Environ Monit Assess
DOI 10.1007/s10661-007-9883-z

Quantifying sample biases of inland lake sampling programs in relation to lake surface area and land use/cover

Tyler Wagner · Patricia A. Soranno · Kendra Spence Cheruvelil · William H. Renwick · Katherine E. Webster · Peter Vaux · Robbyn J. F. Abbitt

Received: 24 May 2007 / Accepted: 26 July 2007

Abstract We quantified potential biases associated with lakes monitored using non-probability based sampling by six state agencies in the USA (Michigan, Wisconsin, Iowa, Ohio, Maine, and New Hampshire). To identify biases, we compared state-monitored lakes to a census population of lakes derived from the National Hydrography Dataset. We then estimated the probability of lakes being sampled using generalized linear mixed models. Our two research questions were: (1) are there systematic differences in lake area and land use/land cover (LULC) surrounding lakes monitored by state agencies when compared to the entire population of lakes? and (2) after controlling for the effects of lake size, does the probability of sampling vary depending on the surrounding LULC features? We examined the biases associated with surrounding LULC because of the established links between LULC and lake water quality. For all states, we found that larger lakes had a higher probability of being sampled compared to smaller lakes. Significant interactions between lake size and LULC prohibit us from drawing conclusions about the main effects of LULC; however, in general lakes that are most likely to be sampled have either high urban use, high agricultural use, high forest cover, or low wetland cover. Our analyses support the assertion that data derived from non-probability-based surveys must be used with caution when attempting to make generalizations to the entire population of interest, and that probability-based surveys are needed to ensure unbiased, accurate estimates of lake status and trends at regional to national scales.

Keywords Lake monitoring · Assessment · Sampling · Bias · Land use · Land cover

T. Wagner (✉)
Quantitative Fisheries Center,
Department of Fisheries and Wildlife,
Michigan State University, 13 Natural Resources Building,
East Lansing, MI 48824, USA
e-mail: wagnerty@msu.edu

P. A. Soranno
Department of Fisheries and Wildlife,
Michigan State University, 13 Natural Resources Building,
East Lansing, MI 48824, USA

K. S. Cheruvelil
Lyman Briggs School of Science,
Department of Fisheries and Wildlife,
Michigan State University, East Lansing, MI 48824, USA

W. H. Renwick · R. J. F. Abbitt
Department of Geography, Miami University, Oxford, OH 45056, USA

K. E. Webster
Department of Biological Sciences, University of Maine, 04469 Orono, ME, USA

P. Vaux
Senator George J. Mitchell Center for Environmental and Watershed Research, University of Maine, Orono, ME 04469, USA

Introduction

The Clean Water Act (CWA) mandates that states assess the overall quality of their aquatic resources [Section 305(b)] and develop a prioritized list of all impaired waterbodies for management action. However, because most historical and many contemporary lake datasets collected by management agencies are based on non-probability-based (NPB) sampling, many current management decisions are based on information derived from these NPB surveys, including assessments that are mandated by the CWA. As a result of the continued use of NPB survey data, it is imperative that researchers and managers attempt to quantify potential biases of NPB survey data to better inform management decisions that are based in whole or in part on such data.

The utility of probability-based (PB) surveys for assessing the regional status and trends of aquatic resources has been well established (Olsen et al. 1999; Peterson et al. 1999). Probability sampling, where every unit (e.g., lake) has a known and nonzero probability of selection, provides a representative sample of the target population and allows for statistically valid extrapolations to the entire population of interest (e.g., all lakes in the Northeast USA, Olsen et al. 1999; Brown et al. 2005). The advantages of PB sampling have been recognized and subsequently adopted in recent times by several state and federal management agencies to meet a variety of natural resource management needs (McDonald 2000; Hayes et al. 2003), including the fulfillment of CWA mandates. The development and implementation of PB surveys is a shift away from the historical reliance on NPB surveys, which are often biased and may result in an inaccurate assessment of the target population (Peterson et al. 1998, 1999).

A few approaches are available to quantify potential biases associated with NPB survey data. A logical first approach is to compare data collected using a NPB survey to those using a PB survey (e.g., Peterson et al. 1999). While this approach is useful for comparing variables that are measured within lakes, such as nutrient concentrations, there is the obvious limitation of requiring PB-based data. There has been only one study that we know of that takes this approach. Using this approach, Peterson et al. (1999) demonstrated that the use of NPB survey data lead to an overestimate of median Secchi depth for northeast lakes. As discussed by Peterson et al. (1999), unreliable estimates from NPB survey data can have a large effect on lake management decisions. Because of the requirements for implementing this approach, as described above, there are still gaps in our knowledge regarding how representative NPB surveys are of the entire population of lakes in a given region.

A second approach, which circumvents the requirements for a PB survey and that we employ here, is to use digital map-based data to obtain information on the entire population of lakes – i.e., a complete census against which NPB datasets can be compared. While this approach is restricted to data that can be collected from maps, it does allow a complete census of lakes within a region as well as examination of variables with known effects on lake dynamics (e.g., lake surface area, land use/cover, etc.).

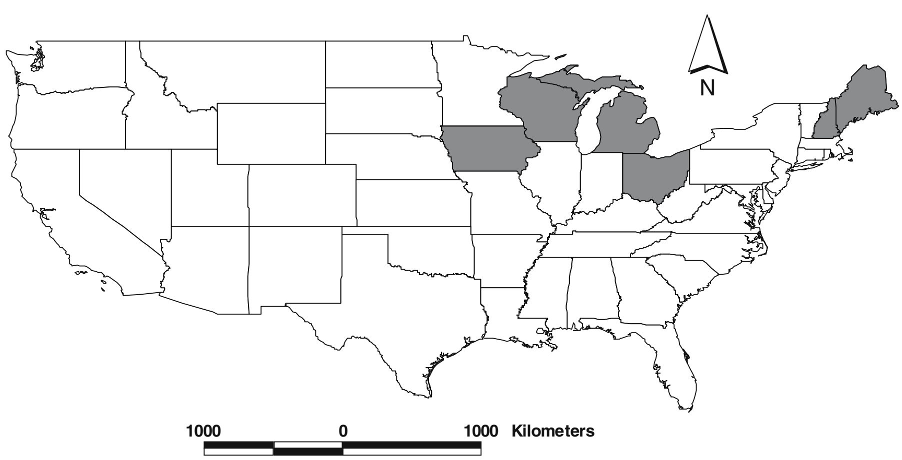

Fig. 1 Map of the USA with states used in the analysis shown in *grey*

In this study, we examine the potential biases in six state-collected NPB datasets for lake size (surface area) and land use/land cover (LULC). There exists some information that NPB sampling may be biased towards large lakes (Peterson et al. 1999); however, this study was limited to northeast US lakes and did not examine potential biases associated with LULC, which has been shown to influence water quality. Specifically we ask: (1) Are there systematic differences in lake area and LULC surrounding lakes sampled by state agencies when compared to the entire population of lakes? and (2) After controlling for the effects of lake size, does the probability of sampling vary depending on the surrounding LULC features? We have two main hypotheses. First, we hypothesize that state NPB sampling programs are biased towards lakes with larger area due to their importance as recreational and municipal water supplies; and second, we hypothesize that urban lakes would have a higher probability of being sampled due to (1) the reasons described for lake size and (2) their increased public visibility and susceptibility to pollution and colonization by invasive species compared to lakes located within natural land cover classes such as forest and wetland areas.

Study area

The study area includes all lakes and reservoirs greater than 0.1 ha surface area located within the US states of Michigan, Wisconsin, Iowa, Ohio, Maine, and New Hampshire (Fig. 1). The geographic area contains a wide range in lake and reservoir area, land use patterns, land cover, and degree of anthro-

Table 2 Percent land use/cover classes based on the 25th, 75th, and 90th percentiles and sample sizes (number of lakes) for each class

Class and land use/cover type	Land use/ cover type (%)	Sample size (n)
Agriculture (total)		
A	<0.80	13,862
B	0.81–49.0	27,705
C	49.1–77.0	8,316
D	>77.1	5,538
Agriculture (row crop)		
A	<0.25	13,825
B	0.25–<28	28,627
C	28–49	7,420
D	>49	5,550
Agriculture (pasture)		
A	<0.11	13,965
B	0.11–15	27,640
C	15–30	8,281
D	>30	5,535
Urban		
A	0.0	24,844
B	0.0–1.4	16,815
C	1.4–18	10,887
D	>18	2,876
Forest		
A	<21	13,867
B	21–76	27,680
C	76–89	8,358
D	>89	5,516
Wetland		
A	<2	13,854
B	2–21	27,718
C	21–40	8,331
D	>40	5,519

Table 1 Sample sizes (n), minimum (min) lake area in hectares, percent of lakes sampled, total area of lakes (ha) for lakes sampled by state agencies and census (NHD) lakes, and the percent of lake area sampled

State	Sampled lakes n (min)	Census lakes n (min)	Percent sampled	Total area (sampled)	Total area (census)	Percent area sampled
Michigan	470 (20.3)	17,595 (0.1)	2.7	134,105	1,581,748	8.5
Wisconsin	426 (2.2)	15,736 (0.1)	2.7	143,532	1,344,526	10.7
Ohio	102 (4.0)	5,404 (0.1)	1.9	45,849	70,664	64.9
Iowa	127 (4.1)	5,550 (0.1)	2.3	31,934	64,610	49.4
Maine	658 (0.1)	6,729 (0.1)	9.8	273,435	685,984	39.9
New Hampshire	721 (0.3)	1,905 (0.3)	37.8	59,273	88,247	67.2
Total	2,504	52,919	4.7	688,128	3,835,779	17.9

pogenic activities in the lake catchments (Table 1). New Hampshire and Maine represent largely forested states with large numbers of natural lakes. Michigan and Wisconsin have the highest density of natural lakes and a wider range of land use including more lakes with agricultural and urban land use. In contrast, Iowa and Ohio have very few natural lakes and mostly agricultural land.

Methods

Sampled lakes database

The NPB study lakes were part of baseline monitoring programs conducted by six state agencies (Michigan Department of Environmental Quality, Wisconsin Department of Natural Resources, Ohio Environmental Protection Agency, Maine Departments of Environmental Protection and Inland Fisheries and Wildlife, and New Hampshire Department of Environmental Services) and Iowa State University. The lakes were sampled predominantly between 1990 and 2003, although the earliest sample dates were in 1975.

Census lakes database

The census population of lakes was extracted from the NHDPlus Version 1.0 dataset created by the US Environmental Protection Agency (USEPA) and the US Geological Survey (USGS). This dataset is based off of a 2005 snapshot of the 1:100,000 medium resolution National Hydrography Dataset (NHD) managed by the USGS. Hydrologic data within the NHD (and subsequently in the NHDPlus) include both stream networks and waterbodies and is organized by eight-digit subbasins (formerly referred to as hydrologic units and/or catalog units). The NHD waterbodies are typically taken from 1:100,000 USGS topographic maps and the waterbody source dates are listed in the NHD metadata (http://nhd.usgs.gov/) as "Publication date of source quadrangle". The six states

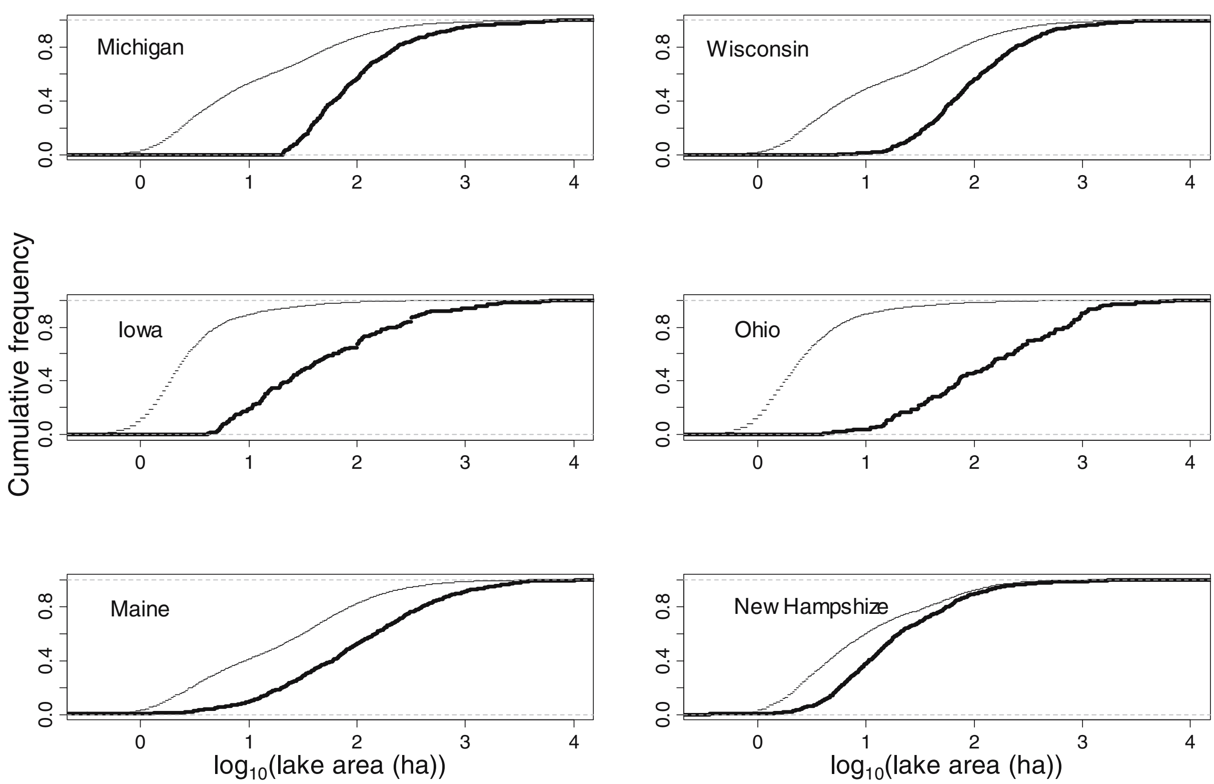

Fig. 2 Lake area cumulative distribution functions for census lakes (*light line*) and lakes sampled by state agencies (*dark line*)

in this study include all or parts of 239 eight-digit subbasins within 161 1:100,000 USGS quadrangles.

Land use/cover data

Land use/cover data were obtained from the National Land Cover Dataset (NLCD, USGS). The NLCD dataset was developed from Landsat satellite imagery captured circa 1992, and has a spatial resolution of 30 m. We examined the following LULC categories which were available in the NLCD dataset because they have relevance to lake water quality: urban, agriculture-row crop, agriculture-pasture, agriculture-total (all agriculture types combined including all other agricultural classes), wetland and forest. To quantify LULC around lakes, we created an equidistant 500 m buffer around each lake and quantified the percent LULC of each class for each lake. We chose this buffer size because it was too costly to estimate topographic catchments around the thousands of lakes included in this study, and 500 m roughly corresponds to the local catchment area (Soranno, unpublished data).

Analyses

We plotted empirical cumulative distribution functions (ECDFs) for lake area and LULC data to compare the distributions from the census population and the sampled population (NPB survey lakes). We also calculated the median value for each dataset/variable combination to examine the biases between the census population and sampled lakes.

Although ECDFs provide important descriptive information to examine differences among datasets, we also wanted to quantify the effects of LULC on the probability of a lake being sampled, while controlling for the effects of lake area. We used a generalized linear mixed model to determine if the surrounding LULC characteristics influence a lake's probability of being sampled, while controlling for the

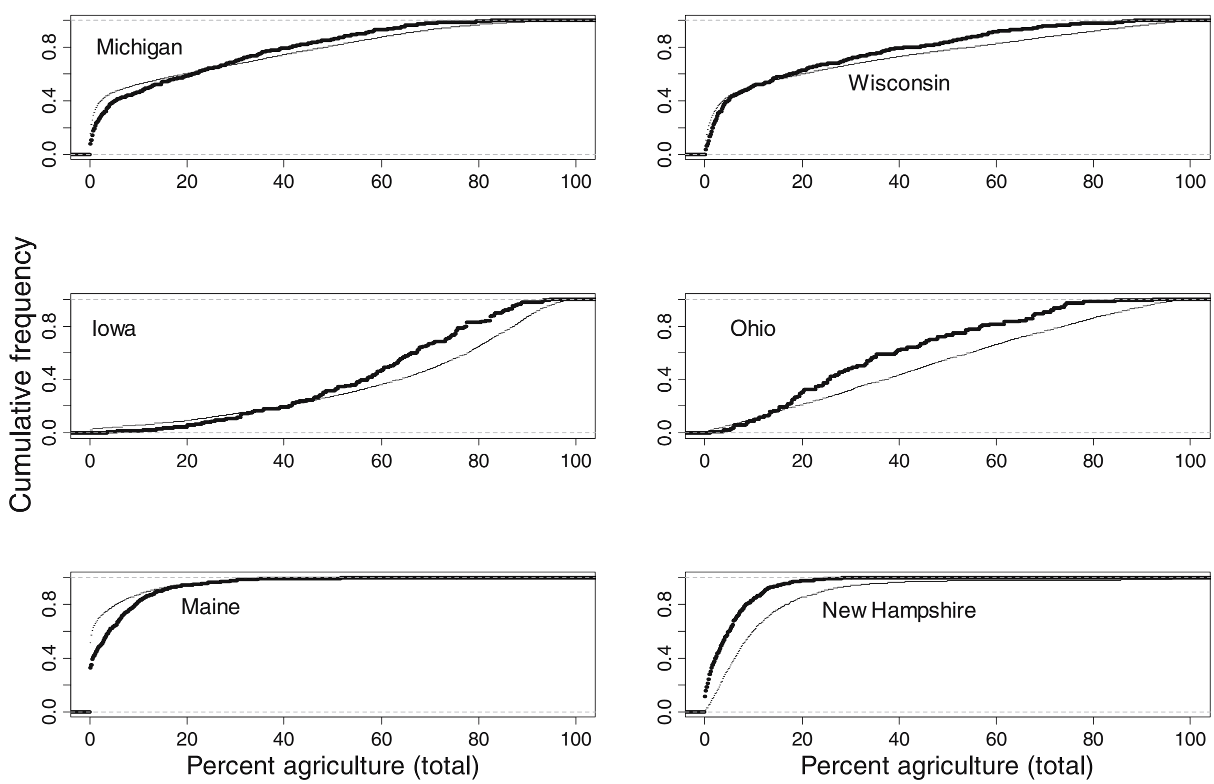

Fig. 3 Percent total agricultural land use in a 500 m buffer cumulative distribution functions for census lakes (*light line*) and lakes sampled by state agencies (*dark line*)

effect of lake size. We first coded each lake as 1 if the lake was sampled and 0 otherwise; this binary response was the dependent variable. Second, we placed all lakes into LULC classes based on the 25th, 75th, and 90th percentiles of the LULC variable of interest (i.e., for agricultural land use, the following classes are: class A = lakes with %LULC <25th percentile, class B = lakes with >25th %LULC <75th percentiles, class C = lakes with >75th %LULC <90th percentiles and class D = lakes with %LULC >90th percentile; Table 2). We used Eq. 1 to examine differences in the average probability of LULC classes being sampled while controlling for lake area.

$$\begin{aligned} \log \mathrm{it}\, P\left(Y_{ij}=1\right) = \beta_0 + \alpha_i \\ + \beta_{1,k}\, LULC + \beta_2 \log_{10}(area) \\ + \beta_{3,1}\, LULC \times \log_{10}(area) \\ + b_i \end{aligned} \tag{1}$$

Where Y_{ij} equals 1 if lake j was sampled in state i and 0 otherwise, j=1,...n_i, with n_i being the number of lakes in state i, and i=1,...6. The parameter β_0 is the fixed intercept. The state-specific random intercept effect is defined as $a_i \sim N\left(0, \sigma^2_{state}\right)$, where σ^2_{state} is the variance estimate representing the variance between states in state-average log odds of a lake being sampled. The state-specific random effect was also included to accommodate the lack of independence among observations within states (Wagner et al. 2006). $\beta_{1,k}$ is the parameter estimate for the LULC classes, k=1,...3, β_2 is the parameter estimate for the effect of $\log_{10}$(area), and $\beta_{3,l}$ is the parameter estimate for the interaction of LULC and lake area, l=1,...3, and b_i is the state-specific random slope effect for the relationship between lake area and the probability of a lake being sampled and is defined as $b_i \sim N\left(0, \tau^2_{state}\right)$, where τ^2_{state} is the variance estimate representing the variance in the slopes between states. We deemed it reasonable to group lakes according to states (i.e., random state effects) because these data were collected by state

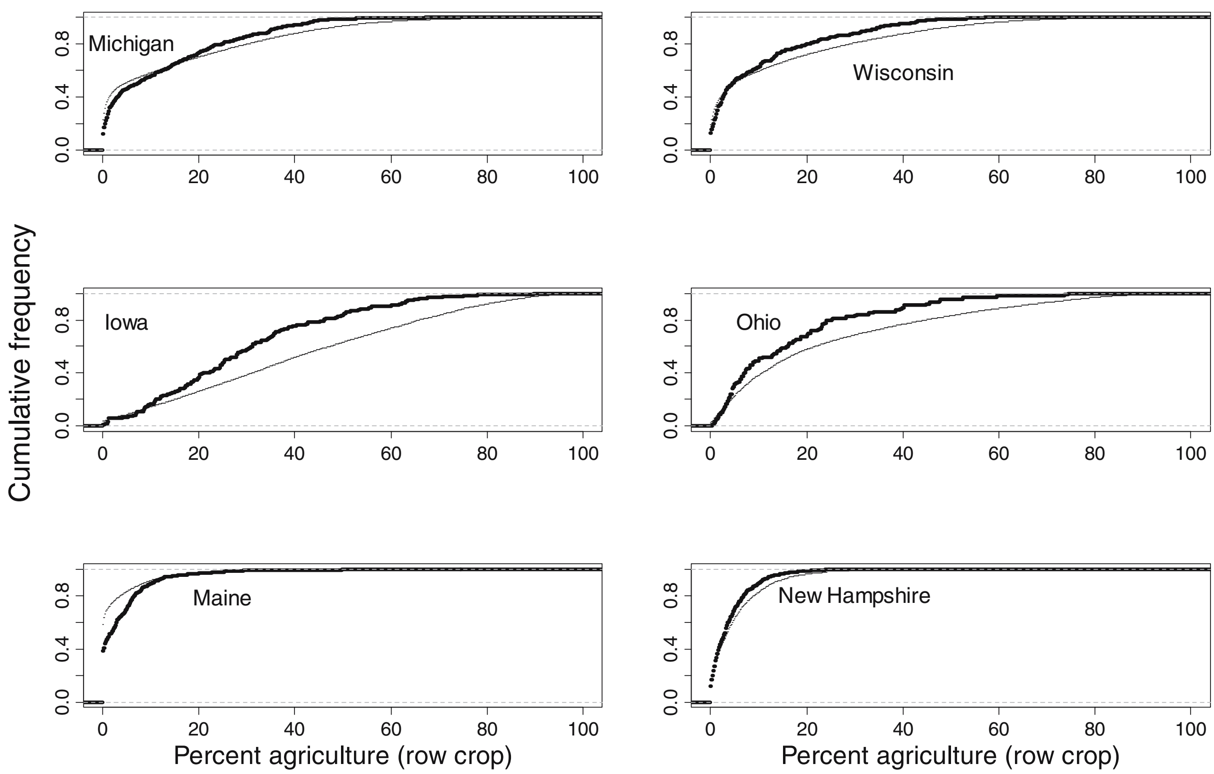

Fig. 4 Percent row crop agricultural land use in a 500 m buffer cumulative distribution functions for census lakes (*light line*) and lakes sampled by state agencies (*dark line*)

agencies, at the state-level. We considered all analyses significant at $P<0.05$.

Results

Summary of the number and total area of sampled lakes among states

The percent of lakes sampled and the percent of total lake area sampled varied among states. The percent of lakes sampled by state agencies ranged from 1.9–37.8% for Ohio and New Hampshire, respectively (Table 1). Although Ohio sampled the smallest percentage of waterbodies, the 1.9% represents 64.9% of the total lake surface area in the state. The percent of total lake area sampled ranged from 8.5% for Michigan to 67.2% for New Hampshire. Both Michigan and Wisconsin had the lowest percentages of total lake surface area sampled; reflecting the large number of waterbodies in those two states (Table 1).

Comparing lake area and LULC between NPB lakes and the entire population of lakes

Empirical cumulative distributions for lake area and the LULC variables show several biases between lakes sampled by state agencies and the census population in all six states (Figs. 2, 3, 4, 5, 6, 7 and 8). Lakes sampled by all state agencies had larger median lake areas compared to census lakes (Table 3). In fact, percent error (assuming that the census median lake size represents the 'true' median size) ranged from 121% for New Hampshire to 6,000% for Ohio. Within a state, lakes sampled by agencies also tended to have higher median values for percent forest land cover (all states) and urban land use (except for New Hampshire, where the medians were nearly identical) within a 500 m buffer of the lakes compared to the census lakes.

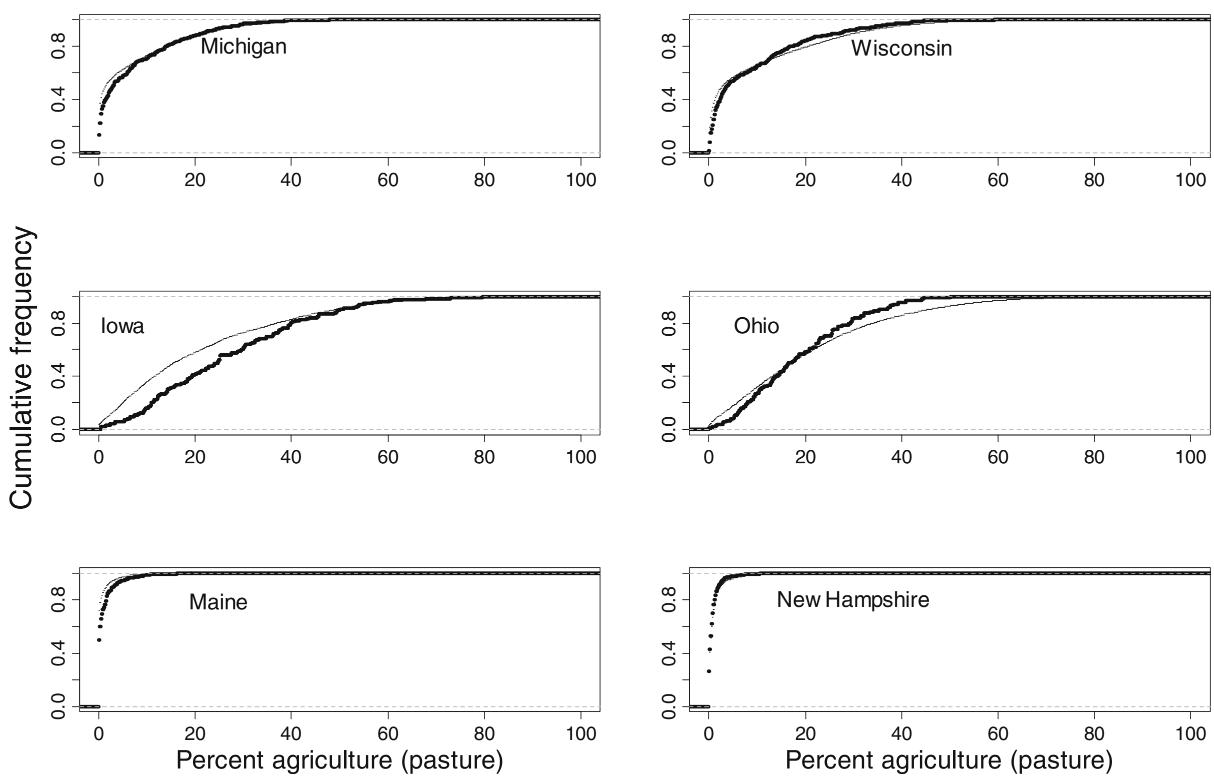

Fig. 5 Percent pasture agricultural land use in a 500 m buffer cumulative distribution functions for census lakes (*light line*) and lakes sampled by state agencies (*dark line*)

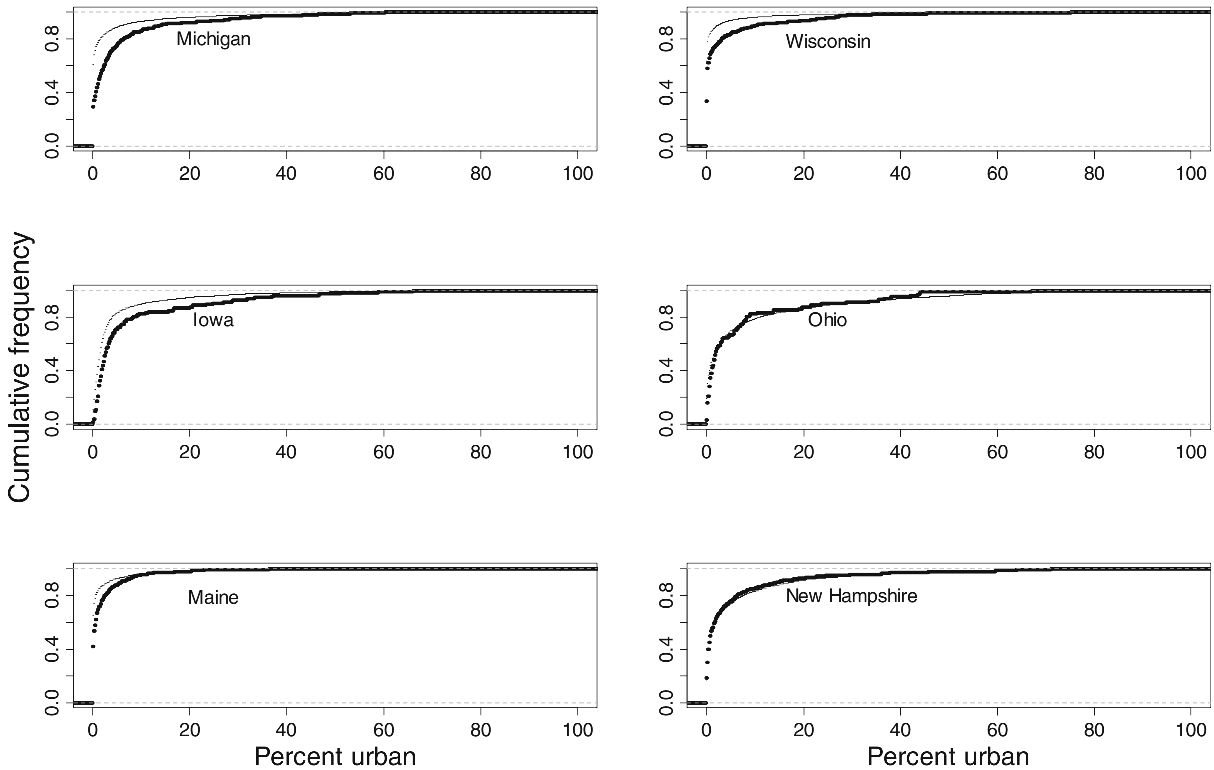

Fig. 6 Percent urban land use in a 500 m buffer cumulative distribution functions for census lakes (*light line*) and lakes sampled by state agencies (*dark line*)

Differences in the other median LULC categories between lakes sampled by states and census lakes were less consistent among states compared to lake area, and the magnitudes of the differences were also quite variable (Table 3). For example, for total agricultural land use, lakes sampled in Michigan, Wisconsin, and Maine had higher median values compared to census lakes; whereas, lakes sampled in Ohio, Iowa, and New Hampshire had smaller median values compared to census lakes. Median percent wetland cover in a 500 m buffer around each lake varied among states for both lakes sampled by agencies and census lakes; however, percent error was relatively small ranging from 16% for Maine to 50% for Ohio.

Land use/land cover and the probability of sampling a lake

All models examining the effects of LULC on the probability that a lake class was sampled, while controlling for the effects of lake area, demonstrated significant interactions (Table 4; Figs. 9, 10, 11, 12, 13 and 14). These significant interactions made it impossible to generalize about the main effect of LULC category; however, there were some general patterns that emerged. For all analyses, the probability of lakes being sampled increased with lake size, as expected. This result implies that if a lake is large, regardless of the surrounding LULC, it has a high probability of being sampled, and if the lake is small it has a low probability of being sampled.

After controlling for the effect of lake area, lakes with high amounts of total agricultural land use tended to have a higher probability of being sampled compared to other percent agriculture classes, a result that was consistent across all states (Fig. 9). Similar patterns were evident for row crop and pasture land use (Figs. 10 and 11), which was expected because they are highly correlated with total agricultural land use (r=0.90 and 0.76, respectively). For all states,

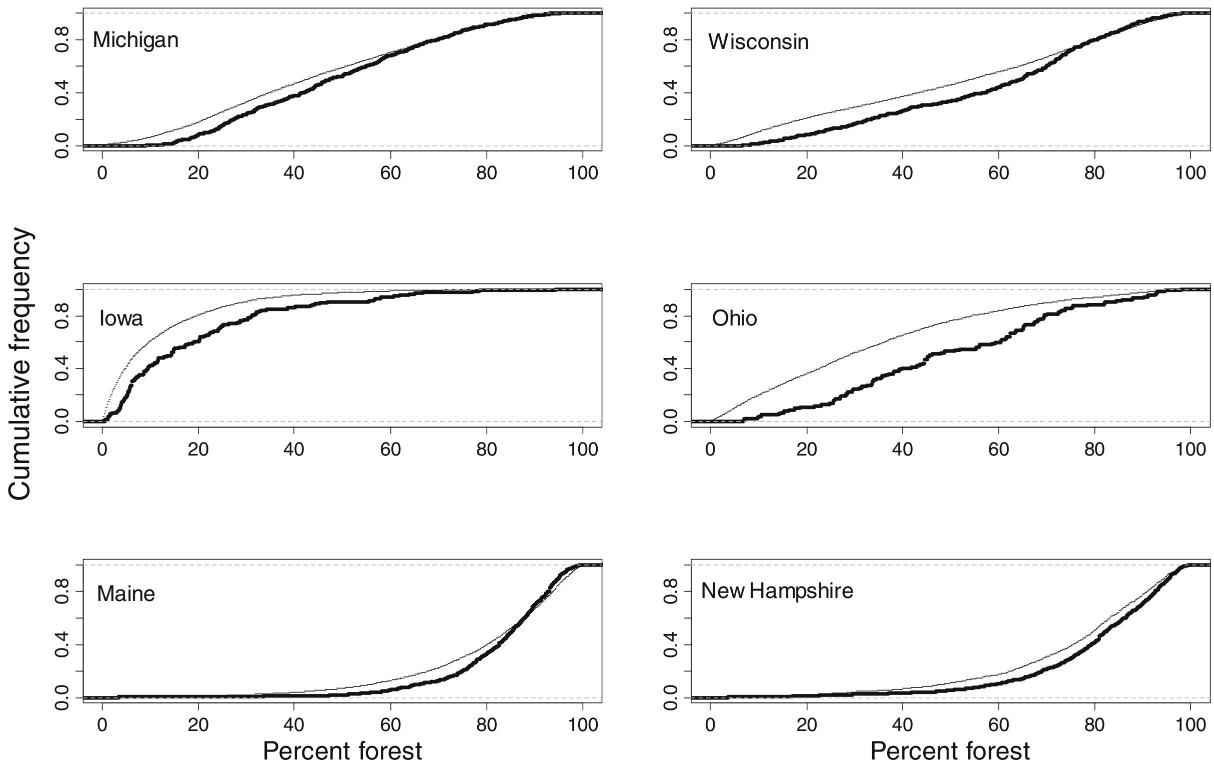

Fig. 7 Percent forest land cover in a 500 m buffer cumulative distribution functions for census lakes (*light line*) and lakes sampled by state agencies (*dark line*)

lakes with either high urban land use or forested land cover tended to have a higher probability of being sampling (Figs. 12 and 13), while lakes with high percent wetland cover had a low probability of being sampled (Fig. 14).

Discussion

Our results demonstrate that small waterbodies are largely underrepresented in all six state NPB sampling programs. Overall, the median lake size for lakes sampled by state agencies (54.1 ha) was approximately nine times larger than the median size of the census population (6.2 ha). Peterson et al. (1999) also found that the median lake size of lakes sampled using a NPB sampling design in the northeast USA was nine times larger than the median lake size estimated from EPA's EMAP program, a PB survey (9.5 ha for the NPB survey vs 85.0 ha for the PB survey; Peterson et al. 1999). This bias towards sampling large lakes was also apparent in our statistical analysis, where the probability of a lake being sampled, although varying by state, increased with increasing lake size. This under-representation of small waterbodies has implications for the regional assessment of a state's waterbodies, particularly as mandated under the CWA, as smaller waterbodies often have different chemical and biological characteristics compared to larger bodies of water. For example, Hanson et al. (2007) found that excluding small lakes during sampling in the Northern Highland Lake District of Wisconsin resulted in biased estimates of organic carbon and inorganic carbon, even though most of the water surface area is contained in relatively few large lakes.

Including small waterbodies in a sampling program is not only necessary to meet the requirements of environmental legislation, but it is also necessary to quantify the structural and functional role of these

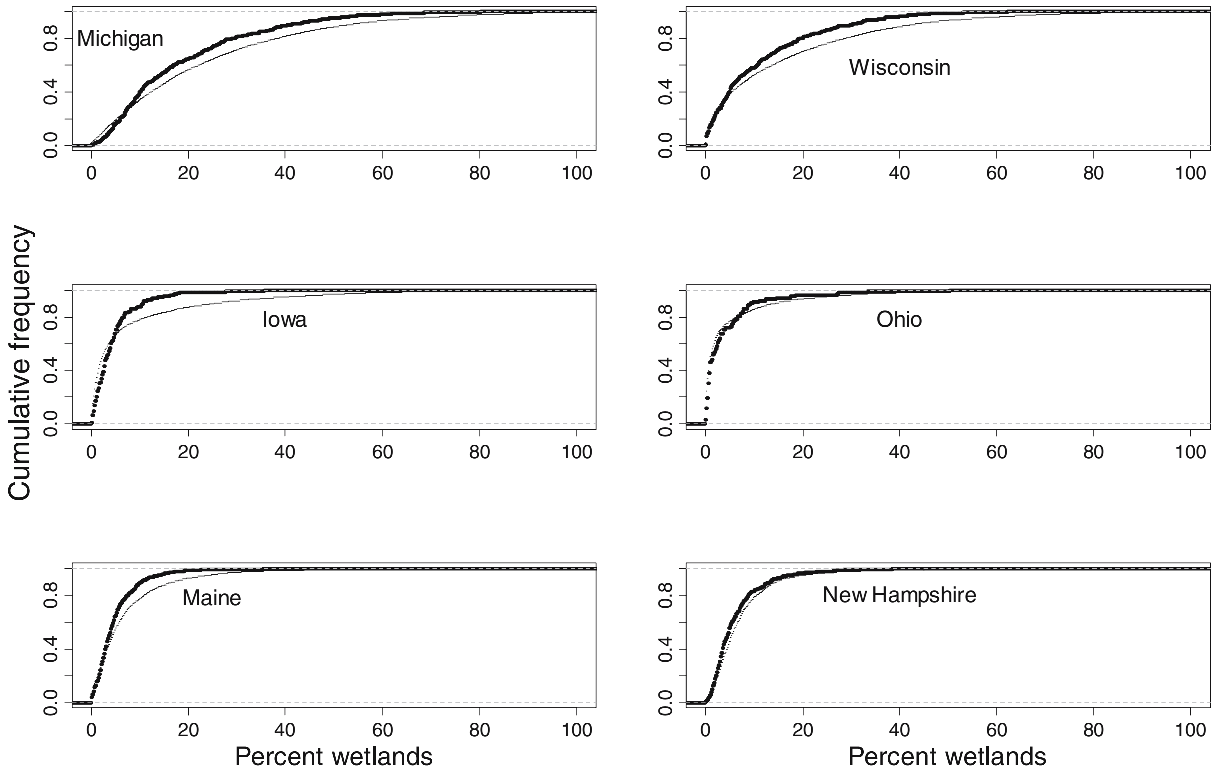

Fig. 8 Percent wetland land cover in a 500 m buffer cumulative distribution functions for census lakes (*light line*) and lakes sampled by state agencies (*dark line*)

aquatic resources at both regional and global scales (Downing et al. 2006). For example, small waterbodies, defined as <10 ha and including artificial waterbodies, account for approximately 20% of the standing water area in the United States (Smith et al. 2002), and the total global lentic area is dominated by small waterbodies <100 ha (Downing et al. 2006). Although each individual waterbody is small by definition, the cumulative effects of them on hydrology, sediment dynamics, geochemistry, and ecology can be quite large and they are rapidly increasing in number (Dahl 2006). For instance, smaller waterbodies can elevate evaporation rates, divert and delay downstream water flow, and modify groundwater flow dynamics (Smith et al. 2002). In areas dominated by artificial lakes (reservoirs) this bias will also

Table 3 Median lake area (ha) and percent land use/land cover in a 500 m buffer around each lake for the sampled lakes and the census lakes datasets (sampled lakes median/census lakes median)

State	Lake area	Agriculture (total)	Agriculture (row crop)	Agriculture (pasture)	Urban	Forest	Wetland
Michigan	79.3/8.2	12.1/7.6	7.2/4.5	2.9/1.4	1.3/0.0	47.3/42.5	13.0/16.6
Wisconsin	84.7/10.6	9.5/9.1	4.5/4.7	3.7/2.8	0.1/0.0	64.6/54.1	6.8/8.9
Ohio	134.2/2.2	32.1/45.6	10.0/15.4	16.8/17.0	1.6/1.3	45.9/28.8	1.5/1.0
Iowa	37.0/2.2	62.0/71.6	25.4/38.7	24.6/15.5	2.4/1.4	14.2/6.7	3.2/2.2
Maine	90.3/18.3	2.3/0.1	1.3/0.0	0.1/0.0	0.2/0.0	85.3/84.2	82.5/79.8
New Hampshire	14.6/6.6	3.5/7.8	2.7/3.2	0.3/0.3	0.8/0.9	82.5/79.8	4.5/5.4

Table 4 Fixed effects *P* values and parameter estimates and standard errors (SE) for random effects for generalized linear mixed models used to examine differences in the probability of sampling lakes in different land use/land cover (LULC) classes after controlling for the effects of lake area

	Fixed effects			Random effects	
Model	LULC	Lake area	LULC × lake area	$\left(\sigma^2_{state}\right)$	$\left(\tau^2_{state}\right)$
Urban	0.0007	0.018	<0.0001	3.1 (1.9)	0.40 (0.25)
Total agriculture	<0.0001	0.0012	<0.0001	2.7 (1.7)	0.30 (0.19)
Agriculture (row crop)	<0.0001	0.0013	0.0009	2.7 (1.7)	0.35 (0.23)
Agriculture (pasture)	0.52	0.0014	<0.0001	3.0 (1.9)	0.34 (0.22)
Wetland	0.43	0.003	<0.0001	3.2 (2.0)	0.41 (0.27)
Forest	0.0012	0.0024	0.1332	2.6 (1.6)	0.44 (0.29)

Dependent variable = 1 if a lake was sampled, 0 otherwise. See Table 2 for description of LULC classes. Random effects estimates are the between-state variance $\left(\sigma^2_{state}\right)$ and random slope variance $\left(\tau^2_{state}\right)$.

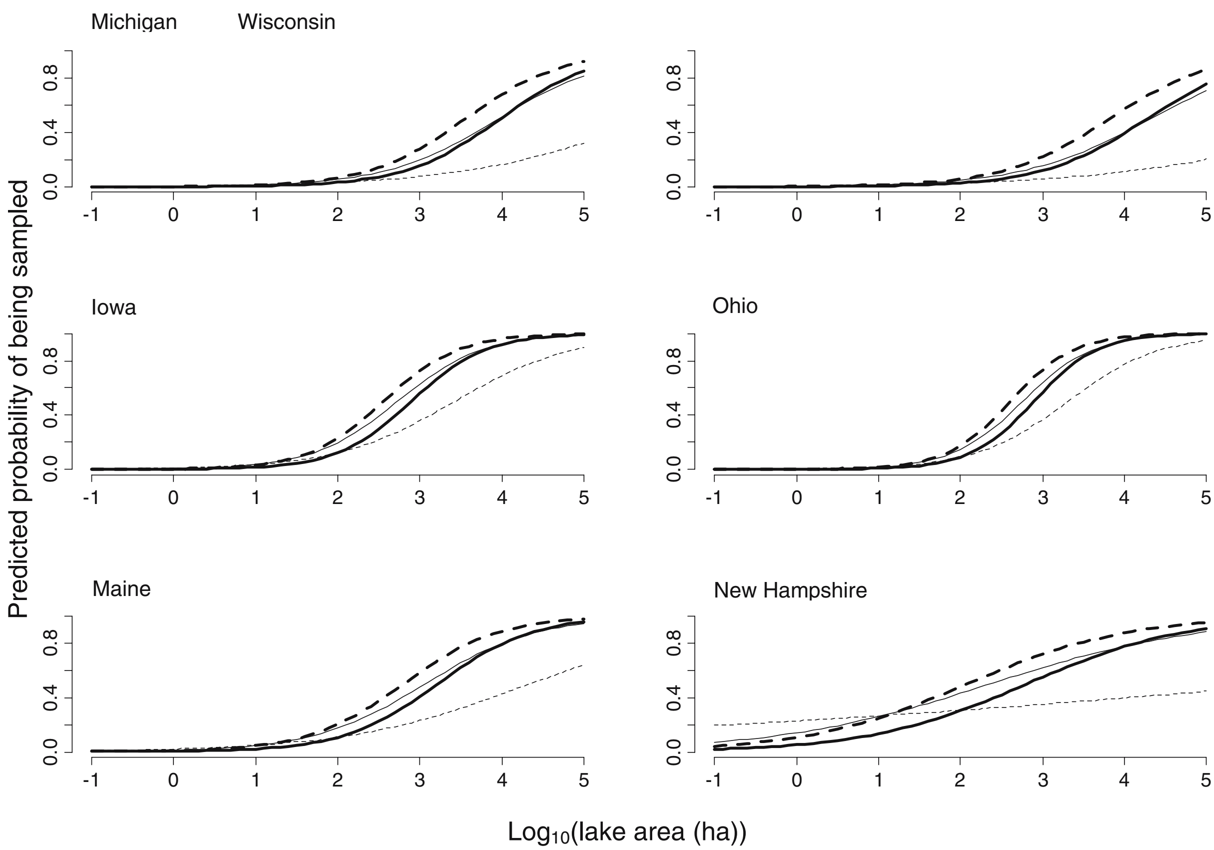

Fig. 9 The predicted probability of sampling lakes that are in different categories based on the amount of total agricultural land use in a 500 m buffer around each lake and lake area. Class A<0.8% (*light dashed line*), B 0.8–49% (*light solid line*), C 49–77% (*dark dashed line*), D>77% (*dark solid line*). Predicted values are based on a generalized linear mixed model with a random state and slope effect (see Eq. 1)

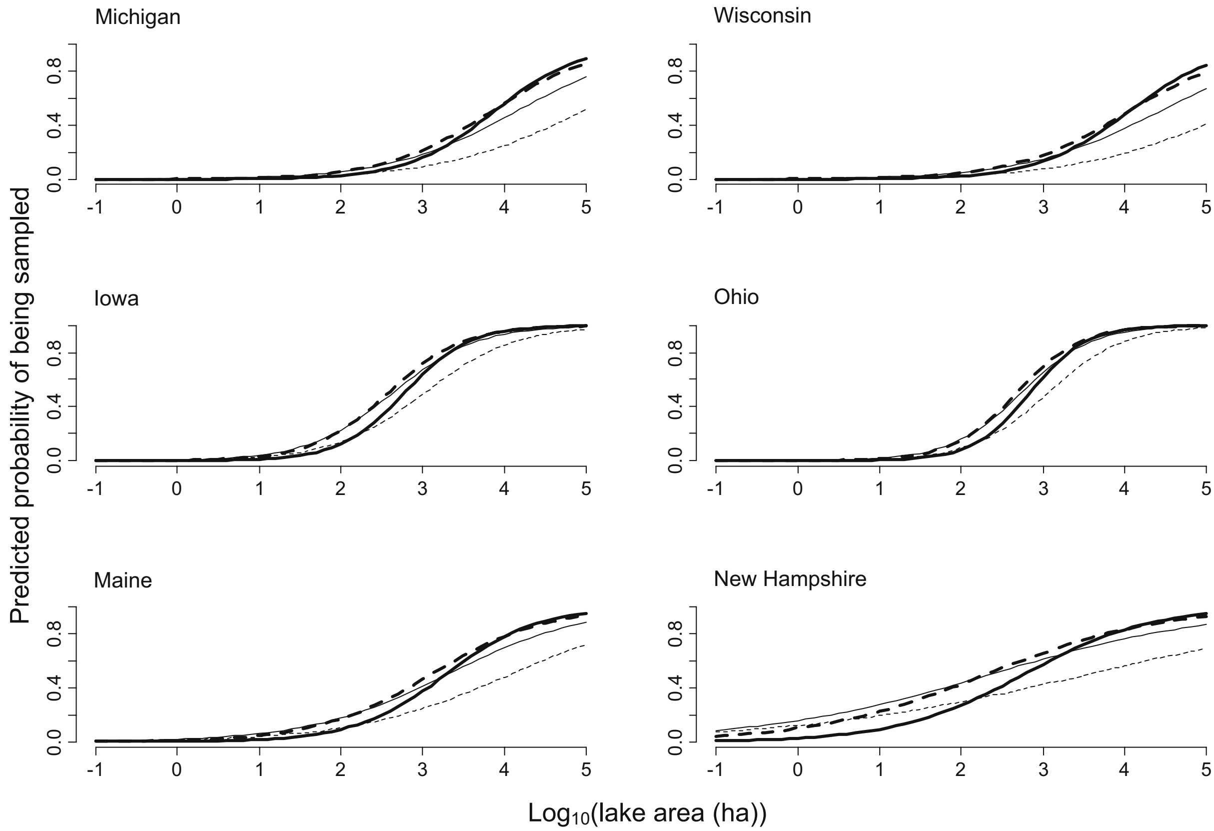

Fig. 10 The predicted probability of sampling lakes that are in different categories based on the amount of row crop agricultural land use in a 500 m buffer around each lake and lake area. Class A <0.25% (*light dashed line*), B 0.25–28% (*light solid line*), C 28–49% (*dark dashed line*), D >49% (*dark solid line*). Predicted values are based on a generalized linear mixed model with a random state and slope effect (see Eq. 1)

underestimate sedimentation rates because small lakes have higher sediment inputs per unit area than do large lakes (Renwick et al. 2005).

Furthermore, small waterbodies provide habitat for many rare species and should be considered of high conservation value, especially in areas of high agricultural land use (Søndergaard et al. 2005; Declerck et al. 2006). For example, because small ponds are often fishless they are an important habitat for aquatic invertebrates and amphibians, with members of both these groups being of high conservation value in the United States, Canada, and Europe (Oertli et al. 2002). In addition, small impoundments can have deleterious effects on biota; for example, through influencing dispersal capabilities of stream fishes and effects on stream habitat conditions (Schrank et al. 2001). However, if management agencies do not include small waterbodies in sampling programs, quantifying their effects on the landscape and biological communities, assessing and monitoring their condition, and documenting species for conservation is not possible. Furthermore, the CWA mandates to maintain and restore biological integrity [Section 101(a)], and the exclusion of smaller waterbodies from monitoring programs may have implications for assessing resource condition based on biological attributes.

Our analysis of LULC demonstrated that, in addition to lake size, the probability of a lake being sampled is also a function of LULC type. In general, our analyses suggest that lakes with more surrounding urban land use had a higher probability of being sampled. However, the overall percent urban land use around these lakes was still low compared to other

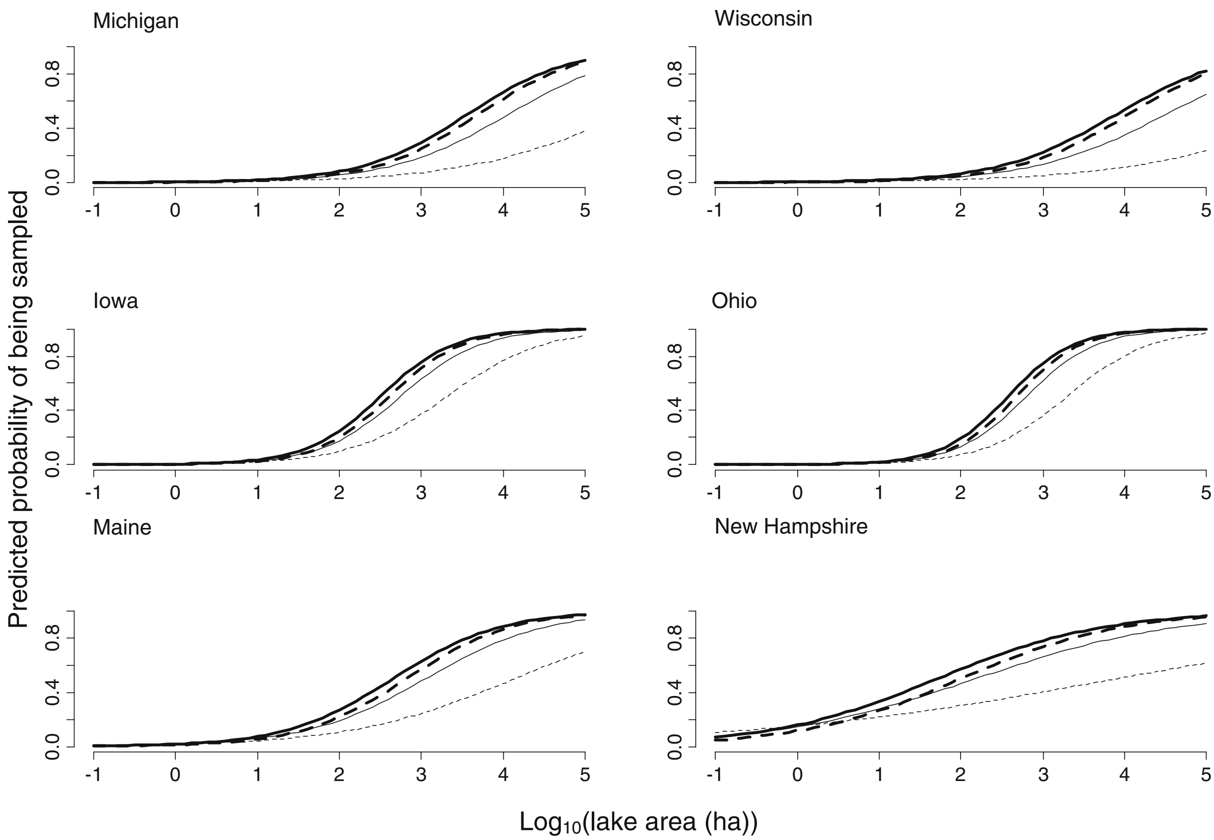

Fig. 11 The predicted probability of sampling lakes that are in different categories based on the amount of pasture agricultural land use in a 500 m buffer around each lake and lake area. Class A <0.11% (*light dashed line*), B 0.11–15% (*light solid line*), C 15–30% (*dark dashed line*), D >30% (*dark solid line*). Predicted values are based on a generalized linear mixed model with a random state and slope effect (see Eq. 1)

land uses (median percent urban ranged from 0.1–2.4% for lakes sampled by state agencies). This result is consistent with our initial hypothesis. Urban lakes are often heavily managed because of their importance as recreational waterbodies and municipal water supplies, especially large urban lakes. Conversely as the amount of wetland cover is higher around a lake, its probability of being sampled is lower. This negative effect of wetlands around a lake on the probability of a lake being sampled was unexpected. One explanation for this pattern, however, is that a loss of wetlands due to anthropogenic activities in urban and agricultural areas is responsible for this negative relationship. Stated another way, as urban and agricultural land use increase so does the probability that a lake will be sampled, and those lakes have low amounts of surrounding wetlands due in part to anthropogenic activity. This idea is supported by the fact that percent wetlands are negatively correlated with agricultural and urban land use ($r=-0.37$ and -0.15, respectively), and appears reasonable because as human population density increases over time the density of wetlands tend to decrease (Gibbs 2000). In fact, not only do the density of wetlands decrease with increasing human population density, but there are also decreases in the size of wetland mosaics, and the proportion of wetlands in the landscape decreased from 5–8% in rural areas to <1% in suburban and urban areas in the New York City region (Gibbs 2000).

The patterns for agricultural land use and forest land cover suggest that lakes with low amounts of agricultural land use and forest cover have a lower probability of being sampled compared to lakes with

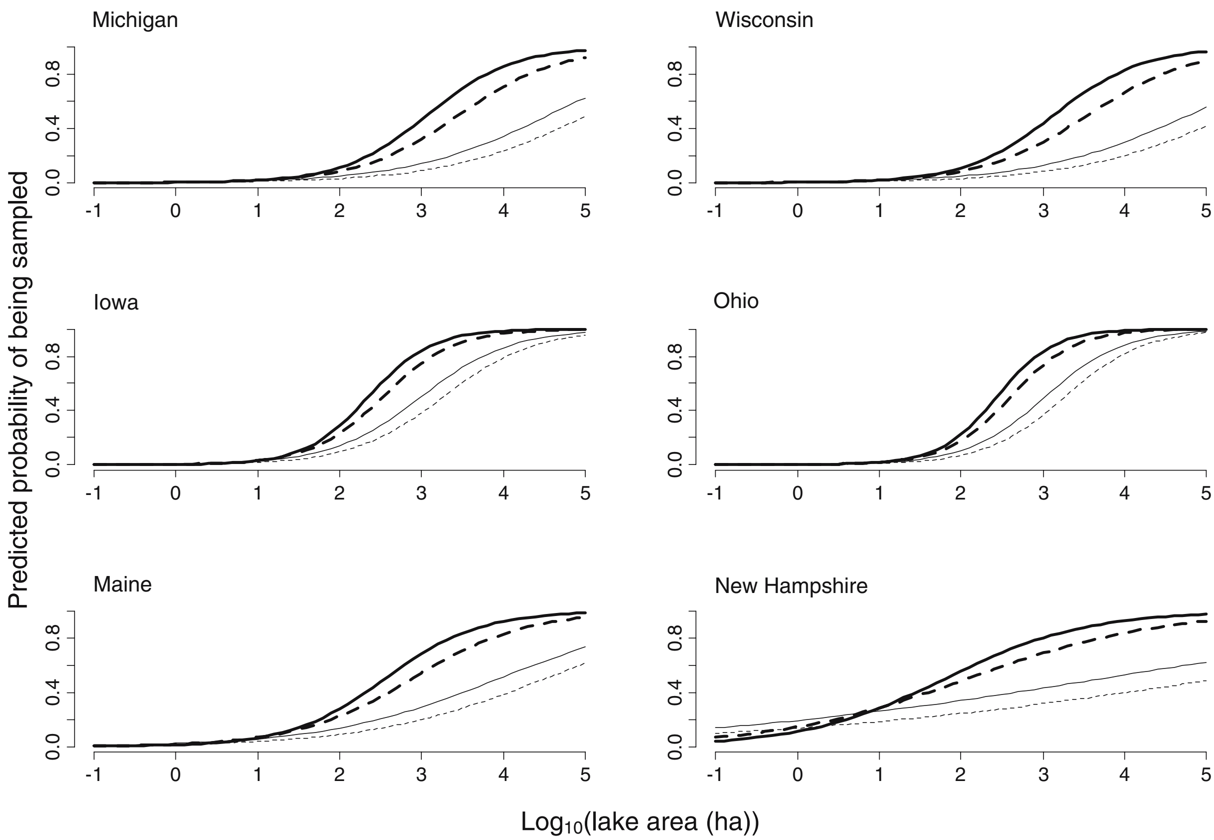

Fig. 12 The predicted probability of sampling lakes that are in different categories based on the amount of urban land use in a 500 m buffer around each lake and lake area. Class A 0.0% (*light dashed line*), B 0.0–1.4% (*light solid line*), C 1.4–18% (*dark dashed line*), D >18% (*dark solid line*). Predicted values are based on a generalized linear mixed model with a random state and slope effect (see Eq. 1)

higher amounts of these LULC types. The fact that we see similar patterns for agriculture land use and forest land cover (i.e., that lakes with low amounts of these LULC types have a lower probability of being sampled) is at first somewhat puzzling given that total agricultural land use and forest land cover are negatively correlated ($r=-0.72$). However, these results reflect the probability of sampling relative to the census population. Thus, it is possible for the probability of sampling a lake to demonstrate similar patterns for two negatively correlated predictor variables (Fig. 15).

Our analyses demonstrated and quantified some of the biases associated with NPB sampling programs used to assess and manage inland lake ecosystems. These biases can lead to inaccurate estimates of regional lake status based on the extrapolation of NPB data. There is a substantial bias towards large waterbodies, which is understandable from the perspective of their importance in many regions. However, smaller waterbodies also serve important functions in the landscape and should be included as part of PB sampling designs. Furthermore, the inclusion of small waterbodies in sampling programs is necessary for states to accurately describe the status and trends of their inland lake ecosystems, as mandated by the CWA.

Our approach of using map-based data provides the advantage of having a ‘census’ population with which to compare state sampled waterbodies. However, we acknowledge that there are limitations inherent in the use of map-based data. For example, the NHD database we used is at a 1:100,000 scale, and at this scale many smaller waterbodies are not

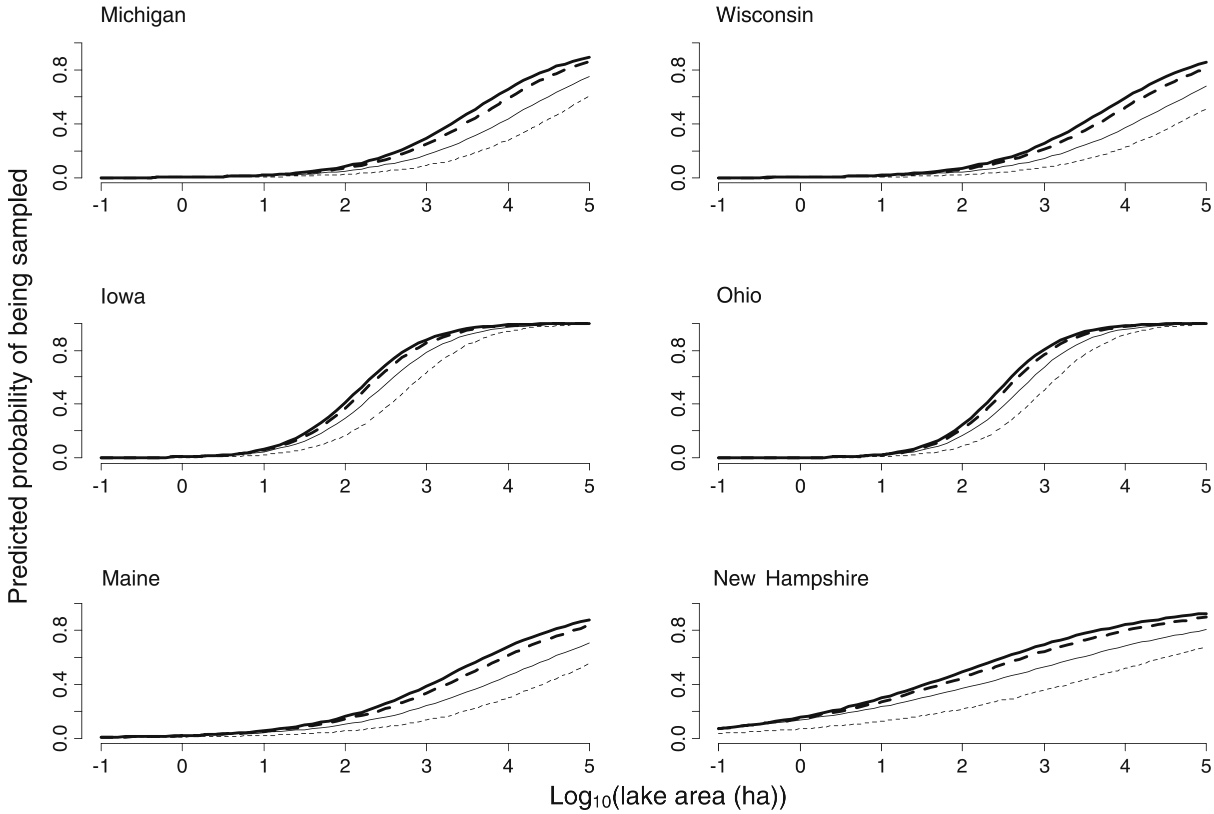

Fig. 13 The predicted probability of sampling lakes that are in different categories based on the amount of forest land cover in a 500 m buffer around each lake and lake area. Class A <21% (*light dashed line*), B 21–76% (*light solid line*), C 76–89% (*dark dashed line*), D >89% (*dark solid line*). Predicted values are based on a generalized linear mixed model with a random state and slope effect (see Eq. 1)

included. The exclusion of smaller waterbodies at the 1:100,000 scale, however, makes our results with regards to lake area conservative estimates. Furthermore, there is a temporal component of potentially changing LULC and numbers of small waterbodies that we are not able to control for in this study. Finally, as previously mentioned, our approach is limited to those variables that can be derived from maps. For example, we were unable to quantify any potential biases in water quality, such as lake total phosphorus concentrations. However, because of the relationships between LULC (and lake size) and many important lake water quality variables (Siver et al. 1999; Hall et al. 1999; Arbuckle and Downing 2001), there are likely state-level biases associated with water quality metrics that have been sampled using NPB surveys and subsequently extrapolated to represent regional conditions. This is not to imply that programs designed to monitor the status of lakes deemed as 'high priority' by managers should be rejected, rather that generalizations made about state water resources using NPB survey data should be avoided.

Acknowledgments We thank Mary Bremigan who provided intellectual contributions to this research, Michigan State University's Remote Sensing and Geographic Information Science Research and Outreach Services for quantification of the land use/land cover data, and the many people who provided help with state agency data: Sherry Martin, Ralph Bednarz, Nancy Nate, Tim Asplund, Jeff DeShon, Zac Andereck, John Downing, Scott Ashley, Jodi Connor and Linda Bacon. This research was supported in part by a grant from U.S. EPA Environmental Protection Agency-Office of Wetlands, Oceans and Watershed National Lakes Assessment Planning Project to P.A.S., K.E.W., M.T.B. and K.S.C. This is publication 2007-04 of the Quantitative Fisheries Center.

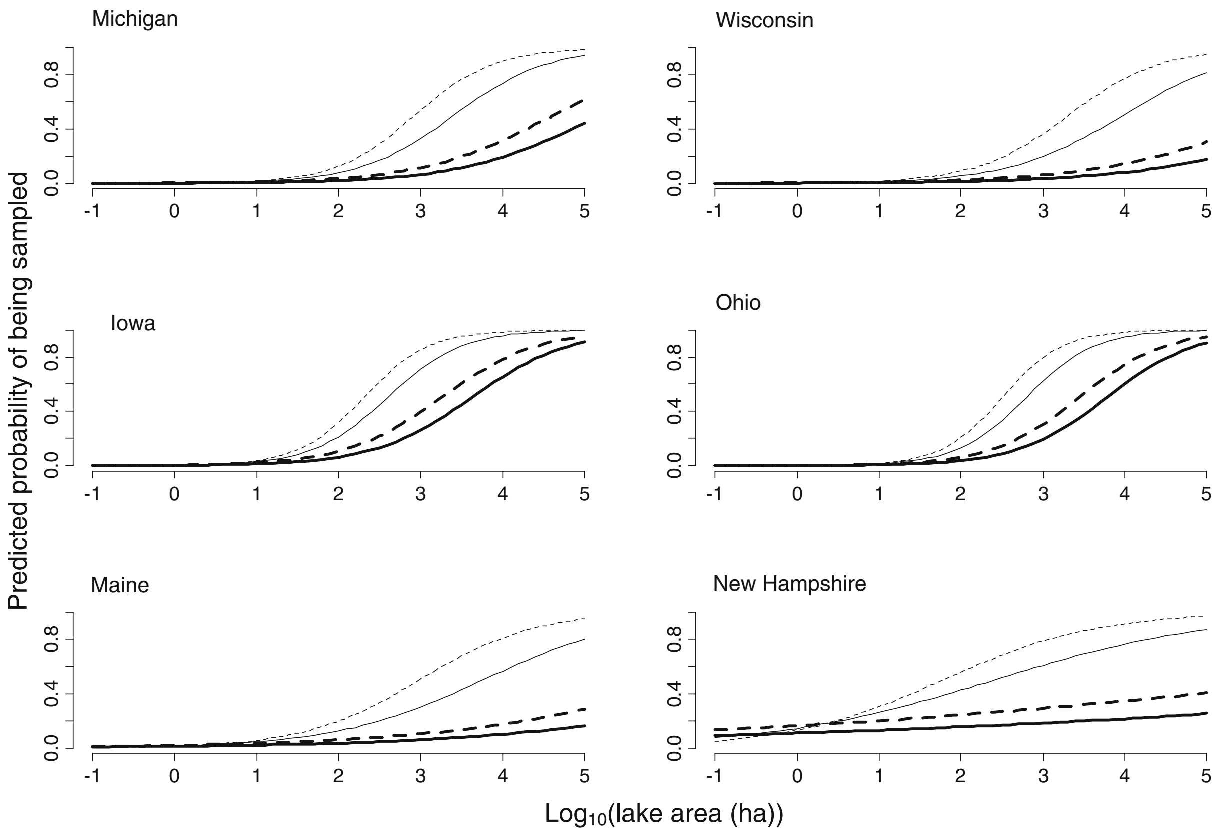

Fig. 14 The predicted probability of sampling lakes that are in different categories based on the amount of wetland land cover in a 500 m buffer around each lake and lake area. Class A <2% (*light dashed line*), B 2–21% (*light solid line*), C 21–40% (*dark dashed line*), D >40% (*dark solid line*). Predicted values are based on a generalized linear mixed model with a random state and slope effect (see Eq. 1)

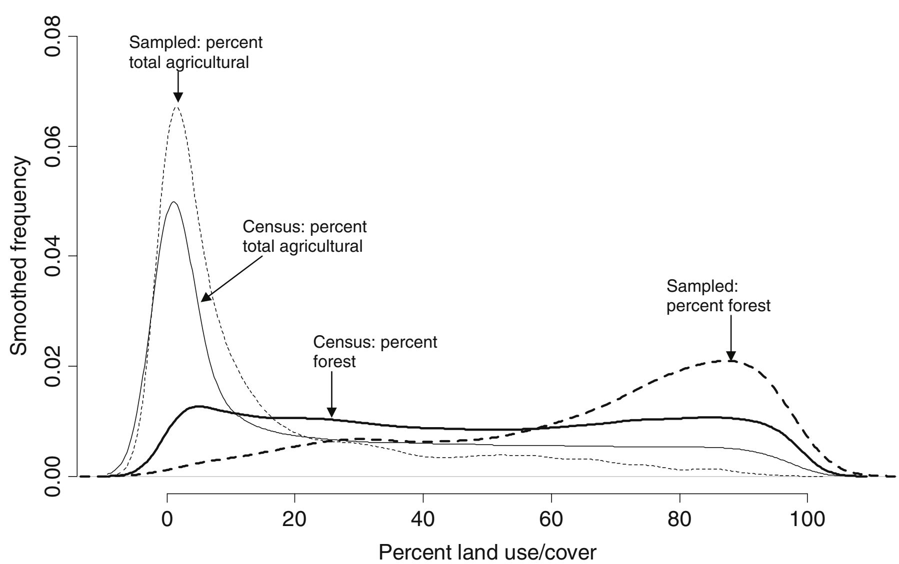

Fig. 15 Smoothed histograms for percent total agricultural land use and forest cover for census and sampled lakes in the entire six-state region. Note that even though percent total agriculture and forest cover are negatively correlated, the distributions of sampled lakes are shifted towards larger values for both land use/cover types

References

Arbuckle, K. E., & Downing, J. A. (2001). The influence of watershed land use on lake N:P in a predominately agricultural landscape. *Limnology and Oceanography, 46*, 970–975.

Brown, B. S., Detenbeck, N. E., & Eskin, R. (2005). How probability survey data can help integrate 305(b) and 303 (d) monitoring and assessment of state waters. *Environmental Monitoring and Assessment, 103*, 41–57.

Dahl, T. E. (2006). *Status and trends of wetlands in the conterminous United States 1998 to 2004* (p. 112). Washington, DC: US Department of the Interior; Fish and Wildlife Service.

Declerck, S., De Bie, T., Ercken, D., Hampel, H., Schrijvers, S., Van Wichelen, J., et al. (2006). Ecological characteristics of small farm ponds: Associations with land use practices at multiple spatial scales. *Biological Conservation, 131*, 523–532.

Downing, J. A., Prairie, Y. T., Cole, J. J., Duarte, C. M., Tranvik, L. J., Striegl, R. G., et al. (2006). The global abundance and size distribution of lakes, ponds, and impoundments. *Limnology and Oceanography, 51*, 2388–2397.

Gibbs, J. P. (2000). Wetland loss and biodiversity conservation. *Conservation Biology, 14*, 314–317.

Hanson, P. C., Carpenter, S. R., Cardille, J. A., Cole, M. T., & Winslow, L. A. (2007). Small lakes dominate a random sample of regional lake characteristics. *Freshwater Biology, 51*, 814–822.

Hall, R. H., Leavitt, P. R., Quinlan, R., Dixit, A. S., & Smol, J. P. (1999). Effects of agriculture, urbanization, and climate on water quality in the northern Great Plains. *Limnology and Oceanography, 44*, 739–756.

Hayes, D., Baker, E., Bednarz, R., Borgeson, D. Jr., Braunscheidel, J., Breck, J., et al. (2003). Developing a standardized sampling program: The Michigan experience. *Fisheries, 28*, 18–25.

McDonald, M. E. (2000). EMAP overview: Objectives, approaches, and achievements. *Environmental Monitoring and Assessment, 64*, 3–8.

Oertli, B., Joye, D. A., Castella, E., Juge, R., Cambin, D., & Lachavanne, J-B. (2002). Does size matter? The relationship between pond area and biodiversity. *Biological Conservation, 104*, 59–70.

Olsen, A. R., et al. (1999). Statistical issues for monitoring ecological and natural resources in the United States. *Environmental Monitoring and Assessment, 54,* 1–45.

Peterson, S. A., Larsen, D. P., Paulsen, S. G., & Urquhart, N. S. (1998). Regional lake trophic patterns in the Northeastern US: Three approaches. *Environmental Management, 22*, 789–801.

Peterson, S. A., Urquhart, N. S., & Welch, E. B. (1999). Sample representativeness: A must for reliable regional lake conditions. *Environmental Science and Technology, 33*, 1559–1565.

Renwick, W. H., Smith, S. V. Bartley, J. D., & Buddemeier, R. W. (2005). The role of impoundments in the sediment budget of the conterminous United States. *Geomorphology, 71*, 99–111.

Siver, P. A., Lott, A. M., Cash, E., Moss, J., & Marsicano L. J. (1999). Century changes in Connecticut, U.S.A., lakes as inferred from siliceous algal remains and their relationships to land-use changes. *Limnology and Oceanography, 44*, 1928–1935.

Smith, S. V., Renwick, W. H., Bartley, J. D., & Buddemeier, R. W. (2002). Distribution and significance of small, artificial water bodies across the United States landscape. *The Science of the Total Environment, 299*, 21–36.

Schrank, S. J., Guy, C. S., Whiles, M. R., & Brock, B. L. (2001). Influence of instream and landscape-level factors on the distribution of Topeka Shiners *Notropis Topeka* in Kansas streams. *Copeia, 2001*, 413–421.

Søndergaard, M., Jeppesen, E., & Peder Jensen, J. (2005). Pond or lake: Does it make any difference? *Archiv fuer Hydrobiologie, 162*, 143–165.

Wagner, T., Hayes, D. B., & Bremigan, M. T. (2006). Accounting for multilevel data structures in fisheries data using mixed models. *Fisheries, 31*, 180–187.