Real-time water quality monitoring and regression analysis to estimate nutrient and bacteria concentrations in Kansas streams

V.G. Christensen, P.P. Rasmussen and A.C. Ziegler
U.S. Geological Survey, 4821 Quail Crest Place, Lawrence, KS, 66049, U.S.A.

Abstract An innovative approach currently is underway in Kansas to estimate and monitor constituent concentrations in streams. Continuous in-stream water-quality monitors are installed at selected U.S. Geological Survey stream-gaging stations to provide real-time measurement of specific conductance, pH, water temperature, dissolved oxygen, turbidity, and total chlorophyll. In addition, periodic water samples are collected manually and analyzed for nutrients, bacteria, and other constituents of concern. Regression equations then are developed from measurements made by the water-quality monitors and analytical results of manually collected samples. These regression equations are used to estimate nutrient, bacteria, and other constituent concentrations. Concentrations then are available to calculate loads and yields to further assess water quality in watersheds. The continuous and real-time nature of the data may be important when considering recreational use of a water body; developing and monitoring total maximum daily loads; adjusting water-treatment strategies; and determining high constituent concentrations in time to prevent adverse effects on fish or other aquatic life.

Keywords Fecal coliform bacteria; nitrogen; phosphorus; real-time monitoring; regression analysis; water quality

Introduction
Historically, the U.S. Geological Survey’s (USGS) stream-gaging network has provided timely water-quantity information to resource managers and others to make informed decisions about floods and water availability. It has not been possible, however, to provide water-quality information in the same manner. Timely water-quality information is useful for many reasons, including assessment of total maximum daily loads (TMDLs) and the effects of urbanization and agriculture on a water supply.

Nutrients are a concern because large inputs of nitrogen and phosphorus compounds into the aquatic environment can cause excessive algal growth. When algal blooms die, concentrations of dissolved oxygen are depleted, which can stress aquatic organisms and may cause taste-and-odor problems in water supplies. Large nutrient concentrations in drinking water also may have adverse physiological effects on humans and may interfere with growth and reproduction of aquatic organisms. Fecal coliform bacteria in surface water are a concern because they indicate the possible presence of other organisms that could cause disease. Therefore, it is desirable to mitigate the introduction of excess nutrients and bacteria in surface water used as public supplies or where sensitive aquatic organisms may be present (Christensen and Pope, 1997).

Estimates of constituent concentrations, loads, and yields are useful to help identify source areas for nutrients and bacteria and to develop mitigation strategies. Estimated concentrations are useful for evaluating a water body with respect to current water-quality criteria. In the past, to determine concentrations of nutrients and bacteria in a stream, it was necessary to manually collect samples and send them to a laboratory for analysis. These analytical methods require at least 24 hours, and when human health is a concern, immediate information is important. Loads, the chemical mass of a constituent transported by a
stream during a given period of time, are particularly important when considering the amount of nutrients entering a marsh, lake, or reservoir. Load estimates also are important to the establishment and monitoring of TMDLs mandated by the Clean Water Act of 1972. Finally, yield estimates, the constituent mass transported per unit area over a given period of time, may be used by resource managers and regulatory authorities to help prioritize efforts with regard to land-use best-management practices.

In response to the need for timely and continuous water-quality information, the USGS, in cooperation with State and other Federal agencies, began using an innovative, continuous, real-time monitoring approach for several Kansas streams. This paper describes results of a study to provide continuous estimates of real-time nutrient and bacteria concentrations for four stream-gaging stations in Kansas (Figure 1). The stations, located in the Kansas River, Rattlesnake Creek, and Little Arkansas River Basins, were chosen to represent various drainage areas, locations, and land uses.

Methods
To develop a more timely method of assessing the quality of water in Kansas streams, continuous in-stream water-quality monitors were installed at four USGS stream-gaging stations to provide real-time measurement of specific conductance, pH, water temperature, dissolved oxygen, turbidity, and total chlorophyll. Periodic water samples were collected manually and analyzed for nutrients, bacteria, and other constituents of concern. The manual samples were collected using depth- and width-integrating techniques (Ward and Harr, 1990) throughout the year and throughout at least 90 percent of the stream’s flow conditions to describe a wide range of seasonal and hydrologic conditions.

Figure 1 Location of selected U.S. Geological Survey stream-gaging stations in the Kansas River, Rattlesnake Creek, and Little Arkansas River Basins
There are advantages to installing the real-time water-quality monitors at USGS stream-gaging stations – field technicians visit the sites regularly; historical water-quantity and quality data may be available; and discharge (Q) is available to use as an explanatory variable in regression equations and as required for load estimates. Discharge at USGS stream-gaging stations is recorded and reported in cubic feet per second (ft³/s). To obtain cubic metres per second (m³/s), multiply discharge in cubic feet per second (ft³/s) by 0.02832. The conversion will not alter the form of the regression equations presented herein; only the units of measurement will be affected.

Linear regression equations were developed, using the least-squares method, from sensor measurements of the water-quality monitors (explanatory variables) and analytical results of manually collected water samples (response variables) for each water-quality constituent at each site. In the least-squares method, the equation chosen is the one in which the sum of the squared errors for all sample points is minimized (Ott, 1988, p. 441). The regression equations were used to estimate nutrient and bacteria concentrations. The measures used to evaluate the equations were the coefficient of determination (R²) and the mean square error (MSE). The R² is the fraction of the variance explained by regression, and the MSE is presented to assess the variance between measured and estimated values. The relative percentage differences (RPDs) between measured and estimated concentrations were calculated using the absolute value of the equation, 

\[ RPD = \left| \frac{B - A}{A} \right| \times 100 \]

where A is the instantaneous measured concentration and B is the instantaneous estimated concentration. Measured loads were based on manually collected samples and the corresponding instantaneous streamflow measurements. Additional information on statistical methods related to this study can be found in Christensen et al. (2000). Current information on continuous, real-time water quality in Kansas may be accessed through the Internet at http://ks.water.usgs.gov/Kansas/qw.

Results

Major sources of nitrogen, phosphorus, and fecal coliform bacteria in Kansas include agricultural activities, such as the pasturing and confined feeding of livestock, and municipal wastewater discharges. Nitrogen and phosphorus may also be the result of the application of synthetic fertilizers, and phosphorus may come from geologic sources. Fecal coliform bacteria analyses were included in this paper because current (2001) and State of Kansas water-quality criteria [2,000 col/100 mL (colonies per 100 millilitres of water) for noncontact recreation and 900 col/100 mL (proposed) for whole-body contact recreation] are based on fecal coliform bacteria densities (Kansas Department of Health and Environment, 2000) and bacteria are the most common cause of stream impairment according to the 1998 305b report for Kansas.

Not all of the regression equations use the same explanatory variables because the equations are site specific. Turbidity is a common explanatory variable in all the equations for all four gaging stations listed in Table 1. The range in turbidity values for each equation is presented in Table 1 to indicate the range for which the equation is valid. Specific conductance and water temperature are included in the equation for total nitrogen at the Rattlesnake Creek station (07142575, Figure 1), and provide an indication of the difference in land use and hydrologic characteristics at this station. The bacteria equation for the Kansas River station (06892350, Figure 1) is simpler than for the other three stations, in which season (day of year) is a significant explanatory variable. In addition, discharge (Q) is significant in some of the nitrogen and bacteria equations. The following discussion will describe some of the land-use and hydrologic characteristics of the four gaging stations and will help explain the differences in explanatory variables used.

Through the least-squares process, certain explanatory variables (such as turbidity) are
selected that have a significant relation to the response variable (nitrogen, phosphorus, or bacteria concentration, for example). However, the explanatory variables are included only if there is a physical basis for their inclusion. Water temperature and turbidity may be related to either total nitrogen or phosphorus because of their secondary relation to time (season) and the application of fertilizers and sediment transport. Specific conductance (which, in most Kansas streams, is high during low flow because of mineralized groundwater) may have an inverse relation to total nitrogen and phosphorus concentrations, which tend to be high during high flow. Finally, turbidity has a relation to both nitrogen and phosphorus because it is a measure of the amount of particulate matter transported by a stream and because total nitrogen and total phosphorus analyses include particulate forms. Turbidity also has a relation to fecal coliform bacteria, and because runoff from a watershed may transport sediment to streams, there may be a relation between bacteria densities and streamflow or possibly to time of year because runoff characteristics may vary with season.

The explanatory variables for the nitrogen and phosphorus equations differed among the three basins. The total nitrogen and phosphorus equations for the Rattlesnake Creek station (07142575, Table 1) include water temperature and specific conductance. Water temperature and specific conductance vary seasonally and can be used to describe the seasonal variation in total nitrogen in Rattlesnake Creek, which in turn may be due to the seasonality of agricultural activities such as fertilizer application. Discharge is a significant variable in the total nitrogen equations for the Kansas River gaging station (06892350, Table 1) and for the upstream Little Arkansas River station (07143672, Table 1). These two stations may be closer to sources of nitrogen in these watersheds, accounting for the better relation between discharge (and runoff) and nitrogen particulate matter.

The regression equations for fecal coliform bacteria also differed among the three basins. The Kansas River Basin, the largest of the three basins, is affected to a greater degree by reservoir releases and point sources, such as wastewater discharge, that may

Table 1 Linear regression equations for the estimation of total nitrogen, total phosphorus, and fecal coliform bacteria at four Kansas stream-gaging stations

<table>
<thead>
<tr>
<th>Station number (Figure 1)</th>
<th>Equation R^2</th>
<th>MSE</th>
<th>N</th>
<th>T range</th>
<th>Median RPD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total nitrogen</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06892250</td>
<td>TN = 0.001867T - 0.0000940Q + 1.08</td>
<td>0.916</td>
<td>0.0293</td>
<td>17</td>
<td>32.2-607</td>
</tr>
<tr>
<td>07142575</td>
<td>TN = 0.003257T + 0.0214VT - 0.0000796SC + 0.515</td>
<td>.764</td>
<td>.0754</td>
<td>18</td>
<td>4.25-301</td>
</tr>
<tr>
<td>07143672</td>
<td>TN = 0.004207T - 0.0000860G + 0.494</td>
<td>.977</td>
<td>.0333</td>
<td>13</td>
<td>10.0-649</td>
</tr>
<tr>
<td>07144100</td>
<td>TN = 0.002467T + 0.656</td>
<td>.952</td>
<td>.106</td>
<td>16</td>
<td>6.30-1,410</td>
</tr>
<tr>
<td><strong>Total phosphorus</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06892250</td>
<td>TP = 0.000606T + 0.186</td>
<td>0.964</td>
<td>0.00213</td>
<td>17</td>
<td>32.2-607</td>
</tr>
<tr>
<td>07142575</td>
<td>Log_{10}TP = 0.00165T + 0.0217VT - 0.000108SC - 1.06</td>
<td>.908</td>
<td>.0175</td>
<td>18</td>
<td>4.25-301</td>
</tr>
<tr>
<td>07143672</td>
<td>TP = 0.001067T + 0.310</td>
<td>.899</td>
<td>.00748</td>
<td>13</td>
<td>10.0-649</td>
</tr>
<tr>
<td>07144100</td>
<td>TP = 0.0008497T + 0.448</td>
<td>.509</td>
<td>.0343</td>
<td>13</td>
<td>6.30-1,410</td>
</tr>
<tr>
<td><strong>Fecal coliform bacteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06892250</td>
<td>FC = 3.41T - 24.4</td>
<td>.620</td>
<td>1,550,000</td>
<td>17</td>
<td>32.2-1,600</td>
</tr>
<tr>
<td>07142575</td>
<td>Log_{10}FC = -0.5027 sin (4(D/365)) + 0.820 cos (4(D/365)) + 0.0113T + 2.20 log_{10}Q + 0.000450SC - 3.71</td>
<td>.734</td>
<td>.204</td>
<td>20</td>
<td>4.25-231</td>
</tr>
<tr>
<td>07143672</td>
<td>Log_{10}FC = -0.129 sin (2(D/365)) - 0.325 cos (2(D/365)) + 0.892 log_{10}T + 0.878</td>
<td>.591</td>
<td>.406</td>
<td>101</td>
<td>0.30-1780</td>
</tr>
<tr>
<td>07144100</td>
<td>Log_{10}FC = -0.169 sin (2(D/365)) + 0.300 cos (2(D/365)) + 0.799 log_{10}T + 0.299 log_{10}Q + 0.474</td>
<td>.591</td>
<td>.561</td>
<td>102</td>
<td>1.44-1230</td>
</tr>
</tbody>
</table>
affect the relation between fecal coliform bacteria and the real-time measurements. The periodic sine and cosine functions are used in the bacteria equations for the other three gaging stations to account for the seasonal cycle of bacteria concentrations. The physical basis for the periodic sine and cosine functions is the seasonal variability of fecal coliform bacteria in streams. Cattle are one of the major sources of fecal coliform bacteria in the three basins described in this paper. During the spring when there is considerable rainfall, runoff from cattle-producing areas may result in large amounts of fecal coliform bacteria reaching streams. As was the case with the nutrients, time (day of year) is not significant in the regression equations for the Kansas River station probably because the Kansas River Basin is affected by many more point sources of wastewater discharges, which are not highly seasonal.

Some of the dependent variables (total phosphorus and fecal coliform bacteria) in the equations for the Rattlesnake Creek and Little Arkansas River stations (07142575, 07143672, and 07144100, Table 1) were transformed to eliminate curvature and achieve a simpler linear equation; therefore, consideration was given to retransformation bias when interpreting the results of the regression analysis. Retransformation has no effect on the form of the equations or on the error associated with the equations. However, retransformation can cause an underestimation of nutrient or fecal coliform bacteria loads when adding individual load estimates over a long period of time. Cohn et al. (1989), Gilroy et al. (1990), and Hirsch et al. (1993) provide additional information on interpreting the results of regression-estimated load estimates.

**Discussion**

The usefulness of regression-estimated concentrations to resource managers, regulatory authorities, and recreational users is illustrated in a graph of continuous regression-estimated data (Figure 2). The regression equation for the Rattlesnake Creek station (07142575, Table 1) was applied to real-time data collected in 2000. Users can see when regression-estimated concentrations of fecal coliform bacteria in Rattlesnake Creek exceeded non-contact and contact criteria. During 2000, noncontact recreational criteria for fecal coliform bacteria were exceeded numerous times in Rattlesnake Creek (Figure 2). Contact

**Figure 2** Measured and regression-estimated fecal coliform bacteria density for the Rattlesnake Creek station (07142575, Figure 1), 2000. Contact recreational water-quality criteria proposed and noncontact recreational criteria established by Kansas Department of Health and Environment (2000)
recreational criteria were exceeded many times during the summer months. Because the data were available in real time, immediate action could be taken to avoid contact with the water until conditions improved, or other water-use decisions regarding wildlife management could be made in time to avoid detrimental effects on aquatic species.

In addition to concentration estimates, load estimates can also be useful for resource managers and regulatory authorities. Load estimates for nutrients can help resource managers determine if possible eutrophication in downstream water bodies is a concern. In addition, regulatory authorities may use load estimates in the development of TMDLs. Annual load estimates for nitrogen and phosphorus at the four stations varied considerably (Figure 3). The estimated loads derived from the regression equations probably are more accurate than if they were estimated on the basis of a few manual samples collected during the year. This is because the data collected with the in-stream monitor are continuous, and significant changes in water quality are not missed.

Yield estimates (Figure 3) are also potentially useful to resource managers. The differences in estimated nutrient yields may help prioritize best-management practices within or among watersheds. Figure 3 shows that regression-estimated total phosphorus yields were largest in the Kansas River Basin (the largest of the four basins), yet the two stations in the Little Arkansas River Basin had the largest regression-estimated total nitrogen yield. The large differences between these two basins in point discharges, sediment trapping in reservoirs, and drainage size (which relates to travel time and the non-conservative nature of nitrogen species) partially explain this difference.

It would not be prudent to consider the estimates presented herein without understanding the error involved. The $R^2$, MSE, and RPD (Table 1) all give an indication of error. It should be noted that the bacteria equations have a higher degree of error than do those for nutrients. The high RPDs not only include the error of regression but also include analytical error, which can be as high as 50 percent for bacteria analysis. In addition, the water-quality monitors have an upper limit of measurement with respect to turbidity. This upper limit can cause an underestimation of bacteria concentration at turbidity values greater than about 1,400 NTU and may have a substantial effect on load and yield estimates. However, the loads and yields estimated using the methods described in this paper are preferable to
the alternative of estimating loads and yields that are based on a few samples collected during the year.

Conclusions
The use of regression equations to estimate nutrient and bacteria concentrations provides timely water-quality information to resource managers that is otherwise not available. The real-time availability of nutrient and bacteria data may be important for several reasons. Water suppliers need timely information to use in adjusting water-treatment strategies. Also, high nutrient or other constituent concentrations may be identified in time to prevent adverse effects on fish or other aquatic life. Fecal coliform bacteria densities are important when determining the safety of recreational activities such as swimming and fishing.

In addition to the utility of the regression equations to estimate concentrations, they are also useful for estimating total maximum daily loads (TMDLs). States are mandated to establish TMDLs for stream segments that have been identified by section 303 (d) of the 1972 Clean Water Act as limited for specific uses due to water-quality concerns. With the development of surrogate relations between continuous water-quality measurements and periodic collection of samples, a more accurate estimation of actual daily loads is probable. These estimated loads may be more reflective of actual loads because of the continuous nature of the in-stream data. In-stream water-quality monitors measure every hour (8,760 measurements per year). On the other hand, loads estimated from manually collected samples are based on discrete samples collected throughout the year, and peaks in concentrations, which frequently occur during peak-flow events, are likely missed, greatly increasing error.

The annual regression-estimated yields for the four gaging stations in this study were significantly different. Estimated yields for these and other stations, determined using the techniques presented in this paper, may be used to evaluate water-quality trends and the effectiveness of land-resource best-management strategies.

This innovative approach complements many other studies (particularly when historical data are available to help refine the regression equations) and utilizes the existing USGS stream-gaging network. The increasing public interest in TMDLs and water quality in general make this approach of regional, national, and international importance.

References