USGS
science for a changing world

Prepared in cooperation with the New Hampshire Department of Health and Human Services and the New Hampshire Department of Environmental Services

Evaluating Associations Between Environmental Variables and *Escherichia Coli* Levels for Predictive Modeling at Pawtuckaway Beach in Nottingham, New Hampshire, From 2015 to 2017

Scientific Investigations Report 2019–5111

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

Cover. Meteorological monitoring station on the beach at Pawtuckaway State Park in Nottingham, New Hampshire; photograph by J. Coles, U.S. Geological Survey.

Evaluating Associations Between Environmental Variables and *Escherichia Coli* Levels for Predictive Modeling at Pawtuckaway Beach in Nottingham, New Hampshire, From 2015 to 2017

By James F. Coles and Kathleen F. Bush

Prepared in cooperation with the New Hampshire Department of Health and Human Services and the New Hampshire Department of Environmental Services

Scientific Investigations Report 2019–5111

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

U.S. Department of the Interior
DAVID BERNHARDT, Secretary

U.S. Geological Survey
James F. Reilly II, Director

U.S. Geological Survey, Reston, Virginia: 2019

Suggested citation:
Coles, J.F., and Bush, K.F., 2019, Evaluating associations between environmental variables and *Escherichia coli* levels for predictive modeling at Pawtuckaway Beach in Nottingham, New Hampshire, from 2015 to 2017: U.S. Geological Survey Scientific Investigations Report 2019–5111, 28 p., https://doi.org/10.3133/sir20195111.

Associated data for this publication:
Coles, J.F., and Bush, K.F., 2019, Data collected at Pawtuckaway Beach in Nottingham, New Hampshire, 2015–2017, including data from *Escherichia coli* (bacteria) samples, and from USGS meteorological and water quality stations: U.S. Geological Survey data release, https://doi.org/10.5066/P9KUIT3W.

ISSN 2328-0328 (online)

Acknowledgments

Amanda McQuaid, David Neils, and Sonya Carlson of the New Hampshire Department of Environmental Services supported sample collection and analysis. Bethany Poulin, formerly with the New Hampshire Environmental Public Health Tracking Program (EPHT), and Lynne Clement of the EPHT supported project management, data analysis, and model development. In addition, EPHT interns Emily Edwards and Kaylee Jackson were instrumental in collecting water samples.

The authors extend special thanks to colleagues who contributed to the success of this project. Richard Kiah, Sanborn Ward, Ian Carlisle, and Joseph Levitt of the U.S. Geological Survey in the New England Water Science Center established the water-quality and meteorological stations at Pawtuckaway Beach, reviewed the data for quality, and ensured the information was made available through the U.S. Geological Survey National Water Information System to provide the public with current conditions at the beach.

Contents

Figures

Tables

Conversion Factors

International System of Units to U.S. customary units

Multiply	By	To obtain
centimeter (cm)	0.3937	inch (in.)
meter (m)	3.281	foot (ft)
kilometer (km)	0.6214	mile (mi)
hectare (ha)	2.471	acre
liter (L)	0.2642	gallon (gal)

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

$$°F = (1.8 \times °C) + 32.$$

Datum

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).

Supplemental Information

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 in milligrams per liter (mg/L).

Abbreviations

CFU	colony forming unit
E. coli	*Escherichia coli*
EPA	U.S. Environmental Protection Agency
EPHT	Environmental Public Health Tracking
MPN	most probable number
NHDES	New Hampshire Department of Environmental Services
NHDHHS	New Hampshire Department of Health and Human Services
NWIS	National Water Information System
USGS	U.S. Geological Survey

Evaluating Associations Between Environmental Variables and *Escherichia Coli* Levels for Predictive Modeling at Pawtuckaway Beach in Nottingham, New Hampshire, From 2015 to 2017

By James F. Coles[1] and Kathleen F. Bush[2]

Abstract

From 2015 through 2017, the U.S. Geological Survey in cooperation with the New Hampshire Department of Health and Human Services and the New Hampshire Department of Environmental Services studied occurrences of high levels of *Escherichia coli* (*E. coli*) bacteria at the Pawtuckaway State Park Beach in Nottingham, New Hampshire. Historic data collected by the New Hampshire Department of Environmental Services indicated that *E. coli* concentrations in the water typically increased through the beach season to levels considered potentially harmful to beachgoers. During the three beach seasons that were studied, *E. coli* samples were collected three to four times per week, and water-quality and meteorological data were collected continuously. The Virtual Beach software was used to generate a predictive model for each year of the study (2015–2017), and the model for each of these years was tested with data from the other two. Additionally, data from all study years were combined to generate a comprehensive model to help identify independent variables that might characterize environmental conditions relative to *E. coli* levels during multiple seasons. The accuracy of the models in predicting the occurrence of high *E. coli* levels was marginal, but the models did provide insights into the likely mechanisms for increased *E. coli* levels during the seasons. Variables most important in explaining high *E. coli* levels were the presence of geese at the beach, the progression of the season, the number of visitors at the beach, and wind vectors relative to beach orientation.

Introduction

Pawtuckaway State Park in Nottingham, New Hampshire, is a 2,200-hectare multiuse recreational facility administered by the New Hampshire Division of Parks and Recreation that was developed in 1966 from land acquired in 1923 (fig. 1; Crawford, 1967). A prominent feature of the park is a 300-hectare lake that was created during the 19th century by the construction of four dams on a small natural pond, Pawtuckaway Pond. Although Pawtuckaway Pond remains the official name of the lake, Pawtuckaway Lake is the name used more commonly, and thus is the name used in this report. The majority of the outflow from the lake forms the Pawtuckaway River, which flows 5.8 kilometers (km) to the Lamprey River and contributes to the Great Bay tidal estuary in eastern New Hampshire. The watershed of Pawtuckaway Lake is primarily forested with some rural and residential development; the lake is classified as a mesotrophic warm-water fishery, with an average depth of about 3 meters (m) and a maximum depth of about 15 m.

A popular feature of the park is the Pawtuckaway State Park Beach (fig. 1). The 170-m-long beach with a general north-south orientation on Pawtuckaway Lake is visited by beachgoers primarily during the summer season (June to August). Attendance is especially high on weekends and favorably warm days, and levels of *Escherichia coli* (*E. coli*) bacteria in the swimming area are often elevated on those days. Bacteria levels are monitored at the beach by the New Hampshire Department of Environmental Services (NHDES); when elevated levels occur, beach advisories are posted to warn of potential health risks associated with wading and swimming in potentially contaminated water. Of the approximately 170 freshwater beaches monitored by the NHDES, Pawtuckaway Beach has been among the highest in number of reported advisories; in annual beach inspection reports, the NHDES reported "clean" water samples at Pawtuckaway 59 percent of the time in 2012, 55 percent of the time in 2013, and 52 percent of the time in 2014 (New Hampshire Department of Environmental Services, 2017).

The factors driving the high levels of *E. coli* at Pawtuckaway Beach during the summer have been unclear. Although the beach received a high number of beachgoers at times during the season, the facility has modern amenities, such as a brick and mortar bathhouse with sanitary restrooms, so sewage discharge into the water was considered unlikely. However, a

[1] U.S. Geological Survey.

[2] New Hampshire Department of Health and Human Services.

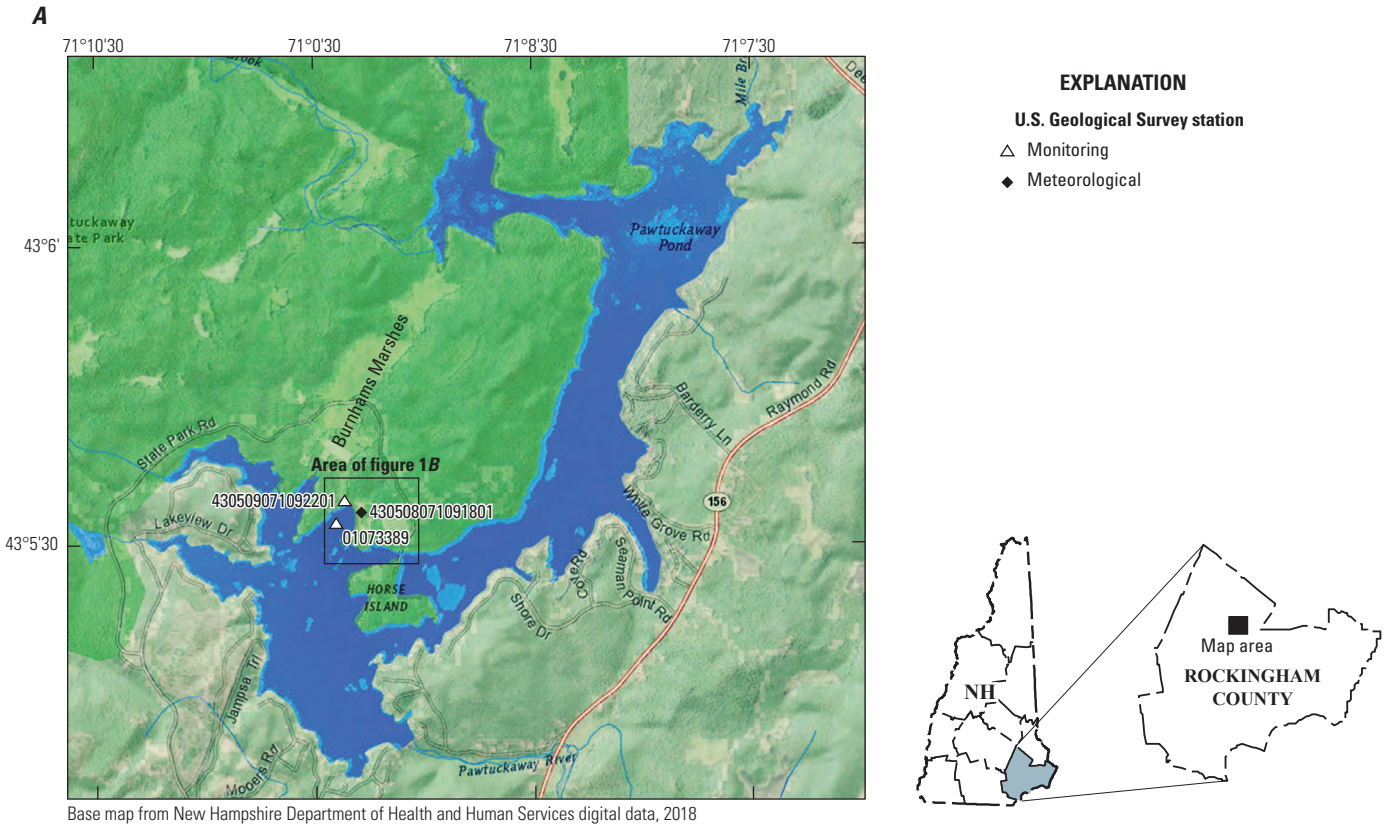

Figure 1. Location of *A*, Pawtuckaway Lake and Pawtuckaway State Park and *B*, Pawtuckaway State Park Beach in Nottingham, New Hampshire, and *C*, meteorological monitoring station on the beach (foreground) and hydrologic monitoring station deployed on a buoy (background) on the lake.

population of *Branta canadensis* (Canadian geese) migrates to the area in and around the beach annually during early summer to nest and raise goslings, then migrates from the area in the fall. The habitation of the goose population at the beach overlaps with summer vacation, when human activity is greatest. Goose droppings have been observed along the beach in high densities at times, and it is highly probable that the geese also defecate in the water; to the extent practical, park staff routinely clear droppings from the beach each morning to improve beach aesthetics and reduce the potential for *E. coli* contamination. The concurrent presence of people and geese is a relevant factor in explaining high bacteria levels at beaches (Kirschner and others, 2004), and was deemed potentially important for Pawtuckaway Beach. In addition, water quality and weather-related factors have helped explain high bacteria counts in other public beaches in the eastern United States (Francy and Darner, 1998), so these variables were also considered in relation to bacteria levels at Pawtuckaway Beach. To evaluate the factors that may affect *E. coli* levels, the NHDES and the New Hampshire Department of Health and Human Services (NHDHHS) began a collaboration in 2014 with the U.S. Geological Survey (USGS) to collect environmental data at Pawtuckaway Beach and assist the NHDHHS in developing models to predict why and when *E. coli* levels are high.

NHDES Beach Inspection Program

In 1993, the NHDES began administering the Beach Inspection Program to monitor levels of bacteria at beaches in New Hampshire (New Hampshire Department of Environmental Services, 2017). Approximately 170 beaches are currently [2019] monitored, beginning around Memorial Day and continuing through Labor Day. Water samples for *E. coli* analysis are collected at least monthly, with some beaches having a more frequent sampling schedule; from 2015 to 2017, Pawtuckaway Beach was sampled about three to four times per week, which is the highest frequency for beaches across the State. The procedure used to collect water samples for analysis depends on beach size. For beaches less than 30 m in length, two separate water samples are collected: one from the left side of the beach and the other from the right side of the beach. For longer beaches, water samples are collected from the left and right sides, and a third from the center of the beach. However, alternative sampling procedures have been implemented when sampling frequency is high, as in the case of Pawtuckaway Beach, which helps reduce analysis costs: combine the three water samples collected across the beach into a single composite or, in fewer cases, collect a single water sample from the middle location.

A beach advisory is triggered when bacteria levels in the beach water samples exceed either of two regulatory thresholds: a level of more than 158 colony forming units per 100 milliliters (CFU/100 mL) in any one sample or 88 CFU/100 mL in any two samples. For the composite samples collected at Pawtuckaway Beach, the level of 158 CFU/100 mL was used to trigger an advisory. The NHDES does not close beaches when high concentrations of bacteria are detected; an advisory notice is physically posted on the beach to inform swimmers of the potential risks associated with high bacteria levels (fig. 2). Even though the NHDES does not usually close beaches, the park manager has that prerogative, which has only been exercised infrequently. The high frequency of beach advisories issued at Pawtuckaway Beach compared with other beaches in New Hampshire in recent years was the main incentive for investigating the causes of high *E. coli* levels at the beach.

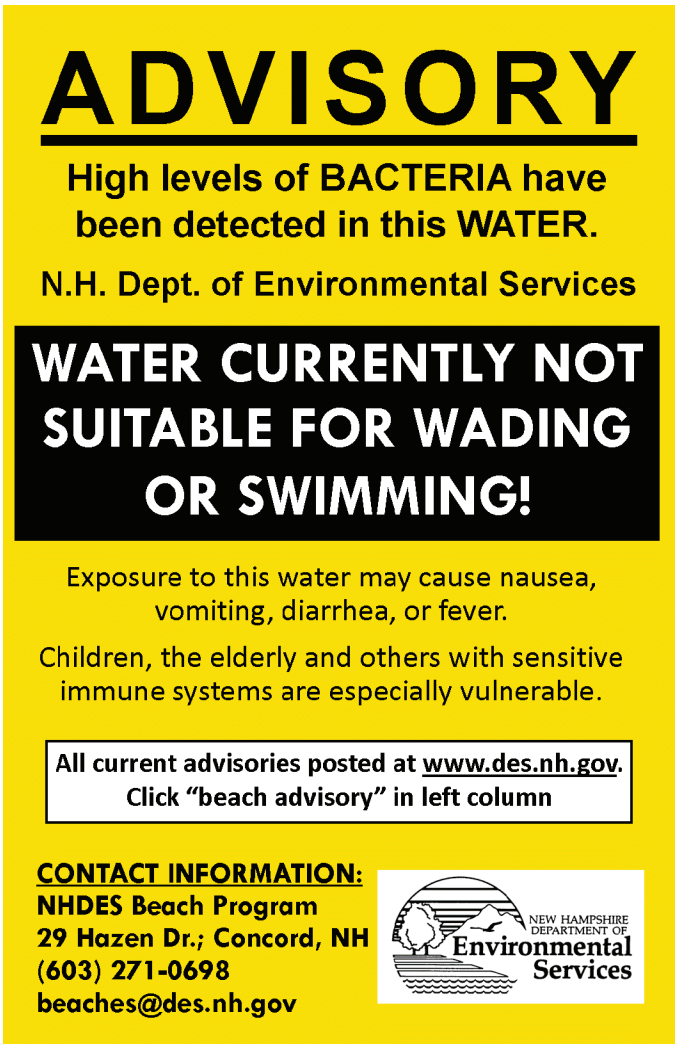

Figure 2. Beach advisory sign, as posted at beaches by the New Hampshire Department of Environmental Services (NHDES) when levels of bacteria are considered unsafe for human health; image courtesy of the NHDES.

Predicting Conditions at Recreational Beaches

In 1997, the USGS started a collaborative project with Federal, State, and local agencies in Ohio to identify environmental factors affecting bacteria levels at public beaches (Francy and Darner, 1998). The work consists of four main components: real-time assessments of water quality; coastal processes; pathogens and source tracking; and data analysis, interpretation, and communication. More recently, the focus has been on the development of predictive models that are built on data from these components; these predictive models relay information to a near real-time information system called NowCast, which predicts water-quality conditions primarily at Great Lake beaches (U.S. Geological Survey, 2018). The NowCast system originated as a water-quality monitoring effort at three Lake Erie beaches in the Cleveland, Ohio, metropolitan area, but has been expanded to include predictive modeling at 45 beaches throughout the Great Lakes region in Illinois, Indiana, Michigan, New York, Ohio, Pennsylvania, and Wisconsin (U.S. Geological Survey, 2010, 2013).

The NowCast system is designed to predict when high levels of *E. coli* are likely to occur at beaches by using models that assimilate environmental data that characterize recent and current conditions at the beaches. The predictive models used in the NowCast system are typically built with the Virtual Beach software program (app. 1; Cyterski and others, 2014), which was specifically developed to help facilitate beach-monitoring programs where high pathogen levels have been reported. Model-based predictions are used to inform the public of potential health risks that are more likely to happen at a beach when the environmental conditions are favorable for high bacteria levels.

The most robust models generated with Virtual Beach for Great Lakes beaches were based on unique combinations of explanatory variables that most commonly included, turbidity, day of the year, wave height, wind direction and speed, antecedent rainfall for various time periods, and changes in lake levels during a 24-hour period (Francy and others, 2013a). In addition to extensive modeling efforts for Great Lake beaches, the USGS investigated the use of predictive modeling on eight inland recreational lakes in Ohio (Francy and others, 2013b). The inland beach models generally were less reliable than the Great Lake models at predicting occurrences of high bacteria levels. Nevertheless, the most relevant explanatory variables used in the inland beach models were rainfall, wind direction and speed, turbidity, and water temperature.

In cases where modeling efforts have been highly successful at predicting when levels of fecal bacteria would rise, the underlying factors were often associated with weather events, especially rainfall (Francy and Darner, 2006; Sampson and others, 2006; Kleinheinz and others, 2009). For example, beaches in densely populated urban areas, such as cities and towns along the Great Lakes, can be affected by *E. coli* contaminated storm water (for example, from animal sources) or effluent from faulty sewer systems; an increase in levels of *E. coli* might be expected after heavy rainfall and when certain wind and current conditions convey the contaminated water to the beach area (Francy and Darner, 1998). Under such circumstances, a predictive model developed for a specific beach could effectively forecast when high levels of bacteria were expected.

Environmental Public Health Tracking Program for New Hampshire

Because the protection of the health of New Hampshire citizens is one of the NHDHHS primary objectives (New Hampshire Department of Health and Human Services, 2016), the high number of beach advisories at Pawtuckaway Beach led the NHDHHS to explore the correlation of environmental conditions and high bacteria levels and the use of environmental data in statistical models to predict when high bacteria levels would likely occur. Through the Environmental Public Health Tracking Program (EPHT Program) for New Hampshire, which is funded by the Centers for Disease Control and Prevention, the NHDHHS monitors and distributes information about environmental hazards and potential health effects related to their exposure (Wall, 2015). Through the EPHT Program, the NHDHHS initiated a study to identify environmental factors related to high *E. coli* levels at Pawtuckaway Beach with assistance from the NHDES and the USGS, which were recruited to collect bacteria and environmental data, respectively. A collaborative agreement was established among these agencies to collect and use the data to generate models predicting the likelihood of high *E. coli* levels at Pawtuckaway Beach and create a web-based portal to inform beach managers and the public of current conditions at the beach.

Objectives and Approach

The study was designed to be implemented in two successive phases in order to address separate and distinct objectives. For phase I, the USGS would provide the NHDHHS near real-time hydrologic and meteorological data collected at Pawtuckaway Beach to help explain and predict high bacteria counts at the beach; the USGS would evaluate and assimilate the data with other beach-related information to be used in developing statistical models that relate environmental conditions with high bacteria levels. Phase II entailed development of a near real-time surveillance tool to provide information about beach conditions to the public and that uses predictive models to determine the likelihood of high beach bacteria levels. The USGS was responsible for meeting the objectives in phase I; the objectives of phase II would be accomplished in conjunction with the NHDES and the EPHT Program.

The USGS addressed phase I objectives by deploying a buoy at Pawtuckaway Lake to collect hydrologic data near the beach and establishing a meteorological station on the beach to collect weather data. Data were recorded

at 15-minute intervals from June through September from 2015 to 2017 and transmitted in near real-time to the USGS National Water Information System (NWIS) database where they were reviewed for quality assurance and made available to the public as real-time information on NWIS (https://waterdata.usgs.gov/nwis). Concurrently, the NHDES collected water samples at the beach three to four times per week during the summer to determine *E. coli* levels and recorded observations of current conditions during the visit that could be relevant to water quality at the beach.

Phase II objectives to analyze the respective datasets from phase I were met through a collaborative effort among the USGS, the NHDES, and the NHDHHS. The datasets were integrated and analyzed using the Virtual Beach software program to explore relations between environmental conditions and bacteria levels and to test models to predict when high bacterial levels might occur at Pawtuckaway Beach. In addition, a beta version of a surveillance tool was developed to provide the public with near real-time conditions at the beach, based on the USGS data available from NWIS (U.S. Geological Survey, 2017a), and the likelihood of high bacteria levels at the beach, based on predictive models developed from the data (Coles and Bush, 2019).

Methods

Environmental and bacteria data collection efforts were coordinated among the USGS, the NHDES, and the NHDHHS at Pawtuckaway Beach for the summer season from 2015 to 2017, which typically occurs from mid-June through mid-September. The USGS had previously collected environmental data at Pawtuckaway Beach from August 12 through September 25, 2014, but this was late in the beach season, and levels of *E. coli* sampled at that time were below the 158 CFU/100 mL exceedance threshold that would trigger an advisory. However, the 2014 data as well as other data collected by the NHDES at Pawtuckaway Beach before this study were useful for identifying patterns in *E. coli* levels that occurred either within a typical beach season or among different years.

Data Collection Procedures

The NHDES has collected water samples for *E. coli* at Pawtuckaway Beach since 1993 through the Beach Inspection Program; the samples are analyzed at the New Hampshire State Water Analysis Laboratory. Before 2016, *E. coli* concentrations were measured using membrane filtration technique (standard method 9222D; U.S. Environmental Protection Agency, 2010); beginning in 2016, concentrations were measured using the colilert-18 method (Warden and others, 2011). Before 2015, water samples for *E. coli* analysis were collected approximately weekly during the beach season for most years; whenever levels exceeded the State standard of 158 CFU/100 mL, samples usually were collected daily until the State standard was met. Also before 2015, sampling typically consisted of collecting separate water grab samples from the left, center, and right areas of the beach, and these were either combined into one composite sample or analyzed separately to provide three *E. coli* values for that day. As part of the quality-assurance procedures implemented by the laboratory, blank and duplicate samples were prepared in the field and submitted to the laboratory for analysis. Blank samples were prepared from deionized water during each sampling visit, and duplicate samples were collected at a frequency of 10 percent of the routine samples. The relative percent difference between the duplicate samples was limited to not more than 75 percent; if this value was exceeded, new samples were collected immediately (Carlson, 2012).

Beginning in 2015, the sampling protocol for Pawtuckaway Beach was changed so that *E. coli* was sampled approximately four times per week during the beach season. As done previously, water samples were collected at the three locations along the beach, but now were combined into a composite sample in the laboratory to create a single *E. coli* sample to adjust for the increased efforts and cost of additional weekly samples. For the 2017 beach season, samples generally were collected only at the center location of the beach, but on some site visits, samples were collected as was done before 2015 at the left, center, and right locations of the beach and were analyzed separately. Site visits were generally made during the morning so that any potential diel variation in *E. coli* would be less of a confounding factor in the analysis. In addition, the number of beachgoers and number of birds on the beach were recorded when the sample was collected because their presence could be a contributing factor in the bacteria levels.

Hydrologic and meteorological data were collected seasonally from 2014 to 2017 at Pawtuckaway State Park by the USGS by deploying a hydrologic monitoring buoy in the lake and establishing a meteorological station on the beach (fig. 1). These instruments became fully operational in transmitting coincident data to NWIS by August 12, 2014 (U.S. Geological Survey, 2017a, b, c). Because the USGS collected data only during the latter part of the 2014 season, during which *E. coli* samples did not exceed the State standard, data from the 2015 sampling season were used to develop the initial predictive models (Coles and Bush, 2019) in Virtual Beach. However, deployment of the USGS stations during 2014 was valuable for establishing data transfer protocols and testing system reliability of the instruments.

The USGS installed two monitoring stations to record data at Pawtuckaway Lake. One station (01073389) recorded hydrologic data, which included air and water temperature, specific conductance, dissolved oxygen (DO), pH, and stage (water level) at the beach. The second station (430508071091801) recorded meteorological data, which included wind speed, wind direction, precipitation, barometric pressure, and relative humidity. In addition, the stage at Burnhams Marshes, a wetland within Pawtuckaway State Park that discharges into the lake about 50 m from the northern

end of the beach (fig. 1), was monitored at USGS station 430509071092201. A rising stage at this station indicated when water flowed into the beach area from the wetland; these periods were identified and evaluated for possible effects of marsh water on *E. coli* levels at the beach. All data collected by the USGS were transmitted directly to the NWIS database (U.S. Geological Survey, 2017a, b, c).

Evaluation of Environmental Datasets

Environmental data that characterized beach conditions were collected by the USGS and were routinely downloaded from NWIS (U.S. Geological Survey, 2017a, b, c). The *E. coli* data were collected by the NHDES (New Hampshire Department of Environmental Services, 2018). For each of the 3 years, the USGS and NHDES data were merged in a single file that ordered environmental data in sequential rows by their 15-minute date and time stamps; each *E. coli* value was assigned to a field in the 15-minute interval that most closely corresponded to, but was also antecedent to, the date and time when the *E. coli* sample was collected.

In building datasets for the models (Coles and Bush, 2019), independent variables were derived from the environmental data as a series of time-interval characterizations that represented environmental conditions relative to when *E. coli* samples were collected. For all independent variables except precipitation, the time-interval characterizations represented the median values for the 3-, 6-, 12-, and 24-hour periods before collecting the *E. coli* sample, the 12-hour period from 08:00 to 20:00 the day before the sample was collected, and the 08:00 to 20:00 period on the day the sample was collected; independent variables for these latter two time periods represented daytime beach conditions the prior day and the same day the *E. coli* sample was collected, respectively. For precipitation, the sum (rather than median) for these same time periods was used. In addition, representations of precipitation included totals for the 24-hour periods that ended exactly 2 and 3 days prior to when the *E. coli* sample was collected, and weighted total precipitation, over the entire 2- and 3-day periods prior to when the *E. coli* sample was collected (Francy and others, 2013a). The time-interval characterizations are denoted in this report as the variables *3h, 6h, 12h, 24h*, *prior_day*, and *same_day*, as defined in this paragraph.

Independent variables representing wind vectors were calculated from wind speed and wind direction to characterize the wind-force component relative to the alongshore (Wind_A) and onshore/offshore (Wind_O) orientation of the beach. The formulae used to derive these independent variables are as follows:

$$Wind\_A = -S \times \cos\frac{(D-B)\times\pi}{180} \text{ and} \quad (1)$$

$$Wind\_O = -S \times \sin\frac{(D-B)\times\pi}{180}, \quad (2)$$

where

S is wind velocity,
D is wind direction in compass degrees,
B is the beach orientation (165 degrees for Pawtuckaway), and
π is the mathematical constant equal to 3.1416.

The Virtual Beach program was used to assess potential associations between logarithmically ($\log_{10}$; referred to as "log" in this report) transformed *E. coli* values and the independent variables for each of the three yearly datasets (app. 1). Pearson correlation coefficients (r), statistical probability values (p), and a series of scatterplots were the initial output products of Virtual Beach that were used to help select preliminary subsets of independent variables potentially related to high *E. coli* levels. Independent variables with absolute r ($|r|$) greater than or equal to 0.365 and with p less than 0.05 were considered to be significant. When an independent variable had several time-interval characterizations with significant correlations, the approach used for selecting candidates for additional analysis was to determine the time interval that was most consistent with other significant independent variables in the dataset.

As a complement to evaluating relations based on correlation coefficients, scatterplots were used to visually assess patterns in the data. For example, a correlation between *E. coli* levels and an independent variable might be relatively weak, but the relation could still be important if the scatterplot revealed a fit between the variables when *E. coli* levels were high. In such a case, *E. coli* might only show an increase over a segment of the independent variable's range, but this type of response can be useful in predicting when *E. coli* might increase to levels that pose a health risk. Scatterplots also were used to indicate when multiple independent variables varied consistently and sequentially during the beach season. Identifying and evaluating such patterns are important so that covariance is not mistakenly interpreted as a cause and effect relation between independent and dependent variables. Using results mainly from the 2015 beach season, examples are provided that describe covariance and how certain independent variables vary with daily progression of the beach season.

Generating and Evaluating Models

The environmental data (USGS) and *E. coli* data (NHDES) collected during the 2015 season were initially used by the NHDHHS to develop four preliminary models with Virtual Beach to predict *E. coli* exceedances at Pawtuckaway Beach. In generating these models in Virtual Beach, means of all daily values (usually recorded at 15-minute intervals) were used to characterize independent variables, and the multiple linear regression option was used because models derived by this method were deemed intuitive in how they function and could be expressed in basic mathematical terms in web-based

applications. During the 2016 and 2017 beach seasons, the near real-time environmental data provided by the USGS were used by the NHDHHS to develop a beta version of a near real-time decision-making tool that could be used to inform the public of conditions at the beach, including water quality, weather conditions, and predicted *E. coli* levels calculated with the four preliminary models. The accuracy and precision of the models were evaluated by the NHDHHS by comparing the predicted *E. coli* values with the observed *E. coli* values that were measured routinely three or four times per week for this study.

Subsequent models generated with Virtual Beach were developed to evaluate the effects of different environmental conditions at Pawtuckaway Beach from one year to another, incorporate park attendance data not previously available, and combine multiyear data to determine if certain independent variables were more strongly correlated with *E. coli* levels when more observations were available for analysis. All resulting models were evaluated at two levels: the initial "training" dataset was used to generate models based on best fit (reliability based only on initial dataset), and additional "testing" data from other years were used to assess the performance of the selected models. In building models, collinearity among independent variables was limited to 80 percent, which is the default value assigned by Virtual Beach for the variance inflation factor, and model fitness was assessed by the predicted residual error sum of squares (PRESS) statistic. Model performance was based on values of specificity (proportion of negatives correctly predicted), sensitivity (proportion of positives correctly predicted), and accuracy (proportion of correct predictions). The reliability of models (potential to make accurate predictions with new data) was tested with the cross-validation function in Virtual Beach and evaluated by the mean squared error of prediction (MSEP) scores using 1,000 trials with 25 percent of the data for testing. In addition to analyses using Virtual Beach, the SYSTAT 12 statistical software was also used to confirm statistical analyses and produce scatterplots (Systat Software, Inc., 2007).

Supplemental Data Analysis

E. coli data that had been collected by the NHDES for Pawtuckaway Beach from 1993 to 2014 were combined with data from the current study and analyzed for trends in *E. coli* levels during successive beach seasons. In addition, data that represented *E. coli* samples collected from three beach locations (left, center, right) during a single sampling event were analyzed to determine if higher *E. coli* levels tended to occur at a particular location. This analysis was relevant to determine if Burnhams Marshes, which is closest to the right side of the beach where samples were collected, was a potential source of *E. coli*.

An investigation that complemented the current study was conducted by the U.S. Environmental Protection Agency (EPA) Region 1 (New England) to assess possible sources of *E. coli* at Pawtuckaway Beach. The premise of the EPA investigation was that the source of *E. coil* in a water sample was more likely anthropogenic, if certain pharmaceuticals, specifically acetaminophen, were also detected in the sample. Water samples were collected on August 11, 2015, from the left, center, and right locations along the beach, from Burnhams Marshes where the outflow enters the lake, and from pore water that filled a 30-centimeter (cm) deep hole dug in sand near the center of the beach. Results from the EPA investigation were provided as supplemental data (Todd Borci, EPA, written commun., October 22, 2015) and are described as they relate to the context and results of the current study.

Results

Following the procedures described in the "Methods" section of this report, about 100 independent variables, including their time-interval characterizations, were derived from the environmental data. Although all independent variables were used in the analyses, the independent variables described in table 1 were either significantly correlated with *E. coli* or related in some way to the other independent variables that were relevant in explaining variations in *E. coli* levels.

Table 1. Independent variables used to model *Escherichia coli* levels for Pawtuckaway State Park Beach in Nottingham, New Hampshire.

Independent variable	Definition
Date	Sequential calendar day of year (CDY)
Mid_Season	CDY to June 22; afterwards 406-CDY
Visitors	Tally of daily park visitors
Geese	Influence of geese removal, 2016 data only
Wind_A	Alongshore wind vector (unitless)
Wind_O	Onshore/offshore wind vector (unitless)
Wind_Spd	Wind speed, in miles per hour
Wind_Dir	Wind direction, in compass degrees
Cond	Specific conductance, in microsiemens per centimeter at 25 degrees Celsius
pH	pH
Rel_Humid	Relative humidity, in percent
DO	Dissolved oxygen, in milligrams per liter
Lake_Stg	Lake stage, in feet
Marsh_Stg	Marsh stage, in feet
Water_T	Water temperature, in degrees Celsius
Precip	Precipitation, in inches
Air_T	Air temperature, in degrees Celsius

During the course of the study, some important variations were documented in data collection procedures and in environmental conditions at Pawtuckaway Beach that are relevant to the results and their interpretation. The number of *E. coli* samples collected annually by the NHDES at Pawtuckaway Beach varied during the years of the beach inspection program, including during this study. Samples had been collected about five times on average during each beach season from 1991 through 2010. This frequency increased to 10 to 12 times during the 2011 through 2013 seasons and 26 times during the 2014 season. For this study, 48, 67, and 43 samples were collected during the 2015, 2016, and 2017 beach seasons, respectively. Also relevant to the results were daily park-attendance data that became available after completion of the regular data collection efforts in 2017. These attendance data represented a headcount of visitors entering the park; although beach attendance was recorded when the *E. coli* samples were collected, attendance counts typically represented early-morning counts that likely underestimated daily beach use, especially on hot days. Consequently, the variable Visitors was used as an independent variable for estimating beach use.

Certain environmental conditions at Pawtuckaway Beach were notably different among the 2015 to 2017 seasons. In spring 2015, as in earlier years, the migratory goose population became established at the beach for the season. On June 30, 2016, the New Hampshire Division of Parks and Recreation extricated approximately 40 geese from the area of the beach, and only a few individual birds remained for that season. Additionally, a drought that affected much of the Northeast in summer 2016 resulted in a 30-cm drop in the stage of Pawtuckaway Lake during the sampling period; for comparison, the seasonal drop in stage was about 10 cm in 2015 and 20 cm in 2017. The goose population became reestablished at the beach in 2017 but began with only three nesting pairs. Thus, for the 2016 season, Geese was a categorical variable created to characterize three conditions of geese at the beach, as follows:

- 2 = presence, assigned to the days before goose removal;
- 1 = recovery, assigned to 10 days, from July 1 to 10, 2016, when some residual influence of geese was likely;
- 0 = absence, assigned to the days after July 10, 2016, when the influence of geese was likely minimal.

Trends in *E. Coli* Levels

Between 1993 and 2017, water samples for *E. coli* analysis were collected concurrently at left, center, and right locations along the beach during 193 site visits. These data represented 579 individual samples (193 × 3) that had a median *E. coli* level of 62 CFU/100 mL; for this study, an *E. coli* level of 100 CFU/100 mL was considered to be relatively high but still below the State advisory level. Of the 193 site visits with three samples, 113 site visits resulted in at least one of the three samples having an *E. coli* level that exceeded 100 CFU/100 mL.

Using the data from those 113 site visits, *E. coli* levels were compared among the three points to determine if high levels tended be more prevalent at a specific location along the beach; particularly, a finding that levels generally were highest at the right location of the beach could indicate that Burnhams Marshes was a potential *E. coli* source because the flow from the marsh entered the lake near the right side of the beach. However, the number of samples where 100 CFU/100 mL was exceeded was relatively balanced among locations: left, 77 site visits; center, 70 site visits; and right, 76 site visits. Furthermore, in evaluating the range of values at each location, the results of a Wilcoxon signed-rank test indicated no significant differences among the three locations (based on $\alpha = 0.05$). However, a pairwise analysis indicated that the center and right locations were similar ($p = 0.995$), but the left location (farthest from the marsh) was somewhat higher than either the center ($p = 0.063$) or right ($p = 0.087$) locations (fig. 3).

From 2011 through 2017 when *E. coli* sampling was conducted routinely on 10 or more days per year, patterns in how *E. coli* levels varied during the beach seasons differed notably. For 2011 through 2014, *E. coli* was sampled a total of 58 times; the combined dataset indicated a distinct "rise-and-fall" parabolic pattern during an "aggregate" beach season where 25 of the samples (43 percent) exceeded 158 CFU/100 mL, mostly during midseason (fig. 4*A*). For 2015, the first full year of data collection for this study, *E. coli* was sampled on 50 days during the beach season, where 20 of the samples (40 percent) exceeded 158 CFU/100 mL, and the rise-and-fall pattern was similar to that of the previous years (fig. 4*B*).

A different pattern was seen in the 67 samples collected during 2016, which was the year the population of geese was removed on June 30 (fig. 4*C*). Only nine samples (13.4 percent) had a concentration that exceeded 158 CFU/100 mL; five of these exceedances occurred during the geese-present period (before goose removal), and two exceedances occurred during the recovery period. During the geese-absent period, when the influence of geese at the beach was considered minimal, two exceedances occurred: 257 CFU/100 mL on July 21, 2016, and 194 CFU/100 mL on August 25, 2016.

During the 2017 beach season, the goose population became re-established at Pawtuckaway Beach, with a few mating pairs arriving in the spring (Amanda McQuaid, NHDES, written commun., November 30, 2018). The pattern in *E. coli* levels seen in 44 samples collected during the 2017 season was similar to the rise-and-fall pattern seen with the 2011–15 data, but with an important difference (fig. 4*D*): the general magnitude of *E. coli* levels was lower in 2017, with a median value of 52 CFU/100 mL compared with 120 CFU/100 mL for the 2011–15 data. Additionally, only eight *E. coli* samples collected in 2017 (18.2 percent) resulted in exceedances, all of which occurred after July 3; in contrast, only four exceedances

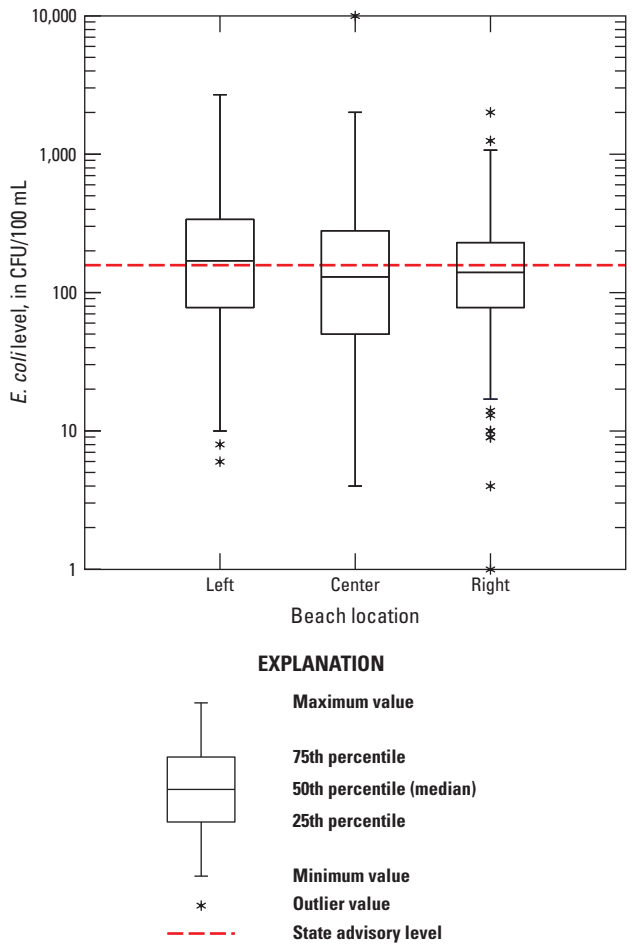

Figure 3. Range of *Escherichia coli* (*E. coli*) concentrations from samples collected at three points (left, center, right) along Pawtuckaway Beach in Nottingham, New Hampshire. The State advisory level for *E. coli* is 158 colony forming units per 100 milliliters (CFU/100 mL), which prompts a beach advisory when exceeded. A Wilcoxon signed-rank test was used to evaluate if median *E. coli* values were significantly different between any two beach locations; a probability value approaches 1.0 as similarity increases between locations.

in *E. coli* levels occurred after July 3 of the previous year (2016, after geese were removed).

E. coli values from 2011 through 2015 (number of samples [n]=108) were log-transformed and plotted together to investigate more closely the rise-and-fall pattern that had occurred consistently during the beach seasons for those years. To interpret the relative difference between raw values and log-transformed values of *E. coli* levels shown in some figures, the exceedance threshold of 158 CFU/100 mL is represented by 2.20 when log-transformed. The pattern was characterized by a second-order polynomial function with a coefficient of determination (r^2) value of 0.44 (fig. 5*A*). The maximum *E. coli* value predicted by the function was 262 CFU/100 mL on July 22 (around midseason) for those years, indicating that *E. coli* levels continually increased daily up to that date, then decreased in the days afterwards. Using this procedure with the 2017 *E. coli* data, a predicted maximum of 184 CFU/100 mL resulted, which was consistent with the above finding that values were overall lower for that year; in addition, the maximum value was predicted on July 28, within a week of the date predicted for the maximum value in the 2011–15 dataset.

A transformation was imposed on the curve (fig. 5*A*) so that the pattern could be expressed with a linear equation (fig. 5*B*) by first reassigning the variable Date (X-axis) whole numbers that represent the calendar day of the year, so that January 1 to December 31 were numbered sequentially from 1 to 365 (for a nonleap year). A new independent variable, Mid_Season, was then derived to represent the number of days from July 22, the approximate midpoint of the beach season when high levels of *E. coli* were predicted. Beginning with a maximum value of 203 corresponding to July 22 (203rd day of year) Mid_Season was derived as follows: $203 - x$, where x is the number of days any date given differed from July 22. For example, Mid_Season = 203 represented July 22 (203 – 0), but Mid_Season = 202 represented both July 21 and 23, because these two dates differed from July 22 by one day (203 – 1), and Mid_Season = 201 represented both July 20 and 24, and so forth. Note in the comparison between independent variables Mid_Season and Date that values are the same from January 1 to July 22 (days 1 to 203); afterwards, values for each successive day decrease by 1 for Mid_Season but increase by 1 for Date.

With this transformation, the resulting regression had an r^2 of 0.43, indicating that *E. coli* levels would more likely be higher on the days nearest to July 22 (fig. 5*B*). This result was observed for the time interval from day 196 to 203, representing the 2-week period from July 15 to July 29 when *E. coli* was sampled 27 times; 19 (70 percent) of the samples had concentrations that exceeded 158 CFU/100 mL. The sequence of days approaching the middle of the beach season, therefore, was regarded as an important factor for predicting exceedances in *E. coli* levels at Pawtuckaway Beach.

EPA *E. Coli* Source Investigation

Results from the lake and pore water investigation by the EPA on August 11, 2015, did not find detectable levels of acetaminophen in samples of lake or pore water collected at any location, indicating that fecal coliforms in the samples would likely be from nonhuman sources. *E. coli*, measured by the EPA as mean probable number, was not detected in the sample from the outflow of Burnhams Marshes, which substantiated the above finding (fig. 3) that the marsh was probably not the source of bacteria at the beach. *E. coli* levels in samples collected from the left, center, and right locations along the beach were relatively low at 12, 25, and 8 mean probable number per 100 milliliters (MPN/100 mL), respectively.

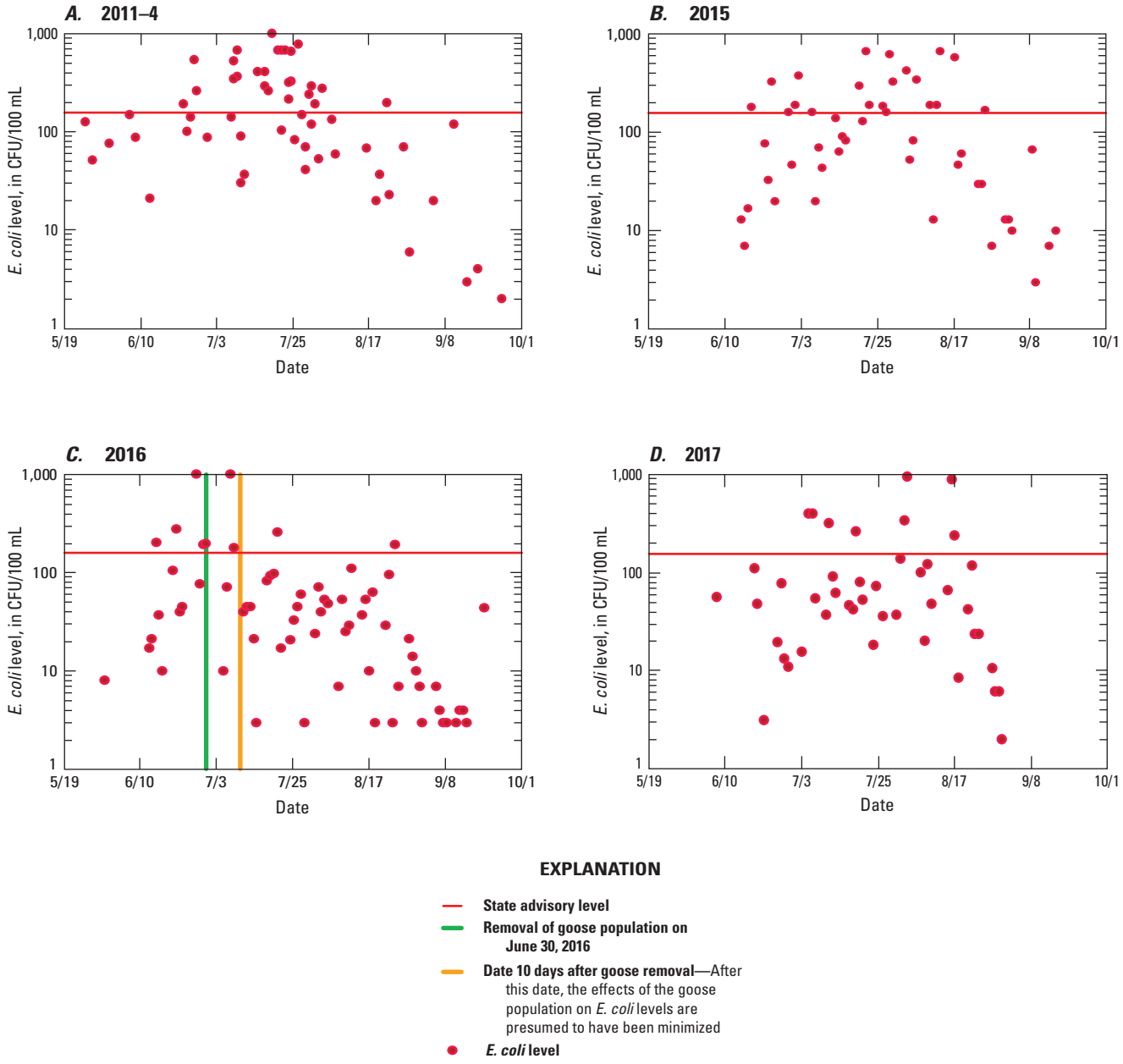

Figure 4. Patterns in *Escherichia coli* (*E. coli*) concentrations for beach seasons *A*, 2011–4, *B*, 2015, *C*, 2016, and *D*, 2017 at Pawtuckaway Beach in Nottingham, New Hampshire. The State advisory level for *E. coli* is 158 colony forming units per 100 milliliters (CFU/100 mL), which prompts a beach advisory when exceeded. Samples with *E. coli* concentrations that exceeded 1,000 CFU/100 mL are represented as the maximum value of 1,000 CFU/100 mL.

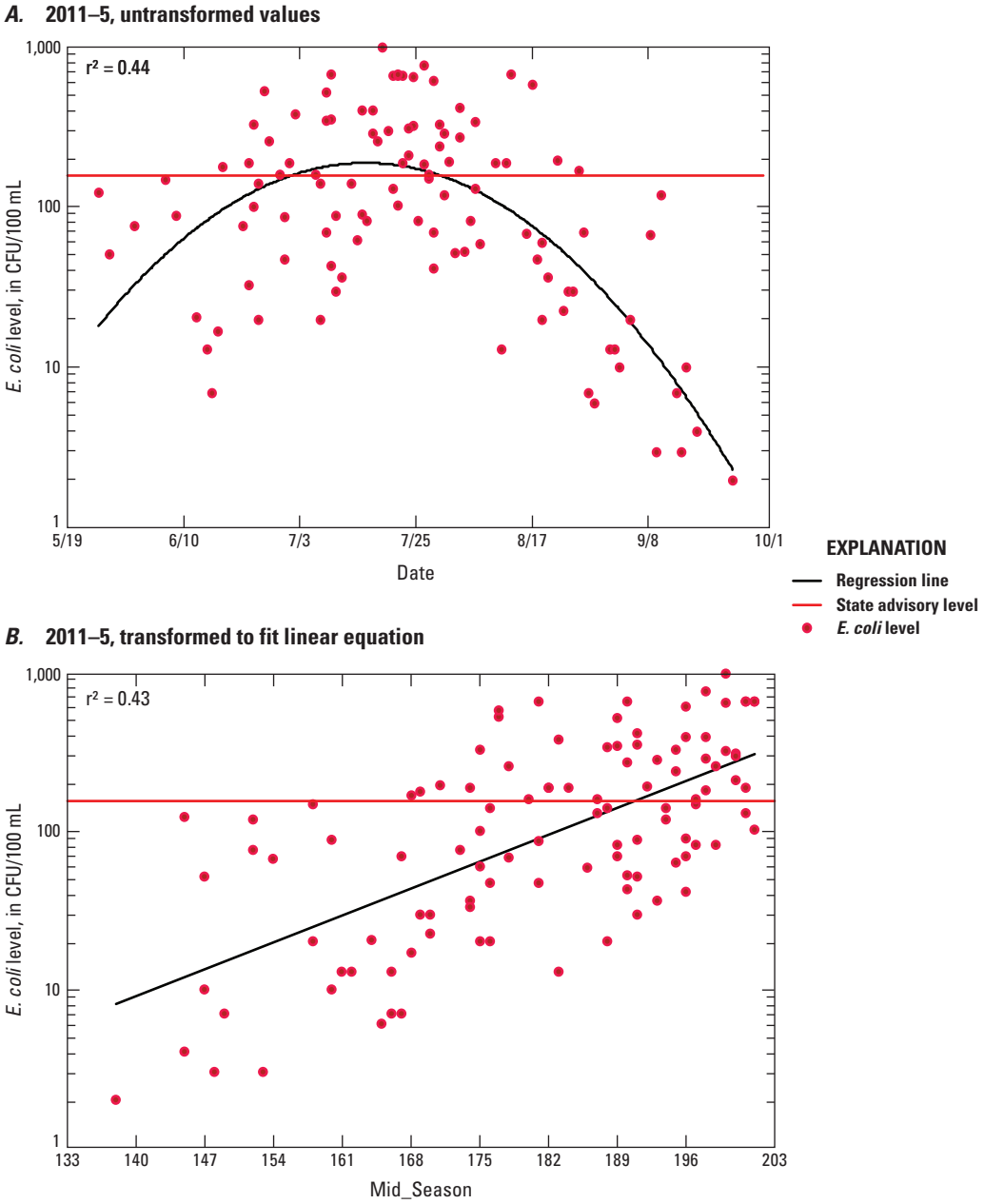

Figure 5. *Escherichia coli* (*E. coli*) concentrations at Pawtuckaway Beach in Nottingham, New Hampshire, from 2011 through 2015. *A*, The rise-and-fall pattern of *E. coli* levels occurring during the beach seasons is characterized by the curved regression line, which is based on a second-order polynomial equation that predicts the highest level occurs on July 22. *B*, Transformation of the relation, showing *E. coli* levels are predicted highest on July 22, considered to be the midseason in this report and represented by the 203rd day of the year. The *E. coli* State advisory level is 158 colony forming units per 100 milliliters (CFU/100 mL), which prompts a beach advisory when exceeded. r^2, coefficient of determination.

However, the pore water in the beach sand was relatively high at 102 MPN/100 mL. Overall, the EPA results indicated that a likely source of *E. coli* at Pawtuckaway Beach was geese, and that pore water in the sand may act as a reservoir for bacteria, which could affect levels of *E. coli* when sediments are disturbed as beachgoers walk along the water's edge (Todd Borci, EPA, written commun., October 22, 2015).

Data Evaluation With Virtual Beach

Independent variables were identified as relevant for each beach season for 2015, 2016, and 2017 (table 2). The data were then analyzed with the Virtual Beach software program to evaluate the performance of selected models (table 3).

2015 Beach Season

Selected independent variables (table 1) from the 2015 beach season data that had statistically significant correlations with *E. coli* levels included Mid_Season and Visitors collected the day before *E. coli* sample collection; Wind_A (alongshore wind), pH, and Rel_Humid (relative humidity) collected the same day as the *E. coli* samples; and Cond (specific conductance), DO (dissolved oxygen), Lake_Stg (lake stage), and Marsh_Stg (marsh stage) collected in the 24-hour period before *E. coli* sample collection. The strongest correlations were with Mid_Season (coefficient of correlation [r] = 0.643) and Visitors (r = 0.554), which showed a similar pattern in how they increased with *E. coli* levels (fig. 6), suggesting that *E. coli* levels increase concurrently with beach-use intensity. The relation among these independent variables was further supported in the correlation between Visitors and Mid_Season (r = 0.508), indicating an increase in park attendance from early to the middle of the beach season.

The alongshore wind vector (Wind_A) had the strongest correlation of any independent variable with *E. coli* levels (r = −0.399) that was not also significantly correlated with either Visitors or Mid_Season. Strong winds, generally from a southerly direction, were indicated as contributing to high *E. coli* levels (fig. 7). A prominent outlier was the sample collected on September 14 (circled in fig. 7); the correlation improves notably when this sample is removed (r = 0.504). This outlier corresponds to a sample near the end of the beach season when beach attendance was low and lower *E. coli* levels (fig. 4) were expected; thus, wind conditions during this time would likely not contribute to high *E. coli* levels without the other contributing factors.

Several independent variables that are significantly correlated with *E. coli* levels were more closely correlated with Date (progression of the beach season), but these independent variables may have had relatively little influence on *E. coli* levels. The variables were Cond, DO, Lake_Stg, and Marsh_Stg, measured during the 24 hours before *E. coli* sample collection. *E. coli* levels were negatively correlated with Cond (r = −0.446), which remained relatively low

Table 2. Selected independent variables resulting from evaluations of *Escherichia coli* data from the 2015, 2016, and 2017 beach seasons for Pawtuckaway State Park Beach in Nottingham, New Hampshire.

[Independent variables are described in table 1. r, correlation coefficient; NA, not applicable; same day, measurement collected the same day as the *E. coli* sample; 24h, measurement collected in the 24-hour period before *E. coli* sample collection; 12h, measurement collected in the 12-hour period before *E. coli* sample collection]

Primary independent variable		Covariable	
Name	**r value**	**Name**	**r value**
2015 beach season			
Mid_Season	0.643	NA	NA
Visitors	0.554	Mid_Season	0.508
Wind_A (same day)	−0.399	NA	NA
	−0.504[1]		
Rel_Humid (same day)	−0.365	Visitors	−0.577
Cond (24h)	−0.446	Date	0.903
DO (24h)	−0.466	Mid_Season	−0.598
		Water_T	−0.500
Lake_Stg (24h)	0.391	Date	−0.969
Marsh_Stg (24h)	0.418	Date	−0.816
pH (same day)	0.407	NA	NA
2016 beach season			
Date	−0.593	NA	NA
Visitors	0.366	Mid_Season	0.600
Geese	0.452	NA	NA
Cond (24h)	−0.571	Date	0.912
DO (24h)	0.575	Date	−0.909
Lake_Stg (24h)	0.601	Date	−0.949
pH (same day)	0.476	Date	−0.600
Rel_Humid (same day)	−0.131[2]	Visitors	−0.405
Water_T (12h)	0.238[2]	NA	NA
Wind_Spd (12h)	0.105[2]	NA	NA
2017 beach season			
Mid_Season	0.435	NA	NA
Visitors	0.429	NA	NA
Wind_O (same day)	0.289[2]	NA	NA
Cond (12h)	0.395	NA	NA
pH (12h)	−0.414	NA	NA
Water_T (same day)	0.574	Mid_Season	0.563
		Visitors	0.564
Wind_Dir (same day)	0.400	NA	NA

[1]The stronger correlation resulted when outlier was removed, as shown in figure 7 of this report.

[2]These r values were not statistically significant but are relevant in explaining results.

Table 3. Evaluation statistics for models generated with *Escherichia coli* data from the 2015, 2016, and 2017 beach seasons for Pawtuckaway State Park Beach in Nottingham, New Hampshire, tested with data from alternative beach seasons and for a model generated with the 2015 and 2017 season data combined.

[Models were generated with data from each of the three beach seasons, then tested against data from alternate beach seasons. Additionally, a model was generated with data combined from all three seasons and covered data for 2015 and 2015, when geese were not removed from the beach. All models were generated and tested with the Virtual Beach software program (Cyterski and others, 2014). r, correlation coefficient]

Evaluation statistic	r value	Evaluation statistic	r value
2015 model generated from 2015 data		2017 model generated from 2017 data	
Coefficient of determination	0.527	Coefficient of determination	0.337
Specificity	0.89	Specificity	0.94
Sensitivity	0.40	Sensitivity	0.25
Accuracy	0.68	Accuracy	0.81
2015 model tested with 2016 data		2017 model tested with 2015 data	
Coefficient of determination	0.209	Coefficient of determination	0.551
Specificity	0.69	Specificity	1.00
Sensitivity	0.38	Sensitivity	0.10
Accuracy	0.65	Accuracy	0.62
2016 model generated from 2016 data		2017 model tested with 2016 data	
Coefficient of determination	0.331	Coefficient of determination	0.229
Specificity	0.98	Specificity	0.89
Sensitivity	0.50	Sensitivity	0.13
Accuracy	0.92	Accuracy	0.79
2015 model tested with 2017 data		Model generated from combined data	
Coefficient of determination	0.266	Coefficient of determination	0.431
Specificity	0.80	Specificity	0.85
Sensitivity	0.25	Sensitivity	0.33
Accuracy	0.70	Accuracy	0.70
2016 model tested with 2017 data			
Coefficient of determination	0.191		
Specificity	0.43		
Sensitivity	0.88		
Accuracy	0.51		

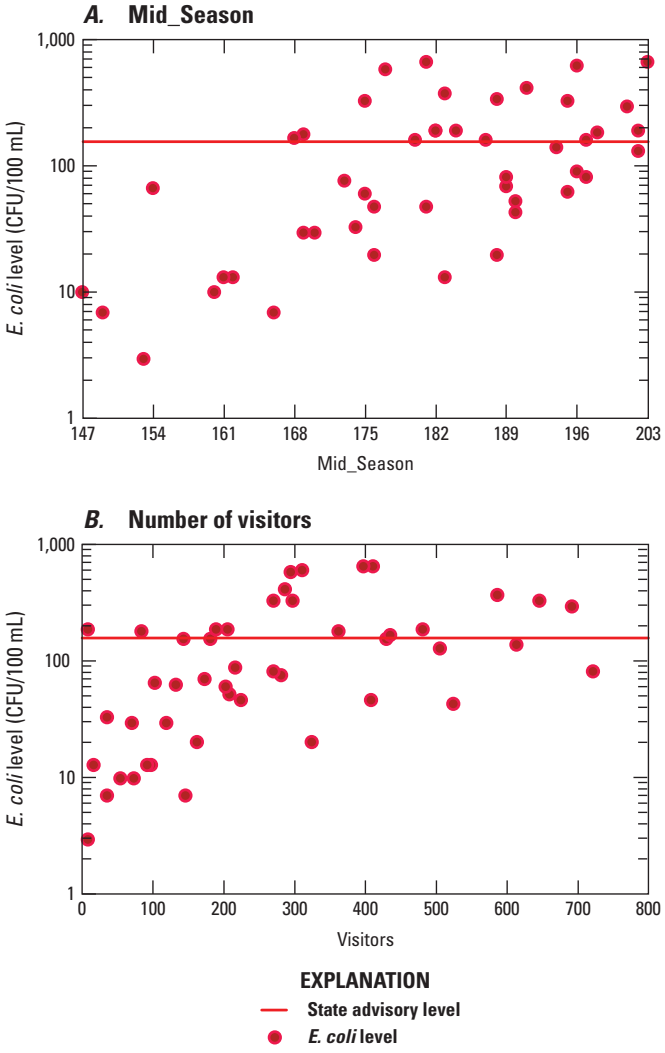

Figure 6. *Escherichia coli* (*E. coli*) concentrations at Pawtuckaway State Park Beach in Nottingham, New Hampshire, relative to the independent variables Mid_Season and Visitors based on data from the 2015 beach season. The comparable patterns as well as the correlation between the variables Mid_Season and Visitors (coefficient of correlation [r] = 0.643) suggest that *E. coli* levels increase as beach use increases. The *E. coli* State advisory level is 158 colony forming units per 100 milliliters (CFU/100 mL), which prompts a beach advisory when exceeded

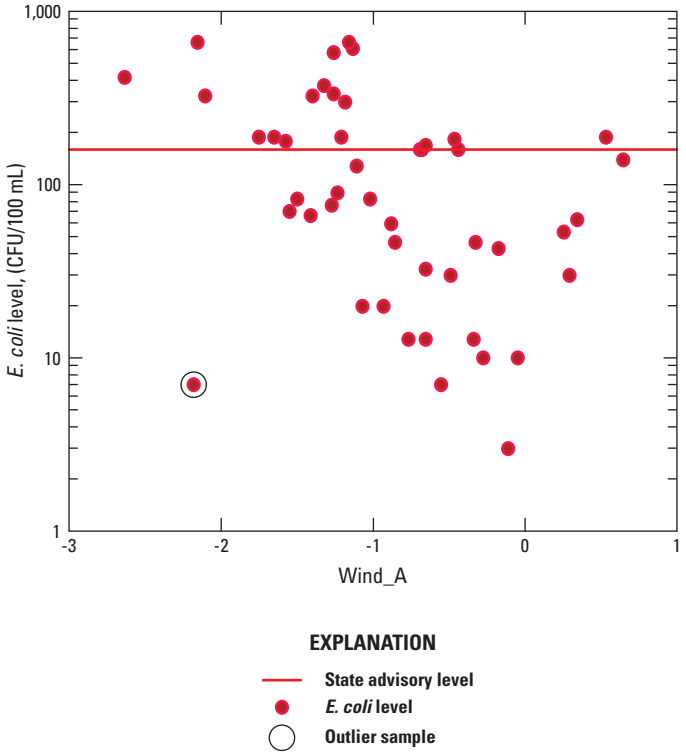

Figure 7. *Escherichia coli* (*E. coli*) concentrations relative to the alongshore wind (Wind_A) independent variable (r = 0.399) for Pawtuckaway State Park Beach in Nottingham, New Hampshire. The horizontal scale corresponds to the alongshore wind relative to beach orientation, where the negative values (unitless) represent a strong southerly wind. Removal of the mid-September outlier (circled) sample results in a stronger correlation (r = 0.504). The *E. coli* State advisory level is 158 colony forming units per 100 milliliters (CFU/100 mL), which prompts a beach advisory when exceeded.

throughout the beach season, ranging from 45 microsiemens per centimeter (µS/cm) at 25 degrees Celsius at the beginning of the season to 50 µS/cm at the end of the season. However, Cond had a strong positive covariance with Date (r = 0.903), and the decrease in *E. coli* levels occurred when conductivity increased from 48 to 50 µS/cm during the latter half of the beach season. This small change in conductivity seems an unlikely factor in the decline in *E. coli* levels, especially considering that *E. coli* was probably declining in response to other factors late in the beach season. A more plausible explanation is the slight increase in conductivity was coincident and thus colinear with the decline in *E. coli* typically observed (fig. 5) during the latter half of the beach season.

A decrease in *E. coli* levels as water levels dropped was indicated by the correlations with Lake_Stg (r = 0.391) and Marsh_Stg (r = 0.418); however, as seen with the change in Cond, drop in stage was more strongly correlated with Date (Lake_Stg, r = −0.969; Marsh_Stg, r = −0.816). Overall, lake and marsh stages continuously dropped during the beach season, a consequence of the summer low-flow period, and the decrease in *E. coli* levels occurred during the latter half of the beach season as Lake_Stg dropped by 12.2 cm (0.4 ft) and Marsh_Stg by 6.1 cm (0.2 ft). As concluded with the correlation between *E. coli* levels and Cond, the decrease in the stage probably was coincident with seasonal progression, and thus the drop in stage probably did not contribute directly to declining *E. coli* levels.

E. coli levels were negatively correlated with DO (r = −0.466), but among the other independent variables, DO was most strongly correlated (negatively) with Mid_Season (r = −0.598) and Water_T (water temperature; r = −0.500), the latter independent variable being relevant for oxygen to remain dissolved in water. Although Water_T was not significantly correlated with *E. coli* levels (r = 0.103), the temperatures increased progressively to about 28 degrees Celsius at the middle of the beach season when *E. coli* levels were highest. The daily median DO levels ranged from 8.6 to 7.7 milligrams per liter during the beach season; although this range is relatively narrow, the lowest DO values occurred around the middle of the beach season, hence the strong correlation between DO and Mid_Season. A tangible explanation is that high *E. coli* levels are associated with warm temperatures and low DO values, conditions indicative of a heterotrophic environment. However, the positive correlation between *E. coli* and DO in 2016 (described below) was contradictory, suggesting that collinearity was a more likely explanation in the 2015 correlation.

E. coli levels were positively correlated with pH (0.407), although the range in pH was small and included only four values from 6.8 to 7.1, which resulted in the pH variable being more categorical than continuous. Only one *E. coli* sample was collected at a pH of 7.1 and had a relatively low value at 67 CFU/100 mL, whereas 22 *E. coli* samples with levels greater than 100 CFU/100 mL were collected at the lower pH values. However, the positive correlation between *E. coli* and pH was strongly influenced by *E. coli* levels at pH 7.0 (median, 300 CFU/100; mean, 316 CFU/100), which were relatively higher than *E. coli* levels at pH 6.9 (median, 83 CFU/100; mean, 152 CFU/100) and 6.8 (median, 45 CFU/100; mean, 78 CFU/100). Thus, these results indicated that pH 7.0 was most conducive to higher *E. coli* levels, but the small differences in pH values that influenced this correlation suggest that the results should not be considered definitive.

E. coli levels were negatively correlated with Rel_Humid (r = −0.365), but among the other independent variables, Rel_Humid was most strongly correlated with Precip (precipitation; r = 0.685) and Visitors (r = −0.577). A direct relation between Rel_Humid and *E. coli* levels is somewhat ambiguous, especially because the *E. coli* samples are collected from an aqueous environment. A possible explanation of the relation, assuming *E. coli* levels are more directly linked to the activity of Visitors, is that beach use would be low on days with rain, which in turn would result in high relative humidity. The correlation between *E. coli* levels and the Precip was relatively weak, however, possibly because a light shower and a downpour could be equally effective in deterring visitors from the beach.

In applying Virtual Beach to generate predictive models to explain *E. coli* levels, Mid_Season was a primary independent variable in nearly all resulting models and, additionally, a quantification of Wind (either as speed, direction, or vector) was usually included. In general, when more than three independent variables were included in a model, the improvement in fit of the model was only slight. Results of the cross-validation procedure indicated the lowest (best) MSEP score was 0.256 for a model that included Mid_Season, Visitors, and Wind_A, independent variables that are described as being correlated with *E. coli* levels. The equation for correlating *E. coli* based on the data from the 2015 beach season is as follows:

$$E.\,coli = -1.908 + \left(0.0007881 \times Visitors\right) + \left(0.01885 \times Mid\_Season\right) - \left(0.2144 \times Wind\_A_{same\_day}\right). \quad (3)$$

The potential performance of this model is shown in a scatterplot of fitted values compared with the observations of *E. coli* (fig. 8); the exceedance threshold of 158 CFU/100 mL for each of the two sets of values is shown as vertical (observed) and horizontal (fitted) lines that divide the plot into four quadrants. The lower left quadrant represents true negative (TN) values where both predicted and observed values (n = 24) were below the exceedance threshold; the upper right quadrant represents true positive (TP) values where both predicted and observed values (n = 8) were above the exceedance threshold; the upper left quadrant represents false positive (FP) values where the predicted values were above the exceedance threshold but were observed below the threshold (n = 3); and the lower right quadrant represents false negative (FN) values where the predicted values were below the exceedance threshold but were observed above the threshold (n = 12).

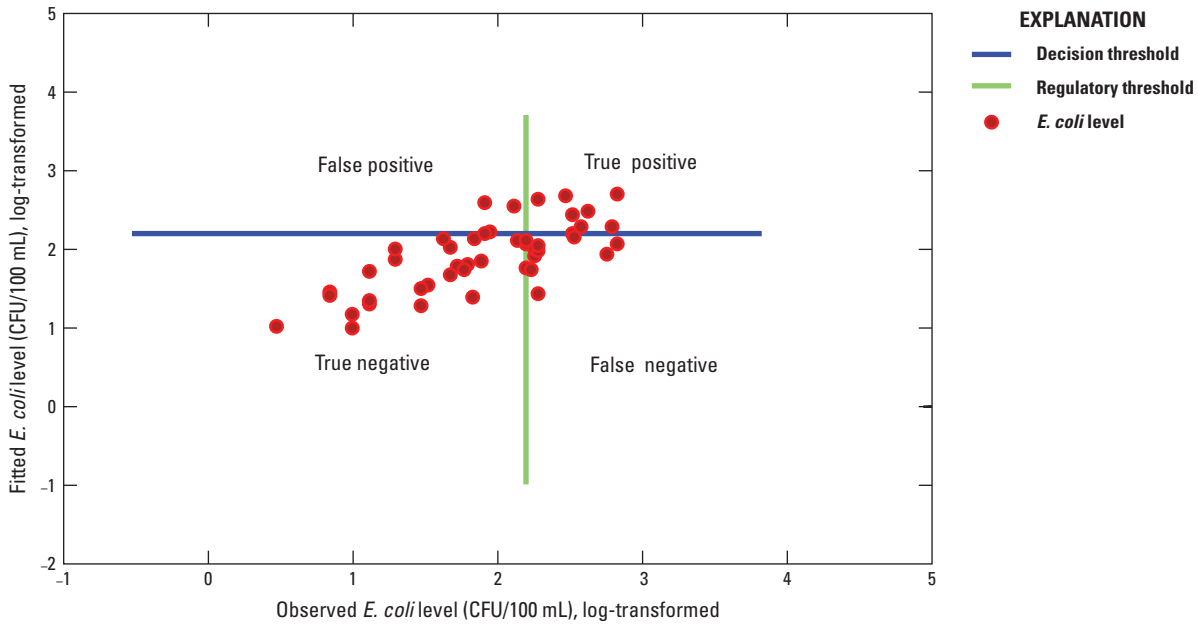

Figure 8. Scatterplot of fitted versus observed values of *Escherichia coli* (*E. coli*) levels for Pawtuckaway State Park Beach in Nottingham, New Hampshire, based on a predictive model generated with beach data from 2015 and incorporates independent variables Visitors, Mid_Season, and Wind_A. The scales for the vertical and horizontal axes indicate logarithmically (log_{10}) transformed *E. coli* levels; the State advisory threshold of 158 CFU/100 mL translates to 2.2 when log-transformed. *E. coli* levels greater than the horizontal decision threshold are exceedances fitted by the model, and levels greater than the vertical regulatory threshold are exceedances observed in samples.

Evaluation statistics for this model include specificity (eq. 4*A*), sensitivity (eq. 4*B*), and accuracy (eq. 4*C*), as follows:

$$\frac{TN}{TN+FP} = 0.89, \tag{4A}$$

$$\frac{TP}{TP+FN} = 0.4, \text{ and} \tag{4B}$$

$$\frac{TP+TN}{n} = 0.68, \tag{4C}$$

where n is the total number of values. Although the accuracy statistic of this model would suggest a correct prediction 68 percent of the time, the low sensitivity is undesirably low because the model would only be 40 percent accurate in correctly predicting when *E. coli* levels are above the exceedance threshold. However, even when evaluating the full set of independent variables (about 100) with the PRESS statistic criterion, no model was generated in Virtual Beach that had a higher sensitivity value.

2016 Beach Season

During 2016, *E. coli* levels had a notably different pattern compared with 2015, both in how *E. coli* levels varied with seasonal progression and in the likelihood that any sample would exceed the advisory level of 158 CFU/100 mL; the likelihood decreased from 40 to 13.4 percent between the two beach seasons. *E. coli* generally decreased continuously with Date (r = −0.593) compared with the parabolic pattern of initial increase and later decrease that was related to Mid_Season in 2015 (fig. 4). This difference in the *E. coli* pattern was likely attributable to the removal of the population of geese in 2016, as indicated by the correlation between *E. coli* levels and Geese (r = 0.452). *E. coli* levels were significantly correlated with number of Visitors (r = 0.366), whereas the correlation with Visitors in 2015 was notably stronger (r = 0.554). However, as in 2015, Visitors was strongly correlated with Mid_Season (r = 0.600) in 2016, which indicates a common pattern of beach use with seasonal progression for the two beach seasons.

In general, collinearity among *E. coli* levels, independent variables, and Date was stronger in 2016 compared with 2015, presumably because *E. coli* levels decreased more linearly during the 2016 beach season compared with the parabolic

pattern in 2015. In the descriptions that follow, independent variables represent the same time-series characterizations as the 2015 results, unless otherwise specified.

E. coli was correlated significantly with Cond (r = −0.571), which increased continually from 50 to 54 µS/cm, but similar to 2015, Cond in 2016 was more strongly correlated (positively) with Date (r = 0.912), indicative of the relevance of seasonal progression. Also, as in 2015, the correlation of *E. coli* with DO was significant (r = 0.575), but with an important difference: the correlation in 2015 was negative (r = −0.466). This inconsistency in the significant responses of *E. coli* with DO for 2015 and 2016 is likely a result of collinearity; both *E. coli* and DO declined with seasonal progression in 2016, as indicated by the strong correlation between DO and Date (r = −0.909). Similarly, the decline in *E. coli* was significantly correlated with drop in Lake_Stg (r = 0.601), whereas Lake_Stg was more strongly correlated with Date (r = −0.949). Decline in *E. coli* was significantly correlated with the decline in pH (r = 0.476), whereas pH was more strongly correlated with Date (r = −0.600).

In summary, results from 2016 indicated that *E. coli* levels declined continually during the beach season (r = −0.593), but several independent variables changed continually during the beach season as well; thus, a significant correlation between *E. coli* and any of these independent variables is not considered a definitive response of *E. coli* to the independent variable. However, an independent variable might still reflect conditions that indirectly relate to *E. coli* levels (such as precipitation [Precip] deters beach use [Visitors], which in turn diminishes *E. coli*) and thus could be important in prediction models. Unlike the 2015 results, *E. coli* levels were not significantly correlated with Rel_Humid or Wind. As in 2015, however, Rel_Humid was significantly correlated with Visitors (r = −0.405), suggesting wet weather would discourage beach use (as described in the "2015 Beach Season" section of this report). Regarding the lack of evidence that *E. coli* levels were not affected by wind in 2016, a possible explanation is that lower *E. coli* levels generally occurred in 2016 because the source of *E. coli* was reduced (geese removed); therefore, the influence of wind would be less important in contributing to high levels of *E. coli*.

The 2016 dataset was used to test the performance of the 2015 model in Virtual Beach by using real-time 2016 environmental data to predict *E. coli* levels, which were then compared with actual (observed) 2016 *E. coli* values (fig. 9). Even though beach conditions were notably different between

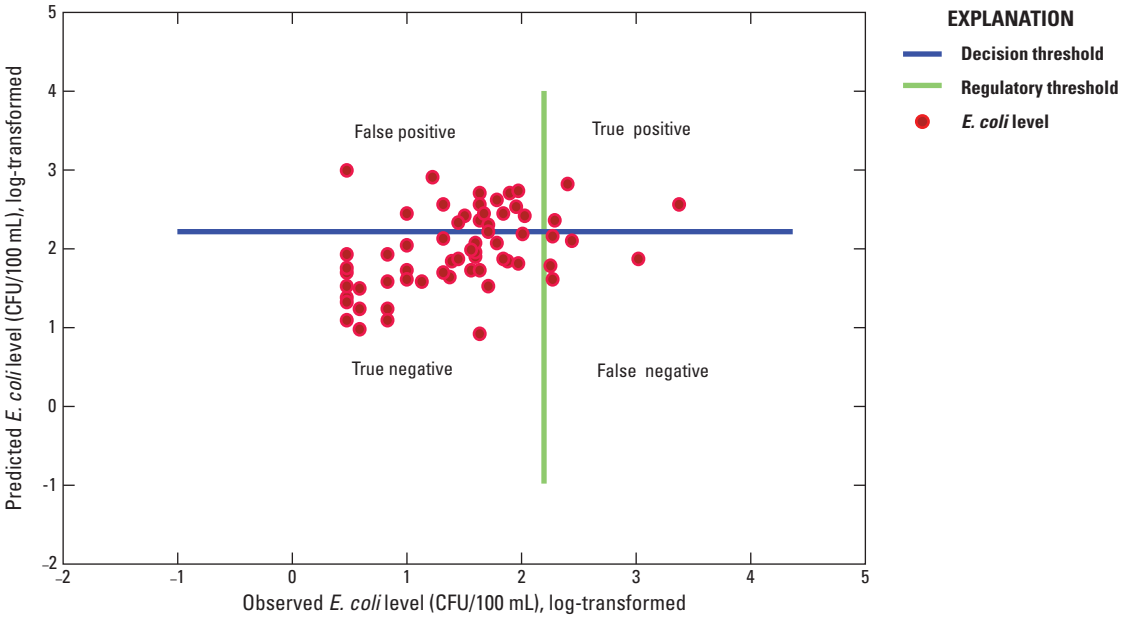

Figure 9. Scatterplot indicating how the 2015 model (originally generated with 2015 data; fig. 8) performed with the 2016 beach season data for Pawtuckaway State Park Beach in Nottingham, New Hampshire, evaluating predicted against observed values. The scales for the vertical and horizontal axes indicate logarithmically (log_{10}) transformed *E. coli* levels; the State advisory threshold of 158 CFU/100 mL translates to 2.2 when log-transformed. *E. coli* levels greater than the horizontal decision threshold are exceedances fitted by the model, and levels greater than the vertical regulatory threshold are exceedances observed in samples.

the 2 years, evaluation of the 2015 model in how well it predicted *E. coli* levels with 2016 data resulted in an accuracy of 0.65, which was only slightly lower in accuracy than the 2015 model with the initial (2015) training data (0.68; eq. 4*C*). In evaluating how well the 2015 model accurately predicted exceedances with the 2016 environmental data, specificity (0.69) and sensitivity (0.38) were somewhat lower than with the 2015 training data. In general, however, the 2015 model performed relatively well with the new 2016 data, which can be explained, in part, because of the fewer exceedances in 2016.

The 2016 dataset was used as new training data to generate additional predictive models in Virtual Beach, and the overall results were models that had very low sensitivity values, ranging from 0 to 0.25, which indicates a high number of false negatives. These results occurred in part because the fundamental nature of a regression model is to reduce variability around the regression line rather than evaluate if values are above or below an exceedance threshold. However, in addition to evaluating model fitness based on the PRESS statistic, an alternative option is to evaluate models based on resulting sensitivity values (that is, to select models with the lowest percentage of false negatives). For the alternative models generated in this fashion, sensitivity increased to 0.50 for the 10 best models, with all models containing Geese as the primary independent variable and a characterization of Water_T (water temperature; most often measured in the 12-hour period before *E. coli* sample collection) as the secondary independent variable. The MSEP values for these models ranged widely from 0.31 to 30.85, with a value of 0.32 for the most "parsimonious" model, containing three independent variables: Geese, Water_T, and Wind_Spd (wind speed; both measured in the 12-hour period before *E. coli* sample collection). The equation for correlating *E. coli* based on the data from the 2016 beach season is as follows:

$$E.\,coli = -3.142 + \left(0.1778 \times Water\_T\right) + \left(1.051 \times Geese\right) - \left(0.01944 \times Wind\_Spd\right). \quad (5)$$

The coefficient of determination was relatively weak with this model ($r^2 = 0.331$) as indicated in the scatterplot of fitted versus observed values (fig. 10), although the accuracy of this model improved to 0.92 compared with the original 2015 model because of higher specificity (0.98) and sensitivity (0.50). However, when models are generated with the primary constraint of a low sensitivity value, independent variables are selected that minimize the number of false negatives, even though other independent variables might better explain the

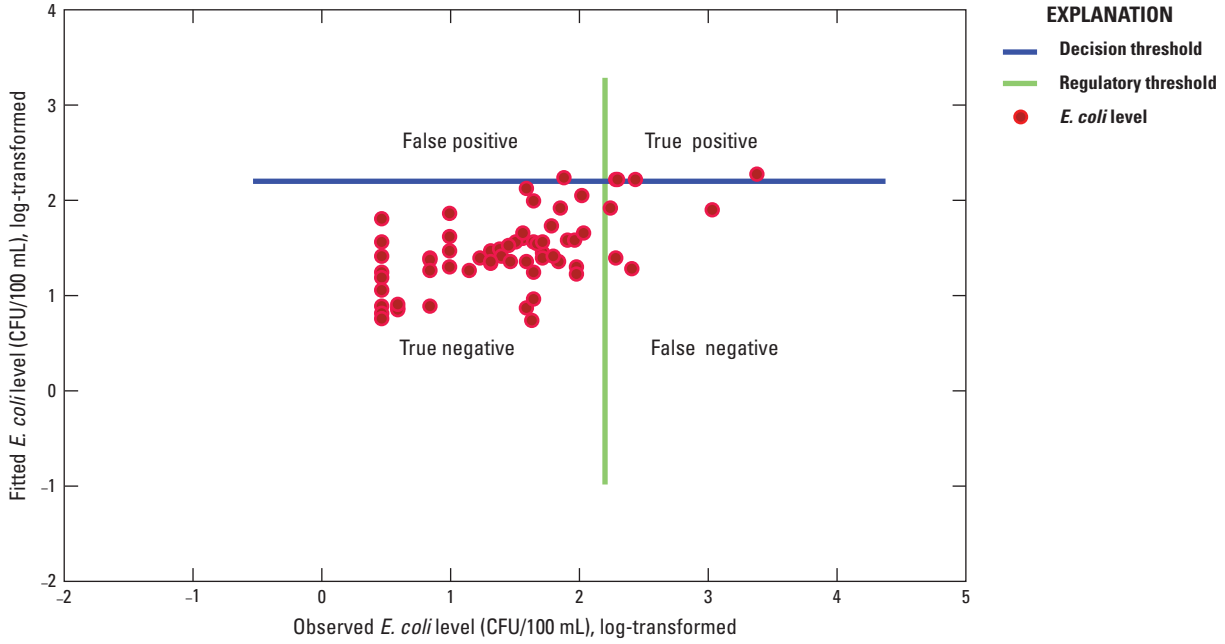

Figure 10. Scatterplot of fitted versus observed values of *Escherichia coli* (*E. coli*) levels based on a model generated with beach data from 2016 for Pawtuckaway State Park Beach in Nottingham, New Hampshire, and incorporates values for the independent variables Geese, Water_T, and Wind_Spd. The scales for the vertical and horizontal axes indicate logarithmically (log_{10}) transformed *E. coli* levels; the State advisory threshold of 158 CFU/100 mL translates to 2.2 when log-transformed. *E. coli* levels greater than the horizontal decision threshold are exceedances fitted by the model, and levels greater than the vertical regulatory threshold are exceedances observed in samples.

relation between environmental conditions and *E. coli* levels. This result was apparent in the 2016 model (eq. 5), where *E. coli* was weakly correlated with Water_T (r = 0.238) and Wind_Spd (r = 0.105) measured in the 12-hour period before *E. coli* sample collection. Thus, although the model indicated relatively high accuracy with the training data used to generate it, the model would likely have a low prediction accuracy with new environmental data, especially because the primary independent variable (Geese) was unique to 2016 and characterized removal of a likely *E. coli* source.

2017 Beach Season

Geese became re-established at Pawtuckaway Beach during the 2017 season, although at lower numbers than in previous years, but with the likely consequence that high *E. coli* levels were related to their presence (fig. 4*D*). As in 2015, *E. coli* levels were positively correlated with Mid_Season (r = 0.435) and Visitors (r = 0.429). However, correlations of *E. coli* with Cond (r = 0.395) and with pH (r = −0.414) were opposite the 2015 correlations with these independent variables; these findings indicated that specific conductance and pH, which had narrow ranges at Pawtuckaway Beach, probably would not be useful in a model that reliably predicts *E. coli* levels at Pawtuckaway Beach.

E. coli was positively correlated with Water_T, same_day (r = 0.574), but this measure of water temperature was also strongly correlated with Mid_Season (r = 0.563) and Visitors (r = 0.564); these results suggest that the apparent relations among these variables are better explained as collinear variance with seasonal progression. *E. coli* was significantly correlated with the Wind_Dir (wind direction, measured on the same day as *E. coli* sample collection; r = 0.400) and is consistent with the finding from 2015 that wind may have contributed to high levels. During the 2017 beach season, the eight *E. coli* samples with levels greater than the exceedance threshold were collected when wind direction ranged from 225 to 245 degrees, indicating the highest *E. coli* levels occurred when the wind was from the southwest. This result, as well as the significant correlation in 2015 between *E. coli* and Wind_A, indicates that winds generally from a southerly to westerly direction likely contribute to high *E. coli* levels.

The performance of the 2015 and 2016 models were tested with the 2017 data in Virtual Beach. The results with the 2015 model were a prediction accuracy of 0.70, which is comparable with the accuracy of the 2015 model with the 2015 dataset (0.68, described above), although specificity (0.80) and sensitivity (0.25) were lower than with the 2015 dataset (0.89 and 0.40, respectively).

Different results occurred with the 2016 model using 2017 data: accuracy was only 0.51, indicating that the 2016 model was correct in predicting exceedances only about half the time. This rather poor outcome occurred because the 2017 data were not especially suited for the 2016 model, the outcome of which was strongly influenced by Geese (as a primary *E. coli* source) that occurred only in 2016. An interesting result of the predictions from the 2016 model with the new data, however, was that sensitivity was relatively high (0.88) because only one false positive resulted; in a sense, the 2016 model generated forecasts that were "overprotective" of public health by predicting that *E. coli* levels would exceed the advisory threshold most days.

New models were generated with the 2017 data in Virtual Beach, and the evaluation statistics indicated that a model similar to the 2015 model was among the best, based on the PRESS evaluation criteria and MSEP score (0.34). The independent variables used in this model were Visitors, Mid_Season, and Wind_O (wind onshore/offshore direction, measured the same day as *E. coli* sample collection), and the accuracy (0.81) was relatively high compared with the 2015 model (0.68). Specificity was among the highest of any model (0.94) but sensitivity was relatively low (0.25), indicating that the model would correctly predict a nonexceedance 94 percent of the time but correctly predict an exceedance only 25 percent of the time. In comparing the 2015 and 2017 models, the primary difference was the independent variable that characterized the wind vector, Wind_A and Wind_O, respectively. The 2015 model incorporated the vector characterizing wind from the south (along the beach), whereas the wind vector in the 2017 model characterized wind from the west (onshore). From a geographical perspective, these two wind vectors are orthogonal, representing winds from the south and west, and together likely indicate that winds generally from the southwest contribute to high *E. coli* levels; this premise is also consistent with the significant correlation in 2017 between *E. coli* levels and *Wind_Dir* (wind from the southwest). The equation for correlating *E. coli* based on the data from the 2017 beach season is as follows:

$$E.\,coli = -2.634 + \left(0.06798 \times Wind\_O\right) + \left(0.02151 \times Mid\_Season\right) + \left(0.0007255 \times Visitors\right). \quad (6)$$

The 2017 model was tested with data from the 2015 and 2016 beach seasons. Results with the 2015 data indicated a relatively strong agreement between the observed and predicted values, based on the coefficient of determination ($r^2 = 0.551$), indicating that most of the *E. coli* variability in 2015 could be explained with the 2017 model. The predictions generated no false positives (specificity = 1), but 18 false negatives (sensitivity = 0.10) were generated, which resulted in an accuracy of 0.62.

Results from using the 2016 data with the 2017 model were notably different. The coefficient of determination was substantially lower ($r^2 = 0.229$), indicating a poor relation between predicted and observed *E. coli* levels, but the accuracy was higher (0.79), indicating greater predictive power with the 2016 data compared with the 2015 data. The basis of this paradox is that the predictive power of a model is more dependent on low percentages of false positives and negatives (thus high accuracy) than the amount of variation (scatter) around the least-squares regression line.

2015–17 Beach Seasons Combined

Although combining data from multiple years to generate a predictive model is generally not endorsed, the intent in this case was to determine if certain independent variables could be identified that have a strong correlation to *E. coli* if a larger dataset were evaluated. The resulting dataset for the combined years represented about 150 records of *E. coli* values with accompanying independent variables. Of the 10 best models generated in Virtual Beach under the constraint of the PRESS evaluation criterion and a maximum of five independent variables, 7 models contained a characterization of Wind, and all models contained Mid_Season, Air_T_3h (air temperature, measured 3 hours before *E. coli* sample collection), and Geese (albeit Geese varied only during 2016). It is notable that air temperature was a predominant independent variable in the models; *E. coli* was weakly correlated with Air_T_3h (r = 0.309), whereas *E. coli* was significantly correlated with Visitors (r = 0.373). The correlation between Air_T and Visitors was also significant and relatively high (r = 0.496), indicating the two independent variables varied similarly with *E. coli* levels.

Among the 10 best-fit models, accuracy ranged from 0.74 to 0.81, but the highest sensitivity value was only 0.29. When models were generated with sensitivity as the evaluation criterion, there was no increase in the sensitivity value. However, specificity in all models was high, with values greater than 0.91, regardless of which evaluation criterion was used (PRESS or sensitivity). Because the 2016 *E. coli* data likely were influenced by the removal of the goose population, different models were generated using only the 2015 and 2017 data and using the PRESS evaluation criterion. The predictive power of these models generally improved, especially values of sensitivity. The cross-validation procedure indicated that one of the most reliable models (MSEP = 0.26) had an accuracy of 0.70, specificity of 0.85, and sensitivity of 0.33, which was the highest among the models generated with the 2015–17 dataset (fig. 11). The independent variables in this model were Mid_Season, Visitors, Wind_O, and Air_T (the last two variables measured on the same day as *E. coli* sample collection). The equation for correlating *E. coli* based on the data from the 2015–17 beach seasons is as follows:

$$E.\,coli = -3.07 + (0.02037 \times Mid\_Season) + (0.03795 \times Air\_T) + (0.04345 \times Wind\_O) + (0.0005244 \times Visitors). \quad (7)$$

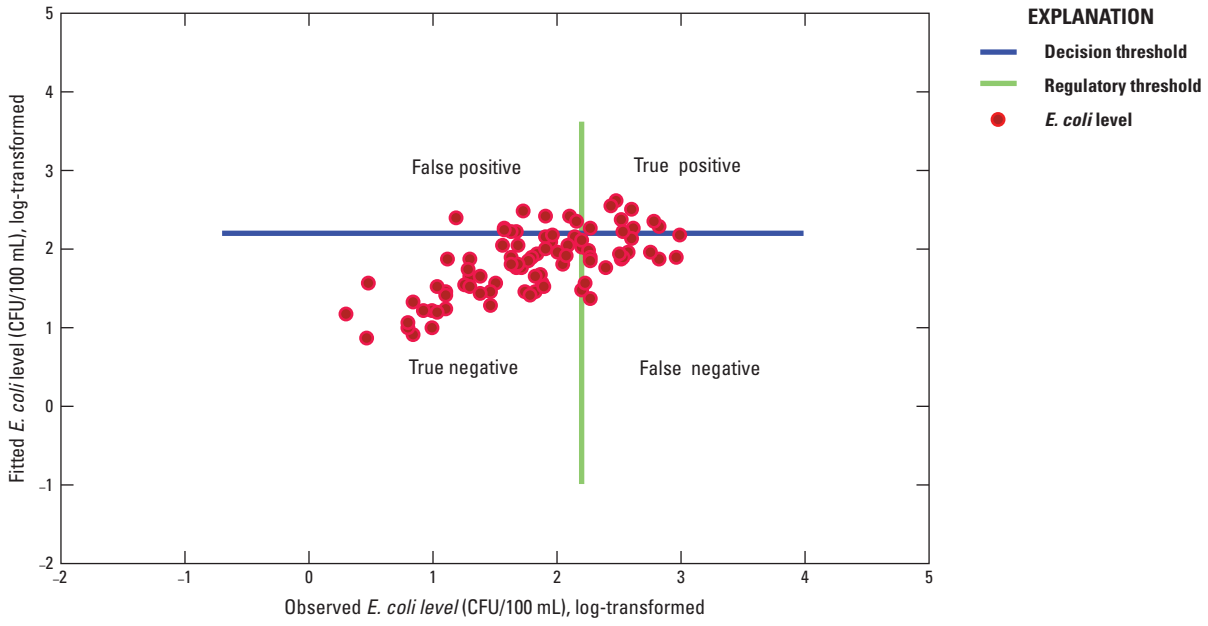

Figure 11. Scatterplot of fitted versus observed values of *Escherichia coli* (*E. coli*) levels at Pawtuckaway State Park Beach in Nottingham, New Hampshire, based on a model generated with beach data from 2015 to 2017 and incorporates independent variables Mid_Season, Air_T, Wind_O, and Visitors. The scales for the vertical and horizontal axes indicate logarithmically (log_{10}) transformed *E. coli* levels; the State advisory threshold of 158 CFU/100 mL translates to 2.2 when log-transformed. *E. coli* levels greater than the horizontal decision threshold are exceedances fitted by the model, and levels greater than the vertical regulatory threshold are exceedances observed in samples.

Discussion

The results of this study provide insights into possible factors associated with high levels of *E. coli* at Pawtuckaway Beach during beach seasons. Compared with more densely developed settings, the heavily forested watershed around Pawtuckaway Beach would likely suggest different mechanisms of *E. coli* contamination. Drainage from Burnhams Marshes, near the northern end of the beach, was initially considered to be a potential source, but results from this study did not support that premise. In reviewing data from 1993 to 2017, where *E. coli* samples were collected concurrently at three locations along the beach, levels in the right-side samples (nearest Burnhams Marshes) were no higher than in samples from the center and left locations (fig. 3). Furthermore, results from the supplemental investigation conducted by the EPA indicated that *E. coli* was not detected in water from Burnhams Marshes.

Low levels of *E. coli* were present in the other water samples collected along the beach for the EPA investigation, but a highly relevant finding was the absence of acetaminophen in any of the samples collected by the EPA, indicating the source of *E. coli* was likely not anthropogenic. As a comparison with the investigation at Pawtuckaway Beach, the EPA conducted a similar investigation later in the day at Silver Lake State Park in New Hampshire, which has a more highly developed watershed; acetaminophen was present in the three water samples collected from Silver Lake, indicating the source was likely anthropogenic (Todd Borci, EPA, written commun., August 11, 2015). The additional finding from the EPA investigation that pore water at Pawtuckaway Beach had higher *E. coli* levels than lake water indicated that pore water could be a reservoir for viable bacteria, even when the source is not immediately present. Other studies have described similar results, as well as indications that sediments below the shallow waters of the beach can function similarly (Whitman and Nevers, 2003; Ishii and others, 2007; Halliday and Gast, 2011). Thus, it is probable that the nearshore sediments at Pawtuckaway Beach could act as a reservoir for *E. coli*, which would become suspended in the water column when disturbed by human or weather-related activities.

The way in which *E. coli* levels varied during the beach season resulted in a distinctive parabolic pattern, except in 2016 when a general decline in *E. coli* levels occurred during the season (fig. 4). The removal of geese from the beach on June 30, 2016, likely was related to this different pattern by eliminating a primary source of *E. coli*. The parabolic function applied to the data indicated that, during a typical beach season, *E. coli* levels would increase until July 22, and begin decreasing afterwards; consequently, the derived variable Mid_Season (with July 22 representing the maximum value) was an important factor in explaining how *E. coli* levels varied during the 2016 beach season (fig. 5). In addition to the likelihood that geese contributed to high *E. coli* levels, the independent variable Visitors was also a likely factor based on the similarity in how *E. coli* levels increased with the Mid_Season and Visitors that suggests collinearity (fig. 6). For the 3 years of the study, Visitors generally increased between the start of the beach season (mid-June) and late in June, after which levels were generally constant until declining in August. Explanations for why *E. coli* levels began declining in late June at Pawtuckaway Beach are unclear, but a sudden decline in visitors appeared not to be a factor. Perhaps a combination of increased beach traffic by visitors and development of goslings would encourage geese to disperse further from the beach, but this conjecture was not investigated.

Results indicated that wind generally from the southwestern quadrant of the compass was related to increased *E. coli* levels (fig. 7). Pawtuckaway Beach is at the eastern end of a cove in a north-to-south alignment, but the cove has a southwestern orientation (fig. 1). Thus, winds associated with high *E. coli* levels were generally into the cove, which is consistent with observations by Pawtuckaway State Park staff that high *E. coil* levels often occurred when visitor numbers were high, and winds were blowing into the cove. Plausible mechanisms that could explain the effect of wind on *E. coli* levels were not investigated in this study, but one possibility is that wind blowing along or onto the beach could contribute to suspending *E. coli*-laden sediments (Francy and others, 2013a). Additionally, the geomorphology of the beach and cove might also contribute to elevated *E. coli* levels, so that winds blowing into the cove could confine waters with *E. coli* to the beach area rather than produce currents conducive to their dispersal.

In evaluating associations between *E. coli* levels and independent variables characterizing beach conditions, relatively strong collinearity was evident among several factors related to progression of the beach season, similar in form to the changes in *E. coli* levels during the beach seasons (fig. 4). Recognizing and interpreting such relations among variables was important because significant correlations could result between *E. coli* levels and independent variables that were not likely processes linked with causality. In all 3 years of the study, incremental changes occurred during the beach season in specific conductivity, DO, pH, lake stage, and water and air temperature. Specific conductance data from 2015 are an example, which had significant correlations with *E. coli* ($r = -0.446$) and Mid_Season ($r = -0.658$); a difference in conductivity of 5 µS/cm during the season was not convincing evidence that a decrease in conductivity was related to an increase in *E. coli* levels. Conversely, an indirect causality between an independent variable and *E. coli* levels is likely to exist in some cases, as in the positive correlations that air temperature had in 2017 with *E. coli* levels and Visitors. Rather than conclude that *E. coli* is responding to an increase in air temperature, a more likely explanation is that warmer weather increases beach use, which in turn promotes increased levels of *E. coli* through mechanisms such as disturbance of sediments.

The multiple linear regression models generated with Virtual Beach for 2015 and 2017 were relatively successful at explaining much of the variation in *E. coli* levels (figs. 8 and 11), but the 2016 model was less successful (fig. 10),

presumably because geese removal from the beach in 2016 was the major factor affecting *E. coli* levels that year. All models were generally effective at predicting true negatives, with specificity values ranging from 0.69 to 1. Conversely, models did not perform particularly well at predicting true positives, with sensitivity values generally below 0.50 (indicating *E. coli* exceedances would be predicted correctly only half of the time). The 2016 model, when tested with the 2017 data, was an exception where the sensitivity value was relatively high (0.88); however, the accuracy of this model was the lowest among all models because it tended to overestimate the likelihood that *E. coli* levels were above the exceedance threshold. Thus, the results of the model were "overprotective" in the sense that predicted *E. coli* levels were potentially unsafe even though the observed levels were below the exceedance threshold.

Although the 2015–7 best-fit models were only marginally successful at correctly predicting exceedances (accuracy ranged from 0.51 to 0.92), the independent variables that consistently were most significant in these models provide insight into environmental conditions associated with exceedances in *E. coli* levels. Mid_Season is a latent independent variable in the models that generally characterizes progression of the beach season as related to the pattern in *E. coli* levels that occurred in most years. Between 2011 and 2015, *E. coli* was sampled 16 times during the week centered on June 22 (Mid_Season), and of these samples, all had concentrations above 100 CFU/100 mL and 14 of these measurements (88.5 percent) exceeded the State advisory level of 158 CFU/100 mL (fig. 5*A*); thus, accuracy in predicting exceedances was strongly dependent on the day of the beach season. A different relation occurred in 2016 (fig. 4*C*), however; a decline in *E. coli* levels generally coincided with the Geese variable, which was a significant factor only applicable in the 2016 model. Wind and Visitors were additional independent variables significant in the models, with the exception of Visitors in the 2016 model, and these two factors likely characterized sediment agitation that suspended *E. coli* in the water column.

Models generated with the Virtual Beach software in similar USGS studies, such as that for the Great Lakes beaches (Francy and others, 2013a), often predicted *E. coli* levels more reliably than did models for the current study. Of the different modeling studies, the difference in the prediction accuracy of *E. coli* exceedances is possibly the result of at least two factors: absence of definitive point sources and the temporary removal of a potential factor that contributed to the contamination. At beaches where localized point sources (such as storm sewers) can contribute to high bacteria levels, reliably predicting exceedances might primarily depend on modeling relatively simple factors, such as quantifying rainfall intensity and duration. For the Pawtuckaway Beach study, however, the goose droppings represented a "nonpoint source" in that they were dispersed somewhat indiscriminately around the beach; thus, predicating exceedances at Pawtuckaway Beach was more complex than could be achieved through models using simple relations between *E. coli* levels and precipitation, for example, and exceedances likely were related to additional environmental variables that were not quantified extensively in this study. In addition, removal of the geese at Pawtuckaway Beach at the beginning of the 2016 beach season presumably reduced *E. coli* levels so that, even when environmental conditions were conducive for elevated levels, the reduced source of *E. coli* in essence voided the other factors in the models.

Regardless of the limitations of the Pawtuckaway Beach models to predict *E. coli* exceedances, the results of the study provided evidence of several factors that were related to high levels of the bacteria at certain times during the season. The migratory goose population that arrived prior to the start of the beach season was recognized to be the primary source of *E. coli*. The levels of *E. coli* generally increased as the goose population grew (hatching of eggs) and continued as the goslings matured during the beach season; these events generally coincided with an increase in the number of beach visitors as the season progressed. Nearshore sediments with high levels of *E. coli* derived from the geese could become suspended in the water column by activities of the visitors and possibly remain in suspension if water turbulence remained high because of winds across or towards the beach. The model generated from the combined dataset for the 3 years, although not intended to be predictive, confirmed that these independent variables were likely relevant to high *E. coli* levels at Pawtuckaway Beach.

Summary

Pawtuckaway State Park Beach in Nottingham, New Hampshire, historically has experienced *Escherichia coli* (*E. coli*) levels that exceeded a State-defined threshold of 158 colony forming units per 100 milliliters (CFU/100 mL), which results in the posting of a health advisory that indicates potentially unsafe conditions for swimmers. A population of geese migrated to and from the beach vicinity each spring and fall to nest and raise goslings and thus was present during the summer when the number of visitors to the beach was highest. However, the extent that the geese were a primary source of the *E. coli* at the beach was unknown. Consequently, the New Hampshire Department of Health and Human Services (NHDHHS) and the New Hampshire Department of Environmental Services (NHDES) were interested in identifying ways to reduce health risks to beachgoers and to better understand causes of high *E. coli* levels at the beach. The two agencies collaborated with the U.S. Geological Survey (USGS) to investigate if certain environmental factors were related to high *E. coli* levels and if models could predict when levels exceed the advisory threshold.

From 2015 through 2017, the USGS investigated the occurrences of high levels of *E. coli* at Pawtuckaway Beach in cooperation with the NHDHHS and NHDES by coordinating an extensive data-collection effort. At the beach and the nearby wetland Burnhams Marshes, the USGS collected water-quality

and meteorological data during the summer seasons for the 3 years; these data are available to the public through the USGS National Water Information System (NWIS) database. During these three beach seasons, the NHDES also collected *E. coli* samples three or four times per week from the beach as part of the Beach Inspection Program. The NHDHHS used the USGS and NHDES data to develop a beta version of a near real-time decision-making tool that could inform the public of conditions at the beach, including water quality, weather conditions, and predicted *E. coli* levels; the tool relied on models generated by the NHDHHS using the Virtual Beach software program, and the accuracy and precision of the models were evaluated by comparing the predicted *E. coli* values with the observed *E. coli* values provided by the NHDES.

The NHDES had sampled *E. coli* levels at Pawtuckaway Beach since 1993, and these historical data were evaluated to identify any possible trends in *E. coli* levels over the beach season. Periodically during routine site visits, separate *E. coli* samples were collected at the left, center, and right sides of the beach. Burnhams Marshes is adjacent to the right side of the beach, and the possibility that its discharge was contributing to high *E. coli* levels was considered; however, a statistical comparison of *E. coli* levels from the three sampling locations indicated that samples collected from the right side of the beach were no higher than at the other two locations. Additional analysis of the historical data indicated that a seasonal trend occurred where *E. coli* levels tended to increase with the early progression of the beach season until about June 22 (about midseason) and levels decreased afterward. Consequently, "Mid_Season" was an independent variable created to represent this beach-season succession that increased progressively up to June 22 and decreased progressively afterwards.

The USGS analyzed the datasets from the 3 years to further explore associations between the environmental data (collected by the USGS) and the *E. coli* data (collected by the NHDES). With the use of Virtual Beach, statistical models were developed for each of the 3 years to identify which independent variables were most strongly associated with high *E. coli* levels for the respective year. In addition, the data from all 3 years were consolidated and analyzed with Virtual Beach to determine if a larger, more comprehensive dataset might indicate stronger associations between *E. coli* levels and the independent variables than would datasets from a single year.

For the 2015 beach season, the pattern of rising and falling *E. coil* levels was observed; among models generated with Virtual Beach, the most significant independent variable selected for the most robust 2015 model was Mid_Season (described above), followed by the number of visitors (representing beach use by people) and a wind vector (wind speed and direction). In 2016, the goose population was removed from the beach vicinity early in the season by the NHDES, and subsequently the previously observed rise and fall pattern in *E. coli* levels was not observed that year. The most significant independent variable selected for the most robust 2016 model was a categorical variable that represented removal of the goose population in 2016 only, followed by wind speed and water temperature. Geese returned to Pawtuckaway Beach for the 2017 season, but the population was smaller than in previous years. The rise and fall pattern in *E. coli* levels reoccurred, but the highest levels were notably lower than previously observed, with fewer exceedances of the advisory threshold. The most significant independent variable selected for the most robust 2017 model was Mid_Season, followed by number of visitors and a wind vector.

Overall, the results of the study indicated that the goose population likely was a primary source of *E. coli* contamination at Pawtuckaway Beach. This finding was supported by a study by the U.S. Environmental Protection Agency to assess *E. coli* levels at various locations along Pawtuckaway Beach and to determine whether acetaminophen was coincidental in samples where *E. coli* occurred. Compared with the lake water, pore water at the beach had higher levels of *E. coli*, suggesting that the pore water could be a reservoir for the bacteria. In addition, acetaminophen was not identified in any of the samples, suggesting that the source of *E. coli* was non-human. The number of visitors and wind were two additional factors that appeared to contribute to elevated *E. coli* levels; a probable explanation is that beachgoers and wind could contribute to suspension of bacteria-laden sediment, and also that winds from a southerly to westerly direction could inhibit water exchange at the beach.

Selected References

Carlson, S., 2012, NHDES beach program—Generic quality assurance project plan: New Hampshire Department of Environmental Services NHDES–WD–13–03, 93 p., accessed January 24, 2019, at https://www.des.nh.gov/organization/divisions/water/wmb/beaches/documents/2012-beach-qapp.pdf.

Coles, J.F., and Bush, K.F., 2019, Data collected at Pawtuckaway Beach in Nottingham, New Hampshire, 2015–2017, including data from *Escherichia coli* (bacteria) samples, and from USGS meteorological and water quality stations. U.S. Geological Survey data release, https://doi.org/10.5066/P9KUIT3W.

Crawford, G.H., 1967, New Hampshire state parks, historic sites, wayside picnic areas: Concord, N.H., New Hampshire State Parks, 9 p.

Cyterski, M., Brooks, W., Galvin, M., Wolfe, K., Carvin, R., Roddick, T., Fienen, M., and Corsi, S., 2014, Virtual beach 3—User's guide: U.S. Environmental Protection Agency EPA/600/R–13/311, 86 p. [Also available at https://www.epa.gov/sites/production/files/2014-02/documents/vb3_manual_final_-_revised.pdf.]

Francy, D.S., and Darner, R.A., 1998, Factors affecting *Escherichia coli* concentrations at Lake Erie public bathing beaches: U.S. Geological Survey Water-Resources Investigations Report 98–4241, 41 p. [Also available at https://doi.org/10.3133/wri984241.]

Francy, D.S., and Darner, R.A., 2006, Procedures for developing models to predict exceedances of recreational water-quality standards at coastal beaches: U.S. Geological Survey Techniques and Methods, book 6, chap. B5, 34 p. [Also available at https://doi.org/10.3133/tm6B5.]

Francy, D.S., Brady, A.M.G., Carvin, R.B., Corsi, S.R., Fuller, L.M., Harrison, J.H., Hayhurst, B.A., Lant, J., Nevers, M.B., Terrio, P.J., and Zimmerman, T.M., 2013a, Developing and implementing predictive models for estimating recreational water quality at Great Lakes beaches: U.S. Geological Survey Scientific Investigations Report 2013–5166, 68 p. [Also available at https://doi.org/10.3133/sir20135166.]

Francy, D.S., Stelzer, E.A., Duris, J.A., Brady, A.M.G., Harrison, J.H., Johnson, H.E., and Ware, M.W., 2013b, Predictive models for *Escherichia coli* concentrations at inland lake beaches and relationship of model variables to pathogen detection: Applied and Environmental Microbiology, v. 79, no. 5, p. 1676–1688. [Also available at https://doi.org/10.1128/AEM.02995-12.]

Halliday, E., and Gast, R.J., 2011, Bacteria in beach sands—An emerging challenge in protecting coastal water quality and bather health: Environmental Science and Technology, v. 45, no. 2, p. 370–379. [Also available at https://doi.org/10.1021/es102747s.]

Ishii, S., Hansen, D.L., Hicks, R.E., and Sadowsky, M.J., 2007, Beach sand and sediments are temporal sinks and sources of *Escherichia coli* in Lake Superior: Environmental Science and Technology, v. 41, no. 7, p. 2203–2209. [Also available at https://doi.org/10.1021/es0623156.]

Kirschner, A.K.T., Zechmeister, T.C., Kavka, G.G., Beiwl, C., Herzig, A., Mach, R.L., and Farnleitner, A.H., 2004, Integral strategy for evaluation of fecal indicator performance in bird-influenced saline inland waters: Applied and Environmental Microbiology, v. 70, no. 12, p. 7396–7403. [Also available at https://doi.org/10.1128/AEM.70.12.7396-7403.2004.]

Kleinheinz, G.T., McDermott, C.M., Hughes, S., and Brown, A., 2009, Effects of rainfall on *E. coli* concentrations at Door County, Wisconsin beaches: International Journal of Microbiology, v. 2009, article 876050, 9 p., accessed March 29, 2019, at https://doi.org/10.1155/2009/876050.

New Hampshire Department of Environmental Services, 2017, Beach inspection program: New Hampshire Department of Environmental Services website, accessed December 2017 at https://www.des.nh.gov/organization/divisions/water/wmb/beaches/index.htm.

New Hampshire Department of Environmental Services, 2018, OneStop—Beaches: New Hampshire Department of Environmental Services dataset, accessed April 9, 2019, at https://www4.des.state.nh.us/DESOnestop/BCHDetail.aspx?ID=221.

New Hampshire Department of Health and Human Services, 2016, DHHS mission statement: New Hampshire Department of Health and Human Services web page, accessed April 9, 2019, at https://www.dhhs.nh.gov/about/mission.htm.

Sampson, R.W., Swiatnicki, S.A., Mcdermott, C.M., and Kleinheinz, G.T., 2006, The effects of rainfall on *Escherichia coli* and total coliform levels at 15 Lake Superior recreational beaches: Water Resources Management, v. 20, no. 1, p. 151–159. [Also available at https://doi.org/10.1007/s11269-006-5528-1.]

Systat Software, Inc., 2007, SYSTAT 12: Systat Software, Inc. website, accessed June 12, 2008, at https://systatsoftware.com/products/systat/.

U.S. Environmental Protection Agency, 2010, EPA microbiological alternate test procedure (ATP) protocol for drinking water, ambient water and wastewater monitoring methods: U.S. Environmental Protection Agency EPA–821–B–10–001, 94 p. [Also available at https://www.epa.gov/sites/production/files/2015-09/documents/micro_atp_protocol_sept-2010.pdf.]

U.S. Geological Survey, 2010, Understanding beach health throughout the Great Lakes—Entering a new era of investigations: U.S. Geological Survey Fact Sheet 2010–3093, 4 p. [Also available at https://doi.org/10.3133/fs20103093.]

U.S. Geological Survey, 2013, Real-time assessments of water quality—Expanding nowcasting throughout the Great Lakes: U.S. Geological Survey Fact Sheet 2013–3069, 4 p., accessed March 14, 2017, at https://doi.org/10.3133/fs20133069.

U.S. Geological Survey, 2017a, USGS 01073389 Pawtuckaway Lake near Nottingham, NH, *in* Water data for the nation: U.S. Geological Survey National Water Information System database, accessed April 10, 2019, at https://doi.org/10.5066/F7P55KJN. [Site information directly accessible at https://waterdata.usgs.gov/nh/nwis/inventory/?site_no=01073389.]

U.S. Geological Survey, 2017b, USGS 430508071091801 Pawtuckaway meteorologic station near Nottingham, NH, *in* Water data for the nation: U.S. Geological Survey National Water Information System database, accessed April 10, 2019, at https://doi.org/10.5066/F7P55KJN. [Site information directly accessible at https://waterdata.usgs.gov/nh/nwis/uv/?site_no=430508071091801.]

U.S. Geological Survey, 2017c, USGS 430509071092201 Burnhams Marshes near Nottingham, NH, *in* Water data for the nation: U.S. Geological Survey National Water Information System database, accessed April 10, 2019, at https://doi.org/10.5066/F7P55KJN. [Site information directly accessible at https://waterdata.usgs.gov/nh/nwis/uv/?site_no=430509071092201.]

U.S. Geological Survey, 2018, NowCast status: U.S. Geological Survey website, accessed April 10, 2019, at https://ny.water.usgs.gov/maps/nowcast/.

Wall, P.A., 2015, Together at last—Exploring health and environmental information on the National Environmental Health Tracking Network: Journal of Environmental Health, v. 78, no. 3, p. 34–36. [Also available at https://www.questia.com/library/journal/1G1-430892345/together-at-last-exploring-health-and-environmental.]

Warden, P.S., DeSarno, M.S., Volk, S.E., and Eldred, B.J., 2011, Evaluation of colilert-18 for detection and enumeration of fecal coliform bacteria in wastewater using the U.S. Environmental Protection Agency alternative test procedure protocol: Journal of AOAC International, v. 94, no. 5, p. 1573–1580. [Also available at https://doi.org/10.5740/jaoacint.11-114.]

Whitman, R.L., and Nevers, M.B., 2003, Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan beach: Applied and Environmental Microbiology, v. 69, no. 9, p. 5555–5562. [Also available at https://doi.org/10.1128/AEM.69.9.5555-5562.2003.]

Appendix 1. The Virtual Beach Modeling Tool

Virtual Beach is a software package originally developed by Frick and others (2008) at the U.S. Environmental Protection Agency Ecosystems Research Division in Athens, Georgia, for building site-specific statistical models to help predict pathogen indicator levels (for example, *Escherichia coli* [*E. coli*]) at recreational beaches. Data requirements are fairly basic for building a predictive model with Virtual Beach, typically consisting of a time-series of data collected during a summer season at the recreational beach of interest, a single dependent "response" variable such as routinely measured *E. coli* levels, and a set of independent variables that characterize environmental conditions at the beach site during the season. The original, version 1 of the Virtual Beach application (VB_1) was designed as a model-building tool that was based primarily on manual analyses of datasets with the use of a multiple linear regression; candidate models were developed through visual inspections of data plots and manipulation of environmental variables (for example, transformations, creating interaction terms), followed by an iterative process of testing and evaluating models. The fitness of models was computed and tracked to allow direct comparisons in order to select a final "best" model that could then estimate fecal indicator levels at a beach by using current or forecasted environmental data for the site.

Version 2 (VB_2) was released in 2010 to provide enhanced utility of basic program functions (for example, visual inspection of data plots, manual transformations of variables, multiple linear regression model building, prediction), but VB_2 also automated and extended functionality in several ways (Cyterski and others, 2014). Among the improvements was an interactive map component that allowed users to locate their site, define the orientation of the beach, and examine nearby potential data sources. The map component integrated the beach location and its geographic orientation (compass bearing) into the program by the user manually delineating the beach outline onto a digital image, available through web-mapping services. The geographic orientation of the beach could then be used with water current and meteorological data to calculate wave, current, and wind vectors that are often important derived environmental variables for predicting *E. coli* levels. Additionally, the map component could identify locations and the availability of local data sources that provide real-time or forecasted data that could be used in the multiple linear regression model to generate predictions; sources of these data included the U.S. Geological Survey National Water Information System (NWIS) and the National Oceanic and Atmospheric Administration National Centers for Environmental Information.

Version 3 of Virtual Beach (VB_3) was developed in conjunction with the USGS at the time of this study and was released in 2013 (Cyterski and others, 2014). It was built on VB_2 by adding additional statistical methods that give users greater flexibility in modeling techniques; in addition to creating multiple linear regression models, VB_3 allows users the option to use partial least squares regression and generalized boosted regression modeling to fit datasets and make predictions. The Pawtuckaway Beach investigation described in this report used VB_3 for modeling the relation between *E. coli* levels and environmental data collected at the beach. The basic operation of VB_3 relies on four main functions that each has a user interface, and is summarized below from Cyterski and others (2014):

Location map.—VB_3 provides a mapping and geographic information system interface for calculating the beach orientation used for computation of orthogonal (alongshore and onshore/offshore) wind, current, or wave vectors for the beach site. Mapping is integrated with web-mapping services.

Global datasheet.—The interface facilitates the import and manipulation of *E. coli* (response variable) and environmental (independent variables) data. In addition to generating wind, current, and wave vectors, data transformations of independent variables can be made, and derived independent variables can be generated that represent the products, means, sums, differences, minimums, and maximums of environmental data.

Methods.—Three options are available for building models: multiple linear regression, partial least squares regression, and generalized boosted regression modeling. Each has its own interface but shares common elements such as a variable-selection tab that allows users to select independent variables for consideration in building models.

Prediction.—The interface allows users to enter or import the independent variable data needed to run the selected model and examine the model predictions and exceedance probabilities. The interface also provides a means of evaluating model performance with the use of time-series and scatterplots of measured response variables (for example, actual *E. coli* values) versus the predicted values.

Interpreting Results

Models are generated by the Virtual Beach software program with the use of an initial, or "training," dataset that contains *E. coli* values and corresponding independent variable data that characterize environmental conditions at a beach. Scatterplots are created by the program to indicate the potential performance of models by displaying the observed (or actual) *E. coli* values on the horizontal (X) axis against the fitted (or estimated) *E. coli* values on the vertical (Y) axis *E. coli* (fig. 1.1). The position of a given data point along the horizontal axis represents the observed *E. coli* value reported in a sample. The position of this data point along the vertical

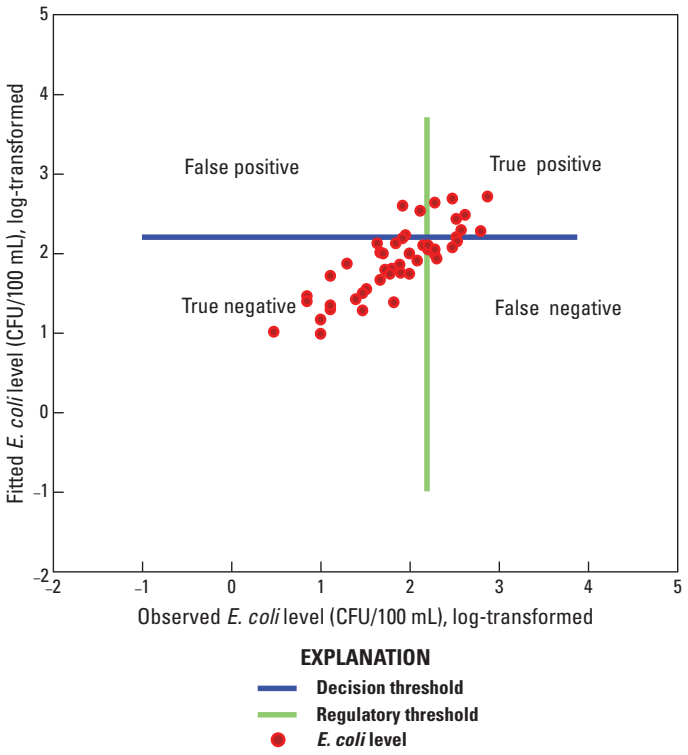

Figure 1.1. Example of a scatterplot created with the Virtual Beach software that indicates the potential of a model to predict *Escherichia coli* (*E. coli*) values from independent variables of environmental data. An initial model is generated from a dataset of *E. coli* values and corresponding independent variables collected during a season. The scales of the scatterplot axes indicate logarithmically (log_{10}) transformed *E. coli* levels; the New Hampshire State advisory level of 158 colony forming units per 100 milliliters (CFU/100 mL), which translates to 2.2 when log-transformed, indicates exceedance thresholds on both axes. Exceedances greater than the decision threshold are values fitted by the model, and exceedances greater than the regulatory threshold are values observed in *E. coli* samples. The distribution of *E. coli* values among the four quadrants formed by the intersection of the thresholds indicates the potential of a model to accurately predict exceedances.

axis represents an estimate of the *E. coli* value as fitted by the model, based on how the model would predict the value from the independent variable data that corresponded to the observed value. Thus, a perfect-fit model would result in a series of points in a straight line with a slope of 1. It is important to note that the fitted *E. coli* values are not strictly "predicted" because only training data are used to generate the model and the values display in the figure; however, predicted *E. coli* values are calculated and displayed when a model is used with additional or "testing" data (example in fig. 9, main text).

An exceedance threshold is displayed on both axes of figures generated by Virtual Beach, indicated by intersecting vertical and horizontal lines that divide the scatterplot into four quadrants. The vertical line represents a "regulatory threshold" that is set by the user (for example, 158 colony forming units per 100 milliliters [CFU/100 mL]); the X-axis values of data points to the right of this line indicate actual "observed" *E. coli* values that exceed the regulatory threshold. The horizontal line represents the "decision threshold," which corresponds to the same value of the regulatory threshold; the Y-axis values of data points above this line are the *E. coli* values fitted by the model from the independent variable data that characterized the observed values. Therefore, when using models to predict *E. coli* levels at a beach, values above the decision threshold can be evaluated by managers in making decisions about issuing an advisory for the beach.

The quadrants formed from the intersecting lines of the regulatory and decision thresholds each have a specific relevance for interpreting figures in Virtual Beach. The lower left quadrant represents true negatives (TN) where both predicted and observed values are below the exceedance threshold; the upper right quadrant represents true positives (TP) where both predicted and observed values are above the threshold. The sum of values that are displayed in these two quadrants (TP and TN) represent observed *E. coli* values that would have been correctly predicted by the particular model. The upper left quadrant represents false positives (FP) where predicted values are above the exceedance threshold but observed values are below the threshold. The lower right quadrant represents false negatives (FN) where predicted values are below the exceedance threshold but observed values are above the threshold. The sum of values that are displayed in these two quadrants (FP and FN) represent observed *E. coli* values that would have been incorrectly predicted by the particular model.

Evaluation statistics for models include specificity (eq. 1.1), sensitivity (eq. 1.2), and accuracy (eq. 1.3), as follows:

$$\frac{TN}{TN+FP}, \tag{1.1}$$

$$\frac{TP}{TP+FN}, \text{ and} \tag{1.2}$$

$$\frac{TP+TN}{n}, \tag{1.3}$$

where n is the total number of *E. coli* values. A perfect model would result in specificity, sensitivity, and accuracy values all equal to 1.0, indicating that all predictions (both above and below the exceedance threshold) were true.

References Cited

Cyterski, M., Brooks, W., Galvin, M., Wolfe, K., Carvin, R., Roddick, T., Fienen, M., and Corsi, S., 2014, Virtual beach 3—User's guide: U.S. Environmental Protection Agency EPA/600/R–13/311, 86 p. [Also available at https://www.epa.gov/sites/production/files/2014-02/documents/vb3_manual_final_-_revised.pdf.]

Frick, W.E., Ge, Z., and Zepp, R.G., 2008, Nowcasting and forecasting concentrations of biological contaminants at beaches—A feasibility and case study: Environmental Science & Technology, v. 42, no. 13, p. 4818–4824. [Also available at https://doi.org/10.1021/es703185p.]

For more information, contact
Director, New England Water Science Center
U.S. Geological Survey
331 Commerce Way, Suite 2
Pembroke, NH 03275
https://www.usgs.gov/centers/new-england-water
dc_nweng@usgs.gov

Publishing support provided by the
Pembroke Publishing Service Center

ISSN 2328-0328 (online)
https://doi.org/10.3133/sir20195111