K. Potter, J. Wolfe, and D. Hoffman. 2008. Conservation effects assessment project research in the Leon River and Riesel watersheds. J. Soil Water Conserv. 63:453–460. doi:10.2489/jswc.63.6.453

Harmel, R.D., and P.K. Smith. 2007. Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic water quality modeling. J. Hydrol. 337:326–336. doi:10.1016/j.jhydrol.2007.01.043

Harmel, R.D., P.K. Smith, and K.W. Migliaccio. 2010. Modifying goodness-of-fit indicators to incorporate both measurement and model uncertainty in model calibration and validation. Trans. ASABE 53:55–63. doi:10.13031/2013.29502

Helsel, D.R. and R.M. Hirsch. 2002. Statistical methods in water resources. U.S. Geological Survey techniques in water resources, book 4. USGS, Reston, VA.

Kleinman, P.J.A., A.N. Sharpley, B.G. Moyer, and G.F. Elwinger. 2002. Effect of mineral and manure phosphorus sources on runoff phosphorus. J. Environ. Qual. 31:2026–2033. doi:10.2134/jeq2002.2026

Larsen, E., J. Grossman, J. Edgell, G. Hoyt, D. Osmond, and S.J. Hu. 2014. Soil biological properties, soil losses and corn yield in long-term organic and conventional farming systems. Soil Tillage Res. 139:37–45. doi:10.1016/j.still.2014.02.002

McFarland, A., A. Saleh, and L. Hauck. 2000. Demonstration of phosphorus best management practices in the North Bosque River Basin. Texas Institute for Applied Environmental Research, Tarleton State Univ., Stephenville, TX.

Menzel, R.G. 1980. Chapter 12: Enrichment ratios for water quality modeling. In: CREAMS A field scale model for chemicals, runoff, and erosion from agricultural management systems. Vol. 3. USDA-SEA-Conserv. Res. Report.

Moriasi, D.N., J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel, and T.L. Veith. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50:885–900. doi:10.13031/2013.23153

Olness, A., S.J. Smith, E.D. Rhoades, and R.G. Menzel. 1975. Nutrient and sediment discharge from agricultural watersheds in Oklahoma. J. Environ. Qual. 4:331–336. doi:10.2134/jeq1975.00472425000400030009x

Osmond, D., C. Bolster, A. Sharpley, M. Cabrera, S. Feagley, A. Forsberg et al. 2017. Southern Phosphorus Indices, water quality data, and modeling (APEX, APLE, and TBET) results: A comparison. J. Environ. Qual. doi:10.2134/jeq2016.05.0200 (in press)

Pierson, S.T., M.L. Cabrera, G.K. Evanylo, H.A. Kuykendall, C.S. Hoveland, M.A. McCann, and L.T. West. 2001a. Phosphorus and ammonium concentrations in surface runoff from grasslands fertilized with broiler litter. J. Environ. Qual. 30:1784–1789. doi:10.2134/jeq2001.3051784x

Pierson, S.T., M.L. Cabrera, G.K. Evanylo, P.D. Schroeder, D.E. Radcliffe, H.A. Kuykendall et al. 2001b. Phosphorus losses from grasslands fertilized with broiler litter. J. Environ. Qual. 30:1790–1795. doi:10.2134/jeq2001.3051790x

Radcliffe, D.E., and M.L. Cabrera. 2007. Modeling phosphorus in the environment. CRC Press, Boca Raton.

Radcliffe, D.E., J. Freer, and O.F. Schoumans. 2009. Diffuse phosphorus models in the United States and Europe: Their usages, scales, and uncertainties. J. Environ. Qual. 38:1956–1967. doi:10.2134/jeq2008.0060

Richardson, C.J., R.S. King, S.S. Qian, P. Vaithiyanathan, R.G. Qualls, and C.A. Stow. 2007. Estimating ecological thresholds for phosphorus in the Everglades. Environ. Sci. Technol. 41:8084–8091. doi:10.1021/es062624w

Schoumans, O.F., F. Bouraoui, C. Kabbe, O. Oenema, and K.C. van Dijk. 2014. Phosphorus management in Europe in a changing world. Ambio 44:108–192.

Schroeder, P.D., D.E. Radcliffe, and M.L. Cabrera. 2004. Rainfall timing and poultry litter application rate effects phosphorus loss in surface runoff. J. Environ. Qual. 33:2201–2209. doi:10.2134/jeq2004.2201

Sen, S., P. Srivastava, P.A. Vadas, and L. Kalin. 2012. Watershed-level comparison of predictability and sensitivity of two phosphorus models. J. Environ. Qual. 41:1642–1652. doi:10.2134/jeq2011.0242

Sharpley, A.N. 1997. Rainfall frequency and nitrogen and phosphorus in runoff from soil amended with poultry litter. J. Environ. Qual. 26:1127–1132. doi:10.2134/jeq1997.00472425002600040026x

Sharpley, A.N. 1980. The enrichment of soil phosphorus in runoff sediments. J. Environ. Qual. 9:521–526. doi:10.2134/jeq1980.00472425000900030039x

Sharpley, A.N., P.J.A. Kleinman, R.W. McDowell, M. Gitau, and R.B. Bryant. 2003. Modeling phosphorus transport in agricultural watersheds: Processes and possibilities. J. Soil Water Conserv. 57:425–439.

Sharpley, A.N., S.J. Smith, W.A. Berg, and J.R. Williams. 1985. Nutrient losses as predicted by annual and monthly soil sampling. J. Environ. Qual. 14:354–360. doi:10.2134/jeq1985.00472425001400030010x

Sistani, K.R., C.H. Bolster, T.R. Way, H.A. Torbert, D.H. Pote, and D.B. Watts. 2010. Influence of poultry litter application methods on the longevity of nutrient and E. coli in runoff from tall fescue pasture. Water Air Soil Pollut. 206:3–12. doi:10.1007/s11270-009-0078-9

Sistani, K.R., H.A. Torbert, T.R. Way, C.H. Bolster, D.H. Pote, and J.G. Warren. 2009. Broiler litter application method and runoff timing effects on nutrient and Escherichia coli losses from tall fescue pasture. J. Environ. Qual. 38:1216–1223. doi:10.2134/jeq2008.0185

Smith, S.J., A.N. Sharpley, J.W. Naney, W.A. Berg, and O.R. Jones. 1991. Water quality impacts associated with wheat culture in the southern plains. J. Environ. Qual. 20:244–249. doi:10.2134/jeq1991.00472425002000010039x

USDA–NRCS. 2012. National I Title 190, Part 302—Nutrient management policy implementation. Rep. NI 190 302. USDA–NRCS, Washington, DC. http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1046179.pdf (accessed 19 Oct. 2016).

Vadas, P.A., C.H. Bolster, and L.W. Good. 2013. Critical evaluation of models used to study agricultural phosphorus and water quality. Soil Use Manage. 29:36–44. doi:10.1111/j.1475-2743.2012.00431.x

Vadas, P.A., W.J. Gburek, A.N. Sharpley, P.J.A. Kleinman, P.A. Moore, M.L. Cabrera, and R.D. Harmel. 2007. A model for phosphorus transformation and runoff loss for surface-applied manures. J. Environ. Qual. 36:324–332. doi:10.2134/jeq2006.0213

Vadas, P.A., L.W. Good, P.A. Moore, Jr., and N. Widman. 2009. Estimating phosphorus loss in runoff from manure and fertilizer for a phosphorus loss quantification tool. J. Environ. Qual. 38:1645–1653. doi:10.2134/jeq2008.0337

Vadas, P.A., B.C. Joern, and P.A. Moore. 2012. Simulating phosphorus dynamics for a phosphorus loss quantification tool. J. Environ. Qual. 41:1750–1757. doi:10.2134/jeq2012.0003

Vadas, P.A., W.E. Jokela, D.H. Franklin, and D.M. Endale. 2011. The effect of rain and runoff when assessing timing of manure application and dissolved phosphorus loss in runoff. J. Am. Water Resour. Assoc. 47:877–886. doi:10.1111/j.1752-1688.2011.00561.x

Vadas, P.A., P.J.A. Kleinman, and A.N. Sharpley. 2004. A simple method to predict dissolved phosphorus in runoff from surface-applied manures. J. Environ. Qual. 33:749–756. doi:10.2134/jeq2004.7490

Vadas, P.A., P.J.A. Kleinman, A.N. Sharpley, and B.L. Turner. 2005. Relating phosphorus to dissolved phosphorus in runoff: A single extraction coefficient for water quality modeling. J. Environ. Qual. 34:572–580. doi:10.2134/jeq2005.0572

Vadas, P.A., L.B. Owens, and A.N. Sharpley. 2008. An empirical model for dissolved phosphorus in runoff from surface-applied fertilizers. Agric. Ecosyst. Environ. 127:59–65. doi:10.1016/j.agee.2008.03.001

White, M.J., R.D. Harmel, and R.L. Haney. 2012. Development and validation of the Texas Best Management Practice Evaluation Tool (TBET). J. Soil Water Conserv. 67:525–535. doi:10.2489/jswc.67.6.525

White, M.J., D.E. Storm, P.R. Busteed, M.D. Smolen, H. Zhang, and G.A. Fox. 2010. A quantitative phosphorus loss assessment tool for agricultural fields. Environ. Model. Softw. 25:1121–1129. doi:10.1016/j.envsoft.2010.03.017

Wilks, D.S. 2011. Statistical methods in the atmospheric sciences, 3rd ed. Elsevier, Oxford, UK.