Abstract
The Agricultural Policy Environmental eXtender (APEX) model is capable of estimating edge-of-field water, nutrient, and sediment transport and is used to assess the environmental impacts of management practices. The current practice is to fully calibrate the model for each site simulation, a task that requires resources and data not always available. The objective of this study was to compare model performance for flow, sediment, and phosphorus transport under two parameterization schemes: a best professional judgment (BPJ) parameterization based on available data and a fully calibrated parameterization based on site-specific soil, weather, event flow, and water quality data. The analysis was conducted using 12 datasets at four locations representing poorly drained soils and row-crop production under different tillage systems. Model performance was based on the Nash–Sutcliffe efficiency (NSE), the coefficient of determination ($r^2$) and the regression slope between simulated and measured annualized loads across all site years. Although the BPJ model performance for flow was acceptable (NSE = 0.7) at the annual time step, calibration improved it (NSE = 0.9). Acceptable simulation of sediment and total phosphorus transport (NSE = 0.5 and 0.9, respectively) was obtained only after full calibration at each site. Given the unacceptable performance of the BPJ approach, uncalibrated use of APEX for planning or management purposes may be misleading. Model calibration with water quality data prior to using APEX for simulating sediment and total phosphorus loss is essential.

Core Ideas
- Uncalibrated, APEX produced unacceptable site-specific sediment and TP estimates.
- Acceptable runoff estimates do not translate to acceptable water quality estimates.
- Distributions of successfully calibrated performance indicator values were normal.
- Calibration of the APEX model with water quality data remains an essential step.

EXCESSIVE PHOSPHORUS (P) and sediment loss from agricultural fields continues to degrade fresh water quality despite decades of efforts to understand loss processes and implement management practices to reduce nonpoint pollution (e.g., Sharpley et al., 1994, 2015; Sims and Kleinman, 2005; Jarvie et al., 2013). Failure of conservation practices to produce expected improvements in water quality has renewed appreciation for the complexity of P movement within landscapes and to waterbodies (Jarvie et al., 2013; Sharpley et al., 2013). Process-based watershed- and field-scale models offer the prospect of integrating knowledge to assess and quantify impacts of conservation and management practices.

The P Index was developed in the mid-1990s to assess risk of P loss from agricultural land (Lemunyon and Gilbert, 1993) and is now an integral part of the NRCS 590 Nutrient Management Standard and other state and federal programs (Sharpley et al., 2003, 2017). However, the diversity in P Index ratings and P management recommendations for similar conditions led Sharpley et al. (2012) and Osmond et al. (2006) to emphasize the need for science-based assessment and improvement of existing P Indices.

Extensive testing of P Indices requires water quality data from field-scale watersheds from a broad range of soils and management scenarios with a sufficient number of monitoring years to estimate long-term average annual losses. Such extensive datasets are rare. Hence, computer models must be a core component of any strategy to assess and improve P Indices (Bolster et al., 2012; Sharpley et al., 2012) by extending limited datasets to simulated edge-of-field losses over a wide range of conditions and to long-term average estimates. Phosphorus Index usefulness is assessed by ensuring...
that P Index ratings and ensuing P management recommendations are directionally consistent with model results. Alternatively, models can be used to assess specific components of a P Index, e.g., the runoff volume, which serves to assess the accuracy of the component and the weights assigned to each component (Bolster et al., 2012). These strategies, however, require that the model used is appropriately parameterized over the range of conditions relevant for P Index validation, i.e., that one has confidence in the model results. The benefits of models can only be achieved if a model parameterization can be successfully used for management scenarios, weather scenarios and/or physiographic areas beyond the calibration space of the model. The greatest benefits of models could be achieved if it was possible to successfully parameterize a model without calibration using water quality data.

The Agricultural Policy Environmental eXtender (APEX; Gassman et al., 2010) model is a daily time-step model that simulates plant growth, water movement, and fate and movement of sediment, nutrients, and pesticides in a catchment. It is especially useful for field-size catchments (30–300 ha) because fields can be described as linked subareas that are simulated separately, thus providing options to simulate features such as filter strips, buffers, and grass waterways. Surface and subsurface flow, and associated nutrient and pesticide fluxes, are routed from each subarea to the outlet of the catchment through a network of channels. Model inputs include the physical characterization of each subarea and of the channels (e.g., slope, Manning coefficients), management information, physical and chemical soil characterization, and daily weather data. Inputs also include control parameters that select methods used to simulate specific processes (e.g., equation used to simulate soil erosion by water), as well as global parameters that define thresholds and rate coefficients for selected processes. A complete description of the inputs is provided in the APEX user’s manual (Steglich and Williams, 2013). Choice of simulation methods and global parameter values can significantly affect results of the model. However, there is little guidance on how to set these parameters for any given situation, and they are usually defined through the process of model calibration and validation.

The APEX model has successfully simulated edge-of-field runoff, sediment and P loss for specific management scenarios after calibration of the model using measured data from that location (Wang et al., 2008; Yin et al., 2009; Kumar et al., 2011; Mudgal et al., 2012; Plotkin et al., 2013; Senaviratne et al., 2013). There are published studies using APEX with limited or no calibration with water quality data to predict water quality outcomes (Harman et al., 2004; Williams et al., 2006; Tuppad et al., 2010). The most extensive use of APEX to date is the Conservation Effects Assessment Project (CEAP; Duriancik et al., 2008). The CEAP project used a detailed farmer survey conducted by the National Agricultural Statistical Service (NASS) and data from the USDA National Resource Inventory to develop regional parameterizations of APEX. A sensitivity analysis identified sensitive parameters at the field scale (Wang et al., 2006). Integration of APEX and the Soil and Water Assessment Tool (SWAT) to simulate the Upper Mississippi River Basin (Wang et al., 2011) provided some basis for the calibration of surface water and sediment transport at the eight-digit watershed level. However, sediment and P-loss estimates generated from these APEX parameterizations were not compared back with measured field-scale data. Thus, there is a need for formal APEX performance evaluation for event or annual runoff and water quality constituents when APEX is not calibrated with edge-of-field data.

The objectives of this paper were to develop a best professional judgment (BPJ) parameterization of APEX0806 for estimating edge-of-field runoff, sediment, and P losses in restricted-layer soils common in Missouri and Kansas and to assess the performance of this parameterization using monitoring data from multiple sites in this region.

Materials and Methods

Processing of Monitoring Data

Four sites including 12 datasets were selected across the Heartland region (Supplemental Material, Section S1; Supplemental Fig. S1) based on the availability of edge-of-field runoff monitoring data (Supplemental Table S1), as well as site-specific soil, management, and weather (Supplemental Table S2). For each dataset, subdaily and daily rainfall and runoff data were first aggregated to event values. For these small (<35 ha) field-scale watersheds, there was no baseflow; events were defined as a day or consecutive days with runoff bracketed by days without runoff. Hourly rainfall records were examined to review rainfall on the day prior to a runoff event to determine whether rain contributed directly to the runoff event. Rainfall and runoff data were further reviewed to exclude events where equipment malfunction was suspected, including instances of runoff greater than rainfall or runoff that was significantly less than or greater than expected, given other data from that location. Annualized sums were calculated by aggregating verified events by year. Event mean, standard deviation, and ranges of annualized sums are tabulated in Supplemental Table S3 for each dataset.

Model Used and Generic Implementation

We acquired the executable and associated code for APEX (version APEX0806; Wang et al., 2012) in October 2012. Subsequently, we modified it to correct identified errors or to increase modeling capability to fit our needs. These changes were made by, or in consultation with, the APEX developers, but we incorporated them in our version so that we would not be affected by other changes in the version maintained by the APEX developers. Most notable were changes in the simulation of contour filter strips and the change in P leaching from the surface (1-cm) layer of the soil profile. These changes are described in Supplemental Material, Section S2.

Best Professional Judgment and Site-Specific Parameterizations

The BPJ parameterization was defined based on our understanding of the hydrological and biological processes that take place in a field, our understanding of APEX, and the parameterization outlined for CEAP, as described by Wang et al. (2011). The BPJ parameterization is intended to be implemented with no field measurement, relying instead on regional and national data for soils, weather, and topography. It did assume that management records were available to define crops and associated field operations.

Delineation Process

For any site, the modeling process starts with subarea delineation based on the knowledge of the site. In APEX, each subarea has homogeneous soil, slope, and management properties. Each
site was delineated as simply as possible while maintaining the spatial variability of soils and management features. As a result, the larger sites had distinct subareas for each soil type, while the smaller sites had only one subarea. Features such as filters, buffers, terraces, and grass waterways were represented as individual subareas to represent different land covers in these areas. Main channels and drainage ways were also considered for the definition of the drainage network and subareas. Delineation was conducted either manually or with the aid of ArcGIS tools (ESRI, Redlands, CA) to process elevation and soil maps, and input files were built with the WinAPEX interface. Characteristics of channel lengths, routing reach lengths, channel top and bottom widths, slopes, slope lengths, Manning coefficients, floodplain width for buffer areas, and fraction of buffer that is bypassed were entered based on observations (channels, channel conditions, buffer conditions) made during site visits. Detailed maps of the subarea delineations for the complex watersheds can be found in Senaviratne et al. (2013) and in Supplemental Material, Section S1.

Management and Soil Properties

Availability of detailed management information was a selection requirement of the sites. Thus, information on crops grown and details on field operations such as tillage, fertilization, and harvest were known for each site and manually entered. Supplemental Table S2 provides information on the sources of data at each site. For the BPJ parameterization, we identified soil map units and used the corresponding soil properties from the SSURGO database (Soil Survey Staff, 2014). Since none of these sites were artificially drained and most soils were poorly drained, most soils were classified in the hydrologic group D. Using the representative value from the SSURGO database, the following soil properties were specified for each soil layer: soil albedo (first layer), soil layer depth, bulk density, sand and silt content, pH, organic carbon content, cation exchange capacity, and saturated conductivity. The number of years of cultivation, which affects mineralization of organic nitrogen and thus crop yields, was set to 100, a value in line with history of agriculture in the region. The mineralization of organic nitrogen and thus crop yields, was set to 100, a value in line with history of agriculture in the region. The number of years of cultivation, which affects mineralization of organic nitrogen and thus crop yields, was set to 100, a value in line with history of agriculture in the region. The number of years of cultivation, which affects mineralization of organic nitrogen and thus crop yields, was set to 100, a value in line with history of agriculture in the region. The number of years of cultivation, which affects mineralization of organic nitrogen and thus crop yields, was set to 100, a value in line with history of agriculture in the region. Availability of detailed management information was a selection requirement of the sites. Thus, information on crops grown and details on field operations such as tillage, fertilization, and harvest were known for each site and manually entered. Supplemental Table S2 provides information on the sources of data at each site. For the BPJ parameterization, we identified soil map units and used the corresponding soil properties from the SSURGO database (Soil Survey Staff, 2014). Since none of these sites were artificially drained and most soils were poorly drained, most soils were classified in the hydrologic group D. Using the representative value from the SSURGO database, the following soil properties were specified for each soil layer: soil albedo (first layer), soil layer depth, bulk density, sand and silt content, pH, organic carbon content, cation exchange capacity, and saturated conductivity. The number of years of cultivation, which affects mineralization of organic nitrogen and thus crop yields, was set to 100, a value in line with history of agriculture in the region. The mineralization of organic nitrogen and thus crop yields, was set to 100, a value in line with history of agriculture in the region. The number of years of cultivation, which affects mineralization of organic nitrogen and thus crop yields, was set to 100, a value in line with history of agriculture in the region. The number of years of cultivation, which affects mineralization of organic nitrogen and thus crop yields, was set to 100, a value in line with history of agriculture in the region. The number of years of cultivation, which affects mineralization of organic nitrogen and thus crop yields, was set to 100, a value in line with history of agriculture in the region.
acceptable for runoff for two-thirds of the sites and is not acceptable for sediment and TP at most sites.

With calibration, all indicators showed improvement in model performance (Table 1). Calibration resulted in a reduced range of values for each indicator, as indicated by a smaller standard deviation, and increased the number of sites where PEC were met. Sites for which sediment $r^2$ and NSE did not meet the PEC were characterized by small sediment loads (average annualized load < 0.8 Mg ha$^{-1}$; Bhandari et al., 2017; Nelson et al., 2017).

Table 1. Characteristics of performance indicators for event runoff, sediment, and total phosphorus simulation at the 12 sites.

<table>
<thead>
<tr>
<th></th>
<th>PBIAS†</th>
<th></th>
<th>$r^2$</th>
<th></th>
<th>NSE§</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BPJ‡</td>
<td>Calibrated</td>
<td>BPJ‡</td>
<td>Calibrated</td>
<td>BPJ‡</td>
</tr>
<tr>
<td>Runoff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>30</td>
<td>18</td>
<td>0.64</td>
<td>0.81</td>
<td>0.29</td>
</tr>
<tr>
<td>Median</td>
<td>24</td>
<td>14</td>
<td>0.69</td>
<td>0.80</td>
<td>0.48</td>
</tr>
<tr>
<td>Min.</td>
<td>2</td>
<td>3</td>
<td>0.24</td>
<td>0.69</td>
<td>−1.62</td>
</tr>
<tr>
<td>Max.</td>
<td>71</td>
<td>36</td>
<td>0.87</td>
<td>0.91</td>
<td>0.81</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>19</td>
<td>11</td>
<td>0.21</td>
<td>0.06</td>
<td>0.67</td>
</tr>
<tr>
<td>No. sites within PEC¶</td>
<td>8</td>
<td>11</td>
<td>9</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>Sediment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1256</td>
<td>36</td>
<td>0.30</td>
<td>0.43</td>
<td>−1494</td>
</tr>
<tr>
<td>Median</td>
<td>757</td>
<td>35</td>
<td>0.26</td>
<td>0.35</td>
<td>−58</td>
</tr>
<tr>
<td>Min.</td>
<td>48</td>
<td>9</td>
<td>0.10</td>
<td>0.25</td>
<td>−17031</td>
</tr>
<tr>
<td>Max.</td>
<td>8205</td>
<td>85</td>
<td>0.81</td>
<td>0.80</td>
<td>0.68</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>2248</td>
<td>22</td>
<td>0.19</td>
<td>0.18</td>
<td>4894</td>
</tr>
<tr>
<td>No. sites within PEC§</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Total phosphorus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>112</td>
<td>22</td>
<td>0.59</td>
<td>0.80</td>
<td>−3.31</td>
</tr>
<tr>
<td>Median</td>
<td>62</td>
<td>15</td>
<td>0.60</td>
<td>0.79</td>
<td>−0.13</td>
</tr>
<tr>
<td>Min.</td>
<td>16</td>
<td>1</td>
<td>0.22</td>
<td>0.61</td>
<td>−28.58</td>
</tr>
<tr>
<td>Max.</td>
<td>552</td>
<td>59</td>
<td>0.97</td>
<td>0.99</td>
<td>0.82</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>156</td>
<td>20</td>
<td>0.23</td>
<td>0.13</td>
<td>8.58</td>
</tr>
<tr>
<td>No. sites within PEC¶</td>
<td>6</td>
<td>12</td>
<td>8</td>
<td>12</td>
<td>3</td>
</tr>
</tbody>
</table>

† PBIAS, percent bias.
‡ BPJ, best professional judgment parameterization.
§ NSE, Nash–Sutcliffe efficiency.
¶ PEC, performance evaluation criteria

Normal probability plots of the 12 estimates of PBIAS, $r^2$, and NSE for the BPJ models imply that these performance evaluation coefficients were not normally distributed (Fig. 1) and had skewed distributions, as indicated by the poor fit to a linear trend line. Small deviations from the linear trend line are indicative of small, random, and independent errors such as small measurement errors, small misrepresentations of the factors and processes that control the variable, proximity of the weather station, expertise of the modeler, etc. Large deviations can indicate misrepresentation of the processes simulated by the model.

Fig. 1. Normal probability plots of percent bias (PBIAS), coefficient of determination ($r^2$), and Nash–Sutcliffe efficiency (NSE) for the best professional judgment model for runoff, sediment, and total phosphorus for 12 watersheds. Sediment PBIAS and NSE are not shown because their range prevents readability of the plots.
Deviations were particularly marked for NSE and PBIAS; distributions for $r^2$ were closer to a normal distribution, especially for TP, as indicated by the good fit to the linear trend line. On the other hand, all values obtained with the calibrated models were approximately normally distributed (Fig. 2). The Shapiro–Wilks statistics confirmed that, among the BPJ models, NSE and PBIAS distributions did not follow a normal distribution. Only $r^2$ values for BPJ runoff and TP met the normality assumption. In contrast, for the calibrated models, all PEC except $r^2$ values for sediment met the normality assumption. Regional event-based evaluation confirmed these results (Table 2), as shown by $|\text{PBIAS}| > 100\%$ and negative NSE for sediment and TP under BPJ. Negative NSE values indicate that the mean of measured loads is a better predictor than results simulated with BPJ. For event data at the regional scale, all PEC for the calibrated models were met (Table 2), except that $r^2$ was marginally $>0.5$ and the regression slope for event sediment loss was $>0.5$ due to underestimation of high sediment loads. Results based on annualized values confirmed the event-based results (Fig. 3): BPJ runoff estimates were acceptable but not sediment or TP. Annualized values obtained

![Normal probability plots of percent bias (PBIAS), coefficient of determination ($r^2$), and Nash–Sutcliffe Efficiency (NSE) obtained with the fully calibrated models for runoff, sediment, and total phosphorus for 12 watersheds.](image)

**Fig. 2.** Normal probability plots of percent bias (PBIAS), coefficient of determination ($r^2$), and Nash–Sutcliffe Efficiency (NSE) obtained with the fully calibrated models for runoff, sediment, and total phosphorus for 12 watersheds.
Table 2. Regional performance indicators at the event temporal scale for runoff, sediment, and total phosphorus (TP).

<table>
<thead>
<tr>
<th>BPJ†</th>
<th>Runoff</th>
<th>Sediment</th>
<th>TP</th>
<th>Calibrated model</th>
<th>Runoff</th>
<th>Sediment</th>
<th>TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBIAS (%)</td>
<td>28</td>
<td>-653</td>
<td>-103</td>
<td>14</td>
<td>24</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>$r^2$</td>
<td>0.59</td>
<td>0.05</td>
<td>0.77</td>
<td>0.74</td>
<td>0.47</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>NSE</td>
<td>0.53</td>
<td>-44</td>
<td>-0.14</td>
<td>0.72</td>
<td>0.43</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Regression slope</td>
<td>0.58</td>
<td>1.42</td>
<td>1.57</td>
<td>0.69</td>
<td>0.35</td>
<td>0.89</td>
<td></td>
</tr>
</tbody>
</table>

† BPJ, best professional judgment parameterization.
‡ PBIAS, percent bias.
§ NSE, Nash–Sutcliffe efficiency.

Fig. 3. Simulated versus measured annualized runoff, sediment, and total phosphorus (P) for 61 site–years across 12 watersheds for the best professional judgment (BPJ) and fully calibrated models, along with values of percent bias (PBIAS), coefficient of determination ($r^2$), and Nash–Sutcliffe efficiency (NSE).
with the calibrated models all passed PEC; performance was very good for runoff and TP and less so for sediment, as indicated by the smaller values of $r^2$, NSE and regression slope.

**Discussion**

**Using APEX with BPJ Parameterization for Event Data at a Specific Site**

Individual site results indicated that using APEX with BPJ parameterization may be appropriate for the purpose of estimating surface runoff volume from cropland underlain by claypan soils but can lead to erroneous results for sediment and TP components. In our dataset, runoff performance indicators using the BPJ parameterization strategy were acceptable for two-thirds of the sites (Table 1). Obviously, it is difficult to set an acceptable fail/pass ratio when using a model without calibration and two-thirds of sites that meet the PEC may be sufficient in some cases; objectives of the specific study would dictate what an acceptable number should be. Nonetheless, in cases where water quality simulation is an objective, correct simulation of runoff is critical (e.g., Arnold et al., 2015).

In this study, in spite of correct runoff simulation for two-thirds of the sites, APEX did not simulate correct sediment loads in all but one case with the BPJ parameterization. At 7 of 12 sites, sediment loss was overestimated by greater than 500%, in large part because of the choice of erosion equation. A comparison of BPJ and calibrated sediment-related parameters revealed significant differences that explain sediment overestimation with BPJ (Supplemental Tables S7 and S8). The Modified Universal Soil Loss Equation (MUSLE) was selected for BPJ because that option was used in the CEAP study (Wang et al., 2011). None of our datasets could be calibrated with the MUSLE, instead requiring the MUSS equation, which is an adaptation of MUSLE to small watersheds with no channel erosion (Williams et al., 2012). While MUSS and MUSLE both calculate sediment loss as a function of runoff volume, runoff peak rate, and drainage area, MUSLE typically estimates greater sediment loss. In addition, sediment routing coefficients were greater with BPJ, which also increases sediment loss. The choice of MUSS for erosion calculation seems appropriate for these watersheds, given that they are small (<35 ha), many of them are in a no-till system, and some have grassed waterways. However, there are no guidelines in the APEX manual or documentation to guide the choice of erosion equation.

Sediment results after calibration were still not satisfactory for some of the Knox, Franklin, and Crawford datasets. Possible factors include the large measurement errors for edge-of-field sediment losses (Harmel et al., 2006), sediment deposition caused by backwater induced by the flume or the weir, and the use of a USLE-type equation, which does not consider the true rainfall distribution during an event and does not account for gully erosion. Given the amount of work required to complete these calibrations, results were disappointing and point out the need for improvement of the algorithms used for estimating edge-of-field sediment loss with APEX.

The BPJ results for TP were also not acceptable; the BPJ NSE values were acceptable for only 25% of sites in spite of acceptable $r^2$ values for two-thirds of the sites. Percent bias was acceptable for half of the sites, and for sites with unacceptable PBIAS, the BPJ overestimated TP loss by an average of 200%. Total P estimates are the sum of dissolved P (DP) and sediment-bound P estimates. Because simulated sediment losses were grossly overestimated, the sediment-bound fractions of TP were overestimated as well. The accuracy of model estimates for particulate P and DP can be evaluated by comparing measured and simulated DP/TP ratios, which was only possible at the Franklin and Crawford sites. At the Franklin sites, where inorganic fertilizer was used, the mean measured ratio of DP/TP was 45%, compared with 16% (or 30% points less) with the BPJ and 63% with the calibrated models. At the Crawford site fertilized with turkey litter, the mean measured ratio was 88%, compared with 36% with the BPJ and 99% with the calibrated models.

While inaccurate initialization of TP concentrations in the soil could explain the overestimation of sediment-bound P, this was not the case here. The APEX model initializes total inorganic P pools as a function of the initial labile P and the PSP, where increasing the PSP decreases the initial amount of total inorganic P in the soil. Calculated values of PSP used for the calibrated models were all <0.5, which was the default PSP for the BPJ model. Therefore, total soil P values in the BPJ models were less than both measured data and that simulated by the calibrated models. Thus, overestimation of sediment loss by the BPJ model, not inaccurate initial soil P concentrations, caused the overestimation of sediment-bound P and thus TP. Although the calibrated models overestimated this DP/TP ratio, likely by underestimating the sediment transport, as indicated by the positive sediment PBIAS for the calibrated models (Table 2), the ratios were closer to measured data.

Overall, when APEX was used with the PBJ parameterization, sediment and TP losses from most locations were overestimated by large amounts, indicating that we cannot use BPJ to generate data and assess the P Index at specific locations. Furthermore, the fact that the BPJ model provided acceptable estimates of runoff yet unacceptable results for sediment and TP loss illustrates that acceptable estimates of runoff do not necessarily translate into acceptable estimates of sediment or TP loss.

**Distribution of Performance Indicators Values**

Performance indicators measure the goodness of fit between the simulated values and the data. Indicator values that meet the PEC reflect the adequacy of the model, including the algorithms and equations embedded in the code and the parameterization used for each site, with the data collected at that site. The fact that distributions of performance indicators ($r^2$, NSE, and PBIAS) for calibrated models approximately fit a normal distribution indicated that variability in performance was caused by small errors in the model, the model parameters, or the data used. One exception to this was the distribution of sediment $r^2$ obtained with the calibrated models, as the assumption of normality had to be rejected ($\alpha = 0.05$). In contrast, most distributions of the same performance indicators obtained with the BPJ parameterization did not follow a normal distribution, indicating a fundamental flaw between simulated values and measured values. Given that it was possible to calibrate the model at each of the sites, we conclude that the model and the data were essentially good but that the BPJ parameterization was not appropriate.

In their development of PEC, Moriasi et al. (2015) found that $r^2$ values in their database of published studies were approximately normally distributed. Nash–Sutcliffe efficiency and
PBIAS, which they also considered, were not. Similarly to these authors, we found that $r^2$ values for runoff and TP were approximately normally distributed with the fully calibrated parameterization among these 12 watersheds. We could not confirm this for sediment, perhaps indicative that several models (5 out of 12) were not well calibrated for sediment. In contrast with the results from Moriasi et al. (2015), both NSE and PBIAS values were normally distributed. Thus, in this study, when the model was performing acceptably, performance indicators values followed a normal distribution. If this can be confirmed with a larger data-set of calibrated models, PEC values identified as outliers relative to the distributions of performance measures may indicate that the model is not simulating processes accurately or that the data used for evaluation are not representative.

**Using APEX with BPJ Parameterization for Regional Analyses**

Results obtained with the results from the 12 sites together, either at the event temporal scale or by annualizing the results, showed that APEX could be used with the BPJ parameterization strategy for the purpose of evaluating runoff from crop-land underlain by claypan soils (Table 2, Fig. 3). Thus, it would be appropriate to use APEX with the BPJ parameterization to evaluate the variability of runoff within this region, as affected by topography or management, for example. Variability of soils was less well represented within this dataset, since all soils were considered claypan soils and were classified in the hydrological group D. In the context of the P Index, APEX using the BPJ parameterization could be used to evaluate the runoff volume component of the index—e.g., by generating runoff values across a large range of slopes, weather scenarios, and management scenarios on claypan soils and other soils with a restrictive layer.

Sediment loss estimates obtained with the BPJ parameter set were still not accurate when considered collectively across the 12 sites (Table 2, Fig. 3). Minimal sediment losses from the Crawford and Franklin sites (four no-till datasets) were so overestimated that estimates were greater than soil loss estimates from any other site, thus making inappropriate the use of this model for regional sediment loss analyses. Consequently, TP losses were not accurate as well because of the overestimation of sediment-bound P. As a result, APEX and the BPJ parameterization could not be used to test whether a P Index, or any other tool, provides directionally correct information because the BPJ parameterization did not provide that information itself when tested across these 12 sites. These results do not necessarily mean that APEX never can be used for locations that have no calibration data. A regional parameterization, based on the calibrated parameter sets obtained for each dataset, was investigated by Nelson et al. (2017) and showed encouraging results for runoff and TP on restricted-layer soils.

**Conclusion**

Performance of the APEX model was assessed for 12 datasets from four sites in Missouri and Kansas using two parameterization strategies: a BPJ parameterization common to all sites and a site-specific parameterization derived through systematic calibration and validation of the model. Each parameterization was assessed for model performance using runoff, sediment, and P-loss monitoring data at each site. The BPJ parameterization resulted in a large range of performance indicators values, some indicating acceptable performance but the majority showing unacceptable performance. The model performance was greatly improved through calibration and the range of indicator values was reduced.

Model performance was also assessed with an annualized time step across the 12 sites. While BPJ performance for runoff simulation was acceptable, performance for sediment and TP was not. Thus, APEX with a BPJ parameterization could not be used as the basis for sediment or TP evaluation or for the evaluation of P Indices. It potentially could be used to evaluate the runoff component of these indices, provided it is successfully tested on soils with different characteristics. Other papers in this special section include analyses of the calibrated parameterizations within and beyond the calibration conditions (Bhandari et al., 2017) and development of a regional parameterization that could provide acceptable performance (Nelson et al., 2017). The results of this analysis emphasize the essential role of calibration of APEX with water quality data to ensure accurate model estimates of TP in edge-of-field runoff from small agricultural watersheds.

**Acknowledgments**

The authors are grateful for partial funding for this project through a USDA Conservation Innovation Grant (Agreement Number 69-3A75-12-193).

**References**


