Synthesis Report: Integrating probability and targeted survey designs in regional stream condition assessments with examples from southern coastal California

prepared for and funded by :

The California State Water Resources Control Board
Nonpoint Source Pollution Prevention Program

Contract # 03-273-250-0

authors:

Andrew C. Rehn and Peter R. Ode
Aquatic Bioassessment Lab
California Dept. of Fish & Game

November 2009

Executive Summary

Ambient water quality monitoring programs have traditionally relied on the use of targeted monitoring designs (in which sampling locations are assigned non-randomly) to assess the condition of aquatic resources. As probability-based surveys (in which site locations are assigned randomly) have become more widely adopted, there has been a growing desire to combine data from targeted sites with data from probability sites when making overall assessments of resource condition. We evaluated the potential for combining datasets and present an approach to combining data from the two types of surveys. We compiled biological, chemical and habitat data from 63 probability sites and 133 targeted monitoring sites in southern coastal California and compared the distribution of analyte values collected under the two designs.

This analysis clearly demonstrates that both targeted and probabilistic sampling designs are necessary for effective non-point source water quality monitoring programs. Probability designs are the only survey type capable of giving an objective overview of the range of conditions present in the resource of interest. By contrast, targeted designs often are needed to fill in critical data gaps and answer site-specific questions. We conclude that the most effective way to 'integrate' probability and targeted surveys is to use probability surveys to produce an unbiased regional framework for interpreting targeted results and to supplement this framework as needed with sites targeted to fill data gaps.

Introduction

Probability and targeted surveys both are widely used throughout California, in most states in the US and in many national programs (US, UK, EU, AU) in bioassessment of streams and rivers. In probability surveys (also known as design-based surveys), sampling sites are selected randomly with provisions to ensure thorough geographic coverage (Stevens and Olsen 2004). Each sampling site represents a specific portion of the total resource of interest, for example, all wadeable stream length in California. The portion of total resource that a site represents is expressed as a statistical weight: sites with higher weights represent a greater extent of resource than sites with lower weights. Because of the statistical nature of site selection, results from the sample population can be extrapolated to the entire population. Probability sampling therefore is well-suited for making unbiased assessments of biological condition and stressor extent across large geographic areas. By contrast, targeted surveys select site locations to address specific questions of interest, for example to evaluate regulatory compliance at point-source discharges, to track trends over time at compliance sites, or to evaluate the effectiveness of restoration projects. Targeted surveys are not statistically based, but this is not detrimental to survey objectives provided that unbiased population estimates are not required. In fact, both survey types are vital to a comprehensive monitoring program and provide complimentary information.

As both types of survey data have become more abundant in California over the last decade, both regionally and statewide, there has been an increased desire to combine data from probability and targeted surveys into 'integrated' condition assessments. This wish is not unique to California, and over the past 20 years a small but growing number of papers have been published on this topic (e.g., Overton et al. 1993; Paul et al. 2008; Stein and Bernstein 2008). Two general approaches are possible, referred to here as the statistical approach and the narrative approach. In the statistical approach, targeted sites are treated as a pseudo-probability sample and are assigned weights based on their physiographic relationship with probability sites (see Overton et al. 1993 for a complete example). First, probability sites are clustered into subsets (that represent corresponding subpopulations) based on a set of frame attributes[1] that are predictive of the indicator variable(s) of interest. Second, targeted sites are assigned to probability subsets and are treated as a simple random sample from the subset to which they belong. Finally, pseudo-probability weights are derived for targeted sites by summing the original weights for probability sites in a given subset and apportioning that sum equally to targeted sites assigned to that subset. Possible reasons for employing a statistical approach may be to increase the size of a probability sample and thereby improve confidence limits around our estimates, or because targeted sites have valuable data not available at probability sites.

Overton et al. (1993) used 'similarity' of frame attributes as a criterion for assigning targeted sites to probability subpopulations. Paul et al. (2008) suggested that similarity of indicator variables between targeted and probability sample populations should also be assessed when evaluating whether the two survey types can be combined into a single estimate. Paul et al. (2008) defined three criteria for evaluating similarity: 1) targeted sites must have been selected to be representative of the area from which they were selected; 2) the cumulative distribution functions (CDFs) of indicator variables from targeted and probability sample populations must not be statistically different; 3) the correlation structure between indicator variables from targeted sites must be equivalent to that from probability sites. However, it is important to note that regardless of how stringent the criteria for assessing similarity between probability and targeted sample populations, the statistical approach can never be carried out in a truly rigorous sense because the theoretical foundation of design-based inference is random selection of sites (Don Stevens, pers. comm.). The use of targeted sites in a pseudo-probability approach involves the augmentation of a statistical protocol that assures site representativeness with an unprovable assumption of representativeness (Overton et al. 1993).

In the narrative approach, results from probability and targeted surveys are derived independently and then are compared, i.e. no statistical weighting of targeted sites is attempted. Instead, the probability design provides a framework for interpreting results from targeted sites in an unbiased regional context (Stein and Bernstein 2008). The requirements to employ a narrative approach are much less stringent than those required to employ a statistical approach: in fact, the only strict requirement is that the two surveys have substantial geographic overlap, but targeted sites do not have to be representative of the region from which they were selected.

The purpose of this report is to provide an example to the California Surface Water Ambient Monitoring Program (SWAMP) of how data from probability and targeted stream surveys might be combined into integrated 305(b) condition assessments, either for particular regions or for the entire state. The current report is restricted to an example of the narrative approach, which in many ways is an extension of work completed by Stein and Bernstein (2008) for the San Gabriel watershed. The statistical approach was not attempted for this report because targeted data sets from southern coastal California did not meet the assumptions required to make it feasible (see the criteria of Paul et al. [2008] listed above).

---

[1] Frame attributes are variables that are known or knowable for all elements of the population, including those that have not been sampled.

Methods

Data sets from probability and targeted surveys in southern coastal California were selected for developing the present examples (Fig. 1). This region was an area of intensification with its own probability design as part of the EPA's western EMAP pilot in 2000-2003, and therefore has a relatively high density of probability sites ($n = 63$)[2]. Data were compiled from targeted surveys that overlapped geographically with probability sites and that collected many of the same indicator variables (biological, chemical and physical) over roughly the same time period as EMAP ($n = 133$ sites, Fig. 1).

Cumulative distribution functions (CDFs), density functions and pie charts for several indicator variables commonly collected in California bioassessments (e.g., IBI score, physical habitat (PHAB) score, and nutrient concentrations) were plotted separately for each survey type. In addition, CDFs and density functions were plotted for 4 land use variables that can be considered predictive of indicator variables: percent urban land use in the local (1km upstream) and full watershed, and percent agricultural land use in the local and full watershed. Nutrient concentrations were log transformed for graphical purposes because they varied over several orders of magnitude. The log of x+1 was used for total nitrogen and turbidity because some sites had a value of '0' for these variables.

The y-axis in CDFs derived from probability surveys represents percent of total stream length since each weighted probability site represents some portion of the total estimated resource[3]. By contrast, the y-axis in CDFs derived from targeted surveys represents percent of sites because each targeted site has equal weight, i.e., each site simply represents 1 part of the total number of sites. To facilitate more direct comparisons of variable distributions between probability and targeted sample populations, additional CDFs were calculated for probability sites as if they were equally weighted. The Kolmogorov-Smirnov test was used to determine whether probability and targeted distributions (CDFs) were statistically different. Density plots were used as an alternative means of displaying and comparing the distributions of indicator variables between probability and targeted sample populations. Nutrient thresholds from either the EPA's western EMAP condition assessments (Stoddard et al. 2005b) or from recent California condition assessments (Ode et al. 2008) were used as examples in CDFs and pie charts.

Results

Results for each indicator variable are presented as a separate panel of figures (Figs 2-7). The CDFs for probability sites (under original and equal weights) and for targeted sites are presented together in the upper left of each panel, density functions are presented in the upper right, and pie charts are presented in the bottom portion of each panel. Some indicators were not measured at all targeted sites (Table 1), so targeted CDFs, density plots and pie charts are based on unequal sample sizes. All indicators were measured at all 63 probability sites.

---

[2] Details of the western EMAP sampling frame and site selection process can be found in Stoddard et al. (2005a,b).
[3] Weights have been based on stream order in statistical surveys used in California to date. Probability sites have different weights depending on how much stream length they represent.

Table 1. Number of targeted sites where each indicator was measured; total targeted $n = 133$.

Indicator variable	n
SoCal IBI	133
Qualitative PHAB score	122
Chloride (mg/L)	86
Sulfate (mg/L)	83
Total Nitrogen (mg/L)	103
Turbidity (NTU)	126

The distributions of all evaluated indicator variables were statistically different between probability and targeted sample populations (Kolmogorov-Smirnov $p < 0.008$ after Bonferroni adjustment for multiple comparisons; Figs 2-7). Targeted sites had lower IBI and PHAB scores, and higher nutrient concentrations than probability sites. Targeted sites also had significantly greater urban land use at both local and watershed scales, but percent agriculture was not different between targeted and probability sample populations at either spatial scale (Figs 8-9). Interestingly, the CDFs for probability sites under equal weights were nearly identical to the CDFs for probability sites under their original weights (Figs 2-9). Since the original weights in the EMAP design were based on stream order, this means that stream order has no effect on the indicator variables evaluated in this report. In other words, weighting by stream order produced an "ignorable design" for southern coastal California (Don Stevens, pers. comm.). It is currently unknown whether other probability survey designs for California in which stream order has been included as a basis of site weights (e.g., the north coast and statewide EMAP designs, CMAP and PSA) have also yielded ignorable designs in terms of indicator variables measured for bioassessment.

Conclusions

Water quality managers and other readers of this report may not be surprised to learn that the distributions of indicator variables commonly collected by bioassessment programs were significantly different between probability and targeted sample populations in southern coastal California (also see Stein and Bernstein 2008). After all, targeted surveys usually *target* streams with known human influences, i.e., places where chemical, physical and biological degradation are more likely to occur than in the overall population of streams. Nonetheless, the question has persisted as to whether the two survey types can be combined into integrated condition assessments like state 305(b) reports, for example by devising a way to assign statistical weights to targeted sites. Some managers have even questioned why regional stream condition assessments can't be derived from targeted surveys alone. The statistical weighting approach for targeted sites in southern coastal California assessments is not feasible because 1) targeted surveys have not been designed to represent the range of conditions that exists across the region, and 2) targeted sample populations do not have the same distribution of indicator variables as probability sample populations. In short, targeted surveys produce a biased picture of resource condition. Our analysis was based on southern coastal datasets, but this is likely true for other regions of the state where probability and targeted surveys overlap geographically. This is not a flaw in targeted surveys: managers need to know what water quality is like in highly influenced systems, and the only way to find out is to sample them, but targeted sites should not be used in place of, or even in conjunction with, probability sites to develop unbiased population estimates for regional stream condition assessments if the above conditions are not met.

In addition, targeted surveys that focus on problem areas leave large portions of the total resource unrepresented. Such surveys do not characterize the highest quality streams, do not identify their geographic locations, and do not tell us what portion of the resource they constitute. Identification and protection of high-quality resources is impossible if all or most of the focus is on streams with known problems.

The other side of this argument is that probability surveys tend not to sample a large number of sites with high levels of human disturbance (see percent urban at local scale in Fig. 8). This is not to say that probability surveys under-represent sites with more disturbance in the watershed--- in theory, probability surveys sample these sites in proportion to their occurrence in the overall population. If such sites are relatively uncommon in an overall population, probability surveys will not likely include very many of them, even though we as managers may be acutely aware of their existence and location. However, many applications of monitoring data, including IBI development (whether the indicator is BMIs, fish, or algae) rely on having a large amount of data from disturbed sites. For example, in IBI development, the characterization of biological response to *complete* disturbance gradients, from the very best sites to the very worst sites, is essential for selecting responsive metrics. The fact that probability and targeted surveys sample different portions of human disturbance gradients is advantageous for the purposes of IBI building, and in that sense the two survey types are complimentary in bioassessment programs. The extent to which the two survey types sample different portions of stressor gradients is perhaps best seen in density plots for chloride, sulfate, and percent urban land use at the local scale (Figs 4, 5 and 8, respectively) where targeted and probability curves are shifted relative to one another along the x-axis.

We conclude that the most effective way to 'integrate' probability and targeted surveys is to use probability results as an unbiased regional framework for interpreting targeted results (also see Stein and Bernstein 2008, who reached a similar conclusion). Different estimates of stressor extent from probability and targeted surveys may lead to different decisions about how to allocate resources. For example, if results from a probability survey indicate that riparian disturbance and sedimentation are stressors with the greatest extent in a particular region, but most of the region's financial resources are spent on nutrient problems downstream of a few point-source discharges, reallocation of resources to emphasize habitat protection may be necessary for protecting overall stream quality.

References

Ode, P.R., A.C. Rehn and J.T. May. 2005. A quantitative tool for assessing the integrity of southern coastal California streams. Environmental Management 35:493-504

Overton, J.M., T.C. Young and W.S. Overton. 1993. Using 'found' data to augment a probability sample: procedure and case study. Environmental Monitoring and Assessment 26:65-83.

Paul, J.F., H.A. Walker, W. Galloway, G. Pesch, D. Cobb, C.J. Strobel, J.K. Summers, M.

Charpentier and J. Heltshe. 2008. Combining existing monitoring sites with a probability survey design: examples from U.S. EPA's National Coastal Assessment. The Open Environmental & Biological Monitoring Journal 1: 16-25.

Stein, E.D. and B. Bernstein. 2008. Integrating probabilistic and targeted compliance monitoring for comprehensive watershed assessment. Environmental Monitoring and Assessment 144:117-129.

Stevens, D.L. and A.R. Olsen. 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99:262-278.

Stoddard, J.L., D.V. Peck, S.G. Paulsen, J. Van Sickle, C.P. Hawkins, A.T. Herlihy, R.M. Hughes, P.R. Kaufmann, D.P. Larsen, G. Lomnicky, A.R. Olsen, S.A. Peterson, P.L. Ringold, and T.R. Whittier. 2005a. *An Ecological Assessment of Western Streams and Rivers*. EPA 620/R-05/005. U.S. Environmental Protection Agency, Office of Research and Development: Washington, D.C.

Stoddard, J.L., D.V. Peck, A.R. Olsen, D.P. Larsen, J. Van Sickle, C.P. Hawkins, R.M. Hughes, T.R. Whittier, G. Lomnicky, A.T. Herlihy, P.R. Kaufmann, S.A. Peterson, P.L. Ringold, S.G. Paulsen and R. Blair. 2005b. *Environmental Monitoring and Assessment Program (EMAP) Western Streams and Rivers Statistical Summary*. EPA 620/R-05/006. U.S. Environmental Protection Agency, Office of Research and Development: Washington, D.C.

Figure 1. Map of probabilistic and targeted site locations in southern coastal California. Regions refer to CA Regional Water Quality Control Boards.

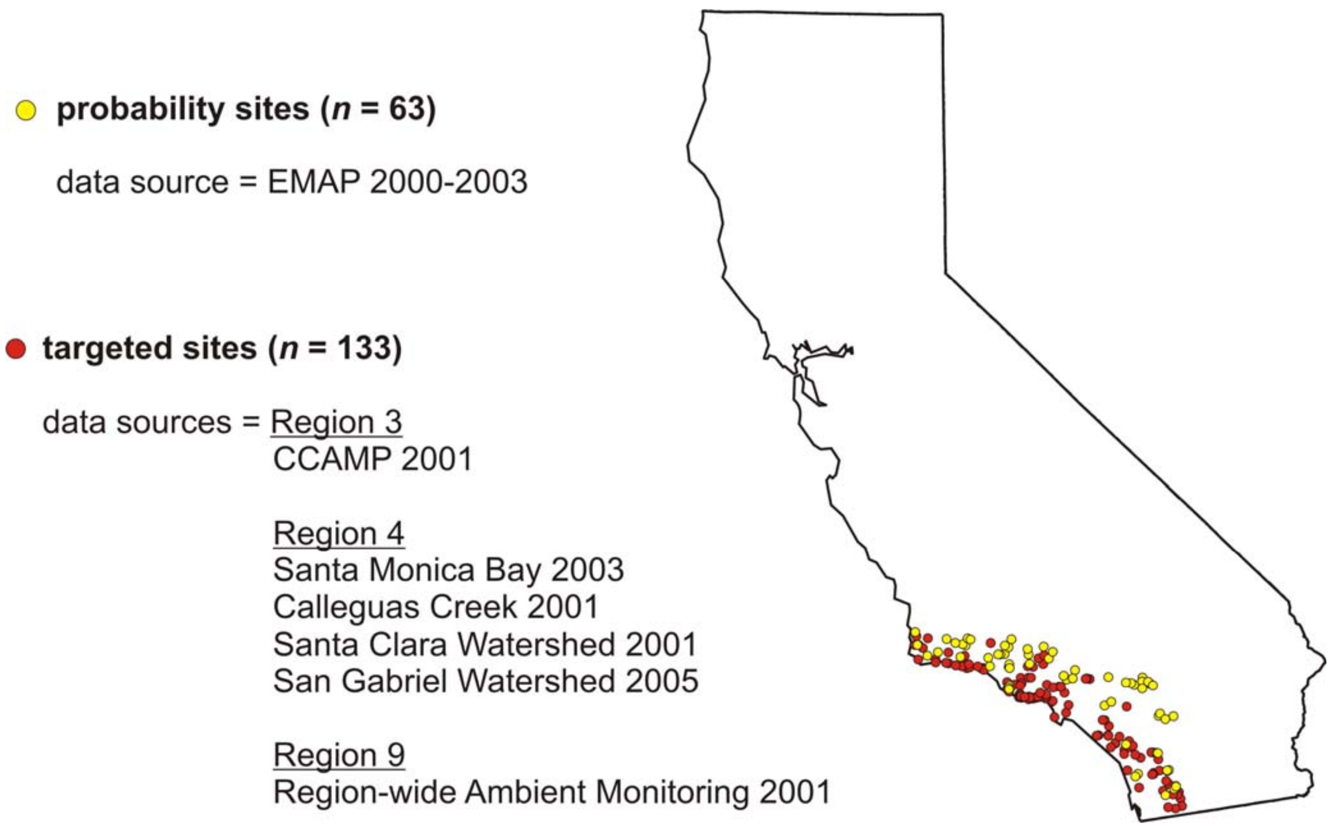

Figure 2. Results for SoCal IBI score. An example threshold between biologically “impaired” and “not impaired” streams (IBI score = ‘39’; see Ode et al. 2005) is shown in CDFs. The *p*-value is from a Kolmogorov-Smirnov test for differences between probability and targeted IBI distributions. The proportion of stream length (for probability sites) or the proportion of sites (for probability sites under equal weights and for targeted sites) in each of the IBI’s 5 condition categories is shown in pie charts.

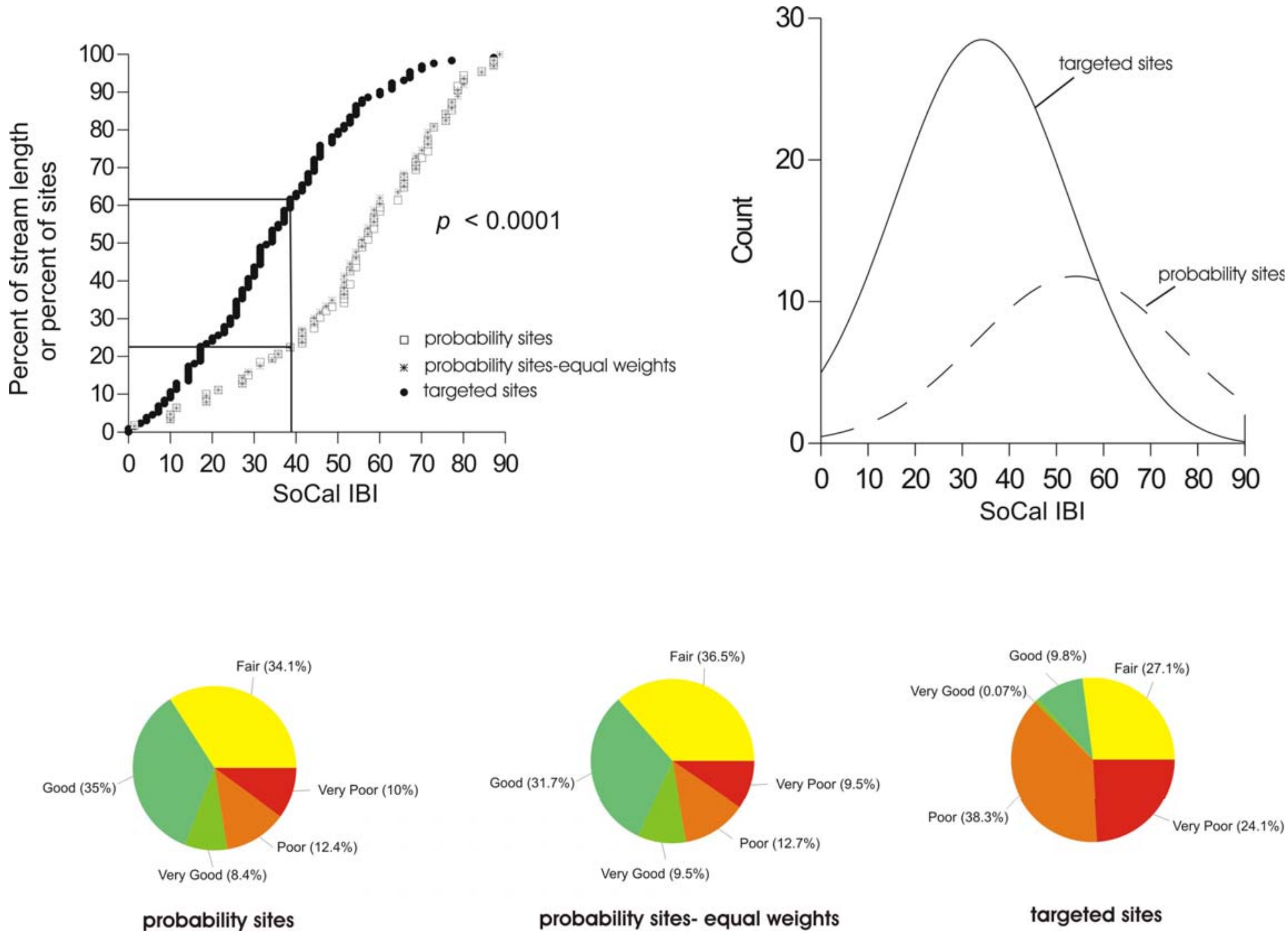

Figure 3. Results for qualitative physical habitat (PHAB) score. An arbitrary threshold (PHAB score = '150') is shown in CDFs. The *p*-value is from a Kolmogorov-Smirnov test for differences between probability and targeted PHAB distributions. The proportion of stream length (for probability sites) or the proportion of sites (for probability sites under equal weights and for targeted sites) above and below this threshold also is shown in pie charts.

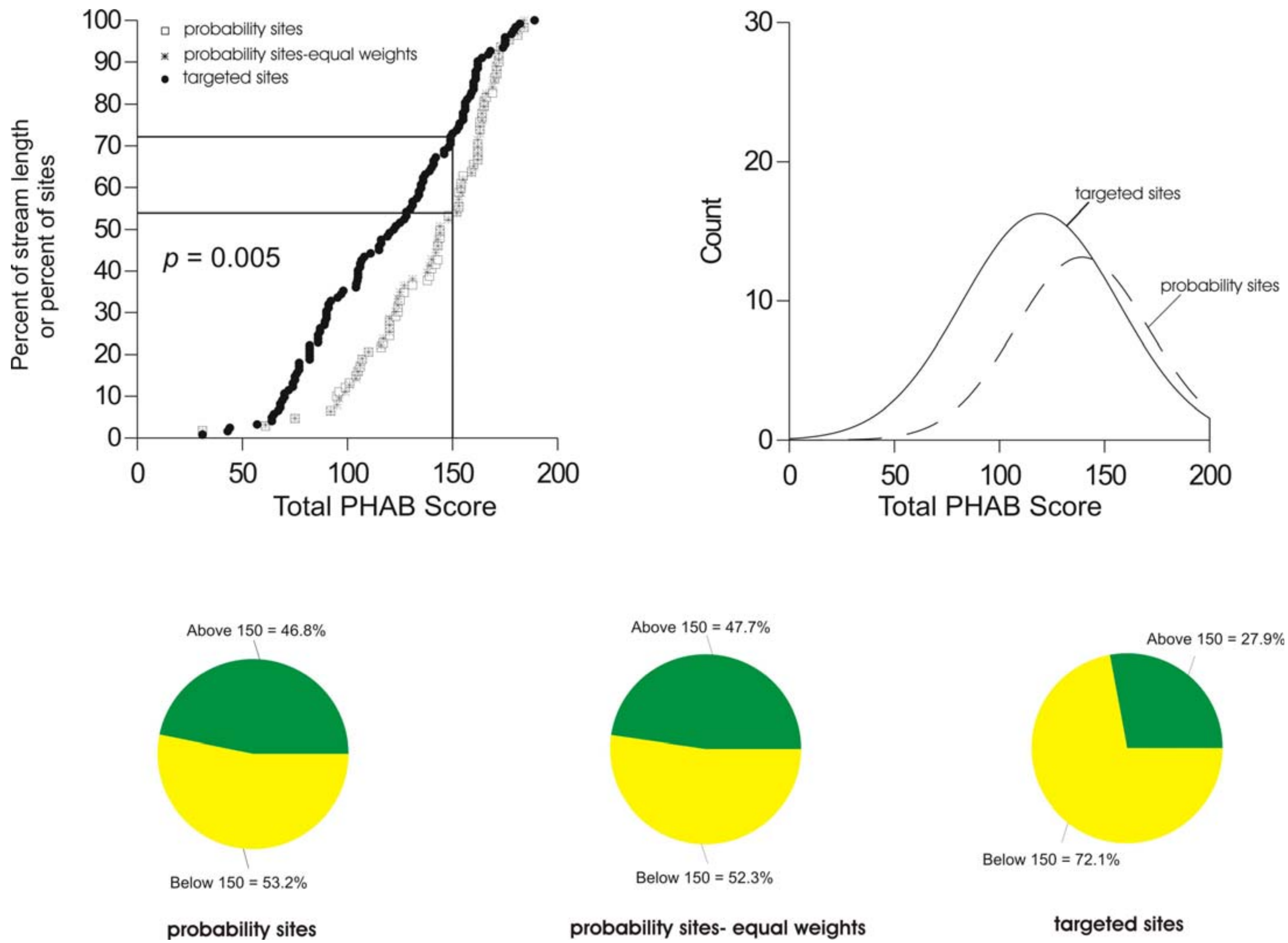

Figure 4. Results for log chloride concentration. The example threshold shown in CDFs (3.57 = 35.5 mg/L) was used to identify least-disturbed reference sites in xeric California for western EMAP. The *p*-value is from a Kolmogorov-Smirnov test for differences between probability and targeted chloride distributions. The proportion of stream length (for probability sites) or the proportion of sites (for probability sites under equal weights and for targeted sites) above and below this threshold also is shown in pie charts.

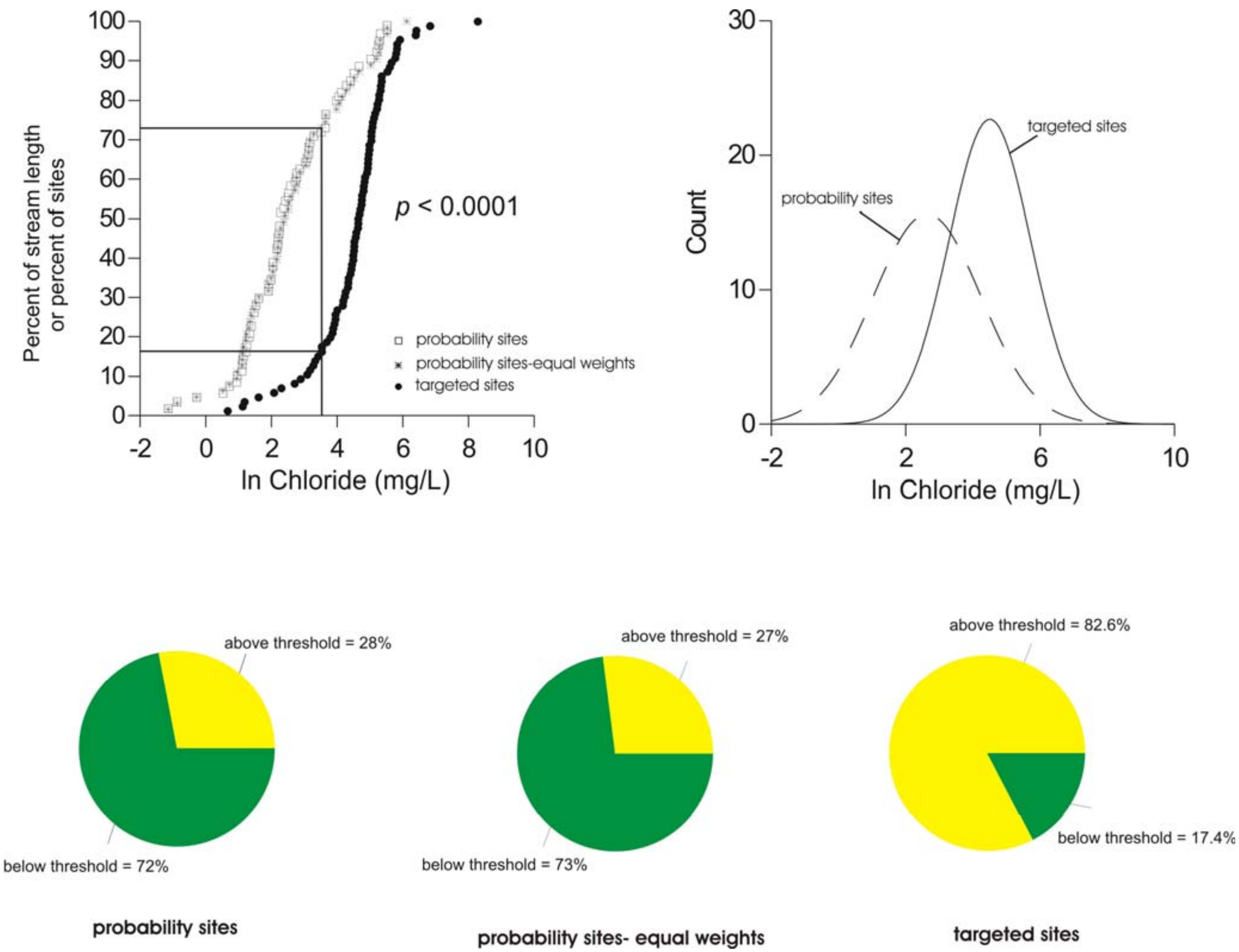

Figure 5. Results for log sulfate concentration. The example threshold shown in CDFs (6.17 = 480 mg/L) was used to identify most-disturbed sites in xeric California for western EMAP. The *p*-value is from a Kolmogorov-Smirnov test for differences between probability and targeted sulfate distributions. The proportion of stream length (for probability sites) or the proportion of sites (for probability sites under equal weights and for targeted sites) above and below this threshold also is shown in pie charts.

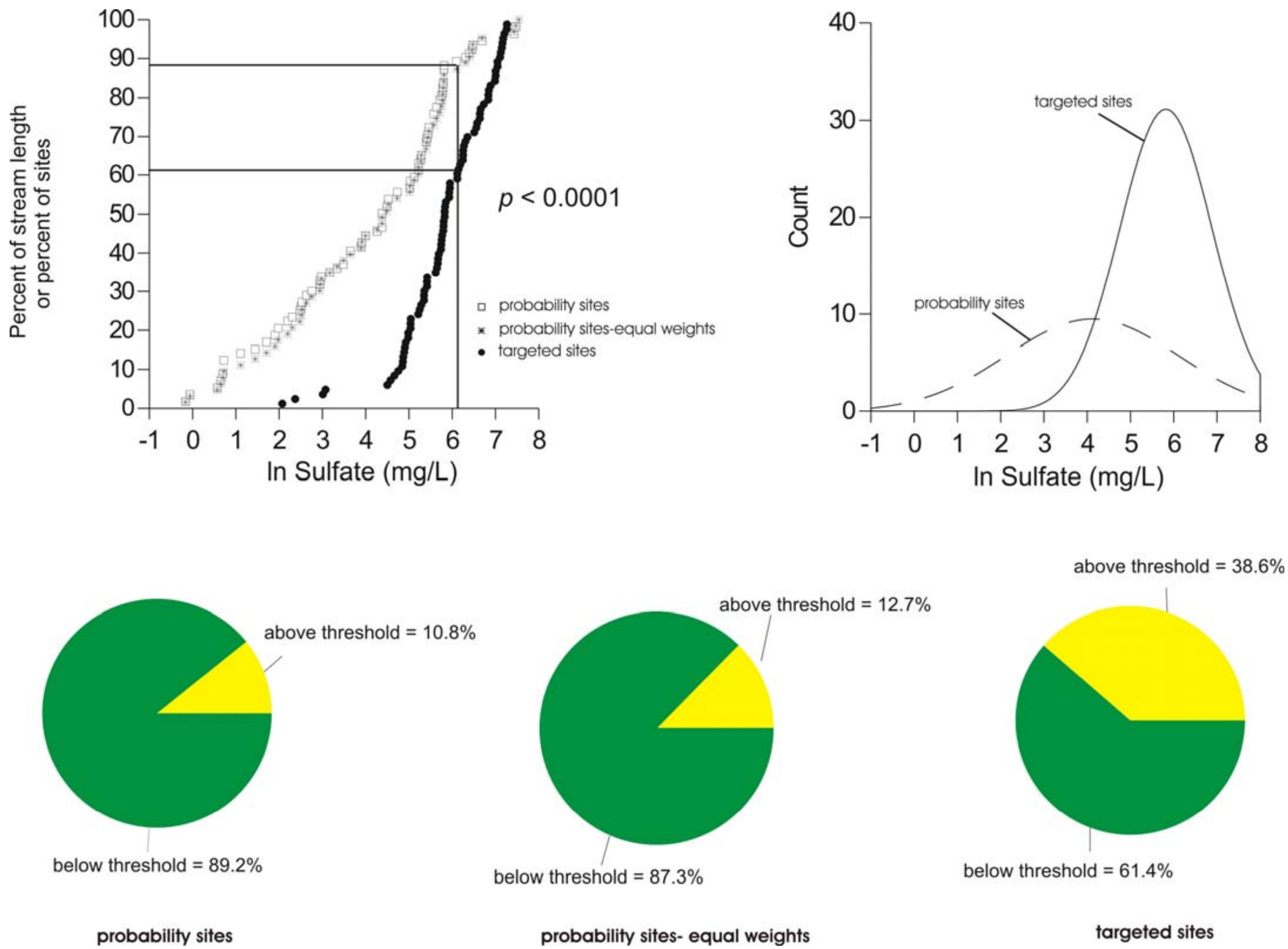

Figure 6. Results for log total nitrogen concentration. The example threshold shown in CDFs (0.47 = 0.6 mg/L) was used to identify most-disturbed sites in xeric California for western EMAP. The *p*-value is from a Kolmogorov-Smirnov test for differences between probability and targeted total nitrogen distributions. The proportion of stream length (for probability sites) or the proportion of sites (for probability sites under equal weights and for targeted sites) above and below this threshold also is shown in pie charts.

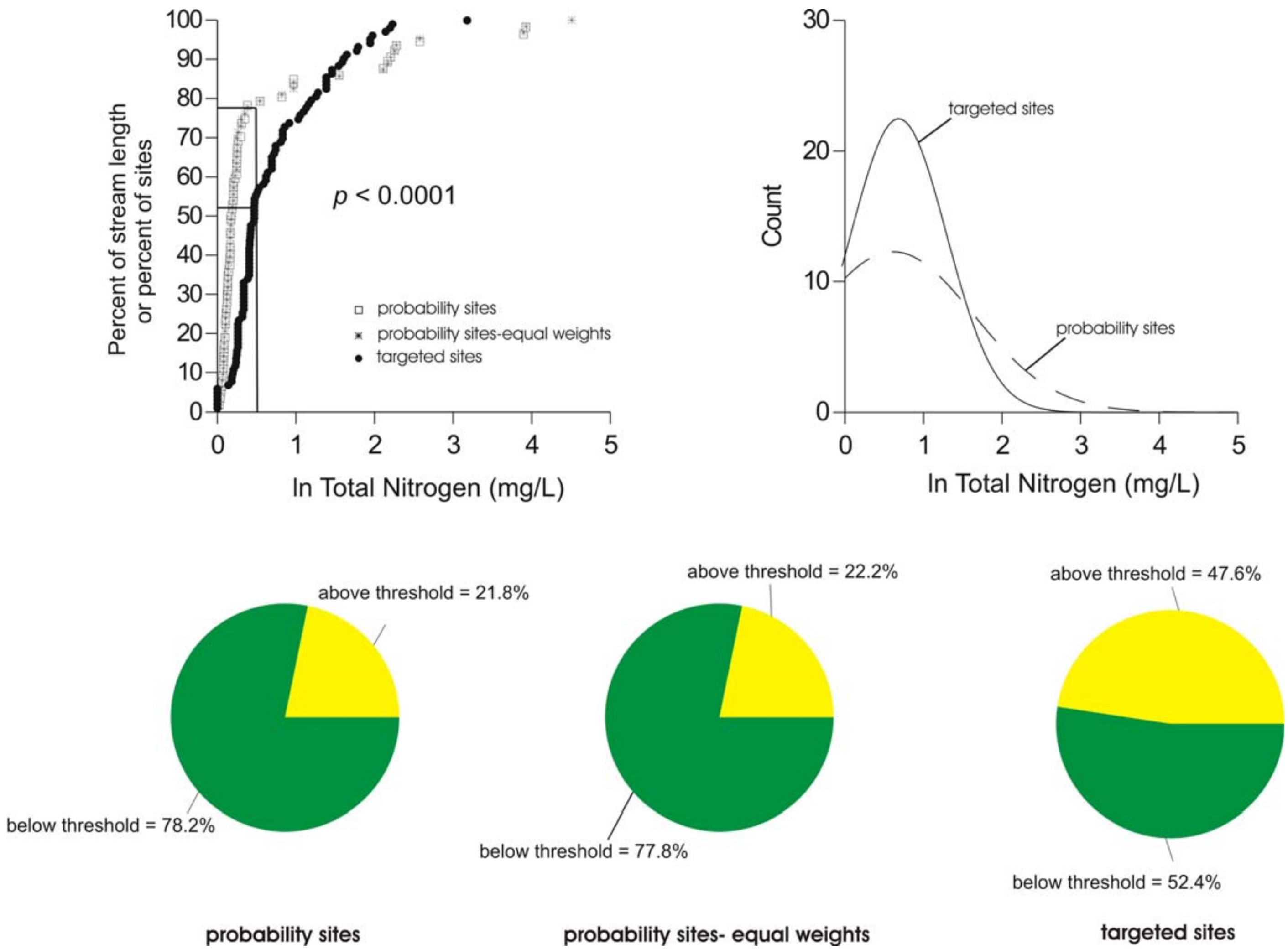

Figure 7. Results for log turbidity. The example threshold shown in CDFs (1.79 = 5 NTU) was used to identify least-disturbed sites in xeric California by Ode et al. (2008). The *p*-value is from a Kolmogorov-Smirnov test for differences between probability and targeted turbidity distributions. The proportion of stream length (for probability sites) or the proportion of sites (for probability sites under equal weights and for targeted sites) above and below this threshold also is shown in pie charts.

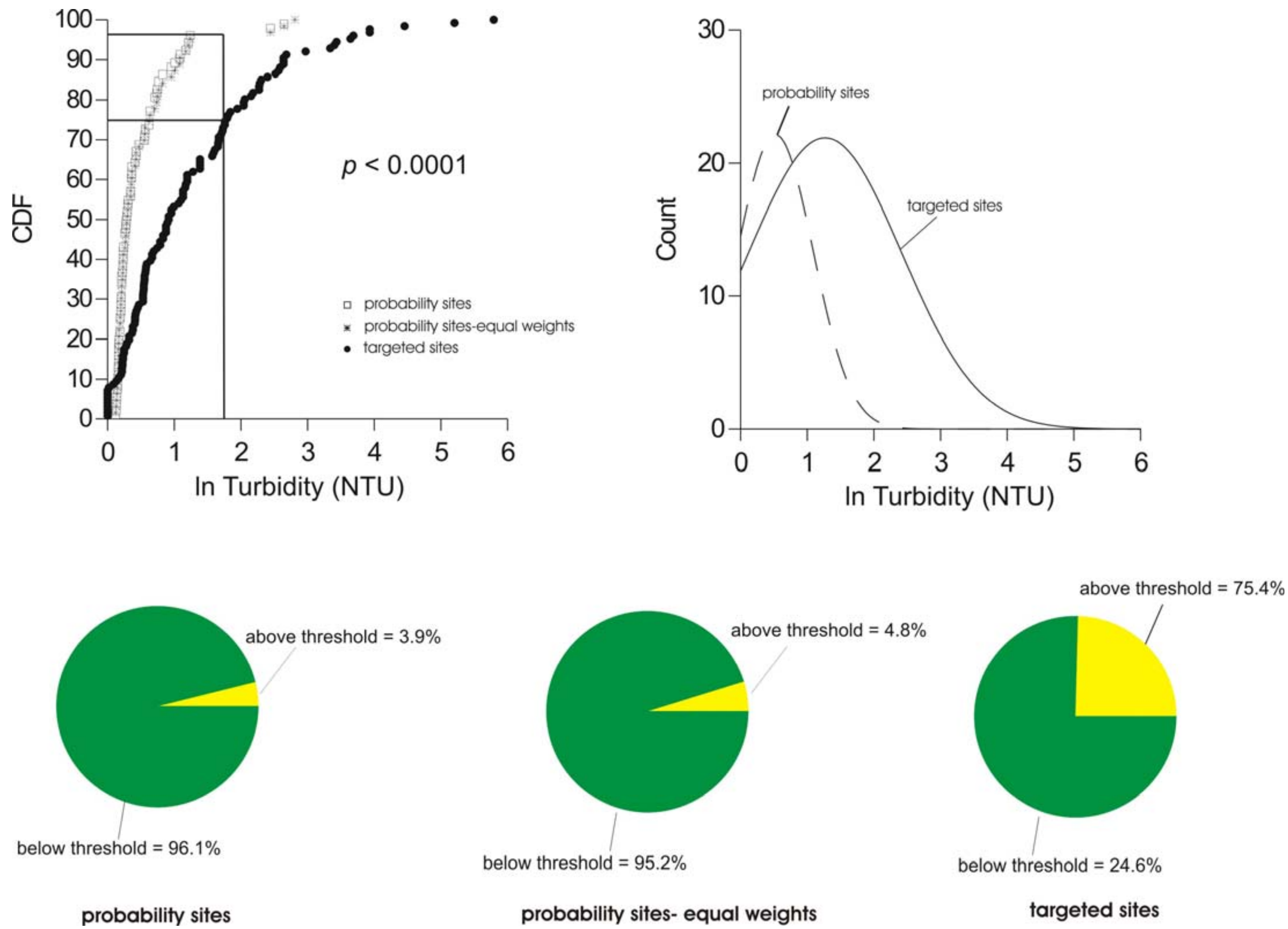

Figure 8. Comparisons of local (1km upstream) urban and agricultural land use between targeted and probability sample populations. The *p*-value is from a Kolmogorov-Smirnov test for differences between probability and targeted turbidity distributions.

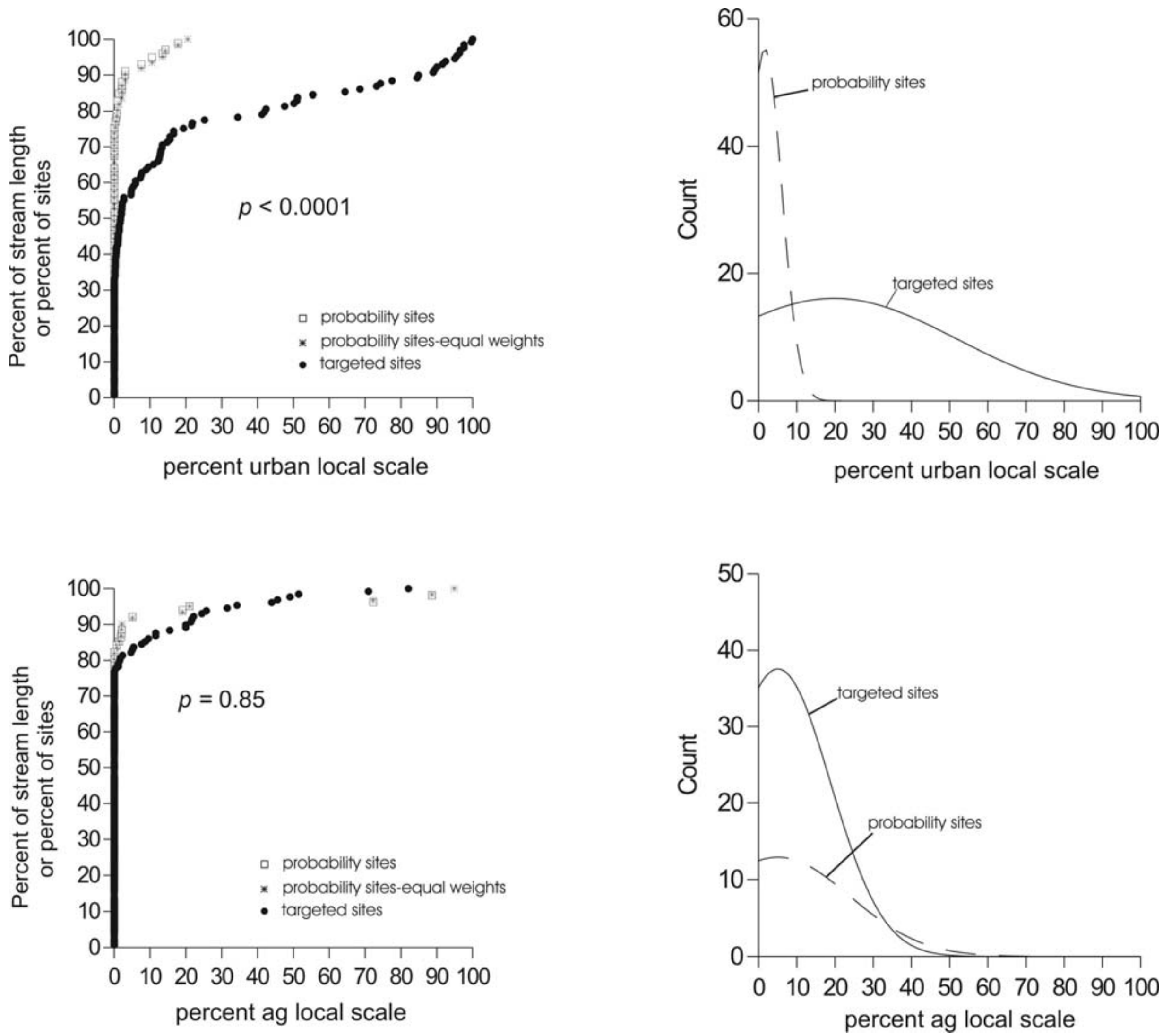

Figure 9. Comparisons of local (1km upstream) urban and agricultural land use between targeted and probability sample populations. The p-value is from a Kolmogorov-Smirnov test for differences between probability and targeted turbidity distributions.