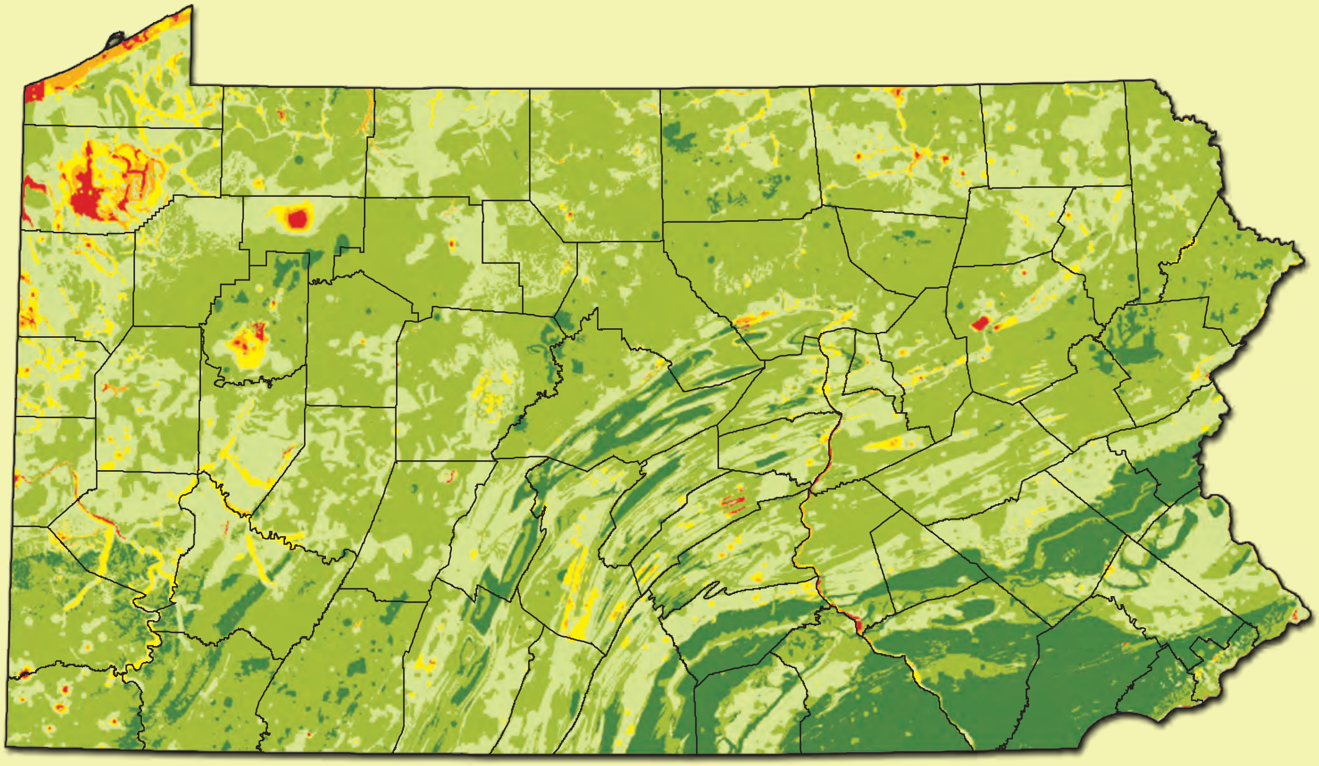
Prepared in cooperation with the
Pennsylvania Department of Health and the
Pennsylvania Department of Environmental Protection

Arsenic Concentrations, Related Environmental Factors, and the Predicted Probability of Elevated Arsenic in Groundwater in Pennsylvania



Scientific Investigations Report 2012–5257

U.S. Department of the Interior
U.S. Geological Survey

Cover: See figure 5 for an explanation of arsenic concentrations in groundwater.

Arsenic Concentrations, Related Environmental Factors, and the Predicted Probability of Elevated Arsenic in Groundwater in Pennsylvania

By Eliza L. Gross and Dennis J. Low

Prepared in cooperation with the
Pennsylvania Department of Health and the
Pennsylvania Department of Environmental Protection

Scientific Investigations Report 2012–5257

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
KEN SALAZAR, Secretary

U.S. Geological Survey
Suzette M. Kimball, Acting Director

U.S. Geological Survey, Reston, Virginia: 2013

Suggested citation:
Gross, E.L., Low, D.J., 2013, Arsenic concentrations, related environmental factors, and the predicted probability of elevated arsenic in groundwater in Pennsylvania: U.S. Geological Survey Scientific Investigations Report 2012–5257, 46 p.

Acknowledgments

This work was performed under the collaborative sub-grant 5U38EH000191-4 from the Pennsylvania Environmental Public Health Tracking Program (PA EPHT) of the Pennsylvania Department of Health, Bureau of Epidemiology, and the Pennsylvania Department of Environmental Protection.

Contents

Figures

Tables

Conversion Factors

SI to Inch/Pound

Multiply	By	To obtain
	Length	
centimeter (cm)	0.3937	inch (in)
millimeter (mm)	0.03937	inch (in)
meter (m)	3.281	foot (ft)
kilometer (km)	0.6214	mile (mi)
	Area	
square centimeter (cm^2)	0.001076	square foot (ft^2)
square meter (m^2)	10.76	square foot (ft^2)
square centimeter (cm^2)	0.1550	square inch (ft^2)
square kilometer (km^2)	0.3861	square mile (mi^2)
	Volume	
liter (L)	0.2642	gallon (gal)
	Mass	
gram (g)	0.03527	ounce, avoirdupois (oz)

Temperature in degrees Celsius (°C) may be converted to degrees Fahrenheit (°F) as follows:

°F=(1.8×°C)+32

Temperature in degrees Fahrenheit (°F) may be converted to degrees Celsius (°C) as follows:

°C=(°F-32)/1.8

Vertical coordinate information is referenced to the North American Vertical Datum of 1988 (NAVD 88).

Horizontal coordinate information is referenced to the North American Datum of 1983 (NAD 83).

Elevation, as used in this report, refers to distance above the vertical datum.

Specific conductance is given in microsiemens per centimeter at 25 degrees Celsius (µS/cm at 25 °C).

Concentrations of chemical constituents in water are given either in milligrams per liter (mg/L) or micrograms per liter (µg/L).

Arsenic Concentrations, Related Environmental Factors, and the Predicted Probability of Elevated Arsenic in Groundwater in Pennsylvania

By Eliza L. Gross and Dennis J. Low

Abstract

Analytical results for arsenic in water samples from 5,023 wells obtained during 1969–2007 across Pennsylvania were compiled and related to other associated groundwater-quality and environmental factors and used to predict the probability of elevated arsenic concentrations, defined as greater than or equal to 4.0 micrograms per liter (µg/L), in groundwater. Arsenic concentrations of 4.0 µg/L or greater (elevated concentrations) were detected in 18 percent of samples across Pennsylvania; 8 percent of samples had concentrations that equaled or exceeded the U.S. Environmental Protection Agency's drinking-water maximum contaminant level of 10.0 µg/L. The highest arsenic concentration was 490.0 µg/L.

Comparison of arsenic concentrations in Pennsylvania groundwater by physiographic province indicates that the Central Lowland physiographic province had the highest median arsenic concentration (4.5 µg/L) and the highest percentage of sample records with arsenic concentrations greater than or equal to 4.0 µg/L (59 percent) and greater than or equal to 10.0 µg/L (43 percent). Evaluation of four major aquifer types (carbonate, crystalline, siliciclastic, and surficial) in Pennsylvania showed that all types had median arsenic concentrations less than 4.0 µg/L, and the highest arsenic concentration (490.0 µg/L) was in a siliciclastic aquifer. The siliciclastic and surficial aquifers had the highest percentage of sample records with arsenic concentrations greater than or equal to 4.0 µg/L and 10.0 µg/L. Elevated arsenic concentrations were associated with low pH (less than or equal to 4.0), high pH (greater than or equal to 8.0), or reducing conditions. For waters classified as anoxic (405 samples), 20 percent of sampled wells contained water with elevated concentrations of arsenic; for waters classified as oxic (1,530 samples) only 10 percent of sampled wells contained water with elevated arsenic concentrations. Nevertheless, regardless of the reduction-oxidation classification, 54 percent of samples with low pH (13 of 24 samples) and 25 percent of samples with high pH (57 of 230 samples) had elevated arsenic concentrations.

Arsenic concentrations in groundwater in Pennsylvania were correlated with concentrations of several chemical constituents or properties, including (1) constituents associated with redox processes, (2) constituents that may have a similar origin or be mobilized under similar chemical conditions as arsenic, and (3) anions or oxyanions that have similar sorption behavior or compete for sorption sites on iron oxides.

Logistic regression models were created to predict and map the probability of elevated arsenic concentrations in groundwater statewide in Pennsylvania and in three intrastate regions to further improve predictions for those three regions (glacial aquifer system, Gettysburg Basin, Newark Basin). Although the Pennsylvania and regional predictive models retained some different variables, they have common characteristics that can be grouped by (1) geologic and soils variables describing arsenic sources and mobilizers, (2) geochemical variables describing the geochemical environment of the groundwater, and (3) locally specific variables that are unique to each of the three regions studied and not applicable to statewide analysis. Maps of Pennsylvania and the three intrastate regions were produced that illustrate that areas most at risk are those with geology and soils capable of functioning as an arsenic source or mobilizer and geochemical groundwater conditions able to facilitate redox reactions. The models have limitations because they may not characterize areas that have localized controls on arsenic mobility. The probability maps associated with this report are intended for regional-scale use and may not be accurate for use at the field scale or when considering individual wells.

Introduction

In many areas worldwide, including Pennsylvania, drinking water is the primary route of human exposure to arsenic (Hopenhayn, 2006). Arsenic data are sparse for groundwater because statewide testing of private wells to determine where concentrations exceed the health-based maximum contaminant level (MCL) of 10.0 micrograms per liter (µg/L) for drinking water, established in 2001 by the U.S. Environmental Protection Agency (USEPA), is not required throughout Pennsylvania (U.S. Environmental Protection Agency, 2006). Domestic wells used for private water supplies in Pennsylvania are not required to be routinely tested for arsenic and other contaminants, so homeowners may not know whether their well water has arsenic concentrations greater than the MCL.

Arsenic is a known carcinogen and consumption of arsenic in drinking water has been linked to multiple health problems, including bladder, lung, prostate, and skin cancers; cardiovascular disease; diabetes; and neurological disfunction (National Research Council, 1999, 2001; Hopenhayn, 2006; Chen and others, 2007; Benbrahim-Tallaa and Waalkes, 2007; Lin and others, 2008). Arsenic is also a potent endocrine disruptor that can cause problems with reproduction and embryotic development (Davey and others, 2007). In 2001, the USEPA decreased the drinking water MCL from 50.0 to 10.0 µg/L in recognition of the health risks associated with arsenic (U.S. Environmental Protection Agency, 2006). Although the USEPA regulates only public water-supply systems, the MCL has general applicability for the consumption of drinking water from private domestic wells.

Arsenic concentrations in Pennsylvania groundwater are difficult to predict on a well-by-well basis because (1) there is considerable local- and regional-scale spatial variability in groundwater quality and (2) arsenic has multiple anthropogenic and natural sources. However, the risk of elevated arsenic concentrations in groundwater is greater in some areas of Pennsylvania than in others (Low and Galeone, 2006). If areas with increased probability for elevated arsenic concentrations could be identified, health monitoring, water-quality monitoring, and educational programs could then be directed where the need is greatest. To address these concerns, the U.S. Geological Survey, in cooperation with the Pennsylvania Department of Health and Pennsylvania Department of Environmental Protection, undertook a study in 2010 to determine areas in Pennsylvania that have increased probability of elevated arsenic concentrations in groundwater using available data describing arsenic concentrations, groundwater chemistry, geology, and other factors.

Purpose and Scope

This report (1) documents arsenic concentrations in groundwater samples collected in Pennsylvania during 1969–2007, (2) describes the relation between arsenic concentrations and reduction-oxidation (redox) conditions and other groundwater-quality variables, and (3) documents the development of logistic regression models to represent the spatial relation between arsenic concentrations in groundwater and anthropogenic and natural factors. The models were developed using existing and constructed geographic information system (GIS) data for Pennsylvania and three intrastate regions (glacial aquifer system, Gettysburg Basin, and Newark Basin). Resulting model coefficients for selected spatial variables were used to produce maps displaying the predicted probability of elevated arsenic concentrations (greater than or equal to 4.0 µg/L) throughout the State and the selected intrastate regions.

Background on Arsenic Occurrence

Arsenic is a naturally occurring trace element in rock, soil, plants, and the aquatic environment. A recent review of occurrence of arsenic in natural waters describes some principal sources and mechanisms of arsenic mobility in groundwater (Smedley and Kinniburgh, 2002). Concentrations of arsenic in groundwater vary greatly owing to the uneven distribution of source materials and dynamic geochemical controls on aqueous arsenic mobility. Although arsenic can be introduced to the environment from anthropogenic sources (such as contaminant releases from industrial facilities or usage as a pesticide for agriculture), it commonly is present as a trace component in naturally occurring minerals, such as sulfides (pyrite), hydrous metal oxides (iron oxides), coal, ironstones, clays, phosphates, silicates, and carbonates. Pyrite and iron oxides are important sources of elevated arsenic in groundwater because they are abundant in aquifers, leading to their dissemination throughout the aquifer matrix or accumulation in fractures, joints, or bedding planes (Smedley and Kinniburgh, 2002).

Arsenic, present as arsenic minerals or as a trace component in other naturally occurring minerals in the soil and aquifers, can be released to or removed from the groundwater as a result of oxidation and reduction, dissolution and precipitation, and surface complexation (sorption) reactions on mineral surfaces. Arsenian pyrite [$Fe(S,As)_2$], arsenopyrite (FeAsS), and (or) other unspecified sulfide minerals in bedrock and surficial sediments are common parent sources for naturally occurring arsenic in the environment (Foster and others, 2003). Substitution of arsenic for sulfur in sulfide minerals can increase their susceptibility to weathering and dissolution when exposed to oxidants (Savage and others, 2000). Arsenic released to solution by sulfide oxidation commonly has a valence state of V or III and forms the protonated oxyanion complexes, arsenate ($H_nAsO_4^{n-3}$) or arsenite ($H_nAsO_3^{n-3}$), respectively (Welch and others, 2000; Smedley and Kinniburgh, 2002; Stollenwerk, 2003). Arsenite is considered the more toxic of the two major oxyanion forms.

Arsenate [As(V)] predominates in oxic groundwaters, whereas arsenite [As(III)] predominates in reducing sulfidic and methanic groundwaters (Welch and others, 2000; Smedley and Kinniburgh, 2002; Stollenwerk, 2003). In strongly reducing waters that are near saturation with sulfide minerals, arsenic sulfide complexes and minerals may form. Mueller and others (2001) noted that the prevalence of arsenite was correlated with low concentrations of dissolved oxygen that reflect strongly reducing conditions (dissolved oxygen less than 0.1 milligrams per liter (mg/L)); arsenate was associated with oxidizing conditions (dissolved oxygen greater than 8 mg/L). The conversion of As(III) to As(V) in oxic waters may be relatively slow and can be measured in years (Eary and Schramke, 1990) with pH, ferric iron, manganese, and bacteria strongly affecting the rate of oxidation. The reduction of As(V) to As(III) under anaerobic conditions is generally much faster than the oxidation of As(III) to As(V).

In groundwater systems, arsenate and arsenite oxyanions commonly form surface complexes (adsorption) on iron oxides and other mineral surfaces (Stollenwerk, 2003). Although As(V) and As(III) adsorb over a wide pH range, As(V) is extensively adsorbed at low pH values and desorbs at alkaline pH; As(III) adsorption increases with pH and peaks at about pH 8 or 9 (Stollenwerk, 2003). In addition to iron oxides, a wide variety of minerals including aluminum oxides and oxyhydroxides, manganese oxides, silica, clays, and carbonates may sorb arsenic, and dissolved organic compounds, phosphate, and other dissolved ions can influence the adsorption of arsenic.

Mobilization of adsorbed arsenic may occur through desorption or dissolution of the host mineral. Arsenic associated with iron oxides tends to be weakly bound on surface sites (adsorbed) and can be released to the groundwater by desorption or by dissolution of the iron oxides (Matisoff and others, 1982; Ayotte and others, 1998; Welch and others, 2000; Smedley and Kinniburgh, 2002; Stollenwerk, 2003; Thomas and others, 2008). Changes in pH and (or) redox conditions can result in the release of arsenic from minerals. Increases in pH can lead to the desorption of arsenate and arsenite. The development of reducing conditions can lead to the reductive dissolution of iron oxides and (or) the reduction of arsenate to arsenite and the consequent desorption of arsenite (reductive desorption) (Stollenwerk, 2003; Thomas, 2007). In general, fine-grained sediments tend to have higher arsenic concentrations than coarse-grained sediments because smaller-sized particles and (or) those with complex shapes have a higher surface-area-to-volume ratio and a more reactive surface area than larger, simply shaped particles (Parks, 1990). The density of sorption sites and potential exposure to reactive waters generally increase with the mineral surface area.

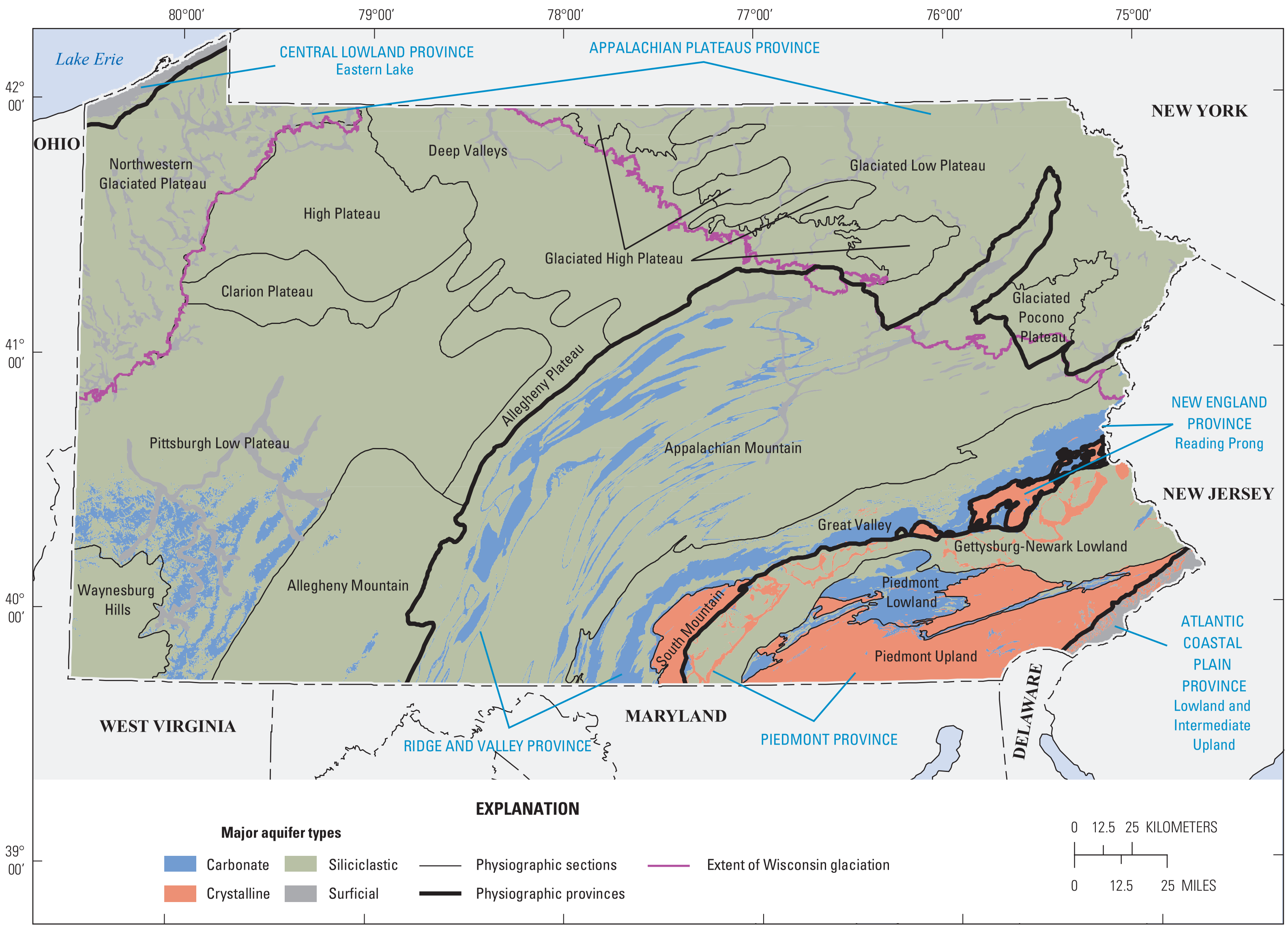
Arsenic concentrations in groundwater may increase (accumulate) with the age of the water. Thomas (2007) reported that arsenic concentrations of 10.0 µg/L or greater were found more frequently in old waters (recharged before 1953) as compared to younger waters (recharged since 1953). Geologic units that have high yields of water are, in general, highly permeable and transmissive, exhibit rapid recharge, and, as a result, consist of relatively young water. This young water will typically be predominantly oxic. In general, water in shallow wells is more likely to be affected by anthropogenic contaminants than water in deeper wells; however, exceptions are numerous because of complexities of groundwater flowpaths.

Description of Study Area

Pennsylvania is a physiographically, geologically, and hydrologically diverse State that covers about 139,859 square kilometers (54,000 square miles). Pennsylvania includes parts of six physiographic provinces, which are subdivided into 20 physiographic sections—(1) Appalachian Plateaus (Allegheny Mountain, Allegheny Plateau, Clarion Plateau, Deep Valleys, Glaciated High Plateau, Glaciated Low Plateau, High Plateau, Glaciated Pocono Plateau, Northwestern Glaciated Plateau, Pittsburgh Low Plateau, Waynesburg Hills), (2) Atlantic Coastal Plain (Lowland and Intermediate Upland), (3) Central Lowland (Eastern Lake), (4) New England (Reading Prong), (5) Piedmont (Gettysburg-Newark Lowland, Piedmont Lowland, Piedmont Upland), and (6) Ridge and Valley (Appalachian Mountain, Great Valley, South Mountain) (Fenneman and Johnson, 1946; Berg and others, 1989) (fig. 1). Land-surface elevations range from sea level (North American Verical Datum of 1988; NAVD 88) (Atlantic Coastal Plain) to 978 meters (3,210 feet) above NAVD 88 (Appalachian Plateaus).

The topography of the Appalachian Plateaus Physiographic Province (hereafter province) varies from deep valleys to glaciated high plateaus with dominant rock types of sandstone, siltstone, and shale and abundant bituminous coal in places. The geologic structure of the Appalachian Plateaus province is complex, varying from horizontal beds to large-amplitude open folds. The Atlantic Coastal Plain province has little relief and consists of unconsolidated sand, gravel, and clay that overlie metamorphic rocks. The Central Lowland province also has little relief with considerable sand and gravel and beach deposits in the Lake Erie area. The New England province includes the steep hills and rounded ridges of the Reading Prong physiographic section (hereafter section), which consists of highly metamorphosed granitic rocks and quartzite. The Piedmont province topography consists of broad, rolling lowlands, narrow valleys, and broad, flat-topped hills. Shale, siltstone, sandstone, and diabase dominate the Gettysburg-Newark Lowland section of the Piedmont province; limestone and dolomite are common in the Piedmont Lowland section. The Piedmont Upland section is dominated by schist, gneiss and quartzite. The geologic structures of the Piedmont province are variable, ranging from half-grabens in the Gettysburg-Newark Lowland section to complex folds and faults elsewhere in the province. The topography of the Ridge and Valley province ranges from narrow to broad valleys with steep uplands or linear ridge and mountain tops. Dominant rock types in the Ridge and Valley province are sandstone, siltstone, and shale, except in the Great Valley section and valleys of the Appalachian Mountain section, which are underlain predominantly by limestone and dolomite rocks. The geologic structure within the Ridge and Valley province is complex with many folds, faults, thrust sheets, nappes, and a major anticlinorium with second- and third-order folds.

Pennsylvania has a complex geological history, which results in many different rock types within the State, with the Pennsylvania Geological Survey recognizing almost 200 different geologic formations or members (Berg and others, 1980; Pennsylvania Geological Survey, 2001). Despite the geologic diversity within the State, the groundwater system in Pennsylvania has been characterized as being representative of four major aquifer types—(1) carbonate bedrock (limestone and dolomite), (2) crystalline bedrock (igneous and metamorphic rocks), (3) siliciclastic bedrock (sandstone, siltstone,



Base from U.S. Bureau of the Census digital data, 1990b 1:100,000
Albers Equal-Area Conic projection: Standard Parallels 40° N and
42° N, Central Meridian -78° W, Latitude of Origin 39°

Physiographic provinces and sections modified from Pennsylvania Geological Survey, 1998 Major aquifer types from Lindsey and Bickford, 1999; Pennsylvania Geological Survey, 2001; Soller & Packard, 1998 Extent of Wisconsin glaciation from Pennsylvania Geological Survey, 1995

Figure 1. Physiographic provinces and sections, major aquifer types, and the extent of the Wisconsin glaciation in Pennsylvania.

conglomerate, and shale), and (4) surficial (fig. 1) (Lindsey and Bickford, 1999; Pennsylvania Geological Survey, 2001; Soller and Packard, 1998). Surficial aquifers consist of unconsolidated material (sand and gravel) overlying bedrock aquifers in depths sufficient to serve as an aquifer, such as glacial outwash, alluvium, and beach deposits. Surficial aquifers shown in figure 1 consist of mapped areas of the State where surficial materials consist of coarse-grained sediment, and these were designated as surficial aquifers for this report. Despite this designation, it is still possible for wells located in other areas of the State within the extent of the Wisconsin glaciation (fig. 1) to be completed in glacial materials.

Carbonate bedrock aquifers are located in some valleys of the Pittsburgh Low Plateau section of the Appalachian Plateaus province, valleys of the Ridge and Valley province, and the Piedmont Lowland section of the Piedmont province. Crystalline bedrock aquifers make up the New England province, the Piedmont Upland section and diabase intrusions of the Gettysburg-Newark Lowland section of the Piedmont province, and South Mountain section of the Ridge and Valley province. The Central Lowland and Atlantic Coastal Plain provinces are predominantly composed of surficial aquifers resulting from beach deposits from Lake Erie and the Atlantic Ocean. Siliciclastic bedrock aquifers can be found throughout the rest of the State and are most prominent in the Appalachian Plateaus province, Gettysburg-Newark Lowland section of the Piedmont province, and the Ridge and Valley province. Surficial deposits are also mostly in the Northwestern Glaciated Plateau and Glaciated Low Plateau sections of the Appalachian Plateaus province and are within the extent of the Wisconsin glaciation. Some surficial deposits also extend into the Appalachian Mountain section of the Ridge and Valley province and the Pittsburgh Low Plateau section of the Appalachian Plateaus province.

Temperature and precipitation vary across the State according to geography and topography. The average annual temperature is 11 degrees Celsius (ºC) (52 degrees Fahrenheit) in southern Pennsylvania and 8 ºC (46 degrees Fahrenheit) in the northern part of the State. The warmest areas correspond to the Atlantic Coastal Plain and Piedmont provinces. Progressive cooling occurs in the higher land-surface elevations of the Ridge and Valley province, and the Appalachian Plateaus province is the coolest area of the State (Cuff and others, 1989). Heat and moisture circulate through the State from the south to the southeast, whereas most of the precipitation-producing weather fronts move from west to east. Precipitation increases from 102 centimeters (cm) (40 inches) at the western border of Pennsylvania to a maximum of 130 cm (51 inches) near the highest land-surface elevations of the Appalachian Plateaus province. The decreasing land-surface elevations in the Ridge and Valley province create rain shadow effects that reduce the average annual precipitation to 97 cm (38 inches). The eastern part of Pennsylvania is strongly affected by airflows coming directly off the Atlantic Ocean, and this contributes to the higher average annual precipitation of 102 to 112 cm (40 to 44 inches), despite the lower land-surface elevation (Cuff and others, 1989).

Land use is primarily forested (65 percent) followed by agricultural (27 percent) (Nakagaki and others, 2007). Forested is the dominant land use in the north-central, northeast, and rugged mountain slopes. Agricultural land use predominates in the valleys of the Ridge and Valley province and much of the Piedmont province. Urban (6 percent) land use is dominant in and around Pennsylvania’s major cities, especially those cities with populations greater than 40,000 (U.S. Bureau of the Census, 2010), of which most are located in the Piedmont and Ridge and Valley provinces.

Previous Studies

Arsenic concentrations in groundwater of Pennsylvania have been documented by a number of previous studies, typically a county-scale or regional-scale study. Cravotta (2008) reports that arsenic concentrations in groundwater discharged to 140 abandoned coal mines in the bituminous and anthracite coalfields of Pennsylvania ranged from less than 0.03 to 64.0 µg/L. Arsenic concentrations were positively correlated with pH, chloride, bromide, and iodide and inversely correlated with dissolved oxygen and redox potential, indicating the potential for arsenic mobilization by desorption or reduction processes, possibly because of interactions with deep, saline groundwater.

Williams and others (1998), who studied the glaciated valleys of Bradford, Potter, and Tioga Counties in Pennsylvania, found a correlation between arsenic and older (recharged before 1953) or briny water and found that arsenic concentrations varied by primary aquifer. Buckwalter and Moore (2007) concentrated their efforts in Warren County, which is also in a glaciated region of Pennsylvania, where almost one-third of the collected samples contained arsenic concentrations that exceeded the MCL of 10.0 µg/L; the maximum was 490.0 µg/L. They also documented that arsenic concentrations exhibited seasonal fluctuations and that arsenic concentrations varied widely, even between adjacent (less than a 76-meter (250-foot) distance) wells. Low and Galeone (2006) collected groundwater samples for analysis for total arsenic in the glaciated region of Pennsylvania within eight counties and found that arsenic concentrations varied greatly over short distances but did not appear to be related to well depth. Thomas (2007) studied the association of arsenic with redox conditions in the glacial aquifer system of the northern United States, which includes the glaciated portion of Pennsylvania and concluded that elevated arsenic concentrations are more commonly detected in older, anoxic groundwaters (recharged before 1953) and that arsenic correlated strongly with constituents linked to redox processes and anions or oxyanions that sorb to iron oxides.

Peters and Burkert (2008) examined groundwater-quality data from over 18,000 wells in the Newark Basin of Pennsylvania. They found that variations in pH were strongly correlated with arsenic concentrations, with the highest concentrations of arsenic associated with pH values greater than 6.4. They concluded that the original source of arsenic in the study area was most likely black and gray shales containing arsenian pyrite and that groundwater concentrations of arsenic are most likely controlled by adsorption/desorption reaction with iron oxides in red mudstone aquifer materials. Senior and Sloto (2006) studied the Newark Basin, sampling 58 wells within the study area to identify areas of elevated arsenic concentrations and characterize the geochemical environment associated with elevated concentrations of arsenic and various constituents. They found that arsenic correlated most strongly and positively with pH, boron, and molybdenum; correlated positively with selenium, uranium, nickel, lithium, fluoride, and strontium; and correlated negatively with total organic carbon, copper, and dissolved oxygen. They concluded that arsenic concentrations may be controlled partly by pH affecting adsorption of arsenate and that the correlation of arsenic with the presence of many trace elements indicates similar geochemical controls and (or) distribution in aquifer materials in the Newark Basin.

Methods of Investigation

Groundwater-quality data from 1969 to 2007 were obtained by the USGS from local, county, private, State, and Federal electronic databases. Spatial data consist of variables representing anthropogenic factors (such as land use and contamination sites) and natural factors (such as geology and climate) (appendix 1). Datasets for most factors were available in geographic information system (GIS) format from various sources, but additional GIS datasets were developed specifically for use as explanatory variables during statistical modeling. Some datasets listed in appendix 1 were not available statewide or pertinent to statewide analysis, so these datasets were only populated for selected intrastate regions (glacial aquifer system, Gettysburg Basin, Newark Basin) (fig. 2). Also, differences in the extent of explanatory variable data coverage caused different explanatory variables to have a different number of sample records associated with them. For example, a total of 5,023 sample records were available statewide, but appendix 1 shows 5,021 sample records available statewide for all of the soil characterization variables and 5,011 sample records available statewide for two of the groundwater geochemistry variables. These differences in data availability were due to the slightly different extents associated with the explanatory datasets, which caused these datasets to have slightly different coverage across Pennsylvania.

Groundwater-Quality Data

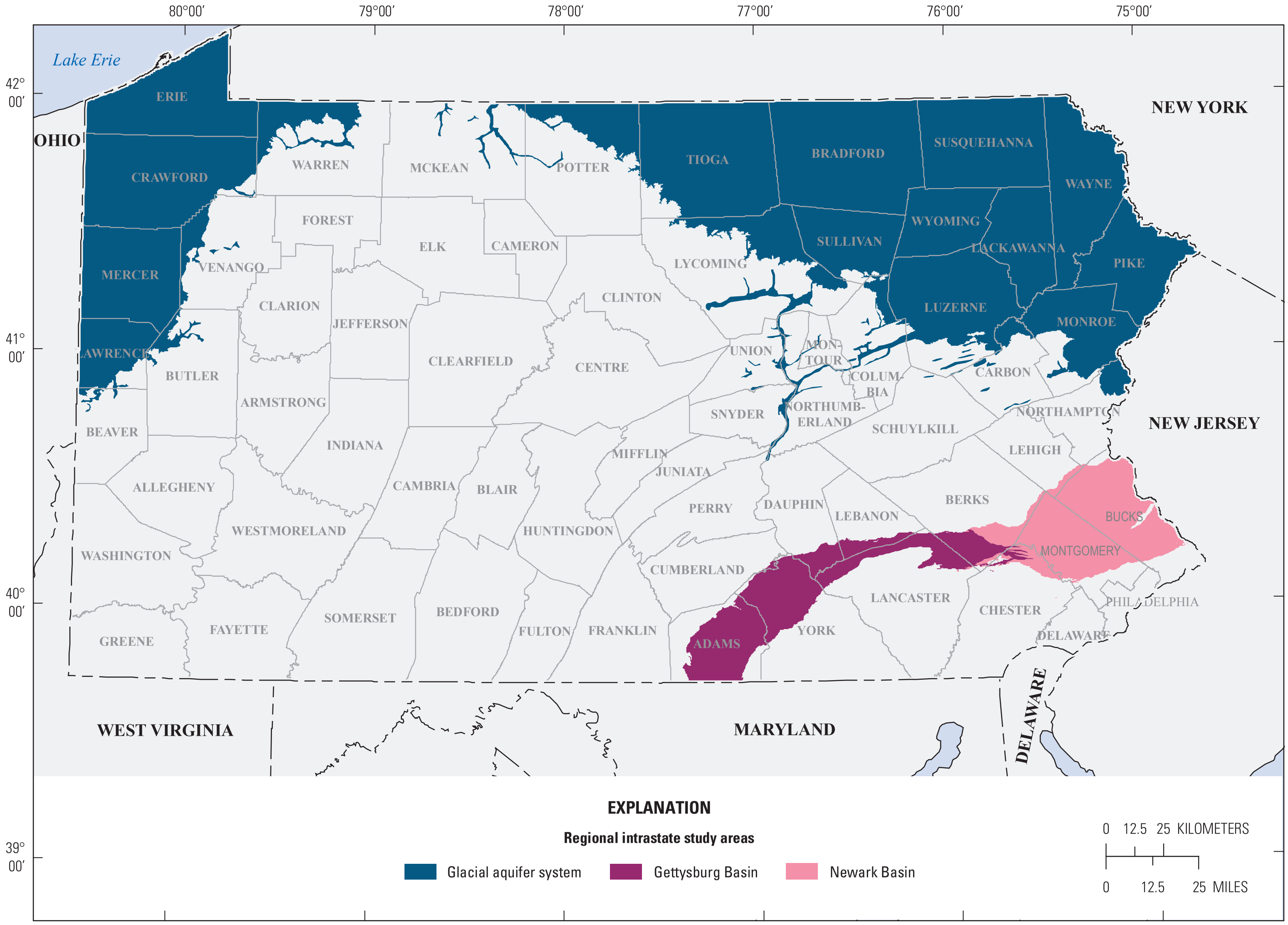
A database consisting of 5,023 groundwater records with reported values for arsenic and various associated field and laboratory measured constituents was created (Low and Chichester, 2006; Low and others, 2008). More than 25,000 groundwater-quality data records were examined initially. This number was reduced to create datasets containing values for arsenic (total and (or) dissolved, 12,781), pH (7,876), and specific conductance (5,931). The pH and specific conductance datasets were used to create explanatory spatial variables, and these datasets include data from areas where arsenic data are not necessarily available. The arsenic dataset of 12,781 was further reduced to 5,023 groundwater records by restricting values to one composite sample per well using three methods: (1) samples with reported censored arsenic data with detection levels greater than 4.0 µg/L were removed from the database; (2) duplicate sites and (or) samples were removed through verification of local well numbers, site identifiers, geographical coordinates (latitude and longitude) within the State boundary, and sample date; and (3) for sites that were sampled repeatedly, data from the analysis that was associated with the highest arsenic concentration were retained. If the arsenic concentrations were identical, then data from the most recent analysis were retained.

Groundwater-quality data associated with the arsenic dataset of 5,023 groundwater records were handled in the following manner: (1) values with remark values coded as “estimated” were assumed to represent actual values; (2) total and dissolved constituent samples were not differentiated in regards to any statistical or chemical analyses, but if total and dissolved values were reported for the same sample, the dissolved value was retained; and (3) field measurements (pH, specific conductance) had precedence over laboratory results. Estimated dissolved solids represents the maximum reported value for residue on evaporation at 105 ºC, residue dried at 180 °C, sum of dissolved constituents, or total dissolved solids. Although water use varied (commercial, monitor, industrial, public) no differentiation or statistical analysis was performed to distinguish among industrial, public, private, or other uses of water.

Groundwater-quality data associated with the arsenic dataset of 5,023 groundwater records compiled for this investigation include up to 53 additional groundwater-quality constituents. These data were reduced to 31 groundwater-quality constituents for statistical analysis because of limited availability of data and differences in data quality among the records analyzed. Samples were analyzed at the USGS National Water Quality Laboratory, the Pennsylvania Department of Environmental Protection Laboratory, Pennsylvania Department of Agriculture, and Pennsylvania State University, as well as a number of private laboratories (Low and Chichester, 2006; Low and others, 2008). As a result, the groundwater-quality data represent multiple project designs and goals, site and well selection criteria, as well as sample collection, sample preservation, analyte detection levels, and quality-assurance/quality-control methods.

Spatial Data

Variables representing anthropogenic and natural factors were compiled and evaluated for statewide and regional study



Base from U.S. Bureau of the Census digital data, 1990b 1:100,000
Albers Equal-Area Conic projection: Standard Parallels 40° N and
42° N, Central Meridian -78° W, Latitude of Origin 39°

County boundaries from U.S. Bureau of the Census, 1990a Glacial aquifer system
boundary from Pennsylvania Geological Survey, 1995; Soller & Packard, 1998
Gettysburg and Newark Basin boundaries from Pennsylvania Geological Survey, 2001

Figure 2. Intrastate regions modeled in Pennsylvania. (Glacial aquifer system, Gettysburg Basin, and Newark Basin)

areas. Arsenic concentrations in groundwater from 5,023 wells within the State were combined with additional potential explanatory data consisting of anthropogenic and natural factors, and a GIS was used to produce a dataset in which each well with a measured arsenic concentration was associated with the explanatory geographic variables. Explanatory variables differed among statewide and regional study areas as a result of differences in anthropogenic activities (land use) and natural conditions (topography, geology, soils) throughout the State.

A raster dataset was created for each factor or variable using a GIS. Raster datasets represent a spatial data model defining space as an array of equally sized cells arranged in rows and columns with each cell containing an attribute value and location coordinates. Original data used for the study consist of previously existing raster or vector (point, line, or polygon) datasets of various resolutions (appendix 1). Previously converted, created, or existing raster datasets were clipped to the State of Pennsylvania's political boundary (U.S. Bureau of the Census, 1990b) and snapped to a common dataset (Nakagaki and others, 2007) to ensure that the cell alignment of each output raster would be the same. The snap dataset was the dataset with the smallest resolution, which was the 30-meter resolution land-cover dataset (Nakagaki and others, 2007). This means that data for each factor or variable were compiled within 30-meter grid cells, thus, creating a spatial layer for each factor or variable lining up with a template representing the State of Pennsylvania consisting of 9,493 rows and 16,508 columns and totaling 156,710,444 cells across the State.

Anthropogenic Factors

Data representing anthropogenic factors include proximity to known sources of contamination, disturbance, and land-cover variables. Variables describing proximity to known sources of contamination illustrate the distance to points or polygons representing areas that are known to be receiving treatment for arsenic contamination or that are in need of arsenic remediation. Disturbance variables describe the distance to points where humans are disturbing natural surroundings by mining or drilling operations. Land-cover variables describe land-use patterns that result from agricultural operations, urban development, or population density.

Pennsylvania municipality boundaries were acquired from the Pennsylvania Department of Transportation (2008). A list of municipalities containing water suppliers receiving arsenic treatment and sites or groundwater in need of remediation owing to arsenic contamination was compiled from the Pennsylvania Bulletin (1997–2009), and only these municipalities were included in a new dataset describing the distance to the nearest municipality receiving arsenic treatment. Toxic chemical release inventory data were obtained from the USEPA (1994), and a similar dataset was created describing the distance to the nearest point representing toxic chemical release inventory sites.

Three datasets indicating anthropogenic disturbances to the landscape were acquired from the Pennsylvania Department of Environmental Protection (2008a, 2008b, 2008c). These datasets describe point locations of (1) underground and surface coal mining operations, (2) industrial mineral mining operations, and (3) drilled oil and gas wells. Spatial datasets were created to represent each of these three datasets by describing the distance to the point location located closest to each raster data cell across the State. Therefore, the raster cell is assigned a value that represents the distance from the raster cell to the closest data point in the dataset of interest (coal mining operations, mineral mining operations, oil and gas wells), whether the closest data point is 30 or 300 meters away. Because underground and surface coal mining operations and drilled oil and gas wells are concentrated in the northwestern and northeastern parts of the State, datasets associated with those parts of the State were considered only for analysis of conditions in the glacial aquifer system (fig. 2).

Land-cover classification data for Pennsylvania were compiled from the 1992 Enhanced National Land Cover Data (NLCD) (Nakagaki and others, 2007), a dataset that has a 30-meter resolution. The agricultural land-cover classification used for this study was created by grouping data by the following classifications: Orchards/Vineyards/Other, Land Use Land Cover (LULC) Orchards/Vineyards/Other, Pasture/Hay, and Row Crops. The urban land-cover classification was created by grouping data by the following classifications: Low intensity residential, High intensity residential, LULC residential, NLCD/LULC forested residential, and Urban/recreational grasses. Focal statistics were used to create agricultural and urban land-cover datasets by calculating the average amount of agricultural or urban land-cover cells within a 500-meter radius of each raster cell within the State. The resulting datasets describe the percentage of agricultural and urban land cover within a 500-meter radius of each raster cell. Population density data were compiled by block group from 1990 Census of Population and Housing data by Price (2003) in people per square kilometer.

Natural Factors

Spatial data describing natural factors that were compiled and evaluated are climate, geology, geophysical, groundwater geochemistry, land cover, identified mineral deposit, soils characterization, and topography variables. Climate variables describe precipitation, temperature, and groundwater recharge. Geology variables describe distance to geologic units with mineral properties that could affect arsenic concentrations in groundwater or the potential of geologic units to act as an arsenic source or mobilizer. Geophysical variables describe average residual total intensity of the earth's magnetic field. Groundwater geochemistry variables describe groundwater corrosivity geologic groupings and pH and specific conductance of groundwater. Land-cover variables that include natural factors describe forested and wetland land-cover patterns. Metals and minerals variables describe the intensity of the earth's magnetic field resulting from the distribution of iron minerals and distance to mapped mineral deposits. Soils characterization variables describe soil water storage, compaction, texture, runoff potential, organic matter, permeability, thickness, and land-surface slope.

Average annual precipitation and minimum and maximum temperature data from 1971 to 2000 were compiled from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM) Climate Group from Oregon State University (2006a, 2006b, 2006c). Minimum and maximum temperature datasets were averaged in order to create a dataset of average temperatures during 1971–2000. Groundwater recharge rates from 1951 to 1980 were taken from Wolock (2003).

Geologic units containing potentially substantial acid-producing sulfide minerals (Pennsylvania Geological Survey, 2005) were obtained from the State bedrock geology spatial dataset (Pennsylvania Geological Survey, 2001). Diabase geologic units also were extracted from the State bedrock geology dataset to form a dataset. A dataset containing point locations for igneous rock samples (Grossman, 1999) was obtained from the USGS National Geochemical Database. Raster datasets were created to represent each of the three previously described datasets by calculating the shortest distance from each cell to the nearest polygon or point location in each dataset. Geophysical data describing average residual total intensity of the earth's magnetic field resulting from variations in earth materials and structure were obtained from Bankey and others (2002).

Hydrogeochemical stream-sediment data from the National Uranium Resource Evaluation (NURE) dataset were obtained from the USGS (2004) and used to create a dataset of estimated arsenic concentrations in stream sediments. Arsenic concentrations were estimated from point data using the inverse distance weighted (IDW) technique, which is an interpolation method used to estimate concentrations by averaging values for sample points within a defined neighborhood. Data on arsenic in stream sediments were not available for the entire State; therefore, this dataset was limited to the Newark Basin as this was the only regional study area within the State with a sufficient amount of data for analysis.

Major aquifers of Pennsylvania (carbonate, crystalline, siliciclastic, and surficial) (fig. 1) were used to create four geologic variables that could be modeled as discrete variables by performing a spatial intersection with mapped geologic units across the State to assign each well to a major aquifer. The discrete major aquifer variables were coded as "one" if a well was located in a particular major aquifer and coded as "zero" if the well was not located in an aquifer. For example, the carbonate major aquifer variable would code all wells spatially intersecting carbonate aquifers as "one" and all wells spatially intersecting the other major aquifers (crystalline, siliciclastic, and surficial) as "zero." Mapped geologic units in Pennsylvania were divided by major aquifers according to their reported primary lithology descriptions (appendix 2) (Pennsylvania Geological Survey, 2001). Carbonate bedrock aquifers consist of primary lithologies of argillaceous dolomite, argillaceous limestone, dolomite, graphitic marble, high-calcium limestone, limestone, limestone conglomerate, marble, and shaly limestone. Crystalline bedrock aquifers consist of primary lithologies of albite-chlorite schist, andesite, anorthosite, chlorite-sericite schist, diabase, feldspathic quartzite, felsic gneiss, granitic gneiss, granitic pegmatite, graphitic felsic gneiss, graphitic gneiss, greenstone schist, mafic gneiss, metabasalt, metadiabase, metagabbro, metarhyolite, oligoclase-mica schist, phyllite, quartzite, serpentinite, and slate. Siliciclastic bedrock aquifers consist of primary lithologies of argillaceous sandstone, argillite, arkosic sandstone, black shale, calcareous sandstone, calcareous shale, graywacke, mudstone, quartz conglomerate, quartzite, sandstone, shale, siliceous sandstone, siltstone, and silty mudstone. Surficial aquifers for mapped geologic units consist of primary lithologies of feldspathic quartz sand, ferruginous clay, gravelly sand, and sand. Additionally, bedrock aquifers in some areas of the State are overlain by unconsolidated material (Lindsey and Bickford, 1999; Soller and Packard, 1998) of sufficient depths to serve as an aquifer. These additional areas of the State where surficial materials consist of coarse-grained sediments were designated as surficial aquifers for this report and are indicated in appendix 2. Table 1 gives a summary of arsenic concentrations in groundwater for the four major aquifer types in Pennsylvania. All four major aquifer types had median arsenic concentrations less than 4.0 µg/L, and the highest arsenic concentration (490.0 µg/L) was in a siliciclastic aquifer. The siliciclastic and surficial aquifers had the highest percentage of sample records with arsenic concentrations greater than or equal to 4.0 µg/L and 10.0 µg/L. Tukey's multiple comparison tests were used to compare mean arsenic concentrations among the four major aquifers. Mean arsenic concentrations among all of the major

Table 1. Summary of arsenic concentrations in groundwater (1969–2007) for the four major aquifer types in Pennsylvania.

[µg/L, micrograms per liter; <, less than]

Major aquifer type	Number of sample records	Median arsenic concentration, in µg/L	Maximum arsenic concentration, in µg/L	Sample records with arsenic concentrations greater than or equal to 4.0 µg/L, in percent	Sample records with arsenic concentrations greater than or equal to 10.0 µg/L, in percent
Carbonate	597	<4.0	217.5	9	3
Crystalline	852	<4.0	60.0	6	2
Siliciclastic	3,112	<4.0	490.0	20	8
Surficial	462	<4.0	293.0	34	20

aquifer types were significantly different (alpha = 0.2), except for the comparison of mean arsenic concentrations between the carbonate and crystalline major aquifer types. Major aquifer variables were included in the statewide analysis but were not included in the analysis of the regional study areas because these consist of a broad characterization of lithologies that do not differentiate well among regions containing only a few geologic units and (or) geologic units categorized as the same major aquifer.

Another dataset was developed for groundwater corrosivity. Geologic units across the State (Pennsylvania Geological Survey, 2001) were ranked according to estimated groundwater corrosivity to indicate the potential for groundwater to intereact with arsenic-bearing minerals and to recreate the dataset resulting from a 1996 study by Langland and Dugas (1996). They evaluated the relations among corrosive groundwater, water chemistry, and geology by the use of a modified version of the Langelier Saturation Index for 11 lithologic units based on the State map unit descriptions, then ranked the units from most to least corrosive according to color codes. Corrosivity rankings were assigned to each lithologic unit grouping according to the ranks used by Langland and Dugas (1996): (1) quartzite, (2) crystalline rocks excluding diabase and quartzite, (3) anthracite-bearing siliciclastic rocks, (4) unconsolidated sediments, (5) predominantly shale with other siliciclastic rocks, (6) predominantly sandstone with other siliciclastic rocks, (7) diabase, (8) mixed siliciclastic rocks with bituminous coal, (9) shale, (10) limestone-bearing siliciclastic rocks, and (11) carbonate rocks.

Groundwater-quality data for pH and specific conductance were obtained from data compilations reported by Low and Chichester (2006) and Low and others (2008) from areas where arsenic data were not necessarily available. The USGS (2004) NURE hydrogeochemical data were also used, which have additional values for specific conductance and pH for groundwater in the entire State; these data, in addition to the Low and Chichester (2006) and Low and others (2008) data compilations, provided additional spatial coverage. These groundwater-quality datasets were combined and used to estimate the continuous spatial distribution of pH and specific conductance across the State using the inverse distance weighted (IDW) technique. IDW is an interpolation technique used for predicting values for unmeasured locations using measured values surrounding the prediction location and is based on the assumption that things that are close to one another are more alike than those that are farther apart. A statewide pH dataset representing pH in groundwater in bedrock and surficial aquifers across the State was created from a total of 13,598 values with 7,876 (3,251 of these had associated arsenic concentrations) of these values from data compilations reported by Low and Chichester (2006) and Low and others (2008) and 5,722 of these values from the USGS (2004) NURE hydrogeochemical data. A statewide specific conductance dataset representing specific conductance in groundwater in bedrock and surficial aquifers across the State was created from a total of 11,652 values with 5,931 (2,514 of these had associated arsenic concentrations) of these values from data compilations reported by Low and Chichester (2006) and Low and others (2008) and 5,721 of these values from the USGS (2004) NURE hydrogeochemical data. The IDW technique was performed for each grouping of geologic units according to the groundwater corrosivity ranking dataset previously described (Langland and Dugas, 1996). Because there are 11 different corrosivity rankings in the State, a separate IDW interpolation was calculated for each grouping of units to establish boundaries in the pH and specific conductance datasets, and each of the interpolated datasets were combined into a statewide raster dataset. Separate pH and specific conductance datasets were created to represent groundwater geochemistry in the glacial aquifer system. For this area, groundwater-quality data on pH and specific conductance from Low and Chichester (2006) and Low and others (2008) were used only if the wells were finished in unconsolidated aquifers and data from NURE hydrogeochemical data (U.S. Geological Survey, 2004) were used if the well types were “dug” or “driven.” Of the data meeting these criteria, only those data from within the defined boundary of the glacial aquifer system (fig. 2) were included as estimates of pH and specific conductance. A pH dataset representing pH in groundwater in the glacial aquifer system in Pennsylvania was created from a total of 619 values with 349 (184 of these had associated arsenic concentrations) of these values from data compilations reported by Low and Chichester (2006) and Low and others (2008) and 270 of these values from the USGS (2004) NURE hydrogeochemical data. A specific conductance dataset representing specific conductance in groundwater in the glacial aquifer system in Pennsylvania was created from a total of 562 values with 293 (192 of these had associated arsenic concentrations) of these values from data compilations reported by Low and Chichester (2006) and Low and others (2008) and 269 of these values from the USGS (2004) NURE hydrogeochemical data.

The forested land-cover classification used for this study consists of deciduous forest, evergreen forest, and mixed forest (Nakagaki and others, 2007). The wetland land-cover classification consists of woody wetlands and emergent herbaceous wetlands (Nakagaki and others, 2007). Focal statistics were used to create forested and wetland land-cover datasets by calculating the average value of each grouped land-cover type in cells within a 500-meter radius of each cell across the State.

Mineral resources data were obtained from the mineral resources data system of the USGS (2007). The mineral resources data were analyzed as a whole and according to metallic and nonmetallic minerals. Raster datasets were created to represent each of the three datasets (distance to all mineral resources, distance to metallic mineral resources, distance to nonmetallic mineral resources) by calculating the shortest distance from each cell to the nearest mineral resource point in each dataset.

Soil characterization criteria were obtained from Wolock (1997), who used the STATSGO soil database (U.S. Department of Agriculture, 1993). The soil data include available water capacity, bulk density, hydrologic soil group, organic matter, permeability, thickness, slope, and texture. Available water capacity describes the amount of water that the soil is able to store. Bulk density is a measure of soil compaction. Hydrologic soil groups define the runoff potential of soil and are described according to the percentage of hydrologic soil group present. Hydrologic soil groups are described as Group A, Group B, Group C, and Group D, in order from low to high runoff potential, with dual hydrologic soil groups (AD, BC, CD) assigned on the basis of drained and undrained conditions. Soil organic matter represents the percentage of organic matter that a soil contains. Permeability is a measurement of the ability of water to flow through the soil. Thickness describes the distance from the surface of the soil to the underlying solid bedrock. Slope, the percentage of soil land-surface slope, describes the potential of precipitation to run off land surfaces or infiltrate into subsurfaces. Texture describes the percentage of sand, silt, or clay that a soil contains.

Land-surface elevation data were retrieved from the USGS (2009) 1-arc second National Elevation Dataset. These data were used to create a slope dataset by calculating the maximum rate of change between each cell within the land-surface elevation dataset and its neighbors. Sinkhole location data were obtained from the Pennsylvania Geological Survey (2007) online sinkhole inventory and database (Kochanov and Reese, 2003). These point data were used to create a raster dataset by calculating the shortest distance from each cell to the nearest sinkhole location. Stream flowline data were obtained from the USGS (2005) high-resolution National Hydrography Dataset and were used to create a stream density raster dataset by calculating the total length of streams within a 500-meter search radius area of each cell.

Statistical Methods

Spearman's rank correlation coefficient (*rho*) (Helsel and Hirsch, 2002), a nonparametric statistical test that uses the data ranks, was used to evaluate the significance of relations and differences among statewide data for arsenic concentrations and 31 groundwater-quality parameters (where sample size exceeded 30). Spearman's *rho* also was used to test for statistical significance in relations among groundwater-quality parameter data (where sample size exceeded 30). Spearman's *rho* is a monotonic correlation test in which a positive value of *rho* indicates that the response variable (Y) increases as the explanatory variable (X) increases, and a negative value of *rho* indicates that the response variable (Y) increases as the explanatory variable (X) decreases. High positive values (lower negative values) of *rho* indicate a strong monotonic correlation. Spearman's *rho* was performed on rank-transformed arsenic concentrations and groundwater-quality parameter data, where ranks were assigned so that non-detects were ranked lower than the lowest value detected or estimated, following the methodology of Gilliom and others (2006). Reported values and censored values less than 4.0 µg/L were ranked as if they were all 3.9 µg/L, which is the lowest rank. Any associated groundwater-quality data constituent with a non-detect data value was ranked as if it had a value of 0 so that the non-detect values would be ranked lower than the lowest value detected or estimated. All other measured and estimated concentrations of arsenic and other groundwater-quality constituent data were ranked according to their nominal values.

Univariate and multivariate logistic regression analyses (Hosmer and Lemeshow, 1989; Helsel and Hirsch, 1992) were used to develop predictive models for the State and for three intrastate regions—glacial aquifer system, Gettysburg Basin, and Newark Basin—for which the model results were used to predict the probability of detecting concentrations of arsenic in groundwater greater than or equal to 4.0 µg/L on the basis of explanatory variables that could affect arsenic values. For the resulting logistic regression analyses, a binary response variable was defined by dividing the measured arsenic concentrations into two groups: concentrations greater than or equal to 4.0 µg/L (exceedances) were classified as "one" and concentrations less than 4.0 µg/L (nonexceedances) as "zero". The threshold of 4.0 µg/L was selected because this value represents the maximum common detection level for the censored arsenic concentration data. For the purposes of this report, arsenic concentrations greater than or equal to 4.0 µg/L are referred to as "elevated."

Univariate logistic regression was used as a first step to test the significance of individual explanatory variables as indicators of elevated arsenic concentrations. Standardized coefficients, which allow common unit comparisons among model variables, indicate the nature of the univariate relation, with positive relations indicated by values greater than zero and inverse relations indicated by values less than zero (Menard, 2002). An alpha level of 0.2 was chosen as the inclusion criteria for selecting explanatory variables to include in multivariate analysis rather than the traditional alpha level of 0.10. Hosmer and Lemeshow (1989) suggest that a traditional alpha level of 0.10 has failed to identify variables known to be important during some logistic regression analyses, and other variables may not be considered important in a model until they are included with other complementary variables.

Stepwise logistic regression was used to create multivariate logistic regression models that predict the probability of elevated arsenic concentrations in groundwater in Pennsylvania. The logistic regression model begins with the intercept; then explanatory variables are added or eliminated through forward and backward selection procedures until changes in variables no longer change the log of the odds ratio (logit). The odds ratio is based on the probability of exceeding the given threshold value, and the log of the odds ratio (logit) transforms a variable constrained between 0 and 1 into a continuous, unbounded variable that is a linear function of the explanatory variables and converts the predicted values of the response variable into probability units (Hosmer and Lemeshow, 1989; Helsel and Hirsch, 1992).

A number of statistical parameters are examined when evaluating multivariate logistic regression models to determine the success of a model (Menard, 2002). Predictive model performance was evaluated using measures such as (1) overall model significance, (2) values and probabilities associated with explanatory variables, (3) model fit statistics, (4) multicollinearity diagnostics, (5) linear regression, and (6) Pearson residuals.

Success and significance of a model is measured by the log-likelihood ratio, which compares observed values with predicted values (Hosmer and Lemeshow, 1989). The most significant model will have the highest log-likelihood ratio, but the degrees of freedom, or number of explanatory variables, are also taken into account. Explanatory variable significance is indicated by p-values less than 0.05 and shows how specific explanatory variables improve the ability of the model to predict the probability of elevated arsenic concentrations in groundwater.

Logistic regression model-fit statistics used in this study include the Hosmer-Lemeshow (H-L) goodness-of-fit test, generalized and maximum rescaled r-square, percent concordance, model sensitivity, and area under the Receiver Operating Characteristic (ROC) curve. The H-L statistic was used to evaluate model calibration by calculating the degree of correspondence between the predicted probabilities exceeding the threshold and the actual concentrations exceeding the threshold. For this test, p-values less than 0.05 indicate that the predicted probabilities are significantly different than the actual concentrations. Therefore, a higher H-L p-value will indicate a well-calibrated model (Hosmer and Lemeshow, 1989). There is no r-square value that can be produced by the logistic regression model that is identical to the r-square value from linear regression; however, some substitutes for the r-square value have been calculated (Hosmer and Lemeshow, 1989). The generalized r-square value (Cox and Snell, 1989) is based on maximizing the log-likelihood and is a generalized method of estimating an r-square value. The maximum-rescaled r-square value (Nagelkerke, 1991) is another method that approximates the linear-regression r-square. Neither of these statistics can be interpreted as the percentage of variance explained by the model, but they can be used to compare one model with another. Logistic regression model results also are described in terms of percent concordance, which is the overall rate of correct classification. This value is the number of observed exceedances predicted by the model as exceedances, plus the number of observed nonexceedances predicted by the model as nonexceedances, divided by the combined number of observed exceedances and nonexceedances (Hosmer and Lemeshow, 1989). Model sensitivity is the number of observed exceedances predicted as exceedances, divided by the total number of observed exceedances. Higher percent concordance and sensitivity values indicate better fitting models. The area under the ROC curve is represented by the c statistic, which is a measure of the model's ability to discriminate between groundwater samples that have arsenic concentrations greater than or equal to 4.0 µg/L and those that do not. The c statistic is a value that varies from 0.5 to 1.0 with higher values indicating better discrimination. Hosmer and Lemeshow (1989) consider values from 0.7 to 0.8 to show acceptable discrimination and values from 0.8 to 0.9 to show excellent discrimination.

Multicollinearity in multivariate regression models is the result of strong correlations among two or more explanatory variables. Models with strongly correlated explanatory variables can produce incorrect signs and magnitudes of regression coefficients, which lead to incorrect conclusions about relations between explanatory and dependent variables (Allison, 2001). Multicollinearity was evaluated using Tolerance and Variance Inflation Factor (VIF) multicollinearity diagnostic statistics, which are based on linear regression analysis of explanatory variables (Allison, 2001). The Tolerance is defined as $1 - r^2$, where r^2 is the coefficient of determination for the regression of one independent variable on all remaining independent variables (Allison, 2001; Menard, 2002). The VIF is equal to the reciprocal of the Tolerance and describes how inflated the variance of coefficient is compared to what it would be if there were no multicollinearity (Allison, 2001). Although there are no formal thresholds to use for the Tolerance or VIF in detecting the presence of multicollinearity, Allison (2001) suggests that Tolerance values less than 0.4 (VIF greater than 2.5) may indicate the presence of multicollinearity.

Calibration for logistic regression predictive models (statewide, glacial aquifer system, Gettysburg Basin, Newark Basin) were evaluated using the degree of correspondence between observed elevated arsenic concentrations in groundwater and the predicted probabilities of elevated arsenic concentrations in groundwater. Data were compiled describing observed detections of elevated arsenic concentrations and the associated predicted probability of an elevated arsenic concentration. Observed detections of elevated arsenic concentrations were calculated by classifying arsenic concentrations less than 4.0 µg/L as "zero" and arsenic concentrations greater than or equal to 4.0 µg/L as "one." Predicted probabilities of elevated arsenic concentrations in groundwater were calculated according to the model results for each of the predictive models (statewide, glacial aquifer system, Gettysburg Basin, Newark Basin). Data were sorted according to the predicted probability of elevated arsenic in groundwater and divided according to deciles. The observed detections of elevated arsenic were averaged for each decile to calculate the percentage of observed detections, and predicted probabilities were also averaged to calculate the average predicted probability for each 10-percent decile. Linear regressions were constructed between the percentage of observed detections of elevated arsenic concentrations and the average predicted probability for each decile. R-square values associated with each regression were used to assess model calibration for the predictive models.

Pearson residuals were calculated for each individual predicted value to determine which predicted values were most poorly fit by the logistic regression models and spatially show the predictive error for each study area. The Pearson residual statistic calculates the difference between observed and estimated probabilities and divides the result by the binomial standard deviation of the estimated probability (Menard, 2002). Cases where values are equal to zero indicate that the probability of arsenic exceeding 4.0 µg/L is exactly what would be expected based on the observation. Cases with positive residual values indicate that the model predicted a lower probability than what was observed (under prediction), whereas cases with negative values indicate that the model predicted a higher probability than what was observed (over prediction). Spatial evaluation of the highest and lowest residual values can indicate specific areas where the model does not do a good job of predicting the probability of elevated arsenic.

Probability Maps

Maps showing the predicted probability of elevated arsenic in groundwater in Pennsylvania and three selected regions within the State were constructed using results from the logistic regression models. Model coefficients for each explanatory variable along with the variable’s value were used to calculate and map probabilities of elevated arsenic in groundwater within the State.

Each variable was compiled for 30-meter grid cells to estimate the probability of elevated arsenic in groundwater in Pennsylvania or in selected regions within Pennsylvania. In order to create the statewide and regional probability maps depicting the predicted probability of elevated arsenic concentrations in groundwater, the values for the 30-meter grid cells for each explanatory variable were used in logistic regression equations within a map-algebra GIS application. For example, the equation for the statewide model was expressed as

$$\text{PREDICTION} = 1 \,/\, (\, 1 + (\text{EXP}(\, - (\, \text{INT} + (\, \text{PE}_1 * [\text{GRANK}]\,) + (\text{PE}_2 * [\text{SPCIDW}]) + (\text{PE}_3 * [\text{AWCAVE}]\,) + (\text{PE}_4 * [\text{PHIDW}])\,)))) \quad , \tag{1}$$

where

PREDICTION is the output dataset describing the predicted probability of detecting elevated arsenic concentrations in groundwater in Pennsylvania,

EXP is the base of the natural logarithm,

INT is the model intercept,

$\text{PE}_1, \text{PE}_2, \text{PE}_3$, and PE_4 represent the parameter estimates for each of the explanatory variables that were in the final statewide logistic regression model, and

GRANK, SPCIDW, AWCAVE, and PHIDW are datasets describing the explanatory variables that were in the final statewide logistic regression model.

This process was used to calculate the probability that arsenic in groundwater exceeds 4.0 µg/L for each cell as a fraction. The units for the predicted probability datasets were then converted from fractions to percentages in order to produce probability datasets with units consistent with those units used for probability datasets produced by similar studies. The resolution of the predicted probability datasets was reduced by aggregating each dataset to keep dataset resolution consistent with the input explanatory dataset that had the lowest resolution for each model. The proportions of the predicted probability datasets were altered by resampling the datasets to correspond with the accuracy of the explanatory dataset having the lowest resolution for each model. For example, the soil characterization variables typically had the lowest resolution (1:250,000) of any variables ending up in the final models, so the predicted probability maps representing these models were aggregated to 1:250,000 because it is the same resolution as the original data used to create the soil characterization variables. The resulting dataset for each of the predicted probability maps represents the percent probability of arsenic concentrations exceeding or equal to 4.0 µg/L in groundwater.

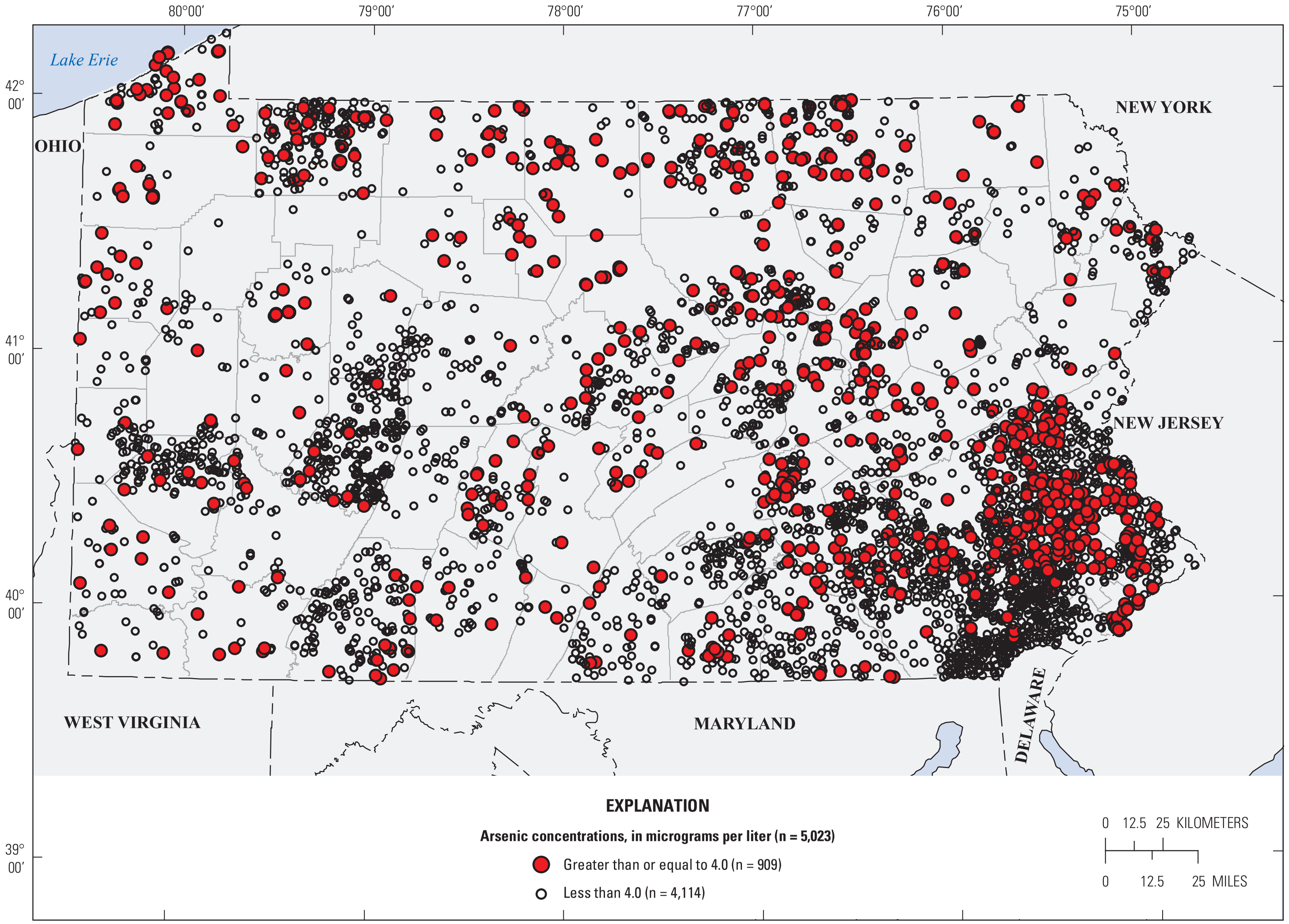
Arsenic Concentrations and Related Factors

This section of the report documents arsenic concentrations and related factors in Pennsylvania using analyses that include only samples with arsenic concentrations. Later sections of the report consider subsets of the data to predict the probability of elevated arsenic concentrations in groundwater.

Arsenic Concentrations

Arsenic concentrations in 5,023 samples of Pennsylvania groundwater ranged from less than the detection level of 0.01 µg/L to 490.0 µg/L. The median concentration was less than 4.0 µg/L. About 82 percent of groundwater samples had arsenic concentrations less than 4.0 µg/L, and about 18 percent of samples had concentrations greater than or equal to the model threshold of 4.0 µg/L. Almost 8 percent of groundwater samples had arsenic concentrations greater than or equal to the USEPA MCL of 10.0 µg/L.

The statewide spatial distribution of the 5,023 samples with measured arsenic concentrations is shown in figure 3. Most of the data are concentrated in the southeastern part of the State because of the large number of groundwater-quality analyses conducted in that area. In most areas, there was a wide range of arsenic concentrations. Table 2 gives a summary of arsenic concentrations in groundwater for the six physiographic provinces in Pennsylvania. When analyzed according to physiographic province, the Central Lowland province had the highest median arsenic concentration (4.5 µg/L) and the highest percentage of sample records with arsenic



Base from U.S. Bureau of the Census digital data, 1990b 1:100,000
Albers Equal-Area Conic projection: Standard Parallels 40° N and
42° N, Central Meridian -78° W, Latitude of Origin 39°

County boundaries from U.S. Bureau of the Census, 1990a

Figure 3. Location of sampling sites and associated reported arsenic concentrations in groundwater in Pennsylvania, 1969–2007. (N, number of samples. Arsenic concentrations are shown as elevated if greater than or equal to 4.0 micrograms per liter. Data from Low and Chichester (2006) and Low and others (2008))

Table 2. Summary of arsenic concentrations in groundwater (1969–2007) for the six physiographic provinces in Pennsylvania.

[µg/L, micrograms per liter; <, less than]

Physiographic province	Number of sample records	Median arsenic concentration, in µg/L	Maximum arsenic concentration, in µg/L	Sample records with arsenic concentrations greater than or equal to 4.0 µg/L, in percent	Sample records with arsenic concentrations greater than or equal to 10.0 µg/L, in percent
Appalachian Plateaus	1,870	<4.0	490.0	19	9
Atlantic Coastal Plain	52	<4.0	50.0	35	19
Central Lowland	76	4.5	293.0	59	43
New England	87	<4.0	20.0	14	6
Piedmont	1,637	<4.0	85.5	15	5
Ridge and Valley	1,301	<4.0	230.0	18	6

concentrations greater than or equal to 4.0 µg/L (59 percent) and greater than or equal to 10.0 µg/L (43 percent) (table 2). No formal statistical test was conducted to evaluate arsenic difference according to physiographic provinces or geologic units in Pennsylvania. Geologic units are listed in appendix 2.

Significance of Groundwater-Quality Properties Related to Arsenic

Spearman's *rho* was used to investigate possible correlations among arsenic and 31 other groundwater-quality properties and constituents: pH, specific conductance, dissolved oxygen, carbon dioxide, alkalinity, major ions (calcium, magnesium, sodium, potassium, iron, manganese, chloride, sulfate, silica, and fluoride), nutrients [nitrate plus nitrite as nitrogen (N), ammonia, phosphate, total organic carbon], and trace elements (aluminum, barium, cadmium, chromium, cobalt, copper, lead, mercury, nickel, selenium, strontium, and zinc). Some of these constituents, such as dissolved oxygen, iron, manganese, sulfate, nitrate, ammonia, and total organic carbon, could be related to the redox environment as described in the next section. Others may have a common source or have mobility similar to that of arsenic, considering variations in redox and pH. Correlations between arsenic and groundwater-quality constituents were examined on a statewide basis, and those groundwater-quality constituents that were statistically significant ($p<0.0001$) and had Spearman's *rho* values of at least 0.10 in the statewide analysis were also examined on a major aquifer basis to determine how correlations vary in the State according to major aquifer type. Correlations may differ by lithology if examined according to geologic unit because geochemical controls on arsenic occurrence and mobility in groundwater may differ locally across the State according to lithology.

For the statewide dataset, a total of 17 constituents and physical properties had correlations with arsenic that were statistically significant ($p<0.0001$), with Spearman's *rho* values of at least 0.10 (table 3). Out of these 17 constituents and physical properties, 1, 0, 10, and 7 had correlations with arsenic that were statistically significant ($p<0.0001$), with Spearman's *rho* values of at least 0.10 for the carbonate, crystalline, siliciclastic, and surficial major aquifers, respectively. In order of decreasing magnitude of Spearman's *rho* for the statewide analysis, cobalt, strontium, nickel, sodium, barium, specific conductance, aluminum, alkalinity, manganese, iron, silica, calcium, pH, sulfate, and total organic carbon were positively correlated with arsenic, whereas dissolved oxygen and nitrate plus nitrite (as N) were inversely correlated with arsenic. Positive and inverse correlations for the major aquifer analyses were the same as the statewide analysis.

The statewide correlation of arsenic with cobalt, strontium, nickel, sodium, barium, aluminum, and calcium indicates that these constituents may occur with arsenic in the geologic environment, and in some cases, they may be mobilized under similar chemical conditions. Correlations of arsenic with these constituents also occurred in the siliciclastic major aquifer; correlations of arsenic with sodium and barium occurred in the surficial major aquifer. This finding indicates that these constituents occur with arsenic in the siliciclastic major aquifer geologic environment and may be mobilized under chemical conditions occurring in this major aquifer.

Statewide correlations with specific conductance, pH, alkalinity, and calcium may reflect greater releases of traces of arsenic with increasing dissolution of major minerals or the tendency for arsenic to be mobilized (desorbed) under alkaline conditions. In uncontaminated water, and except for carbonate aquifers, higher specific conductance values often are associated with older water, which has a higher residence time in fractures and other secondary openings than in soils and sediments and greater contact with minerals, leading to increased hardness, alkalinity, and arsenic concentrations in the water. Correlations with specific conductance and calcium occurred for the siliciclastic major aquifer, and correlations with specific conductance and alkalinity occurred for the surficial major aquifer.

Positive correlations of arsenic with iron, total organic carbon and negative correlations of arsenic with dissolved oxygen and nitrate plus nitrite (as N) are consistent with the observation that arsenic may be released by iron oxides

Table 3. Spearman's *rho* correlations for the 17 groundwater-quality constituents that have the best correlations with arsenic in groundwater in Pennsylvania and major aquifer types.

[p less than 0.0001 and *rho* greater than or equal to 0.10; NS, not signficant (p equal to or greater than 0.0001 and (or) *rho* less than 0.10)]

Groundwater-quality constituent	Statewide		Carbonate major aquifer		Crystalline major aquifer		Siliciclastic major aquifer		Surficial major aquifer	
	Number of samples	Spearman's *rho*	Number of samples	Spearman's *rho*	Number of samples	Spearman's *rho*	Number of samples	Spearman's *rho*	Number of samples	Spearman's *rho*
Direct correlation										
Cobalt	791	0.287	36	NS	102	NS	592	0.206	61	NS
Strontium	939	0.252	43	NS	78	NS	725	0.254	93	NS
Nickel	1,131	0.237	109	NS	129	NS	797	0.284	96	NS
Sodium	3,546	0.180	477	NS	685	NS	2,136	0.173	248	0.377
Barium	2,851	0.171	399	NS	467	NS	1,729	0.145	256	0.450
Specific conductance	2,514	0.155	222	NS	451	NS	1,561	0.150	280	0.275
Aluminum	1,794	0.155	185	NS	158	NS	1,285	0.225	166	NS
Alkalinity	2,411	0.155	346	NS	518	NS	1,390	NS	157	0.390
Manganese	3,541	0.153	477	NS	730	NS	2,086	NS	248	0.447
Iron	3,938	0.140	519	NS	789	NS	2,351	NS	279	0.418
Silica	2,378	0.134	301	NS	612	NS	1,330	0.230	135	NS
Calcium	3,445	0.121	456	NS	683	NS	2,079	0.119	227	NS
pH	3,251	0.113	421	NS	453	NS	2,109	NS	268	NS
Sulfate	3,837	0.103	502	0.183	742	NS	2,294	0.124	299	NS
Total organic carbon	2,096	0.101	352	NS	500	NS	1,146	NS	98	NS
Inverse correlation										
Dissolved oxygen	639	-0.268	38	NS	205	NS	338	NS	58	NS
Nitrate plus nitrite (as N)	3,802	-0.108	509	NS	768	NS	2,256	NS	269	-0.235

under reducing conditions (Smedley and Kinniburgh, 2002). Specifically, the inverse correlation of arsenic with dissolved oxygen and nitrate plus nitrite (as N) and the direct correlation with iron and manganese indicate that reductive dissolution of iron and manganese oxides may be a mechanism that releases arsenic into groundwater. This may be especially true in the surficial major aquifer because arsenic had a direct correlation with iron and manganese and an inverse correlation with nitrate plus nitrite (as N), and these correlations were higher for the surficial major aquifer than those correlations that resulted from the statewide analysis (table 3). Smedley and Kinniburgh (2002) state that arsenic can be released or desorbed from mineral oxides in aquifer materials under alkaline (high pH) conditions, as can other anion-forming elements such as silica and sulfate, which explains the direct correlation of these constituents with arsenic statewide and in the siliciclastic major aquifer.

Redox Conditions

Because arsenic can be mobilized in groundwater as a result of redox reactions and pH conditions, redox-pH matrices describing redox conditions and pH in groundwater were created (table 4) to further evaluate geochemical controls within the State. The frequency and number of samples with arsenic concentrations greater than or equal to 4.0 µg/L were summarized for specified redox and pH classifications. Redox conditions of groundwater were classified on the basis of concentrations of dissolved oxygen, nitrate, sulfate, manganese, and iron. This redox classification system, based on the work of McMahon and Chapelle (2008), used four redox categories for this report: (1) anoxic (dissolved oxygen, nitrate, or sulfate less than 0.5 mg/L, manganese greater than 50.0 µg/L, and iron greater than 100.0 µg/L), (2) mixed (dissolved oxygen, nitrate, or sulfate greater than or equal to 0.5 mg/L and either manganese greater than 50.0 µg/L or iron greater than 100.0 µg/L), (3) oxic (dissolved oxygen, nitrate, or sulfate greater than or equal to 0.5 mg/L, manganese less than 50.0 µg/L, and iron less than 100.0 µg/L), and (4) unknown

Table 4. Redox-pH matrix showing the frequency of arsenic in 5,023 groundwater samples collected statewide in Pennsylvania, by redox classification and range of arsenic concentrations.

[Values within colored blocks are the percentage and, in parentheses, the number of samples that exceed the threshold of 4 micrograms per liter relative to the total number of samples for the redox-pH class. Redox classifications are based on thresholds of McMahon and Chapelle (2008) and modified to four classes because of data limitations: (1) anoxic (dissolved oxygen, nitrate, or sulfate less than 0.5 milligrams per liter (mg/L), manganese greater than 50.0 micrograms per liter (µg/L), and iron greater than 100.0 µg/L), (2) mixed (dissolved oxygen, nitrate, or sulfate greater than or equal to 0.5 mg/L and either manganese greater than 50.0 µg/L or iron greater than 100.0 µg/L), (3) oxic (dissolved oxygen, nitrate, or sulfate greater than or equal to 0.5 mg/L, manganese less than 50.0 µg/L, and iron less than 100.0 µg/L), and (4) unknown (data for dissolved oxygen, nitrate, or sulfate and more than one of the other parameters were missing). %, percent; n, number]

Redox classification		**pH, less than 4.0**	**pH, greater than or equal to 4.0 and less than 6.0**	**pH, greater than or equal to 6.0 and less than 8.0**	**pH, greater than or equal to 8.0**	**Unknown**	**All samples in redox classification:**
		< 1 % (n = 24/5023)	**5 % (n = 268/5023)**	**54 % (n = 2729/5023)**	**5 % (n = 230/5023)**	**35 % (n = 1772/5023)**	
Anoxic	8 % (n = 405/5023)	75 (n = 3/4)	18 (n = 2/11)	18 (n = 38/208)	30 (n = 13/43)	18 (n = 25/139)	20 (n = 81/405)
Mixed	3 % (n = 129/5023)	0 (n = 0/2)	0 (n = 0/11)	15 (n = 4/27)	100 (n = 1/1)	8 (n = 7/88)	9 (n = 12/129)
Oxic	30 % (n = 1530/5023)	0 (n = 0/1)	1 (n = 1/82)	13 (n = 107/809)	24 (n = 10/42)	7 (n = 42/596)	10 (n = 160/1530)
Unknown	59 % (n = 2959/5023)	59 (n = 10/17)	14 (n = 23/164)	16 (n = 269/1685)	23 (n = 33/144)	21 (n = 204/949)	18 (n = 539/2959)
All samples in pH classification		54 (n = 13/24)	10 (n = 26/268)	15 (n = 418/2729)	25 (n = 57/230)	16 (n = 278/1772)	

EXPLANATION

Percentage of samples with arsenic concentrations greater than or equal to 4.0 micrograms per liter

⬜	No data
🟦	0 to less than 1
🟩	1 to less than 10
▨	10 to less than 20
🟧	20 to less than 50
🟥	50 or more

(data for dissolved oxygen, nitrate, or sulfate and more than one of the other constituents were missing). As a result of the small number of sites with measured dissolved oxygen values (12.7 percent of the dataset), much of the redox classifications (oxic, anoxic, mixed, and unknown) were dependent upon the presence and concentration of nitrate and (or) sulfate and manganese and (or) iron. Five general pH classifications were also considered: (1) pH less than 4.0, (2) pH greater than or equal to 4.0 and less than 6.0, (3) pH greater than or equal to 6.0 and less than 8.0, (4) pH greater than or equal to 8.0, and (5) pH concentration unknown.

The frequency of redox conditions of groundwater samples in Pennsylvania was 8 percent of samples were anoxic, 3 percent were mixed, 30 percent were oxic, and 59 percent were unknown (table 4). The frequency of pH classifications of groundwater samples was as follows: 1 percent of samples had pH values less than 4.0, 5 percent had pH values ranging from greater than or equal to 4.0 to less than 6.0, 54 percent had pH values ranging from greater than or equal to 6.0 to less than 8.0, 5 percent had pH values greater than or equal to 8.0, and 35 percent of samples had unknown pH values. In table 4, the percentage of samples within each pH and redox classification with arsenic concentrations greater than or equal to 4.0 µg/L is color coded to highlight those classifications with the greatest frequency of samples with elevated arsenic concentrations.

As groundwater became more reducing, elevated arsenic concentrations were detected more frequently (table 4). For example, 10 percent of oxic waters had elevated arsenic concentrations, as compared to 20 percent of anoxic waters. Also, for a given pH range, more samples with anoxic redox classification more often had elevated arsenic concentrations than samples with oxic or mixed classification. This finding is consistent with the release of arsenic from iron oxides by reductive dissolution of iron oxides or reductive desorption (reduction of arsenate and consequent desorption of arsenite). For oxic and mixed redox classifications, the percentage of samples with elevated arsenic concentrations increased with pH, which is consistent with pH control of arsenate and arsenite adsorption by iron oxides. Regardless of redox classification, samples with pH of less than 4.0 or pH of greater than or equal to 8.0 constituted the highest percentages of samples with elevated arsenic concentrations, indicating that both high and low pH may be associated with elevated arsenic. At pH values less than 4.0, iron oxides may not be stable, which is indicated by elevated iron concentrations and low sulfate and iron ratios. Redox/pH matrices show that as waters become more strongly reduced, strongly acidic, or strongly alkaline, arsenic concentrations tend to increase. Elevated pH may provide a geochemical trigger to release arsenic from the iron oxides. Iron and manganese are highly soluble under reducing conditions, which are common in confined aquifers or where groundwater is in contact with organic material. In addition, dissolution of iron oxides may occur in strongly acidic waters.

Predicted Probability of Elevated Arsenic Concentrations in Groundwater

Logistic regression models were created to predict the probability of elevated arsenic concentrations in groundwater in the State of Pennsylvania and to further improve those predictions for three regions within the State: (1) glacial aquifer system, (2) Gettysburg Basin, and (3) Newark Basin (fig. 2). For the purposes of this report, the predictions for the statewide map represent both surficial and bedrock aquifers, which may differ for any given area, because wells known to be completed in bedrock materials and wells known to be completed in glacial and (or) surficial materials were used for the statewide study. The glacial aquifer system consists of the portion of the State encompassed by the extent of the Wisconsin glaciation (Pennsylvania Geological Survey, 1995) (fig. 1) and those areas in the northern half of the State overlain by surficial material consisting of coarse-grained sediments (Soller and Packard, 1998) (fig. 1). Only wells known to be completed in glacial materials were used for the glacial aquifer system model, and these wells were also included in the statewide model analysis. The Gettysburg and Newark Basins are located in the Gettysburg-Newark Lowland section of the Piedmont province (fig. 1), with the Gettysburg Basin consisting of the southwestern part of the section and the Newark Basin consisting of the northeastern part of the section. The purpose of creating models for these three regions was to create a model specifically for the glaciated part of the State and to evaluate whether predictions could be improved in these areas of Pennsylvania by including locally specific variables that are not representative of or available across the entire State. These regional models were created using explanatory variables that were available or statistically significant only for certain regions within the State.

Model development included univariate and multivariate logistic regression analyses. Univariate relations between elevated arsenic concentrations and explanatory variables were evaluated and are summarized in appendix 3; coefficients indicate the nature of the univariate relation. An alpha level of 0.20 was chosen as the inclusion criterion for selecting explanatory variables for the multivariate analysis because Hosmer and Lemeshow (1989) indicate that a traditional alpha level of 0.10 failed to identify variables known to be important during some logistic regression analyses. The number of explanatory variables selected for multivariate analyses varied according to study area: 30 out of 40 variables statewide, 32 out of 40 variables for the glacial aquifer system, 23 out of 38 variables for the Gettysburg Basin, and 20 out of 40 variables for the Newark Basin. Coefficients from the final statewide and regional multivariate logistic regression models were used to construct maps illustrating the probability of arsenic concentrations equaling or exceeding 4.0 µg/L in groundwater in Pennsylvania and in selected regions within the State.

Statewide

The results of the final statewide model are summarized in table 5 and show the overall fit for the statewide model is statistically significant with a Wald Chi-Square probability value of less than 0.0001. Explanatory variables included in the statewide model are the siliciclastic major aquifer type (SIL), surficial major aquifer type (SURF), estimated groundwater specific conductance (SPCIDW), estimated groundwater pH (PHIDW), and average soil available water capacity (AWCAVE). The standardized regression coefficient (table 5) allows for comparisons among variables in the model in common units where units differed among variables (Menard, 2002). This statistic showed that presence or absence of the siliciclastic and surficial major aquifer types were the most significant variables in the model. SIL, SURF, SPCIDW, and PHIDW showed positive correlations with elevated arsenic concentrations. When the siliciclastic or surficial aquifer types are present, arsenic concentrations increase, which illustrates the potential of primarily siliciclastic and surficial geologic units to affect elevated arsenic concentrations. The positive correlation between elevated arsenic concentrations and the SPCIDW and PHIDW variables describes how groundwater geochemistry can be an indicator of elevated arsenic concentrations in groundwater. Also, the positive correlation between elevated arsenic concentrations and the SPCIDW and PHIDW variables illustrates how groundwater geochemical properties, such as groundwater specific conductance and pH, can increase the potential for arsenic mobilization, which is consistent with other studies where increased pH was shown to mobilize arsenic (Ayotte and others, 2006). Increases in pH can occur due to ion exchange with sodium, and high specific conductance in groundwater may indicate high amounts of salt (Schlottmann and others, 1998). The only variable with a negative relation with elevated arsenic concentrations was AWCAVE (appendix 3 and table 5), indicating that if a soil has the capability of storing smaller amounts of water, the potential for elevated arsenic concentrations in groundwater increases. This relation suggests that the probability of elevated arsenic concentrations increases beneath sandy and loamy soils, which are more prone to leaching and may show how soils that poorly store water may serve as pathways for arsenic to enter groundwater (U.S. Department of Agriculture, 2008). This is a problem especially if arsenic is present on the land surface from contaminated soils (such as from land application of pesticides, mining and ore processing operations, or waste disposal) or mineral deposits because arsenic is more prone to leach from these soils into groundwater. Therefore, groundwater vulnerability to elevated arsenic concentrations in Pennsylvania may be characterized as a function of aquifer type and its associated groundwater geochemistry parameters (pH and specific conductance), along with the ability of overlying soils to store water.

Overall model fit is good with an H-L p-value of 0.2751 (p-values greater than 0.05 indicate good model fit). The generalized r-square value was 0.0709, and the maximum-rescaled r-square value was 0.1161. Percent concordance was 69.2 percent, and model sensitivity was 22.1. Model discrimination was acceptable, as indicated by a c statistic of 0.695. Diagnostic statistics indicated a lack of multicollinearity for all explanatory variables for the statewide model. Multicollinearity diagnostic statistics indicate no detection of multicollinearity among model variables. Tolerance values were greater than 0.4 and VIF values were less than 2.5.

Linear regressions between the actual probability of elevated arsenic concentrations and the average predicted probabilities were constructed according to 10 percent deciles calculated with the statewide model calibration data. The results of the regression confirm good model calibration statewide in Pennsylvania, with an r-square of 0.9647 (fig. 4).

Spatial distribution of predicted probabilities of elevated arsenic concentrations resulting from the statewide model varies across the State and is expressed as a percent probability on a map (fig. 5). The map shows the predicted probability of elevated arsenic in groundwater for both surficial and bedrock aquifers, which may differ in some areas of the State but are not differentiated in figure 5. Probabilities of elevated arsenic concentrations of 20 to 50 percent are more widespread in those parts of the State underlain by siliciclastic and surficial aquifers, which also have the highest potential to act as an arsenic source or mobilizer. Areas with less than 20 percent predicted probability of elevated arsenic concentrations, such as the southwestern part of the State, have soils with high percentages of available water capacity, which can also be indicative of other geologic factors because soil development is related to underlying bedrock lithology, except in glaciated areas. Within units having siliciclastic or surficial major aquifers, high predicted probabilities (greater than 30 percent) of elevated arsenic concentrations are most prominent in areas with high groundwater specific conductance and pH.

The Pearson residual statistic was calculated for each well to determine which observations were most poorly fit by the model and to spatially show the model's predictive error. This statistic calculates the difference between observed and estimated probabilities and divides the result by the binomial standard deviation of the estimated probability (Menard, 2002). Resulting Pearson residuals have a mean equal to 0 and a standard deviation equal to 1. For the 5,009 wells, Pearson residual values include 347 values greater than 2.0 and 3 values less than -2.0. Large positive values account for 7 percent of the dataset, whereas large negative values account for less than 1 percent of the dataset. Spatial locations of Pearson residual values associated with outliers indicate areas where the model is not entirely correct and does not fit the dependent data (fig. 5). The northeastern and north-central parts of the State have the fewest poor predictions, whereas the southeastern part of the State has a large cluster of poor predictions. Some poor predictions are spread through the northwestern tier and central parts of the State. These poor predictions can be attributed to the uneven spatial distribution of arsenic in rock and associated groundwater and the inability of the variables in the model to completely capture the diversity of

Table 5. Summary of statistics for statewide and regional logistic regression models predicting the probability of arsenic exceeding 4.0 micrograms per liter statewide and in three regions in Pennsylvania.

[--, no data; <, less than; m, meter; %, percent; µS, microsiemens; cm, centimeter; g, grams; cm³, cubic centimeter; hr, hour; H-L, Hosmer-Lemeshow goodness-of-fit test; VIF, variance inflation factor]

Model	Statistically significant explanatory variables		Variable results			Model fit						Multicollinearity diagnostics	
	Explanatory variable	Description	Parameter estimate	Wald Chi-square probability (p-value)	Standardized regression coefficient	H-L (p-value)	Generalized r-square	Maximum rescaled r-square	Percent concordant	Sensitivity (%)	c (area under the Receiver operating characteristic curve)	Tolerance	VIF
Statewide	Intercept	Model intercept	-4.1987	<0.0001	-	0.2751	0.0709	0.1161	69.2	22.1	0.695	-	-
	SIL	Siliciclastic major aquifer type (1 = siliciclastic, 0 = all other)	0.8697	<0.0001	0.2327							0.62584	1.59785
	SURF	Surficial major aquifer type (1 = surficial, 0 = all other)	1.4432	<0.0001	0.2289							0.7445	1.34319
	SPCIDW	Estimated groundwater specific conductance (µS/cm)	0.000701	<0.0001	0.2071							0.95841	1.04339
	PHIDW	Estimated groundwater pH	0.4189	<0.0001	0.1722							0.93767	1.06648
	AWCAVE	Average soil available water capacity (%)	-11.3568	<0.0001	-0.1417							0.74158	1.34847
Glacial aquifer system	Intercept	Model intercept	-15.8949	0.0004	-	0.4477	0.2314	0.3297	77.8	55.6	0.780	-	-
	BDAVE	Average soil bulk density (g/cm³)	6.5455	0.0165	0.3576							0.95209	1.05032
	WETAVE	Average wetlands land cover within a 500-m radius (%)	0.0392	0.0115	0.2896							0.76741	1.30309
	SPCIDWGLAC	Estimated groundwater specific conductance using only glacial data (µS/cm)	0.00242	0.0013	0.2793							0.78779	1.26937
	PERMAVE	Average soil permeability (cm/hr)	0.1318	0.0164	0.2039							0.82167	1.21703
	PHIDWGLAC	Estimated groundwater pH using only glacial data	0.5392	0.0288	0.2017							0.88717	1.12718
Gettysburg Basin	Intercept	Model intercept	-12.4639	0.0056	-	0.9244	0.2428	0.3655	83.3	65.9	0.833	-	-
	SANDAVE	Average soil sand content (%)	0.237	0.0406	0.6875							0.94178	1.06182
	ELEV	Land surface elevation above sea level (m, North American Vertical Datum of 1988)	-0.0182	0.0027	-0.4521							0.86271	1.15913
	TRI	Distance to toxic chemical release inventory sites (m)	-0.00025	0.0053	-0.3705							0.93638	1.06794
	PHIDW	Estimated groundwater pH	1.0117	0.0096	0.3542							0.86822	1.15178
	SPCIDW	Estimated groundwater specific conductance (µS/cm)	0.0029	0.0473	0.2439							0.75589	1.32295
Newark Basin	Intercept	Model intercept	-6.6069	0.0143	-	0.5738	0.1331	0.1860	72.7	74.1	0.729	-	-
	PHIDW	Estimated groundwater pH	1.4425	<0.0001	0.3734							0.98324	1.01705
	SPCIDW	Estimated groundwater specific conductance (µs/cm)	0.00115	0.0225	0.2809							0.97962	1.02080
	MUN	Distance to municipalities containing water suppliers with arsenic treatment and sites or groundwater in need of remediation because of arsenic contamination (m)	-0.00011	0.0113	-0.1846							0.87983	1.13659
	AWCAVE	Average soil available water capacity (%)	-42.294	0.0141	-0.1737							0.89884	1.11254

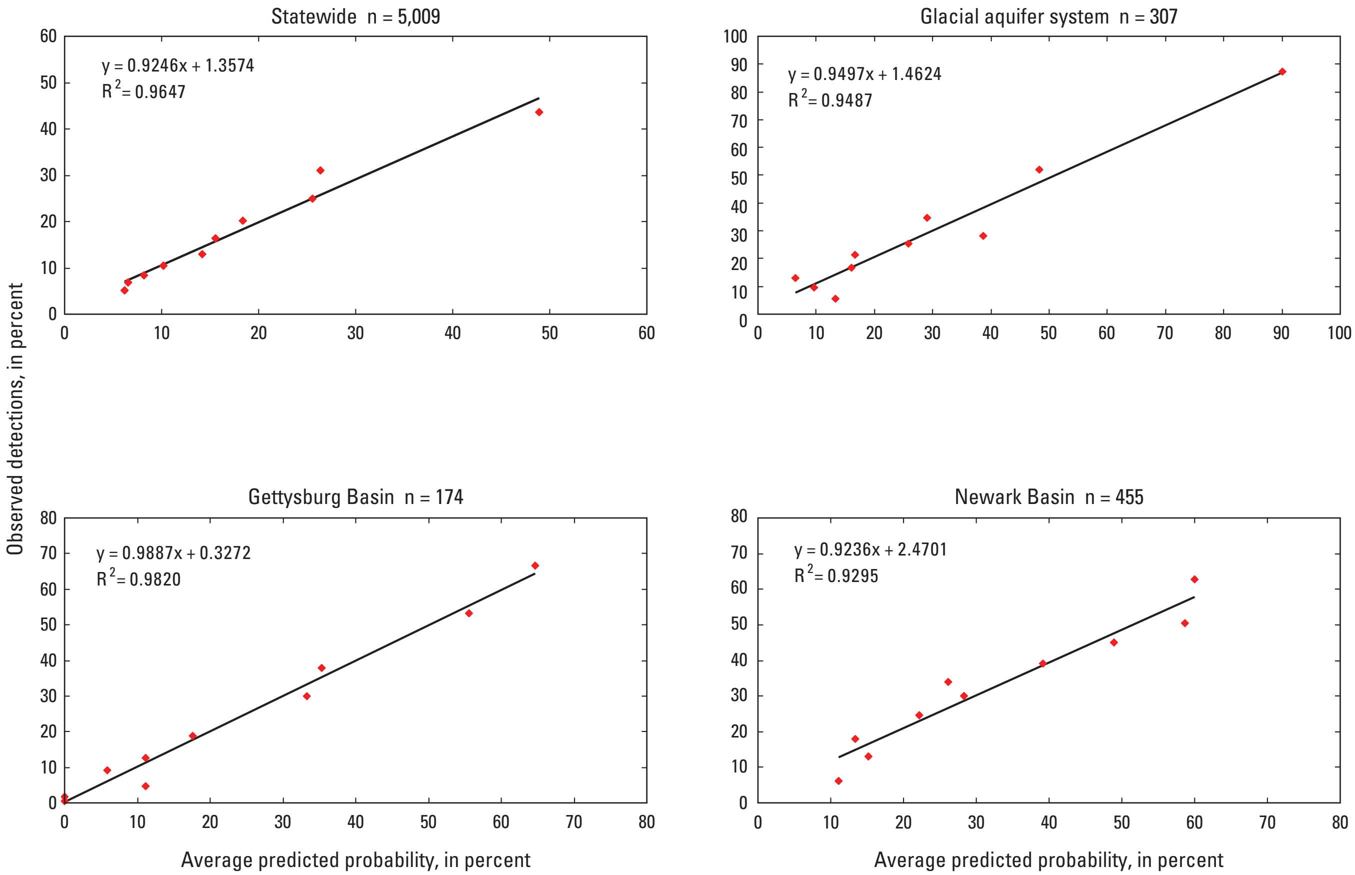
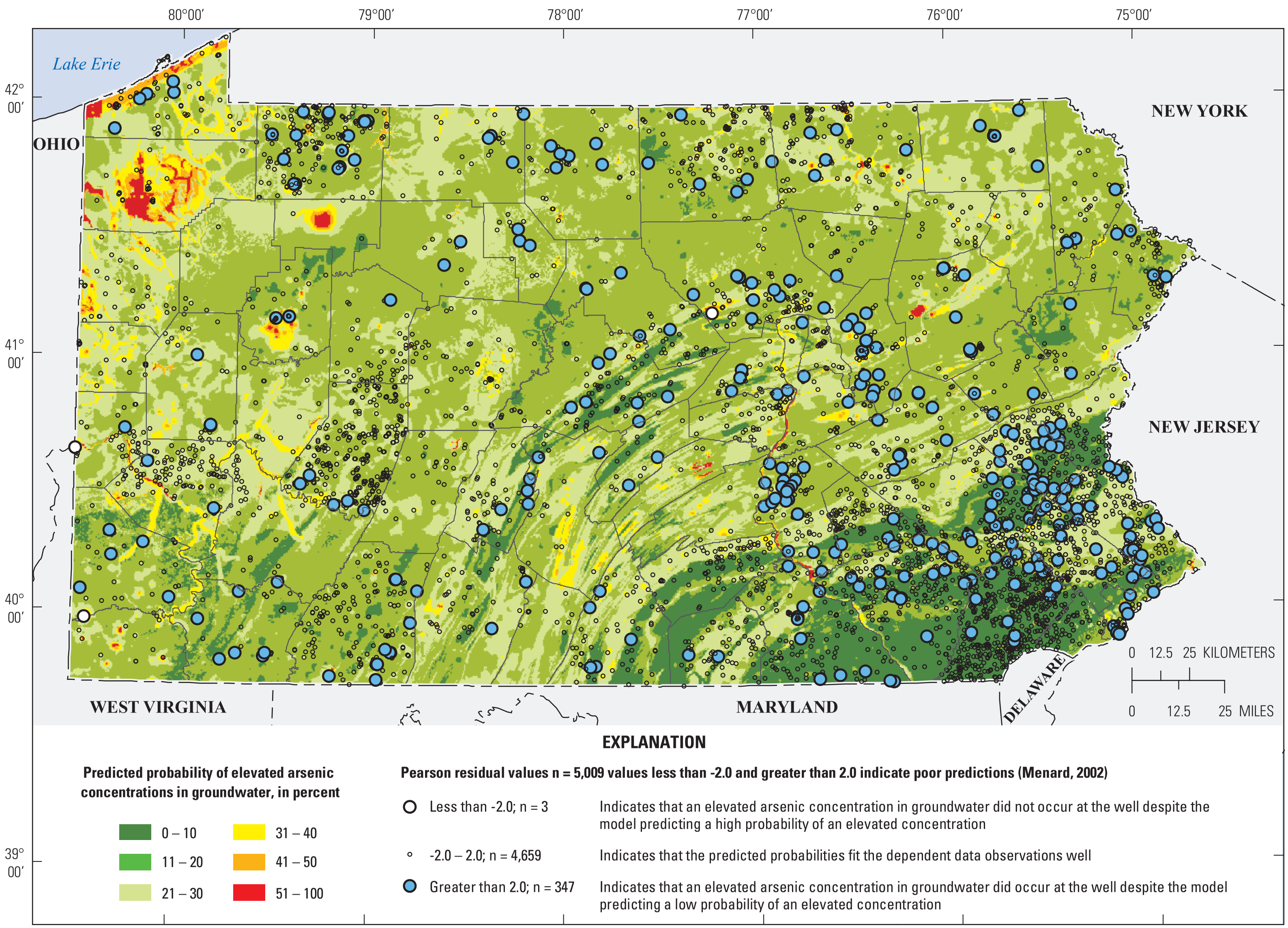


Figure 4. Percentage of observed detections of elevated arsenic concentrations in relation to the average predicted probability of detecting elevated arsenic concentrations in groundwater statewide and in three regions in Pennsylvania.

the geologic, soil, and geochemical properties controlling arsenic in the State; additional variables or separate models for regions would be needed to seamlessly characterize elevated concentrations of arsenic across the entirety of Pennsylvania.

Glacial Aquifer System

The results of the glacial aquifer system model, which includes the late Wisconsinian glacial border (Pennsylvania Geological Survey, 1995) and areas in the northern half of Pennsylvania that are overlain by surficial material consisting of coarse-grained sediments (Soller and Packard, 1998) (fig. 2), are summarized in table 5. The overall fit for the model is statistically significant with a Wald Chi-Square probability value of 0.0004. Explanatory variables included in the glacial aquifer system model are average soil bulk density (BDAVE), average wetlands land cover within a 500-meter radius (WETAVE), specific conductance in groundwater estimated using only glacial data (SPCIDWGLAC), average soil permeability (PERMAVE), and groundwater pH estimated using only glacial data (PHIDWGLAC). Results of the standardized regression show that average soil bulk density is the most significant variable in the model, and all variables showed positive correlations with elevated arsenic concentrations. High soil bulk density is related to a greater degree of soil compaction and may be an indicator of sandy soils because total pore space in sandy soils commonly is less than total pore space in clay or silt (U.S. Department of Agriculture, 2008). Arsenic concentrations increase as average soil permeability increases (appendix 3 and table 5), which indicates that areas of sandy soils with high permeability estimates tend to have higher concentrations of arsenic in groundwater than areas with other soil types. Areas in the glacial aquifer system with sandy or coarse-grained deposits are typically associated with glacial valley-fill deposits that overlay fine-grained deposits of till or clay in some places; sediment texture typically is not uniform with depth in glaciated regions (Thomas, 2007). The sandy soils may also exist as a thin veneer that overlies fine-grained till or clay deposits that form the valley walls and (or) terraces. Extensive clay deposits (glacial lake sediments) are found in many glacial valleys and where surface depressions exist the clay deposits encompass wetlands such as peat bogs. Average wetlands land cover within a 500-meter radius (WETAVE) has the second highest



Base from U.S. Bureau of the Census digital data, 1990b 1:100,000
Albers Equal-Area Conic projection: Standard Parallels 40° N and
42° N, Central Meridian -78° W, Latitude of Origin 39°

County boundaries from U.S. Bureau of the Census, 1990a

Figure 5. Areas with predicted probability that arsenic concentrations in groundwater in surficial and bedrock aquifers statewide in Pennsylvania will be greater than or equal to 4.0 micrograms per liter and sampling sites with associated Pearson residual values showing predictive error for measured arsenic concentrations.

standardized coefficient in the model, which means that the greater the percentage of wetlands within a 500-meter radius, the higher the probability of an elevated arsenic concentration in groundwater. Because many of the glacial valley-fill deposits identified by the bulk density and permeability of soils are overlain by wetlands, underlying till or clay may be creating a confining layer close to the land surface. This finding suggests that arsenic is being mobilized in these areas in part because of the confining layers that create the wetlands in the glacial valleys. Wetlands, which relate to low dissolved oxygen and redox, could be the result of the upward flow of groundwater. These poorly drained areas can lead to high pH and arsenic desorption. The positive correlation between elevated arsenic concentrations and SPCIDWGLAC and PHIDWGLAC variables describes groundwater geochemistry as an indicator of elevated arsenic concentrations in groundwater in the glacial aquifer system. Increases in pH can cause arsenic to become mobilized in water. Increases in specific conductance indicate older, less oxic water, which can result in the mobilization of arsenic. In inundated areas such as wetlands, groundwater is anoxic and organic rich, and flow is restricted, which allows arsenic to be released. The glacial aquifer system model illustrates that groundwater geochemical properties, such as high groundwater specific conductance and pH, may be acting as geochemical triggers and releasing arsenic from the aquifer matrix to the groundwater in the glacial aquifer system (Thomas, 2007).

Model fit was very good with an H-L p-value of 0.4477. The glacial aquifer system model had a generalized r-square value of 0.2314 with a maximum-rescaled r-square value of 0.3297. According to the percent concordant, 77.8 percent of responses were correctly predicted, and the sensitivity value was 55.6. Model discrimination was acceptable with a c statistic of 0.780. Diagnostic statistics indicated a lack of multicollinearity for all explanatory variables because the lowest Tolerance value was 0.76741 (greater than 0.4), and the highest VIF value was 1.30309 (less than 2.5).

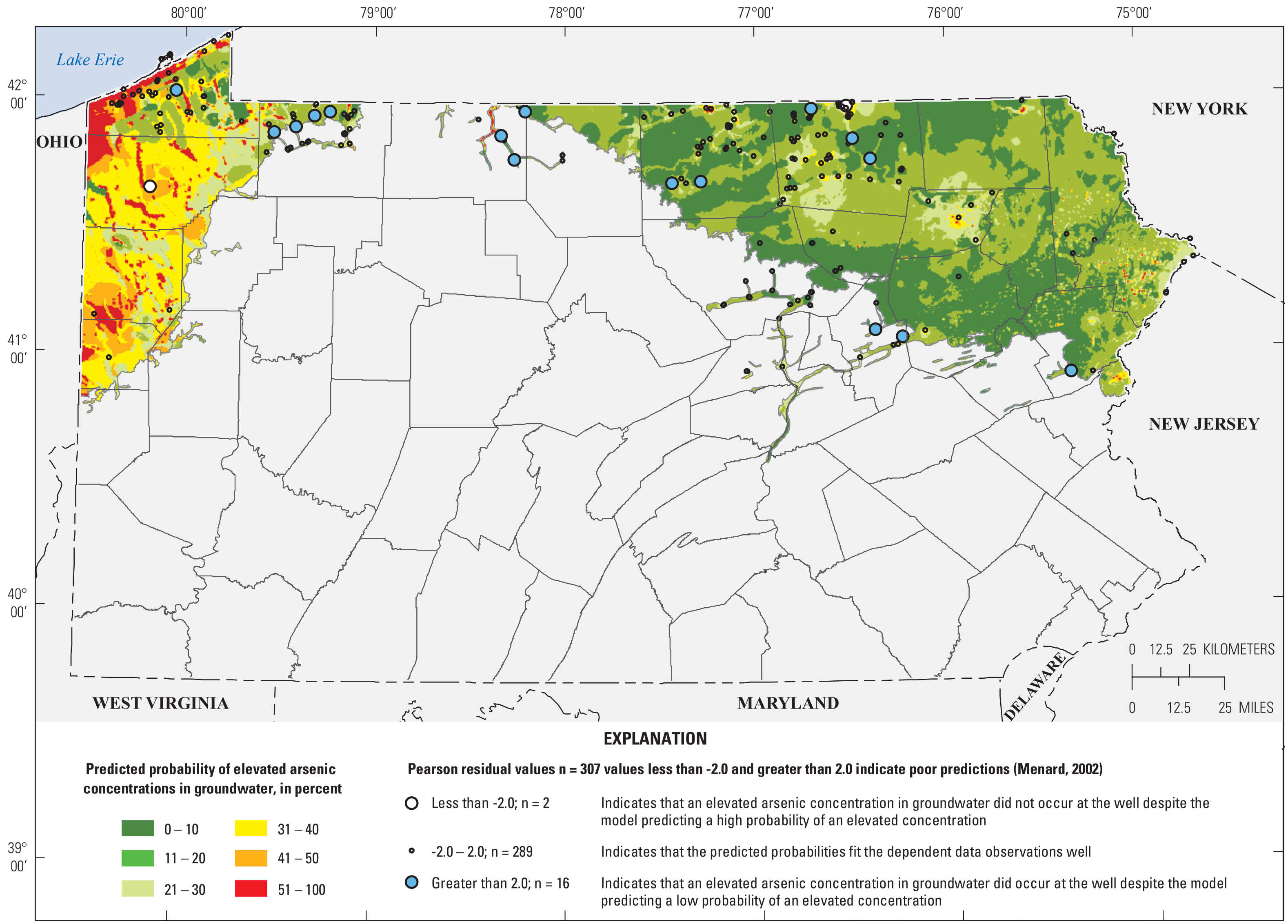
Linear regressions between the actual probability of elevated arsenic concentrations and the average predicted probabilities were constructed according to 10 percent deciles calculated with the glacial aquifer system model calibration data. Regression results confirm good model calibration with an r-square of 0.9487 (fig. 4).

Predicted probabilities of elevated arsenic concentrations produced by the glacial aquifer system model were calculated using arsenic concentration data that were known to represent the glacial units in order to create a model distinguishing between the glacial units and the underlying bedrock. Predicted probabilities for the glacial aquifer system model have a different spatial distribution than probabilities of elevated arsenic resulting from the statewide model that represents underlying bedrock (figs. 5–6). The predicted probabilities of elevated arsenic in groundwater are much greater in the northwestern than in the northeastern glacial tier. High predicted probabilities (greater than 50 percent) of elevated arsenic concentrations are most prominent in areas with high estimated groundwater specific conductance and pH, which were calculated only for those wells finished in the glacial aquifer system. These areas are located along Lake Erie in the northwestern glacial tier and in areas with glacial valley-fill deposits that correspond to high percentages of wetlands. Probabilities of elevated arsenic concentrations of 20 to 50 percent are more widespread in those areas with high average soil bulk density and high soil permeability; most of these areas are located in the northwestern glacial tier. Areas with the lowest estimated specific conductance and pH, such as the northeastern glacial tier, have less than 20 percent predicted probability of elevated arsenic concentrations.

The Pearson residual statistic was calculated for each well and plotted on a map to determine which observations and spatial areas were most poorly fit by the model and to spatially show the model's predictive error. Results indicate Pearson residuals have a mean equal to 0 and standard deviation equal to 1. For the 307 wells, Pearson residual values include 16 values greater than 2.0 and two values less than -2.0. Large positive values account for 5 percent of the dataset, whereas large negative values account for less than 1 percent of the dataset. Spatial locations of Pearson residual outliers appear to be randomly distributed throughout the region (fig. 6).

Gettysburg Basin

Results for the Gettysburg Basin model, which represents south-central Pennsylvania, show the overall model fit is statistically significant with a Wald Chi-Square probability value of 0.0056 (table 5). Explanatory variables in the model include average soil sand content (SANDAVE), land-surface elevation above sea level (NAVD 88) (ELEV), distance to toxic chemical release inventory sites (TRI), average estimated groundwater pH (PHIDW), and estimated groundwater specific conductance (SPCIDW). Standardized regression coefficients show that average soil sand content is the most significant variable in the model. SANDAVE, PHIDW, and SPCIDW were positively correlated with elevated arsenic concentrations, whereas ELEV and TRI were negatively correlated. As average soil sand content increases, arsenic concentrations also increase, and as land-surface elevation above NAVD 88 decreases, arsenic concentrations increase. Soils with high sand content seem to be found mostly in the parts of the region with lower land-surface elevations, which seems to correspond to hornfels bordering diabase intrusions with associated mineral deposits of hydrothermal origin, which Senior and Sloto (2006) also found to have locally elevated arsenic concentrations. Sandy soils and low land-surface elevations may be indicators of the underlying geologic environment acting as an arsenic source or mobilizer, causing elevated arsenic concentrations in groundwater within the Gettysburg Basin. The positive correlation between elevated arsenic concentrations and SPCIDW and PHIDW variables indicate groundwater geochemistry conditions that facilitate arsenic mobilization in groundwater. Geochemical conditions may be further mobilizing arsenic



Base from U.S. Bureau of the Census digital data, 1990b 1:100,000
Albers Equal-Area Conic projection: Standard Parallels 40° N and 42° N, Central Meridian -78° W, Latitude of Origin 39°

County boundaries from U.S. Bureau of the Census, 1990a
Glacial boundary from Pennsylvania Geological Survey, 1995 and Soller & Packard, 1998

Figure 6. Areas with predicted probability that arsenic concentrations in groundwater in the glacial aquifer system in Pennsylvania will be greater than or equal to 4.0 micrograms per liter and sampling sites with associated Pearson residual values showing predictive error for measured arsenic concentrations in wells.

from aquifer materials because high predicted probabilities of elevated arsenic in groundwater also correspond to high estimated groundwater specific conductance and pH. This finding is consistent with the results of the study by Senior and Sloto (2006) in the Newark Basin. With decreased distances to toxic chemical release inventory sites, arsenic concentrations increase. This relation indicates elevated arsenic concentrations occur close to facilities that have been known to release toxic chemicals directly to air, water, or land. These facility locations represent anthropogenic contamination sites that could have released chemicals containing arsenic or chemicals containing constituents that affect redox environment or pH and, hence, promote arsenic mobilization. Senior and Sloto (2006) also noted elevated arsenic concentrations in groundwater near a few contamination sites but did not directly attribute elevated arsenic concentrations in groundwater to contaminant sources. The Gettysburg Basin model illustrates that predictions of elevated arsenic concentrations in groundwater can be refined by considering soils prone to leaching, proximity to contaminated sites, and topography in conjunction with previously associated groundwater geochemical properties (pH, redox, specific conductance).

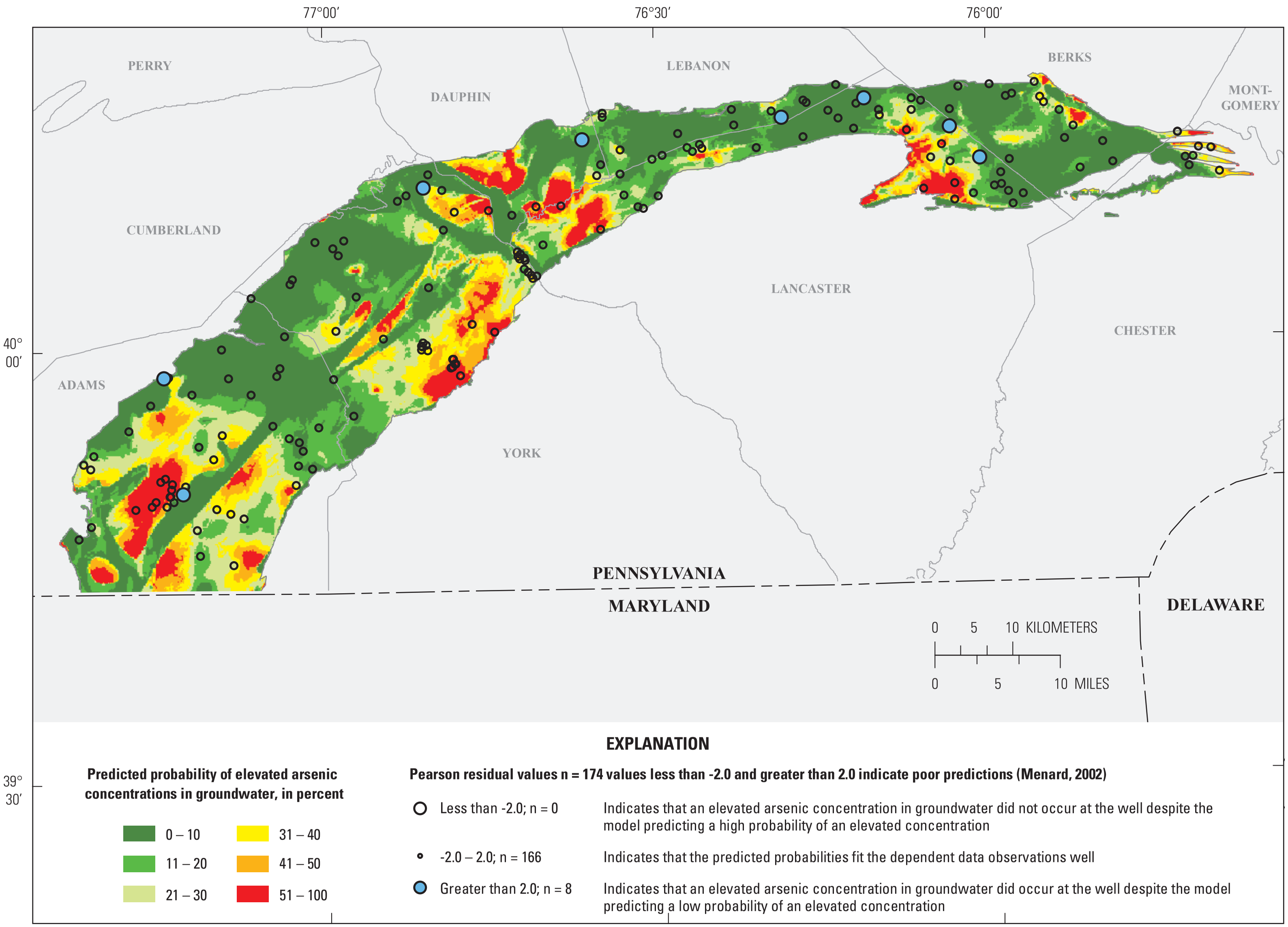
An H-L p-value of 0.9244 indicates that model fit was very good. The generalized r-square value was 0.2428, and the maximum-rescaled r-square value was 0.3655. These were the highest r-square values for any of the models, indicating that these data fit the Gettysburg Basin model better than data for the other models fit those models. The percent concordant value indicates that 83.3 percent of responses were correctly predicted, and model sensitivity was 65.9. Model discrimination was excellent with a c statistic of 0.833. No multicollinearity was detected among model variables because the lowest Tolerance value was greater than 0.4 and the highest VIF value was less than 2.5. Linear regressions between the percentage of observed detections of elevated arsenic concentrations and the average predicted probabilities confirm good model calibration, with an r-square of 0.9820 (fig. 4).

Elevated arsenic concentration predicted probabilities resulting from the Gettysburg Basin model (fig. 7) were higher for the southern part of the region than the probabilities that resulted from the statewide model (fig. 5). Predicted probabilities of elevated arsenic concentrations greater than 50 percent are focused in areas with high sand soil content and low land-surface elevation; many of these areas are located along the southern boundary of the region. Probabilities of elevated arsenic concentrations between 20 to 50 percent surround the areas that represent probabilities greater than 50 percent, which is mainly attributed to the 122 toxic chemical release inventory sites within the basin because as distances to these sites decrease, predicted probabilities of elevated arsenic concentrations increase. Areas that have under 20 percent predicted probability of elevated arsenic concentrations largely correspond to those parts of the region with the lowest estimated groundwater specific conductance and pH.

Results for the Pearson residual statistic include a mean equal to 0 and a standard deviation equal to 1. For the 174 wells, Pearson residual values include 8 values greater than 2.0 and 0 values less than -2.0; large positive values account for 5 percent of the dataset. The spatial locations for the eight Pearson residual values associated with outliers are randomly spread throughout the region (fig. 7).

Newark Basin

The Newark Basin model results, which represent part of southeastern Pennsylvania northeast of the Gettysburg Basin, have geologic variables that are similar to the Gettysburg Basin, which means that these two models have similar model variables. These two areas were considered separately for this study because the Newark Basin has often been considered separately for other studies and has similar but different geology (Peters and Burkert, 2008; Senior and Sloto, 2006). The Newark Basin has an overall model fit that is statistically significant with a Wald Chi-Square probability value of 0.0143 (table 5). Explanatory variables for the Newark Basin model include estimated groundwater pH (PHIDW), estimated groundwater specific conductance (SPCIDW), distance to municipalities containing water suppliers receiving arsenic treatment and sites or groundwater in need of remediation because of arsenic contamination (MUN), and average soil available water capacity (AWCAVE). Standardized regression coefficients show that estimated groundwater pH was the most significant variable in the model. PHIDW and SPCIDW had positive correlations with elevated arsenic concentrations, whereas MUN and AWCAVE had negative correlations (table 5). Correlations between arsenic and PHIDW, SPCIDW, and AWCAVE indicate relations are similar to those for the statewide and Gettysburg Basin models. High specific conductance and pH in groundwater are more than likely providing a geochemical environment that facilitates the mobilization of arsenic from mineral deposits bordering diabase intrusions. Senior and Sloto (2006) found in their study of the Newark Basin that wells completed near diabase in the hornfels commonly had elevated groundwater concentrations of arsenic and that areas with pH of 8 or greater were associated with a geochemical environment favorable for arsenic mobilization. Also, the lowest predicted probabilities of arsenic correspond to those parts of the region with the highest soil available water capacity, which are underlain by the diabase intrusions. The MUN variable was not part of the statewide model; it represents areas with known arsenic contamination in groundwater and soils. The MUN variable’s negative correlation with arsenic shows that arsenic concentrations increase as distance to municipalities with treatment decreases. These municipalities not only represent areas containing geochemical and geologic properties favorable for arsenic mobilization, but they also contain sites that may have been affected by anthropogenic contamination. As with the toxic chemical release inventory sites, it is possible for contaminants, such as arsenic, to enter soils and groundwater from these sites that are in need of remediation. As with the Gettysburg Basin, results of the



Base from U.S. Bureau of the Census digital data, 1990b 1:100,000
Albers Equal-Area Conic projection: Standard Parallels 40° N and
42° N, Central Meridian -78° W, Latitude of Origin 39°

County boundaries from U.S. Bureau of the Census, 1990a
Gettysburg Basin from Pennsylvania Geological Survey, 2001

Figure 7. Areas with predicted probability that arsenic concentrations in groundwater in the Gettysburg Basin in Pennsylvania will be greater than or equal to 4.0 micrograms per liter and sampling sites with associated Pearson residual values showing predictive error for measured arsenic concentrations.

Newark Basin model illustrate that geochemical properties of groundwater, soils properties, and proximity to areas containing water suppliers with arsenic treatment or sites in need of remediation because of arsenic contamination describe an environment favorable for arsenic mobilization.

Overall model fit was very good, indicated by an H-L p-value of 0.5738. The generalized r-square value was 0.1331, and the maximum-rescaled r-square value was 0.1860. The percent concordant value indicates that 72.7 percent of responses were correctly predicted, and model sensitivity was 74.1. A c statistic of 0.729 shows acceptable model discrimination. Diagnostic statistics indicate a lack of multicollinearity for all explanatory variables for the Newark Basin model. Because the lowest Tolerance value was 0.87983 (all values are greater than 0.4) and the highest VIF value was 1.13659 (all values were less than 2.5), these multicollinearity diagnostic statistics indicate that there is no multicollinearity among model variables. Linear regressions between the percentage of observed detections of elevated arsenic concentrations and the average predicted probabilities confirm good model calibration, with an r-square of 0.9295 (fig. 4).

Predicted probabilities of elevated arsenic concentrations resulting from the Newark Basin model (fig. 8) are higher than those probabilities that resulted from the statewide model (fig. 5). Predicted probabilities of elevated arsenic concentrations greater than 30 percent are distributed mainly in the central and northern parts of the region and correspond to high estimated groundwater specific conductance and pH and to proximity to municipalities containing water suppliers with arsenic treatment and sites or groundwater in need of remediation because of arsenic contamination. Predicted probabilities less than 30 percent primarily correspond to high percentages of soil available water capacity. These areas are located in the southeastern part of the Newark Basin and in the northern part of the basin where they correspond to diabase intrusions.

Pearson residual results show a mean equal to 0 and standard deviation equal to 1. For the 455 wells, Pearson residual values include 0 values less than -2.0 and 14 values greater than 2.0; large positive values account for 3 percent of the dataset. Spatial locations of Pearson residual values associated with outliers indicate that most poor predictions are located in the southeastern part of the region (fig. 8).

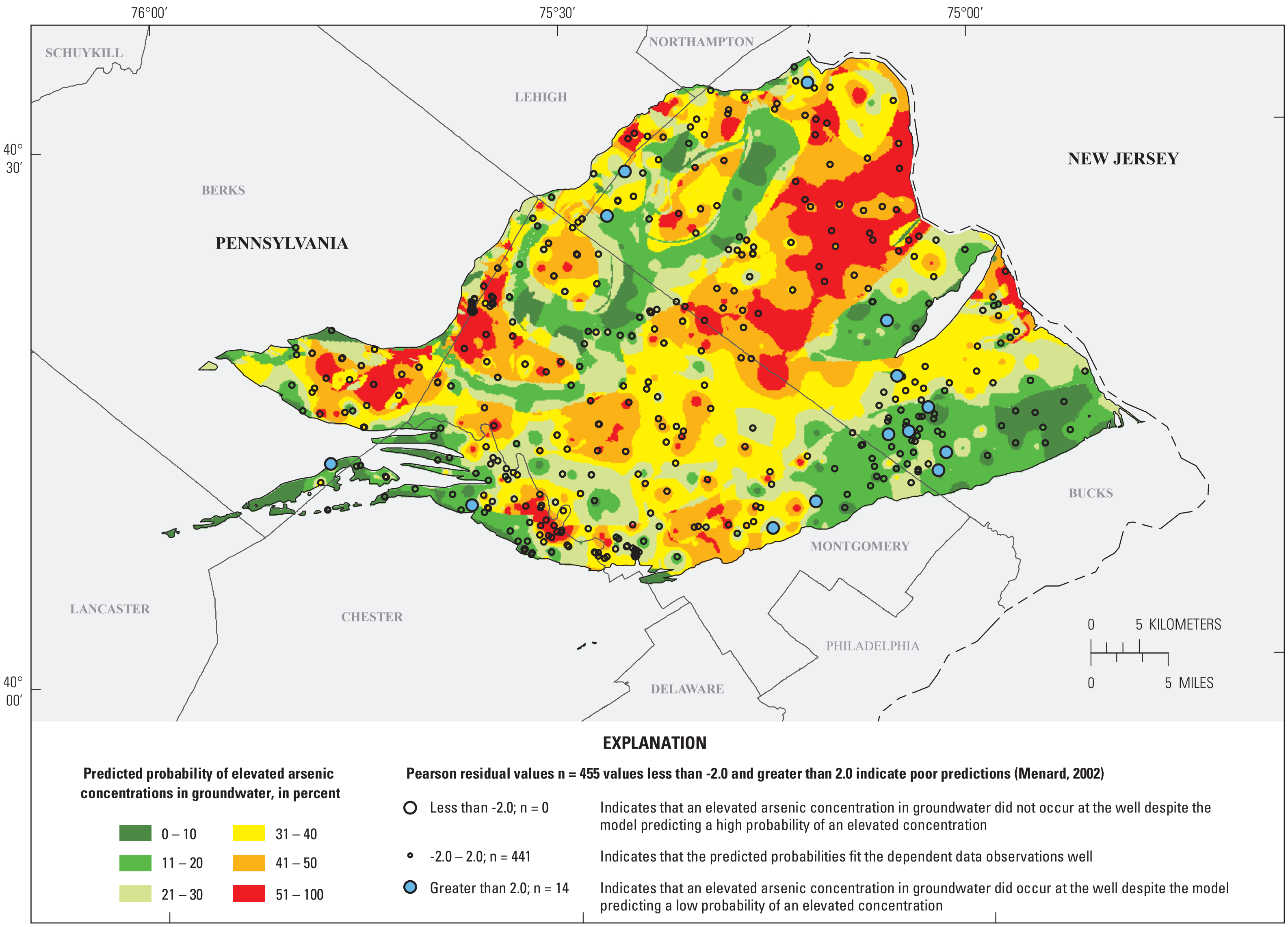
Limitations and Uses of Arsenic Models and Probability Maps

Probability maps developed for this study show the predicted probability of arsenic concentrations greater than or equal to 4.0 µg/L in groundwater in Pennsylvania and three regions within the State, including the glacial aquifer system, the Gettysburg Basin, and the Newark Basin (figs. 5–8). The probability maps do not show actual arsenic contamination of groundwater but rather depict areas that have the potential or likelihood of having groundwater with elevated arsenic concentrations. There is inherent uncertainty associated with each of the maps, which stems from data quality and availability in the well database and GIS-based explanatory variables and the estimation errors in the logistic regression coefficients.

Certain areas of the predictive maps may have elevated arsenic that are not characterized by the statewide or regional models and are indicated by areas with many Pearson residuals indicating poor predictions (values less than -2.0 and greater than 2.0). These areas have greater uncertainty associated with their relative predicted probabilities. The occurrence of Pearson residuals representing poor predictions resulting from the statewide and regional models supports the concept that regional models across the entire State or regional models encompassing smaller regions than those analyzed in this study may be needed to improve predicted probabilities. Higher resolution spatial data and additional data on arsenic concentrations in groundwater and associated variables would improve the predictive power of the models. Further study and additional independent data are also needed to validate the results of each logistic regression model.

In addition, regional differences in numerous factors, such as lithology and geochemical controls, affecting arsenic mobility are not accounted for in the statewide model and may reduce the predictive power of the statewide model. For example, the major aquifer type variable present in the statewide analysis is a broad characterization of dominant lithologies and may mask localized mineralogical differences in geologic units within the major aquifers. An example of a geochemical control not accounted for can be seen in table 4 where it shows how elevated concentrations of arsenic tend to occur at low pH (less than 4.0) and high pH (greater than or equal to 8.0), but the statewide Spearman's *rho* correlation indicates a positive correlation between arsenic concentrations and pH values. This positive correlation between arsenic concentrations and pH values masks the relation of elevated arsenic concentrations to low pH, which may occur in fewer places statewide but still results in arsenic concentrations of concern. Differences in variables that were significant among statewide and regional models also illustrate factors that are not able to be accounted for in the statewide model. It is likely that the regional models for geologic units and major aquifers in other physiographic provinces, such as the Appalachian Plateaus, Atlantic Coastal Plain, Piedmont Uplands, or Ridge and Valley, that were not separately analyzed for this study, would also differ and provide better predictive ability in those regions than the statewide model.

By providing an improved understanding of spatial controls on arsenic statewide and in three regions in Pennsylvania, the probability maps developed for this report may help resource managers to prioritize areas for groundwater-quality monitoring or implement alternative management practices. However, the probability maps associated with this report are intended for regional-scale use and have limitations for use at the field-scale or when considering individual wells. The maps are not appropriate at any scale larger than 1:250,000, which is the smallest scale of any of the explanatory variables



Base from U.S. Bureau of the Census digital data, 1990b 1:100,000
Albers Equal-Area Conic projection: Standard Parallels 40° N and
42° N, Central Meridian -78° W, Latitude of Origin 39°

County boundaries from U.S. Bureau of the Census, 1990a
Newark Basin from Pennsylvania Geological Survey, 2001

Figure 8. Areas with predicted probability that arsenic concentrations in groundwater in the Newark Basin in Pennsylvania will be greater than or equal to 4.0 micrograms per liter and sampling sites with associated Pearson residual values showing predictive error for measured arsenic concentrations in wells.

used in the models. The threshold accuracy value for a scale of 1:250,000, as established by National Map Accuracy Standards, is 127 meters (417 feet) in ground units. Concentrations of arsenic in groundwater and in a well are affected by many field-scale complexities that are not accounted for in the models. For example, the models do not account for preferential groundwater flow paths in bedrock units. As a result, although a well may be located in a region with a high predicted probability of elevated arsenic, the well may yield water with low arsenic concentrations as a result of complexities that cannot be represented in the regional-scale models that were developed for this study. Therefore, the models and maps are not appropriate for predicting the risk of elevated arsenic in individual wells.

In addition to the explanatory variables included in each of the four models, there may be other explanatory variables that can affect concentrations of arsenic in groundwater. For example, the regional models do not take geology into account, which means that these regions need to be studied in greater detail. Also, exclusion of an explanatory variable from the model does not mean that the variable does not affect concentrations of arsenic. For example, data on dissolved oxygen concentrations could have been used to define the redox environment, but sufficient data were not available to create a dissolved oxygen dataset for use in the analysis. In addition, for two or more explanatory variables that exhibit strong correlations with one another, all of these variables were not included in the same model because including one variable usually accounts for the effects of the other correlated variables. For example, soil available water capacity and soil thickness have a strong positive correlation, so only one of the variables was included. Additionally, characterization of lithologies other than by the selected major aquifer types may result in different correlations between elevated arsenic concentrations in groundwater and geologic units. Some geologic units from each major aquifer type may be more likely than other geologic units to have elevated arsenic concentrations, and so the more subtle differences in mineralogy of these units that affects arsenic occurrence and mobility are not captured in the models.

Summary and Conclusions

Data on arsenic concentrations in groundwater were available for 5,023 wells—monitoring, domestic, public supply, commercial, irrigation, and industrial wells—across Pennsylvania. Arsenic was detected at a concentration of 4.0 micrograms per liter (µg/L) or greater in 18 percent of samples. Arsenic concentrations that met or exceeded the U.S. Environmental Protection Agency maximum contaminant level of 10.0 µg/L were measured in about 8 percent of samples; the highest arsenic concentration was 490.0 µg/L.

Comparison of arsenic concentrations in groundwater by physiographic province indicates that the Central Lowland province in northwestern Pennsylvania has the highest median arsenic concentration (4.5 µg/L) and highest percentage of sample records with arsenic concentrations greater than or equal to 4.0 µg/L (59 percent) and greater than or equal to 10.0 µg/L (43 percent). Evaluation of arsenic concentrations in groundwater of four major aquifer types in Pennsylvania (carbonate, crystalline, siliciclastic, and surficial) showed that all aquifer types have median arsenic concentrations less than 4.0 µg/L, and the highest arsenic concentration (490.0 µg/L) was in a siliciclastic aquifer. The siliciclastic and surficial aquifers have the highest percentage of sample records with arsenic concentrations greater than or equal to 4.0 µg/L and 10.0 µg/L. Elevated arsenic concentrations, which are those arsenic concentrations greater than or equal to 4.0 µg/L, were most commonly found in strongly reducing groundwater statewide. Arsenic concentrations were elevated in 10 percent of oxic waters, whereas 20 percent of anoxic waters had elevated arsenic concentrations. Also, for a given pH range, samples classified as anoxic reductive-oxidative (redox) water more typically had elevated arsenic concentrations than samples classified as oxic or mixed. The relatively large percentage of anoxic and high pH waters with elevated concentrations of arsenic indicates that mechanisms such as reductive desorption or dissolution are occurring, releasing arsenic from iron oxides. However, elevated arsenic concentrations were also found in reducing waters with low pH, indicating that other mechanisms, such as reductive dissolution of iron oxides, also may result in mobilization of arsenic in groundwater.

Arsenic concentrations were correlated with concentrations of several chemical constituents, including (1) constituents linked to redox processes (specific conductance, alkalinity, iron, pH, total organic carbon, dissolved oxygen, and nitrate plus nitrite (as N)), (2) constituents that may be mobilized under chemical conditions similar to those that mobilize arsenic (cobalt, strontium, nickel, sodium, barium, aluminum, and calcium), and (3) anions or oxyanions that sorb to iron oxides (silica and sulfate).

Logistic regression was used to develop spatial statistical models predicting the probability of detecting arsenic concentrations greater than or equal to 4.0 µg/L in groundwater statewide and in three intrastate regions. Although, the statewide and regional (glacial aquifer system, Gettysburg Basin, and Newark Basin) models consisted of slightly different variables, the results have common characteristics that can be grouped as (1) geologic and soils variables that describe arsenic sources and mobilizers (major aquifer types, soil available water capacity, soil bulk density, soil permeability, and soil sand content), (2) geochemical variables that describe the geochemical environment of the groundwater (specific conductance and pH), and (3) locally specific variables that are unique to each of the three regions studied and are not applicable to the statewide analysis (wetlands land cover, land-surface elevations above the North American Vertical Datum of 1988 (NAVD 88), distance to toxic chemical release inventory sites, and distance to municipalities containing water suppliers with arsenic treatment and sites or groundwater in need of remediation because of arsenic contamination). These models have limitations because they may not characterize areas that have different controls on arsenic mobility and should not be used to estimate probabilities of elevated arsenic in groundwater at the field-scale or when considering individual wells.

The occurrence of arsenic in groundwater of Pennsylvania is largely the result of mobilization of arsenic from natural sources in the aquifer and the geochemical conditions of groundwater. For this investigation, the potential for natural geologic features to be sources of arsenic was established on the basis of selected geologic characteristics (primary lithology to determine major aquifer type). Also, the data reviewed for this investigation indicate that variations in arsenic concentrations in Pennsylvania groundwater generally are related to variations in redox conditions, pH, and specific conductance at State and regional scales. As demonstrated in this report, these groundwater-quality variables, which affect the release, transport, and attenuation of arsenic in the aquifer, may be considered in conjunction with siliciclastic and surficial major aquifer types to indicate geologic conditions where groundwater may contain elevated arsenic. Health officials may consider testing or advising homeowners in geologic settings within siliciclastic or surficial major aquifer types to test their groundwater for arsenic, particularly if other available data, such as proximity to known areas of elevated arsenic or arsenic contaminant release sites or geochemical conditions (low redox, high pH, high specific conductance) indicate a geochemical environment that favors the mobilization of arsenic.

References Cited

Allison, P.D., 2001, Logistic regression using the SAS system: Theory and application: Cary, North Carolina, SAS Institute Inc., 288 p.

Ayotte, J.D., Nielsen, M.G., and Robinson, B.R., Jr., 1998, Relation of arsenic concentrations in ground water to bedrock lithology in eastern New England: Geological Society of America Annual Meeting, Toronto, Ontario, p. A-58.

Ayotte, J.D., Nolan, B.T., Nuckols, J.R., Cantor, K.P., Robinson, G.R. Jr, Baris, Dalsu, Hayes, Laura, Karagas, M.R., Bress, William, Silverman, D.T., and Lubin, J.H., 2006, Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment: Environmental Science & Technology, v. 40, no.11, p. 3578–3585.

Bankey, V., Cuevas, A., Daniels, D., Finn, C.A., Hernandez, I., Hill, P., Kucks, R., Miles, W., Pilkington, M., Roberts, C., Roest, W., Rystrom, V., Shearer, S., Snyder, S., Sweeney, R., Velez, J., Philips, J.D., Ravat, D., 2002, Digital data grids for the magnetic anomaly map of North America: U.S. Geological Survey Open-File Report 02–414, raster digital data.

Benbrahim-Tallaa, Lamia, and Waalkes, M.P., 2007, Inorganic arsenic and human prostate cancer: Environmental Health Perspectives, accessed May 4, 2009, at *http://dx.doi.org/doi:10.1289/ehp.10423*

Berg, T.M., Barnes, J.H., Sevon, W.D., Skema, V.W., Wilshusen, J.P., and Yannacci, D.W., 1989, Physiographic provinces of Pennsylvania: Pennsylvania Geological Survey, 4th ser., Map 13 (color), scale 1:2,000,000, 8.5 x 11.

Berg, T.M., Edmunds, W.E., Geyer, A.R., Glover, A.D., Hoskins, D.M., MacLachlan, D.B., Root, S.I., Sevon, W.D., and Socolow, A.A., comps., 1980, Geologic map of Pennsylvania: Pennsylvania Geological Survey, 4th ser., Map 1, scale 1:250,000, 3 sheets.

Buckwalter, T.F., and Moore, M.E., 2007, Ground-water resources and the hydrologic effects of petroleum occurrence and development, Warren County, Northwestern Pennsylvania: U.S. Geological Survey Scientific Investigations Report 2006–5263, 86 p.

Chen, Yu, Santella, R.M., Kibriya, M.G., Wang, Qiao, Kapil, Maya, Verret, W.J., Graziano, J.H., and Ahsan, Habibul, 2007, Association between arsenic exposure from drinking water and plasma levels of soluble cell adhesion molecules: Environmental Health Perspectives, accessed May 4, 2009, at *http://dx.doi.org/doi:10.1289/ehp.10277*.

Cox, D.R., and Snell, E.J., 1989, Analysis of binary data (3rd ed.): Boca Raton, Fla., CRC Press, 236 p.

Cravotta, C.A., III, 2008, Dissolved metals and associated constituents in abandoned coal-mine discharges, Pennsylvania, USA – 1. Constituent concentrations and correlations: Applied Geochemistry, v. 23, p. 166–202.

Cuff, D.J., Young, W.J., Muller, E.K., Zelinsky, W., and Abler, R.F., 1989, The Atlas of Pennsylvania: Philadelphia, Pa., Temple University Press, 288 p.

Davey, J.C., Nomikos, A.P., Wungjiranirum, Manida, Sherman, J.R., Ingram, Liam, Batki, Cavus, Lariviere, J.P., and Hamilton, J.W., 2007, Arsenic as an endocrine disruptor—receptor-mediated gene regulation and thyroid hormone-mediated amphibian tail metamorphosis: Environmental Health Perspectives, accessed May 4, 2009, at *http://dx.doi.org/doi:10.1289/ehp.10131*.

Eary, L.E., and Schramke, J.A., 1990, Rates of inorganic oxidation reactions involving dissolved oxygen, *in* Melchior, D.C., and Bassett, R.L., eds., Chemical modeling of aqueous systems II: American Chemical Society Symposium 416, p. 379–396.

Fenneman, N.M., and Johnson, D.W., 1946, Physical divisions of the United States: U.S. Geological Survey, scale 1:7,000,000, 1 sheet.

Foster, A.L., Brown Jr. G.E., Parks, G.A., 2003, XAFS study of As(V) and Se(IV) sorption complexes on hydrous Mn oxides: Geochimica et Cosmochimica Acta, v. 67, no. 11, p. 1937–1953.

Gilliom, R.J., Barbash, J.E., Crawford, C.G., Hamilton, P.A., Martin, J.D., Nakagaki, Naomi, Nowell, L.H., Scott, J.C., Stackelberg, P.E., Thelin, G.P., and Wolock, D.M., 2006, The quality of our Nation's waters—Pesticides in the Nation's streams and ground water, 1992–2001: U.S. Geological Survey Circular 1291, 172 p.

Grossman, J.N., 1999, Geochemistry of igneous rocks in the U.S. (major elements) from the PLUTO database: U.S. Geological Survey, digital data, accessed September 29, 2008 at *http://tin.er.usgs.gov/pluto/igneous/*.

Helsel, D.R., and Hirsch, R.M., 1992, Statistical methods in water resources: New York, Elsevier Science Publishing Company Inc., Studies in Environmental Science 49, 522 p.

Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 510 p.

Hopenhayn, Claudia, 2006, Arsenic in drinking water—impact on human health: Elements, v. 2, p. 103–107.

Hosmer, D.W., Jr., and Lemeshow, Stanley, 1989, Applied logistic regression: New York, John Wiley & Sons, 307 p.

Kochanov, W.E., and Reese, S.O., 2003, Density of mapped karst features in south-central and southeastern Pennsylvania: Pennsylvania Geological Survey, 4th ser., Map 68, Scale 1:300,000, 1 pl.

Langland, M.J., and Dugas, D.L., 1996, Assessment of severity and distribution of corrosive ground water in Pennsylvania: U.S. Geological Survey Open-File Report 95–377, 2 p.

Lin, Wei, Wang, S.-L., Wu, H.-J., Chang, K.-H., Yeh, Peter, Chen, C.-J., and Guo, H.-R., 2008, Associations between arsenic in drinking water and pterygium in southwestern Taiwan: Environmental Health Perspectives, accessed May 4, 2009, at *http://dx.doi.org/doi:10.1289/ehp.11111*.

Lindsey, B.D., and Bickford, T.M., 1999, Hydrogeologic framework and sampling design for an assessment of agricultural pesticides in ground water in Pennsylvania: U.S. Geological Survey Water-Resources Investigations Report 99–4076, 44 p.

Low, D.J., and Chichester, D.G., 2006, Ground-water-quality data in Pennsylvania—A compilation of computerized [electronic] databases, 1979–2004: U.S. Geological Survey Data Series 150, 22 p. 1 CD.

Low, D.J., Chichester, D.C., and Zarr, L.F., 2008, Selected ground-water-quality data in Pennsylvania, 1979–2006: U.S. Geological Survey Data Series 314, 22 p., 1 CD.

Low, D.J., and Galeone, D.G., 2006, Reconnaissance of arsenic concentrations in ground water from bedrock and unconsolidated aquifers in eight northern-tier counties of Pennsylvania: U.S. Geological Survey Open-File Report 2006–1376, 35 p.

Matisoff, G.C., Khourey, C.J., Hall, J.F., Varnes, A.W., and Strain, W.H., 1982, The nature and source of arsenic in northeastern Ohio ground water: Ground Water, v. 20, no. 4, p. 446–456.

McMahon, P.B., and Chapelle, F.H., 2008, Redox processes and water quality of selected principal aquifer systems: Ground Water, v. 46, no. 2, p. 259–271.

Menard, S.W., 2002, Applied Logistic Regression Analysis, (2nd. ed.): Thousand Oaks, Calif., Sage Publications, Inc., 120 p.

Mueller, S.H., Verplanck, P.L., and Goldfarb, R.J., 2001, Ground-Water studies in Fairbanks, Alaska - A better understanding of some of the United States' highest natural arsenic concentrations: U.S. Geological Survey Fact Sheet 111-01, 2 p.

Nagelkerke, N.J.D., 1991, A note on the general definition of the coefficient of determination: Biometricka, v. 66, p. 403–412.

Nakagaki, N., Price, C.V., Falcone, J.A., Hitt, K.J., Ruddy, B.C., 2007, Enhanced national land cover data 1992 (NLCDe 92): U.S. Geological Survey, digital data, accessed March 31, 2008, at *http://water.usgs.gov/GIS/metadata/usgswrd/XML/nlcde92.xml*.

National Research Council, 1999, Arsenic in drinking water: Washington, D.C., National Academy Press, 273 p.

National Research Council, 2001, Arsenic in drinking water: 2001 Update: Washington, D.C., National Academy Press, 225 p.

Parks, G.A., 1990, Surface energy and adsorption at mineral-water interfaces: An introduction, *in* Hochella, M.F., Jr., and White, A.F., eds., Mineral-water interface geochemistry: Washington, D.C., Mineralogical Society of America, p. 133–175.

Pennsylvania Bulletin, 1997–2009, Water suppliers receiving arsenic treatment and sites/groundwater in need of remediation due to arsenic contamination: Pennsylvania Bulletin, online database, accessed April 2, 2008, at *http://www.pabulletin.com/index.asp*.

Pennsylvania Department of Environmental Protection, 2008a, Coal mining operations: Pennsylvania Department of Environmental Protection, digital data, accessed May 15, 2008, at *http://www.pasda.psu.edu*.

Pennsylvania Department of Environmental Protection, 2008b, Industrial mineral mining operations: Pennsylvania Department of Environmental Protection, digital data, accessed April 2, 2008, at *http://www.pasda.psu.edu*.

Pennsylvania Department of Environmental Protection, 2008c, Oil & gas locations: Pennsylvania Department of Environmental Protection, digital data, accessed April 2, 2008, at *http://www.pasda.psu.edu*.

Pennsylvania Department of Transportation, 2008, PennDOT—Pennsylvania municipality boundaries 2008: Pennsylvania Department of Transportation, digital data, accessed January 23, 2009, at *http://www.pasda.psu.edu*.

Pennsylvania Geological Survey, 1995, Late Wisconsinan glacial borders 1:500,000: Pennsylvania Geological Survey, digital data, accessed January 23, 2009, at *http://www.pasda.psu.edu*.

Pennsylvania Geological Survey, 1998, Preliminary landform subdivisions of Pennsylvania, 1998: Pennsylvania Geological Survey, digital data, accessed June 23, 2008, at *http://www.pasda.psu.edu*.

Pennsylvania Geological Survey, 2001, Bedrock geology of Pennsylvania: Pennsylvania Geological Survey, digital data, accessed June 23, 2008, at *http://www.dcnr.state.pa.us/topogeo/map1/bedmap.aspx*.

Pennsylvania Geological Survey, 2005, Geologic units containing potentially significant acid-producing sulfide minerals: Pennsylvania Geological Survey, 4th ser., Open-File Report OFMI 05–01.1, 9 p.

Pennsylvania Geological Survey, 2007, Sinkhole inventory and online database: Pennsylvania Geological Survey, online database, accessed January 2007, at *http://www.dcnr.state.pa.us/topogeo/hazards/sinkhole/*.

Peters, S.C., and Burkert, Lori, 2008, The occurrence and geochemistry of arsenic in groundwaters in the Newark basin of Pennsylvania: Applied Geochemistry 23, p. 85–98.

Price, C., 2003, 1990 population density by block group for the conterminous United States: U.S. Geological Survey, digital data, accessed June 11, 2008, at *http://water.usgs.gov/GIS/metadata/usgswrd/XML/uspopd90x10g.xml*.

PRISM Group at Oregon State University, 2006a, United States average monthly or annual maximum temperature, 1971–2000: PRISM Group at Oregon State University, digital data, accessed May 19, 2009, at *http://www.prism.oregonstate.edu/products/matrix.phtml?vartype=tmax&view=data*.

PRISM Group at Oregon State University, 2006b, United States average monthly or annual minimum temperature, 1971–2000: PRISM Group at Oregon State University, digital data, accessed May 19, 2009, at *http://www.prism.oregonstate.edu/products/matrix.phtml?vartype=tmin&view=data*.

PRISM Group at Oregon State University, 2006c, United States average monthly or annual precipitation, 1971–2000: PRISM Group at Oregon State University, digital data, accessed May 19, 2009, *at http://www.prism.oregonstate.edu/products/matrix.phtml?vartype=ppt&view=data*.

Savage, K.S., Tingle, T.N., O'Day, P.A., Waychunas, G.A., and Bird, D.K., 2000, Arsenic speciation in pyrite and secondary weathering phases, Mother Lode Gold District, Tuolumne County, California: Applied Geochemistry, v. 15, p. 1219–1244.

Schlottmann, J.L., Mosier, E.L., and Breit, G.N., 1998, Arsenic, chromium, selenium, and uranium in the central Oklahoma aquifer, Oklahoma: Results of Investigations, 1985: U.S. Geological Survey Water-Supply Paper 2357-A, p. 119–179.

Senior, L.A., and Sloto, R.A., 2006, Arsenic, boron, and fluoride concentrations in ground water in and near diabase intrusions, Newark Basin, southeastern Pennsylvania: U.S. Geological Survey Scientific Investigations Report 2006–5261, 105 p.

Smedley, P.L., and Kinniburgh, D.G., 2002, A review of the source, behavior, and distribution of arsenic in natural waters: Applied Geochemistry, v. 17, no. 5, p. 517–568.

Soller, D.R., and Packard, P.H., 1998, Digital representation of a map showing the thickness and character of Quaternary sediments in the glaciated United States east of the Rocky Mountains: U.S. Geological Survey Digital Data Series DDS-38, 1 CD.

Stollenwerk, K.G., 2003, Geochemical process controlling transport of arsenic in groundwater—A review of adsorption, *in* Welch, A.H., and Stollenwerk, K.G., eds., Arsenic in ground water – Geochemistry and occurrence: Boston, Kluwer Academic Publishers, chap. 3, p. 67–100.

Thomas, M.A., 2007, The association of arsenic with redox conditions, depth, and ground-water age in the Glacial Aquifer System of the northern United States: U.S. Geological Survey Scientific Investigations Report 2007–5036, 26 p.

Thomas, M.A., Diehl, S.F., Pletsch, B.A., Schumann, T.L., Pavey, R.R., and Swinford, E.M., 2008, Relation between solid-phase and dissolved arsenic in the ground-water system underlying northern Preble County, Ohio: U.S. Geological Survey Scientific Investigations Report 2008–5205, 56 p.

U.S. Bureau of the Census, 1990a, 1990 county & county equivalent areas: U.S. Bureau of the Census, digital data, accessed June 11, 2008, at *http://www.census.gov/geo/www/cob/co1990.html*.

U.S. Bureau of the Census, 1990b, 1990 state & state equivalent areas: U.S. Bureau of the Census, digital data, accessed June 11, 2008, at *http://www.census.gov/geo/www/cob/st1990.html*.

U.S. Bureau of the Census, 2010, DP-1 profile of general population and housing characteristics, 2010: U.S. Bureau of the Census, table, accessed June 20, 2012, at *http://factfinder2.census.gov/*.

U.S. Department of Agriculture, 1993, Soil survey manual: U.S. Department of Agriculture, Handbook no. 18, 437 p.

U.S. Department of Agriculture, Natural Resources Conservation Service, 2008, Soil quality indicators: available water capacity and bulk density: U.S. Department of Agriculture, Soil Quality Physical Indicator Information Sheet Series Sheet, accessed March 17, 2010, at *http://soils.usda.gov/sqi/assessment/assessment.html*.

U.S. Environmental Protection Agency, 1994, Toxic Chemical Release Inventory (TRI) data for sites in Pennsylvania: U.S. Environmental Protection Agency, digital data, accessed November 21, 2008, at *http://www.pasda.psu.edu*.

U.S. Environmental Protection Agency, 2006, 2006 Edition of the drinking water standards and health advisories, accessed September 12, 2007, at *http:/www.epa.gov/waterscience/criteria/drinking/dwstandards.html*.

U.S. Geological Survey, 2004, National Uranium Resource Evaluation (NURE) hydrogeochemical and stream sediment reconnaissance data: U.S. Geological Survey, digital data, accessed September 29, 2008, at *http://tin.er.usgs.gov/nure/sediment/*.

U.S. Geological Survey, 2005, National Hydrography Dataset (NHD) - high-resolution: flowline features clipped to PA state border: U.S. Geological Survey, vector digital data, accessed June 19, 2008, at *http://nhd.usgs.gov*.

U.S. Geological Survey, 2007, Mineral resources data system: U.S. Geological Survey, digital data, accessed October 1, 2008, at *http://tin.er.usgs.gov/mrds*.

U.S. Geological Survey, 2009, 1-Arc second National Elevation Dataset: U.S. Geological Survey, raster digital data, accessed September 8, 2009, at *http://seamless.usgs.gov/index.php*.

Welch, A.H., Westjohn, D.B., Helsel, D.R., and Wanty, R.B., 2000, Arsenic in ground water of the United States—Occurrence and geochemistry: Ground Water, v. 38, no. 4, p. 589–604.

Williams, J.H., Taylor, L.E., and Low, D.J., 1998, Hydrogeology and groundwater quality of the glaciated valleys of Bradford, Tioga, and Potter Counties, Pennsylvania: Pennsylvania Geological Survey, 4th ser., Water Resource Report 68, 89 p.

Wolock, D.M., 1997, STATSGO soil characteristics for the conterminous United States: U.S. Geological Survey Open-File Report 656, digital data, accessed May 28, 2008, at *http://water.usgs.gov/GIS/metadata/usgswrd/XML/muid.xml*.

Wolock, D.M., 2003, Estimated mean annual natural groundwater recharge in the conterminous United States: U.S. Geological Survey, digital data, accessed May 28, 2008, at *http://water.usgs.gov/GIS/metadata/usgswrd/XML/rech48grd.xml*.

Appendixes 1, 2, and 3

Appendix 1. Summary of statewide and regional anthropogenic and natural factors used as explanatory variables in logistic regression models for elevated arsenic concentrations in groundwater in Pennsylvania and number of sample records.

Appendix 2. Summary of arsenic concentrations in groundwater (1969–2007) for the 193 geologic units in Pennsylvania with major aquifer type and primary lithology.

Appendix 3. Results of univariate logistic regression analyses with logistic regression standardized coefficients and individual p-values of independent variables related to the detection of elevated concentrations of arsenic in groundwater samples collected statewide and in three regions in Pennsylvania.

Appendix 1. Summary of statewide and regional anthropogenic and natural factors used as explanatory variables in logistic regression models for elevated arsenic concentrations in groundwater in Pennsylvania and number of sample records.

[-, variable not assessed for this study area; GIS, geographic information system; m, meter; km², square kilometer; %, percent; ºF, degrees Fahrenheit; µg/L, micrograms per liter; µS, microsiemens; cm, centimeter; nT, nanoteslas; g, grams; cm³, cubic centimeter; hr, hour; µS/cm, microsiemens per centimeter at 25 degrees Celsius; mm/yr, millimeters per year]

Explanatory variable	Description	Data source(s)	Original data GIS format	Original data resolution	Number of sample records			
					Statewide	Glacial aquifer system	Gettysburg Basin	Newark Basin
		Anthropogenic factors						
Contamination								
MUN	Distance to municipalities containing water suppliers with arsenic treatment and sites or groundwater in need of remediation because of arsenic contamination (m)	Developed for this study from Pennsylvania Bulletin (1997 to 2009) and Pennsylvania Department of Transportation (2008)	Polygon	1:24,000	5,023	309	175	455
TRI	Distance to toxic chemical release inventory sites (m)	U.S. Environmental Protection Agency (1994)	Point	1:100,000	5,023	309	175	455
Disturbance								
COAL	Distance to underground and surface coal mining operations (m)	Pennsylvania Department of Environmental Protection (2008a)	Point	1:24,000	-	309	-	-
INDMIN	Distance to industrial mineral mining operations (m)	Pennsylvania Department of Environmental Protection (2008b)	Point	1:24,000	5,023	309	175	455
OILGAS	Distance to oil and gas locations (m)	Pennsylvania Department of Environmental Protection (2008c)	Point	1:24,000	-	309	-	-
Land cover								
AGAVE	Average agricultural land cover within a 500-m radius (%)	Nakagaki and others (2007)	Raster	1:24,000	5,023	309	175	455
POP	Population density (people/km²)	Price (2003)	Raster	1:250,000	5,023	309	175	455
URBAVE	Average urban land cover within a 500-m radius (%)	Nakagaki and others (2007)	Raster	1:24,000	5,023	309	175	455
		Natural factors						
Climate								
PRE	Average precipitation from 1971 to 2000 (mm/yr)	PRISM Group at Oregon State University (2006c)	Raster	1:250,000	5,023	309	175	455
RECH	Average groundwater recharge from 1951 to 1980 (mm/yr)	Wolock (2003)	Raster	1:250,000	4,995	281	175	455
TEMP	Average temperature from 1971 to 2000 (ºF)	PRISM Group at Oregon State University (2006a, 2006b)	Raster	1:250,000	5,023	309	175	455

Appendix 1. Summary of statewide and regional anthropogenic and natural factors used as explanatory variables in logistic regression models for elevated arsenic concentrations in groundwater in Pennsylvania and number of sample records.—Continued

[-, variable not assessed for this study area; GIS, geographic information system; m, meter; km^2, square kilometer; %, percent; ºF, degrees Fahrenheit; µg/L, micrograms per liter; µS, microsiemens; cm, centimeter; nT, nanoteslas; g, grams; cm^3, cubic centimeter; hr, hour; µS/cm, microsiemens per centimeter at 25 degrees Celsius; mm/yr, millimeters per year]

Explanatory variable	Description	Data source(s)	Original data GIS format	Original data resolution	Number of sample records			
					Statewide	Glacial aquifer system	Gettysburg Basin	Newark Basin
		Natural factors—Continued						
Geology								
ACID	Distance to geologic units containing acid-producing sulfide minerals (m)	Pennsylvania Geological Survey (2001, 2005)	Polygon	1:250,000	5,023	309	-	455
CARB	Carbonate major aquifer type (1 = carbonate, 0 = all other)	Pennsylvania Geological Survey (2001)	Polygon	1:250,000	5,023	-	-	-
CRYST	Crystalline major aquifer type (1 = crystalline, 0 = all other)	Pennsylvania Geological Survey (2001)	Polygon	1:250,000	5,023	-	-	-
DIA	Distance to diabase geologic units (m)	Pennsylvania Geological Survey (2001)	Polygon	1:250,000	-	-	175	455
PLUTIG	Distance to igneous rock locations (m)	Grossman (1999)	Point	1:62,500	-	-	175	455
SIL	Siliciclastic major aquifer type (1 = siliciclastic, 0 = all other)	Pennsylvania Geological Survey (2001)	Polygon	1:250,000	5,023	-	-	-
SURF	Surficial major aquifer type (1 = surficial, 0 = all other)	Pennsylvania Geological Survey (2001)	Polygon	1:250,000	5,023	-	-	-
Geophysical								
MAG	Average residual total intensity of the earth's magnetic field (nT)	Bankey and others (2002)	Raster	1:250,000	5,023	309	175	455
Groundwater geochemistry								
CORR	Grouping of geologic units according to estimated groundwater corrosivity (1 to 11 based on Langelier Saturation Index)	Developed for this study from Langland and Dugas (1996) and Pennsylvania Geological Survey (2001)	Polygon	1:250,000	5,023	309	175	455
NUREAS	Estimated arsenic concentration derived from stream-sediment data (µg/L)	U.S. Geological Survey (2004)	Point	1:24,000	-	-	-	455
PHIDW	Estimated groundwater pH	Developed for this study from Langland and Dugas (1996), Low and Chichester (2006), Pennsylvania Geological Survey (2001), and U.S. Geological Survey (2004)	Point	1:24,000	5,011	-	174	455

Appendix 1. Summary of statewide and regional anthropogenic and natural factors used as explanatory variables in logistic regression models for elevated arsenic concentrations in groundwater in Pennsylvania and number of sample records.—Continued

[-, variable not assessed for this study area; GIS, geographic information system; m, meter; km², square kilometer; %, percent; ºF, degrees Fahrenheit; µg/L, micrograms per liter; µS, microsiemens; cm, centimeter; nT, nanoteslas; g, grams; cm³, cubic centimeter; hr, hour; µS/cm, microsiemens per centimeter at 25 degrees Celsius; mm/yr, millimeters per year]

Explanatory variable	Description	Data source(s)	Original data GIS format	Original data resolution	Number of sample records			
					Statewide	Glacial aquifer system	Gettysburg Basin	Newark Basin
		Natural factors—Continued						
Groundwater geochemistry—Continued								
PHIDWGLAC	Estimated groundwater pH using only glacial data	Developed for this study from Langland and Dugas (1996),Low and Chichester (2006), Pennsylvania Geological Survey (1995, 2001), Soller and Packard (1998), and U.S. Geological Survey (2004)	Point	1:24,000	-	309	-	-
SPCIDW	Estimated groundwater specific conductance (µS/cm)	Developed for this study from Langland and Dugas (1996), Low and Chichester (2006), Pennsylvania Geological Survey (2001), and U.S. Geological Survey (2004)	Point	1:24,000	5,011	-	174	455
SPCIDWGLAC	Estimated groundwater specific conductance using only glacial data (µS/cm)	Developed for this study from Langland and Dugas (1996), Low and Chichester (2006), Pennsylvania Geological Survey (1995, 2001), Soller and Packard (1998), and U.S. Geological Survey (2004)	Point	1:24,000	-	309	-	-
Land cover								
FORAVE	Average forested land cover within a 500-m radius (%)	Nakagaki and others (2007)	Raster	1:24,000	5,023	309	175	455
WETAVE	Average wetlands land cover within a 500-m radius (%)	Nakagaki and others (2007)	Raster	1:24,000	5,023	309	175	455
Metals and minerals								
METMIN	Distance to metallic mineral resources (m)	U.S. Geological Survey (2007)	Point	1:24,000	5,023	309	175	455
MIN	Distance to metallic and nonmetallic mineral resources (m)	U.S. Geological Survey (2007)	Point	1:24,000	5,023	309	175	455
NONMETMIN	Distance to nonmetallic mineral resources (m)	U.S. Geological Survey (2007)	Point	1:24,000	5,023	309	175	455
Soil characterization								
AWCAVE	Average soil available water capacity (%)	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
BDAVE	Average soil bulk density (g/cm³)	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
CLAYAVE	Average soil clay content (%)	Wolock (1997)	Raster	1:250,000	5,021	307	175	455

Appendix 1. Summary of statewide and regional anthropogenic and natural factors used as explanatory variables in logistic regression models for elevated arsenic concentrations in groundwater in Pennsylvania and number of sample records.—Continued

[-, variable not assessed for this study area; GIS, geographic information system; m, meter; km^2, square kilometer; %, percent; ºF, degrees Fahrenheit; µg/L, micrograms per liter; µS, microsiemens; cm, centimeter; nT, nanoteslas; g, grams; cm^3, cubic centimeter; hr, hour; µS/cm, microsiemens per centimeter at 25 degrees Celsius; mm/yr, millimeters per year]

Explanatory variable	Description	Data source(s)	Original data GIS format	Original data resolution	Number of sample records			
					Statewide	Glacial aquifer system	Gettysburg Basin	Newark Basin
		Natural factors—Continued						
Soil characterization—Continued								
HGA	Average soil hydrologic soil group A content (%)*	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
HGAD	Average soil hydrologic soil group AD content (%)*	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
HGB	Average soil hydrologic soil group B content (%)*	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
HGBC	Average soil hydrologic soil group BC content (%)*	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
HGBD	Average soil hydrologic soil group BD content (%)*	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
HGC	Average soil hydrologic soil group C content (%)*	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
HGCD	Average soil hydrologic soil group CD content (%)*	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
HGD	Average soil hydrologic soil group D content (%)*	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
OMAVE	Average soil organic matter content (% by weight)	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
PERMAVE	Average soil permeability (cm/hr)	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
ROCKDEPAVE	Average soil thickness (cm)	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
SANDAVE	Average soil sand content (%)	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
SILTAVE	Average soil silt content (%)	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
SLOPEAVE	Average soil land-surface slope (%)	Wolock (1997)	Raster	1:250,000	5,021	307	175	455
Topography								
ELEV	Land-surface elevation above sea level (m, North American Vertical Datum of 1988)	U.S. Geological Survey (2009)	Raster	1:24,000	5,023	309	175	455
SINK	Distance to nearest sinkhole (m)	Kochanov and Reese (2003) and Pennsylvania Geological Survey (2007)	Point	1:24,000	-	-	175	455
SLOPE	Slope derived from elevation (%)	U.S. Geological Survey (2009)	Raster	1:24,000	5,023	309	175	455
STREAMD	Average stream density within a 500-m radius (km^2)	U.S. Geological Survey (2005)	Line	1:100,000	5,023	309	175	455

*Hydrologic soil groups are described as Group A, Group B, Group C, and Group D, in order from low to high runoff potential, with dual hydrologic soil groups (AD, BC, CD) assigned on the basis of drained and undrained conditions.

Appendix 2. Summary of arsenic concentrations in groundwater (1969–2007) for the 193 geologic units in Pennsylvania with major aquifer type and primary lithology.

[* indicates geologic unit partially overlain by surficial materials consisting of coarse grained sediments (Soller and Packard, 1998; Lindsey and Bickford, 1999) and that some sample records associated with these geologic units may be categorized with the surficial major aquifer type; µg/L, micrograms per liter; <, less than; --, geologic unit with no sample records]

Major aquifer type	Primary lithology[1]	Geologic unit name[1]	Number of sample records	Median arsenic concentration, in µg/L	Maximum arsenic concentration, in µg/L	Number of sample records with arsenic concentrations greater than or equal to 4.0 µg/L	Number of sample records with arsenic concentrations greater than or equal to 10.0 µg/L
Carbonate	Argillaceous dolomite	Vintage Formation	8	<4.0	<4.0	0	0
Carbonate	Argillaceous limestone	Chambersburg Formation	4	<4.0	4.4	1	0
Carbonate	Argillaceous limestone	Hershey and Myerstown Formations, undivided	3	<4.0	12.0	1	1
Carbonate	Argillaceous limestone	Hershey Formation through Annville Formation, undivided	0	--	--	0	0
Carbonate	Argillaceous limestone	Limestone of Martinsburg Formation	3	<4.0	8.7	1	0
Carbonate	Argillaceous limestone	Pleasant Hill Formation	1	<4.0	<4.0	0	0
Carbonate	Dolomite	Allentown Formation	39	<4.0	10.0	4	1
Carbonate	Dolomite	Bellefonte and Axemann Formations, undivided	0	--	--	0	0
Carbonate	Dolomite	Bellefonte Formation	17	<4.0	6.4	1	0
Carbonate	Dolomite	Gatesburg Formation	7	<4.0	<4.0	0	0
Carbonate	Dolomite	Ledger Formation	54	<4.0	20.0	4	4
Carbonate	Dolomite	Leithsville Formation	24	<4.0	40.0	1	1
Carbonate	Dolomite	Lower (Middle?) Cambrian rocks, undivided	0	--	--	0	0
Carbonate	Dolomite	Mines Member of Gatesburg Formation	8	<4.0	5.9	1	0
Carbonate	Dolomite	Nittany and Stonehenge/Larke Formations, undivided	12	<4.0	10.0	3	2
Carbonate	Dolomite	Nittany Formation	7	<4.0	<4.0	0	0
Carbonate	Dolomite	Ontelaunee Formation	6	<4.0	10.0	2	1
Carbonate	Dolomite	Pinesburg Station Formation	6	<4.0	5.0	2	0
Carbonate	Dolomite	Richland Formation	3	<4.0	<4.0	0	0
Carbonate	Dolomite	Rickenbach Formation	9	<4.0	<4.0	0	0
Carbonate	Dolomite	Snitz Creek Formation	6	<4.0	<4.0	0	0
Carbonate	Dolomite	Tomstown Formation	7	<4.0	<4.0	0	0
Carbonate	Dolomite	Zooks Corner Formation	25	<4.0	<4.0	0	0
Carbonate	Graphitic marble	Franklin Marble	0	--	--	0	0
Carbonate	High-calcium limestone	Annville Formation	0	--	--	0	0
Carbonate	High-calcium limestone	Valentine Member of Benner Formation	0	--	--	0	0
Carbonate	Limestone	Axemann Formation	6	<4.0	<4.0	0	0
Carbonate	Limestone	Beekmantown Group	14	<4.0	40.0	1	1
Carbonate	Limestone	Benner Formation through Loysburg Formation, undivided	13	<4.0	4.0	1	0
Carbonate	Limestone	Buffalo Springs Formation	12	<4.0	7.0	2	0
Carbonate	Limestone	Coburn Formation through Loysburg Formation, undivided	4	<4.0	10.0	1	1
Carbonate	Limestone	Coburn Formation through Nealmont Formation, undivided	20	<4.0	<4.0	0	0
Carbonate	Limestone	Conestoga Formation	59	<4.0	67.0	6	3
Carbonate	Limestone	Epler Formation	42	<4.0	5.0	3	0
Carbonate	Limestone	Keyser and Tonoloway Formations, undivided*	46	<4.0	110.0	8	4
Carbonate	Limestone	Keyser Formation through Clinton Group, undivided	0	--	--	0	0
Carbonate	Limestone	Keyser Formation through Mifflintown Formation, undivided*	9	<4.0	217.5	1	1
Carbonate	Limestone	Limestone of Hamburg sequence	4	<4.0	<4.0	0	0
Carbonate	Limestone	Millbach and Schaefferstown Formations, undivided	2	<4.0	<4.0	0	0

Appendix 2. Summary of arsenic concentrations in groundwater (1969–2007) for the 193 geologic units in Pennsylvania with major aquifer type and primary lithology.—Continued

[* indicates geologic unit partially overlain by surficial materials consisting of coarse grained sediments (Soller and Packard, 1998; Lindsey and Bickford, 1999) and that some sample records associated with these geologic units may be categorized with the surficial major aquifer type; µg/L, micrograms per liter; <, less than; --, geologic unit with no sample records]

Major aquifer type	Primary lithology[1]	Geologic unit name[1]	Number of sample records	Median arsenic concentration, in µg/L	Maximum arsenic concentration, in µg/L	Number of sample records with arsenic concentrations greater than or equal to 4.0 µg/L	Number of sample records with arsenic concentrations greater than or equal to 10.0 µg/L
Carbonate	Limestone	Millbach Formation	5	<4.0	<4.0	0	0
Carbonate	Limestone	Monongahela Group*	19	<4.0	24.1	1	1
Carbonate	Limestone	Rockdale Run Formation	16	<4.0	6.1	1	0
Carbonate	Limestone	Shadygrove Formation	6	<4.0	<4.0	0	0
Carbonate	Limestone	Shriver, Mandata, Corriganville, and New Creek Members of Old Port Formation, undivided	0	--	--	0	0
Carbonate	Limestone	Snitz Creek and Buffalo Springs Formations, undivided	4	<4.0	<4.0	0	0
Carbonate	Limestone	St. Paul Group	9	<4.0	<4.0	0	0
Carbonate	Limestone	Stonehenge Formation	29	<4.0	5.8	3	0
Carbonate	Limestone	Stonehenge/Larke Formation	4	<4.0	8.8	1	0
Carbonate	Limestone	Warrior Formation	2	<4.0	<4.0	0	0
Carbonate	Limestone	Zullinger Formation	7	<4.0	<4.0	0	0
Carbonate	Limestone conglomerate	Limestone fanglomerate	7	<4.0	4.3	1	0
Carbonate	Marble	Cockeysville Marble	19	<4.0	<4.0	0	0
Carbonate	Marble	Wakefield Marble	0	--	--	0	0
Carbonate	Shaly limestone	Jacksonburg Formation	8	6.2	20.7	5	2
Crystalline	Albite-chlorite schist	Octoraro Formation	66	<4.0	10.0	3	2
Crystalline	Andesite	Jonestown Volcanic Suite	3	<4.0	<4.0	0	0
Crystalline	Anorthosite	Anorthosite	7	<4.0	<4.0	0	0
Crystalline	Chlorite-sericite schist	Peters Creek Schist	47	<4.0	7.0	3	0
Crystalline	Diabase	Diabase	92	<4.0	60.0	17	7
Crystalline	Feldspathic quartzite	Setters Quartzite	28	<4.0	6.1	1	0
Crystalline	Felsic gneiss	Felsic and intermediate gneiss	106	<4.0	<4.0	0	0
Crystalline	Felsic gneiss	Felsic gneiss	34	<4.0	23.9	2	1
Crystalline	Felsic gneiss	Felsic to mafic gneiss	58	<4.0	20.0	4	1
Crystalline	Granitic gneiss	Granitic gneiss and granite	0	--	--	0	0
Crystalline	Granitic pegmatite	Pegmatite	1	<4.0	<4.0	0	0
Crystalline	Graphitic felsic gneiss	Graphitic felsic gneiss	18	<4.0	<4.0	0	0
Crystalline	Graphitic gneiss	Graphitic felsic gneiss	28	<4.0	5.0	2	0
Crystalline	Greenstone schist	Greenstone schist	0	--	--	0	0
Crystalline	Mafic gneiss	Banded mafic gneiss	16	<4.0	<4.0	0	0
Crystalline	Mafic gneiss	Hornblende gneiss	21	<4.0	20.0	6	2
Crystalline	Mafic gneiss	Mafic gneiss	15	<4.0	<4.0	0	0
Crystalline	Metabasalt	Metabasalt	1	<4.0	<4.0	0	0
Crystalline	Metabasalt	Sams Creek Metabasalt	2	<4.0	<4.0	0	0
Crystalline	Metadiabase	Metadiabase	0	--	--	0	0
Crystalline	Metagabbro	Metagabbro	0	--	--	0	0
Crystalline	Metarhyolite	Metarhyolite	5	<4.0	<4.0	0	0
Crystalline	Oligoclase-mica schist	Glenarm Wissahickon formation	134	<4.0	4.0	1	0
Crystalline	Oligoclase-mica schist	Wissahickon Formation	28	<4.0	4.0	1	0

Appendix 2. Summary of arsenic concentrations in groundwater (1969–2007) for the 193 geologic units in Pennsylvania with major aquifer type and primary lithology.—Continued

[* indicates geologic unit partially overlain by surficial materials consisting of coarse grained sediments (Soller and Packard, 1998; Lindsey and Bickford, 1999) and that some sample records associated with these geologic units may be categorized with the surficial major aquifer type; µg/L, micrograms per liter; <, less than; --, geologic unit with no sample records]

Major aquifer type	Primary lithology[1]	Geologic unit name[1]	Number of sample records	Median arsenic concentration, in µg/L	Maximum arsenic concentration, in µg/L	Number of sample records with arsenic concentrations greater than or equal to 4.0 µg/L	Number of sample records with arsenic concentrations greater than or equal to 10.0 µg/L
Crystalline	Phyllite	Harpers Formation	1	<4.0	<4.0	0	0
Crystalline	Phyllite	Marburg Schist	16	<4.0	<4.0	0	0
Crystalline	Quartzite	Antietam and Harpers Formations, undivided	24	<4.0	10.0	1	1
Crystalline	Quartzite	Antietam Formation	2	<4.0	<4.0	0	0
Crystalline	Quartzite	Chickies Formation	55	<4.0	40.0	9	3
Crystalline	Quartzite	Hardyston Formation	14	<4.0	20.0	3	3
Crystalline	Quartzite	Montalto Member of Harpers Formation	1	<4.0	<4.0	0	0
Crystalline	Quartzite	Weverton and Loudoun Formations, undivided	0	--	--	0	0
Crystalline	Serpentinite	Ultramafic rocks	27	<4.0	<4.0	0	0
Crystalline	Slate	Peach Bottom Slate and Cardiff Conglomerate, undivided	2	<4.0	10.0	1	1
Siliciclastic	Argillaceous sandstone	Rockwell Formation	1	<4.0	<4.0	0	0
Siliciclastic	Argillaceous shale	Girard Shale*	41	<4.0	293.0	16	4
Siliciclastic	Argillite	Lockatong Formation	54	<4.0	40.0	19	4
Siliciclastic	Arkosic sandstone	New Oxford Formation	45	<4.0	85.5	15	10
Siliciclastic	Arkosic sandstone	Sedimentary strata at Jacksonwald and Aspers	0	--	--	0	0
Siliciclastic	Arkosic sandstone	Stockton Formation	141	<4.0	69.0	23	4
Siliciclastic	Black shale	Marcellus Formation*	4	<4.0	7.0	1	0
Siliciclastic	Calcareous sandstone	Decker Formation through Poxono Island Formation, undivided*	2	<4.0	<4.0	0	0
Siliciclastic	Calcareous shale	Elbrook Formation	25	<4.0	<4.0	0	0
Siliciclastic	Calcareous shale	Onondaga and Old Port Formations, undivided*	37	<4.0	20.0	5	2
Siliciclastic	Calcareous shale	Onondaga Formation	1	<4.0	<4.0	0	0
Siliciclastic	Calcareous shale	Wills Creek Formation through Mifflintown Formation, undivided	2	<4.0	<4.0	0	0
Siliciclastic	Calcareous shale	Wills Creek Formation*	36	<4.0	22.0	6	1
Siliciclastic	Graywacke	Graywacke and shale of Martinsburg Formation	22	<4.0	20.0	7	1
Siliciclastic	Graywacke	Graywacke of Hamburg sequence	10	<4.0	<4.0	0	0
Siliciclastic	Mudstone	Brunswick Formation	257	<4.0	60.9	104	43
Siliciclastic	Mudstone	Lock Haven Formation*	197	<4.0	178.0	59	30
Siliciclastic	Mudstone	Sherman Creek Member of Catskill Formation*	47	4.0	49.8	24	12
Siliciclastic	Quartz conglomerate	Gettysburg conglomerate	6	<4.0	<4.0	0	0
Siliciclastic	Quartz conglomerate	Hammer Creek conglomerate	19	<4.0	7.6	4	0
Siliciclastic	Quartz conglomerate	New Oxford conglomerate	3	<4.0	<4.0	0	0
Siliciclastic	Quartz conglomerate	Quartz fanglomerate	10	<4.0	13.6	2	1
Siliciclastic	Quartz conglomerate	Stockton conglomerate	0	--	--	0	0
Siliciclastic	Quartzite	Tuscarora Formation*	1	<4.0	<4.0	0	0
Siliciclastic	Sandstone	Allegheny and Pottsville Formations, undivided*	1	<4.0	<4.0	0	0
Siliciclastic	Sandstone	Allegheny Formation*	233	<4.0	490.0	30	12
Siliciclastic	Sandstone	Bald Eagle Formation	8	<4.0	6.0	2	0
Siliciclastic	Sandstone	Berea Sandstone through Riceville Formation, undivided*	8	<4.0	180.0	2	1
Siliciclastic	Sandstone	Berea Sandstone through Venango Formation, undivided*	17	16.0	58.0	14	10

Appendix 2. Summary of arsenic concentrations in groundwater (1969–2007) for the 193 geologic units in Pennsylvania with major aquifer type and primary lithology.—Continued

[* indicates geologic unit partially overlain by surficial materials consisting of coarse grained sediments (Soller and Packard, 1998; Lindsey and Bickford, 1999) and that some sample records associated with these geologic units may be categorized with the surficial major aquifer type; µg/L, micrograms per liter; <, less than; --, geologic unit with no sample records]

Major aquifer type	Primary lithology[1]	Geologic unit name[1]	Number of sample records	Median arsenic concentration, in µg/L	Maximum arsenic concentration, in µg/L	Number of sample records with arsenic concentrations greater than or equal to 4.0 µg/L	Number of sample records with arsenic concentrations greater than or equal to 10.0 µg/L
Siliciclastic	Sandstone	Berry Run and Sawmill Run Members of Catskill Formation, undivided	0	--	--	0	0
Siliciclastic	Sandstone	Burgoon Sandstone through Cuyahoga Group, undifferentiated*	1	4.0	4.0	1	0
Siliciclastic	Sandstone	Burgoon Sandstone*	19	<4.0	9.3	3	0
Siliciclastic	Sandstone	Catskill Formation*	247	<4.0	188.0	82	37
Siliciclastic	Sandstone	Clarks Ferry Member of Catskill Formation	0	--	--	0	0
Siliciclastic	Sandstone	Corry Sandstone through Riceville Formation, undivided*	28	<4.0	64.4	5	3
Siliciclastic	Sandstone	Duncannon Member of Catskill Formation*	20	<4.0	4.0	1	0
Siliciclastic	Sandstone	Foreknobs Formation	15	6.3	12.6	8	3
Siliciclastic	Sandstone	Greene Formation	7	<4.0	10.0	2	1
Siliciclastic	Sandstone	Hammer Creek Formation	43	<4.0	20.7	8	3
Siliciclastic	Sandstone	Huntley Mountain Formation*	30	<4.0	10.0	14	1
Siliciclastic	Sandstone	Juniata and Bald Eagle Formations, undivided	1	<4.0	<4.0	0	0
Siliciclastic	Sandstone	Juniata Formation	9	<4.0	10.0	3	1
Siliciclastic	Sandstone	Llewellyn Formation*	20	<4.0	11.4	2	1
Siliciclastic	Sandstone	Long Run and Walcksville Members of Catskill Formation, undivided*	43	<4.0	15.0	7	1
Siliciclastic	Sandstone	Long Run Member of Catskill Formation*	7	<4.0	10.0	2	1
Siliciclastic	Sandstone	Lower members of Gatesburg Formation, undivided	2	<4.0	<4.0	0	0
Siliciclastic	Sandstone	Packerton Member of Catskill Formation	0	--	--	0	0
Siliciclastic	Sandstone	Pocono and Rockwell Formations, undivided	0	--	--	0	0
Siliciclastic	Sandstone	Pocono Formation*	5	<4.0	<4.0	0	0
Siliciclastic	Sandstone	Poplar Gap and Packerton Members of Catskill Formation, undivided	18	<4.0	20.0	3	1
Siliciclastic	Sandstone	Poplar Gap Member of Catskill Formation	4	<4.0	10.0	1	1
Siliciclastic	Sandstone	Pottsville Formation*	89	<4.0	210.0	23	11
Siliciclastic	Sandstone	Ridgeley Formation through Coeymans Formation, undivided*	0	--	--	0	0
Siliciclastic	Sandstone	Ridgeley Member of Old Port Formation	0	--	--	0	0
Siliciclastic	Sandstone	Shawangunk Formation*	0	--	--	0	0
Siliciclastic	Sandstone	Shenango Formation through Cuyahoga Group, undivided*	11	<4.0	4.0	1	0
Siliciclastic	Sandstone	Shenango Formation through Oswayo Formation, undivided*	7	<4.0	6.0	1	0
Siliciclastic	Sandstone	Shenango Formation through Riceville Formation, undivided*	13	<4.0	17.0	5	3
Siliciclastic	Sandstone	Shenango Formation*	8	<4.0	11.0	2	2
Siliciclastic	Sandstone	Spechty Kopf Formation*	2	<4.0	<4.0	0	0
Siliciclastic	Sandstone	Towamensing Member of Catskill Formation*	8	<4.0	<4.0	0	0
Siliciclastic	Sandstone	Walcksville and Towamensing Members of Catskill Formation, undivided	2	<4.0	<4.0	0	0
Siliciclastic	Sandstone	Walcksville Member of Catskill Formation*	2	<4.0	4.0	1	0
Siliciclastic	Sandstone	Washington Formation*	14	<4.0	6.0	3	0
Siliciclastic	Sandstone	Waynesburg Formation*	16	<4.0	21.0	4	2
Siliciclastic	Shale	Bloomsburg and Mifflintown Formations, undivided*	37	<4.0	57.0	18	8
Siliciclastic	Shale	Bloomsburg Formation*	2	4.3	4.7	2	0
Siliciclastic	Shale	Casselman Formation*	136	<4.0	9.0	5	0

Appendix 2. Summary of arsenic concentrations in groundwater (1969–2007) for the 193 geologic units in Pennsylvania with major aquifer type and primary lithology.—Continued

[* indicates geologic unit partially overlain by surficial materials consisting of coarse grained sediments (Soller and Packard, 1998; Lindsey and Bickford, 1999) and that some sample records associated with these geologic units may be categorized with the surficial major aquifer type; µg/L, micrograms per liter; <, less than; --, geologic unit with no sample records]

Major aquifer type	Primary lithology[1]	Geologic unit name[1]	Number of sample records	Median arsenic concentration, in µg/L	Maximum arsenic concentration, in µg/L	Number of sample records with arsenic concentrations greater than or equal to 4.0 µg/L	Number of sample records with arsenic concentrations greater than or equal to 10.0 µg/L
Siliciclastic	Shale	Clinton Group*	25	<4.0	7.0	4	0
Siliciclastic	Shale	Cocalico Formation	20	<4.0	30.5	5	1
Siliciclastic	Shale	Conemaugh Group	7	<4.0	<4.0	0	0
Siliciclastic	Shale	Glenshaw Formation*	426	<4.0	200.0	20	11
Siliciclastic	Shale	Hamburg sequence rocks	86	<4.0	12.0	8	1
Siliciclastic	Shale	Hamilton Group*	109	<4.0	28.0	18	4
Siliciclastic	Shale	Heidlersburg Member of Gettysburg Formation	6	<4.0	6.0	1	0
Siliciclastic	Shale	Kinzers Formation	9	<4.0	<4.0	0	0
Siliciclastic	Shale	Mahantango Formation*	18	<4.0	4.5	2	0
Siliciclastic	Shale	Martinsburg Formation	107	<4.0	161.0	16	9
Siliciclastic	Shale	Mauch Chunk Formation*	91	<4.0	230.0	16	7
Siliciclastic	Shale	Northeast Shale*	3	<4.0	<4.0	0	0
Siliciclastic	Shale	Onondaga Formation through Poxono Island Formation, undivided	0	--	--	0	0
Siliciclastic	Shale	Reedsville Formation	21	<4.0	4.2	2	0
Siliciclastic	Shale	Shale and graywacke of Hamburg sequence	31	<4.0	20.0	4	2
Siliciclastic	Shale	Waynesboro Formation	5	<4.0	<4.0	0	0
Siliciclastic	Siliceous sandstone	Buttermilk Falls Limestone through Esopus Formation, undivided*	1	<4.0	<4.0	0	0
Siliciclastic	Siltstone	Beaverdam Run Member of Catskill Formation*	2	4.5	5.0	2	0
Siliciclastic	Siltstone	Brallier and Harrell Formations, undivided*	15	<4.0	23.0	3	1
Siliciclastic	Siltstone	Buddys Run Member of Catskill Formation*	11	<4.0	19.0	4	1
Siliciclastic	Siltstone	Chadakoin Formation*	129	<4.0	230.0	49	32
Siliciclastic	Siltstone	Cuyahoga Group*	8	<4.0	21.0	3	1
Siliciclastic	Siltstone	Irish Valley Member of Catskill Formation*	20	<4.0	85.0	5	2
Siliciclastic	Siltstone	Scherr Formation	8	<4.0	5.0	2	0
Siliciclastic	Siltstone	Trimmers Rock Formation*	72	<4.0	7.0	6	0
Siliciclastic	Siltstone	Venango Formation*	117	<4.0	68.0	31	19
Siliciclastic	Silty mudstone	Gettysburg Formation	47	<4.0	14.0	14	3
Surficial	Feldspathic quartz sand	Pensauken and Bridgeton Formations, undifferentiated	3	<4.0	11.0	1	1
Surficial	Ferruginous clay	Patapsco(?) Formation	0	--	--	0	0
Surficial	Gravelly sand	Bryn Mawr Formation	0	--	--	0	0
Surficial	Gravelly sand	Trenton Gravel	38	<4.0	50.0	16	9
Surficial	Sand	Sands of Presque Isle	27	103.0	216.0	27	27

[1] Data from Pennsylvania Geological Survey, 2001.

Appendix 3. Results of univariate logistic regression analyses with logistic regression standardized coefficients and individual p-values of independent variables related to the detection of elevated concentrations of arsenic in groundwater samples collected statewide and in three regions in Pennsylvania.

[NS, variable not selected for multivariate analysis; --, variable not assessed for this study area; <, less than; values not enclosed in parentheses are logistic regression standardized coefficients; values enclosed in parentheses are p-values; m, meter; km², square kilometer; %, percent; ºF, degrees Fahrenheit; µg/L, micrograms per liter; µS, microsiemens; cm, centimeter; µS/cm, microsiemens per centimeter at 25 degrees Celsius; nT, nanoteslas; g, grams; cm³, cubic centimeter; mm/yr, millimeters per year; hr, hour]

Explanatory variable	Description	Univariate logistic regression results			
		Statewide	Glacial aquifer system	Gettysburg Basin	Newark Basin
	Anthropogenic factors				
MUN	Distance to municipalities containing water suppliers with arsenic treatment and sites or groundwater in need of remediation because of arsenic contamination (m)	-0.0505 (0.0160)	-0.1297 (0.0755)	0.2747 (0.0057)	-0.2951 (<0.0001)
TRI	Distance to toxic chemical release inventory sites (m)	NS	-0.3107 (0.0006)	-0.2893 (0.0100)	NS
COAL	Distance to underground and surface coal mining operations (m)	--	NS	--	--
INDMIN	Distance to industrial mineral mining operations (m)	-0.0540 (0.0106)	0.3967 (<0.0001)	NS	NS
OILGAS	Distance to oil and gas locations (m)	--	-0.3066 (0.0068)	--	--
AGAVE	Average agricultural land cover within a 500-m radius (%)	-0.1037 (<0.0001)	-0.2530 (0.0006)	-0.1502 (0.1332)	0.1044 (0.0619)
POP	Population density (people/km²)	NS	NS	0.1873 (0.0365)	NS
URBAVE	Average urban land cover within a 500-m radius (%)	0.0571 (0.0028)	NS	0.4095 (<0.0001)	NS
	Natural factors				
PRE	Average precipitation from 1971–2000 (mm/yr)	-0.1073 (<0.0001)	NS	-0.1905 (0.0619)	NS
RECH	Average groundwater recharge from 1951–1980 (mm/yr)	-0.1292 (<0.0001)	NS	NS	NS
TEMP	Average temperature from 1971–2000 (ºF)	-0.1066 (<0.0001)	0.2391 (0.0006)	0.2950 (0.0061)	-0.1130 (0.0423)
ACID	Distance to geologic units containing acid-producing sulfide minerals (m)	0.1041 (<0.0001)	0.4306 (<0.0001)	--	0.1099 (0.0451)
CARB	Carbonate major aquifer type (1 = carbonate, 0 = all other)	-0.1647 (<0.0001)	--	--	--
CRYST	Crystalline major aquifer type (1 = crystalline, 0 = all other)	-0.2768 (<0.0001)	--	--	--
DIA	Distance to diabase geologic units (m)	--	--	0.2645 (0.0039)	-0.1460 (0.0133)
PLUTIG	Distance to igneous rock locations (m)	--	--	NS	-0.1999 (0.0009)
SIL	Siliciclastic major aquifer type (1 = siliciclastic, 0 = all other)	0.1344 (<0.0001)	--	--	--

Appendix 3. Results of univariate logistic regression analyses with logistic regression standardized coefficients and individual p-values of independent variables related to the detection of elevated concentrations of arsenic in groundwater samples collected statewide and in three regions in Pennsylvania.—Continued

[NS, variable not selected for multivariate analysis; --, variable not assessed for this study area; <, less than; values not enclosed in parentheses are logistic regression standardized coefficients; values enclosed in parentheses are p-values; m, meter; km², square kilometer; %, percent; ºF, degrees Fahrenheit; µg/L, micrograms per liter; µS, microsiemens; cm, centimeter; µS/cm, microsiemens per centimeter at 25 degrees Celsius; nT, nanoteslas; g, grams; cm³, cubic centimeter; mm/yr, millimeters per year; hr, hour]

Explanatory variable	Description	Univariate logistic regression results			
		Statewide	Glacial aquifer system	Gettysburg Basin	Newark Basin
SURF	Surficial major aquifer type (1 = surficial, 0 = all other)	0.1509 (<0.0001)	--	--	--
MAG	Average residual total intensity of the earth's magnetic field (nT)	NS	-0.1232 (0.0963)	-0.1723 (0.1119)	0.1418 (0.0093)
CORR	Grouping of geologic units according to estimated groundwater corrosivity (1 to 11 based on Langelier Saturation Index)	0.0805 (0.0001)	-0.3546 (<0.0001)	0.2126 (0.0422)	0.2975 (<0.0001)
NUREAS	Estimated arsenic concentration derived from stream sediment data (µg/L)	--	--	--	0.1056 (0.0550)
PHIDW	Estimated groundwater pH	0.1283 (<0.0001)	--	0.3748 (0.0016)	0.3500 (<0.0001)
PHIDWGLAC	Estimated groundwater pH using only glacial data	--	0.2958 (0.0002)	--	--
SPCIDW	Estimated groundwater specific conductance (µS/cm)	0.1752 (<0.0001)	--	0.3619 (0.0006)	0.3624 (0.0037)
SPCIDWGLAC	Estimated groundwater specific conductance using only glacial data (µS/cm)	--	0.5250 (<0.0001)	--	--
FORAVE	Average forested land cover within a 500-m radius (%)	NS	-0.2439 (0.0013)	-0.2359 (0.0337)	-0.0740 (0.1908)
WETAVE	Average wetlands land cover within a 500-m radius (%)	0.1409 (<0.0001)	0.5217 (<0.0001)	NS	0.1171 (0.0414)
METMIN	Distance to metallic mineral resources (m)	0.0700 (0.0003)	-0.2371 (0.0012)	0.1825 (0.0674)	NS
MIN	Distance to metallic and nonmetallic mineral resources (m)	NS	-0.1354 (0.0727)	NS	NS
NONMETMIN	Distance to nonmetallic mineral resources (m)	NS	NS	-0.1620 (0.1335)	NS
AWCAVE	Average soil available water capacity (%)	-0.2133 (<0.0001)	-0.2979 (0.0003)	NS	-0.2028 (0.0015)
BDAVE	Average soil bulk density (g/cm³)	0.1268 (0.0001)	0.5296 (0.0001)	NS	NS
CLAYAVE	Average soil clay content (%)	-0.2421 (<0.0001)	-0.1699 (0.0206)	NS	NS
HGA	Average soil hydrologic soil group A content (%)[1]	0.0784 (<0.0001)	0.1009 (0.1402)	NS	-0.1833 (0.1339)
HGAD	Average soil hydrologic soil group AD content (%)[1]	0.0714 (<0.0001)	0.1180 (0.0721)	--	--

Appendix 3. Results of univariate logistic regression analyses with logistic regression standardized coefficients and individual p-values of independent variables related to the detection of elevated concentrations of arsenic in groundwater samples collected statewide and in three regions in Pennsylvania.—Continued

[NS, variable not selected for multivariate analysis; --, variable not assessed for this study area; <, less than; values not enclosed in parentheses are logistic regression standardized coefficients; values enclosed in parentheses are p-values; m, meter; km^2, square kilometer; %, percent; ºF, degrees Fahrenheit; µg/L, micrograms per liter; µS, microsiemens; cm, centimeter; µS/cm, microsiemens per centimeter at 25 degrees Celsius; nT, nanoteslas; g, grams; cm^3, cubic centimeter; mm/yr, millimeters per year; hr, hour]

Explanatory variable	Description	Univariate logistic regression results			
		Statewide	Glacial aquifer system	Gettysburg Basin	Newark Basin
HGB	Average soil hydrologic soil group B content (%)[1]	-0.1696 (<0.0001)	0.2881 (<0.0001)	NS	-0.0777 (0.1582)
HGBC	Average soil hydrologic soil group BC content (%)[1]	NS	--	--	--
HGBD	Average soil hydrologic soil group BD content (%)[1]	0.1021 (<0.0001)	-0.2740 (0.0010)	NS	NS
HGC	Average soil hydrologic soil group C content (%)[1]	0.0724 (0.0003)	-0.1702 (0.0195)	0.4309 (0.0166)	0.1114 (0.0554)
HGCD	Average soil hydrologic soil group CD content (%)[1]	0.0940 (<0.0001)	NS	--	--
HGD	Average soil hydrologic soil group D content (%)[1]	NS	NS	-0.2575 (0.0638)	NS
OMAVE	Average soil organic matter content (% by weight)	0.0726 (<0.0001)	-0.2107 (0.0068)	-0.0705 (0.4501)	NS
PERMAVE	Average soil permeability (cm/hr)	0.1595 (<0.0001)	0.3858 (<0.0001)	0.2934 (0.0305)	NS
ROCKDEPAVE	Average soil thickness (cm)	-0.0373 (0.0633)	0.3020 (0.0056)	NS	-0.1159 (0.0619)
SANDAVE	Average soil sand content (%)	0.2124 (<0.0001)	0.2967 (0.0002)	0.8370 (0.0118)	NS
SILTAVE	Average soil silt content (%)	-0.0942 (<0.0001)	-0.2157 (0.0073)	NS	NS
SLOPEAVE	Average soil land-surface slope (%)	-0.1324 (<0.0001)	-0.2921 (0.0009)	NS	NS
ELEV	Land-surface elevation above sea level (m, North American Vertical Datum of 1988)	NS	-0.1933 (0.0059)	-0.3632 (0.0016)	0.0843 (0.1245)
SINK	Distance to nearest sinkhole (m)	--	--	NS	NS
SLOPE	Slope derived from elevation (%)	-0.0512 (0.0182)	-0.2902 (0.0010)	-0.1555 (0.1817)	NS
STREAMD	Average stream density within a 500-m radius (km^2)	NS	-0.2598 (0.0008)	0.1510 (0.1155)	0.1137 (0.0378)

---

[1] Hydrologic soil groups are described as Group A, Group B, Group C, and Group D, in order from low to high runoff potential, with dual hydrologic soil groups (AD, BC, CD) assigned on the basis of drained and undrained conditions.

Prepared by the West Trenton Publishing Service Center

For additional information, write to:

Director
U.S. Geological Survey
Pennsylvania Water Science Center
215 Limekiln Road
New Cumberland, Pa. 17070

or visit our Web site at:
http://pa.water.usgs.gov/

Gross and Low—**Arsenic Concentrations, Environmental Factors, and the Probability of Elevated Arsenic in Groundwater in Pa.**—Scientific Investigations Report 2012–5257