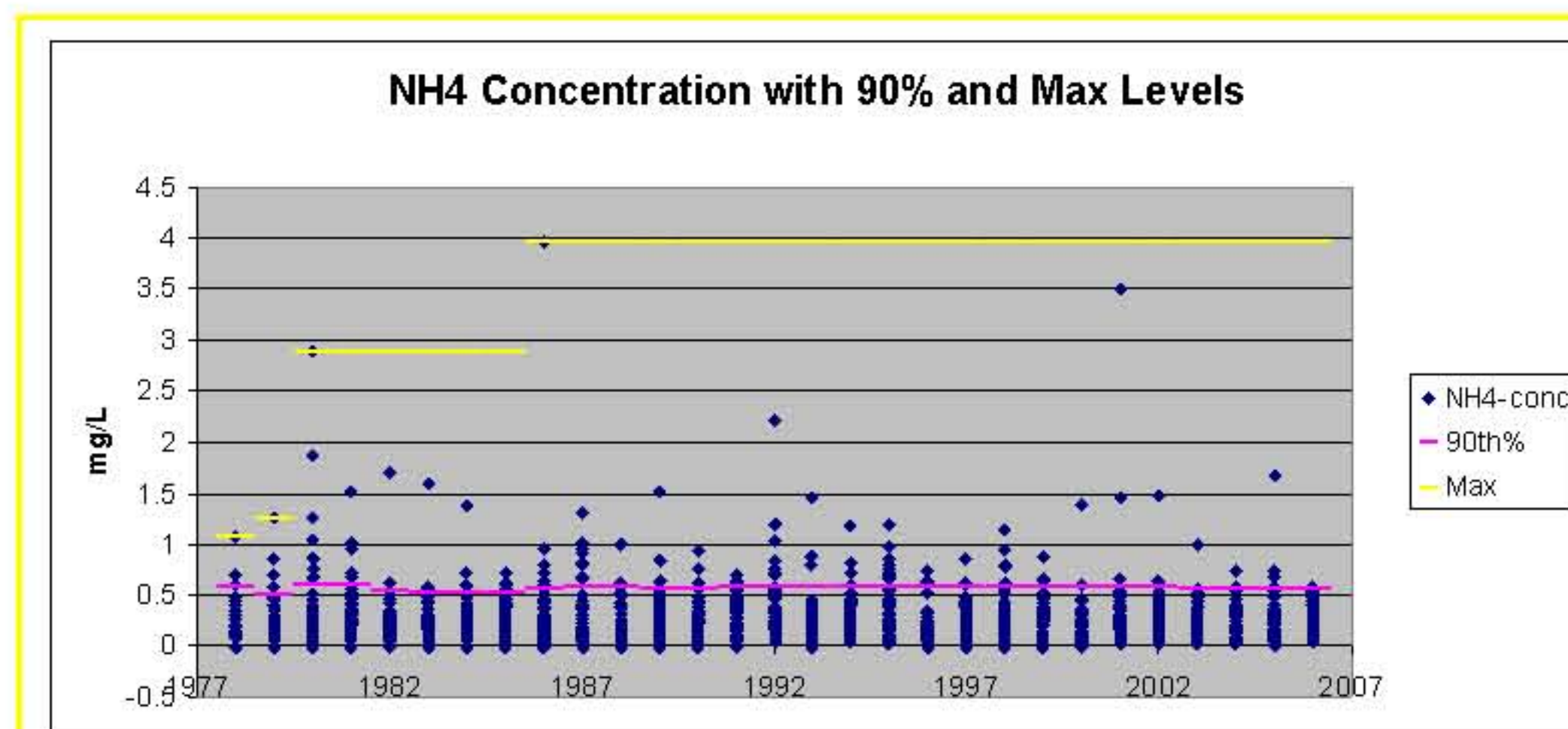


NADP/NTN Data Validation Coding and Use of Site History: Comparison of utilizing analyte concentration vs. analyte deposition

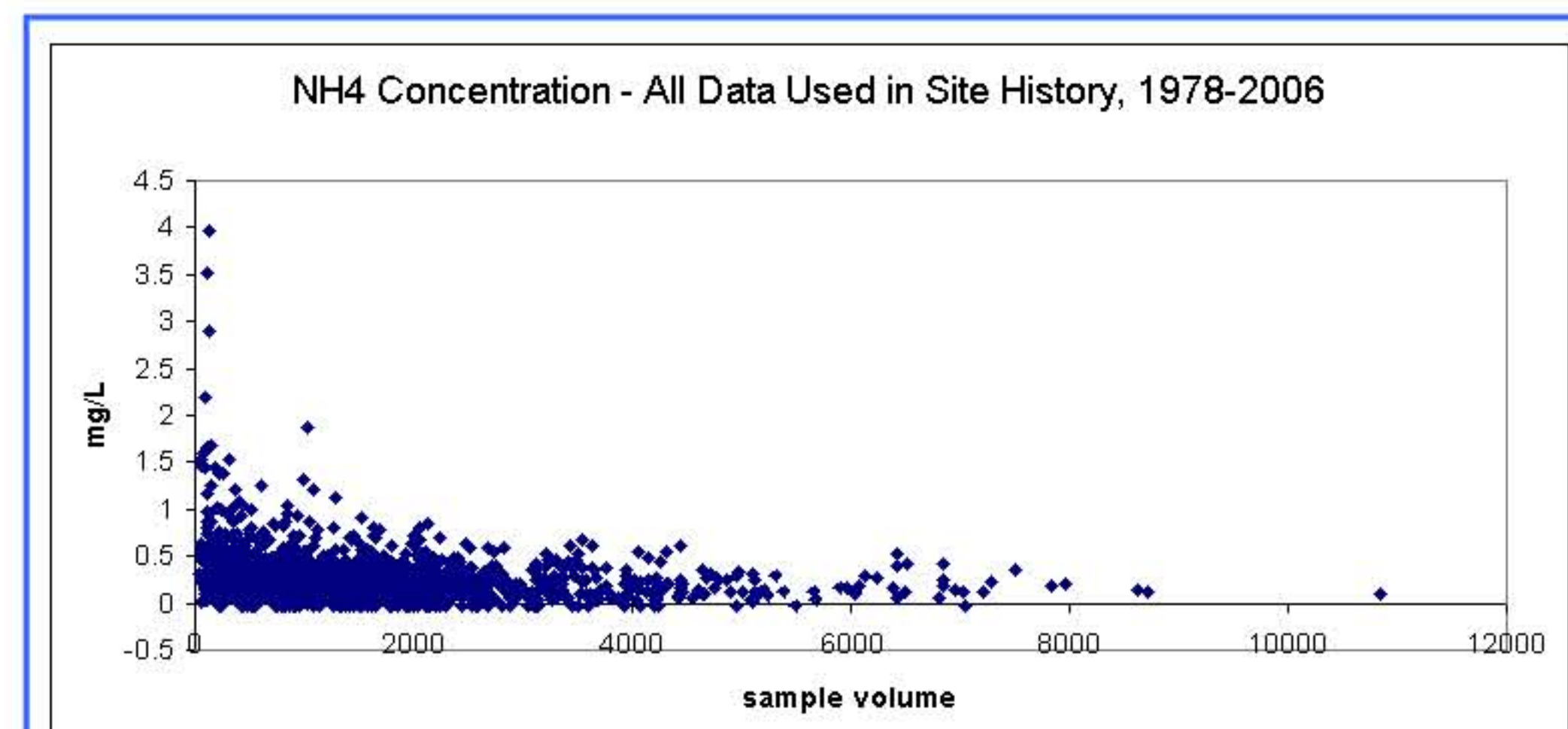
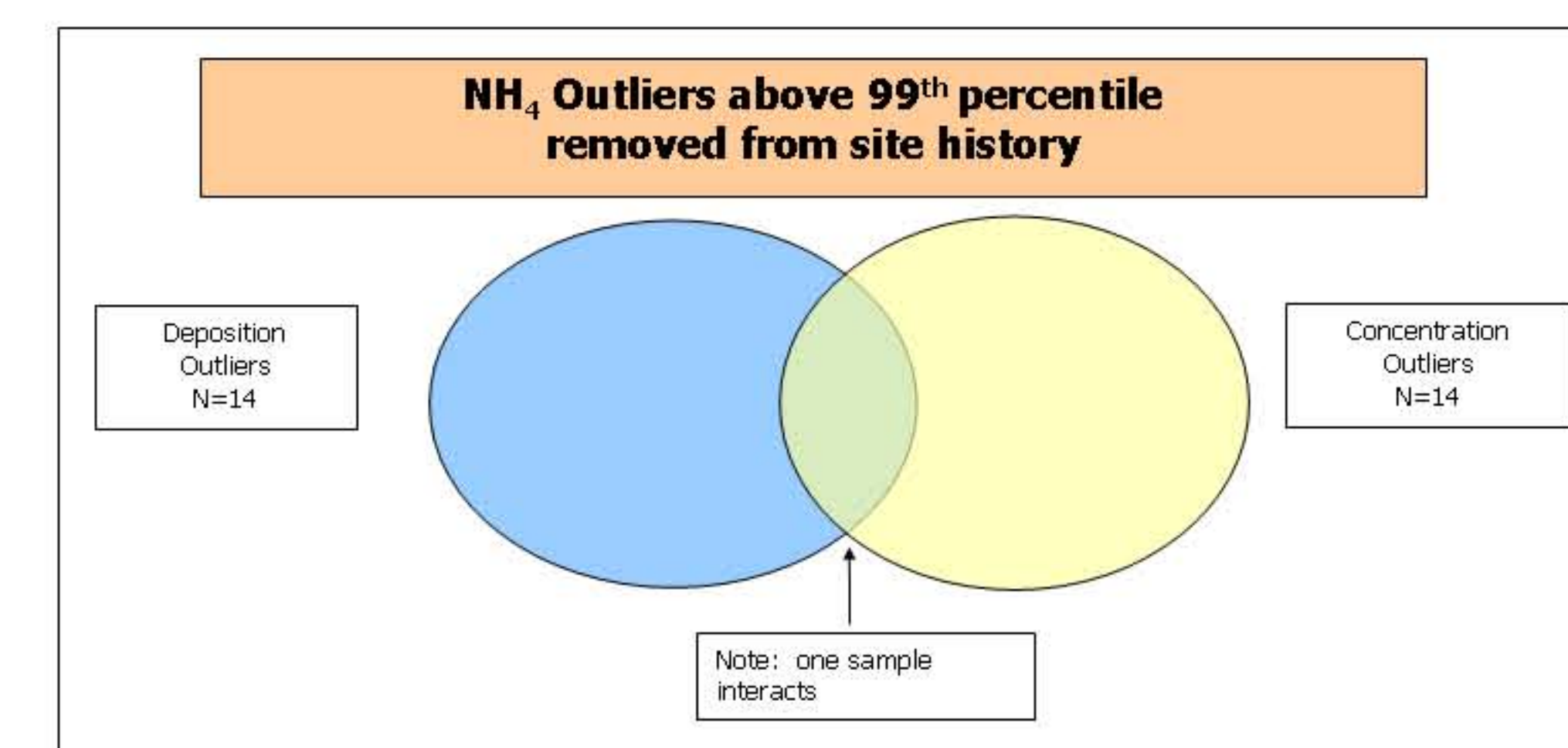
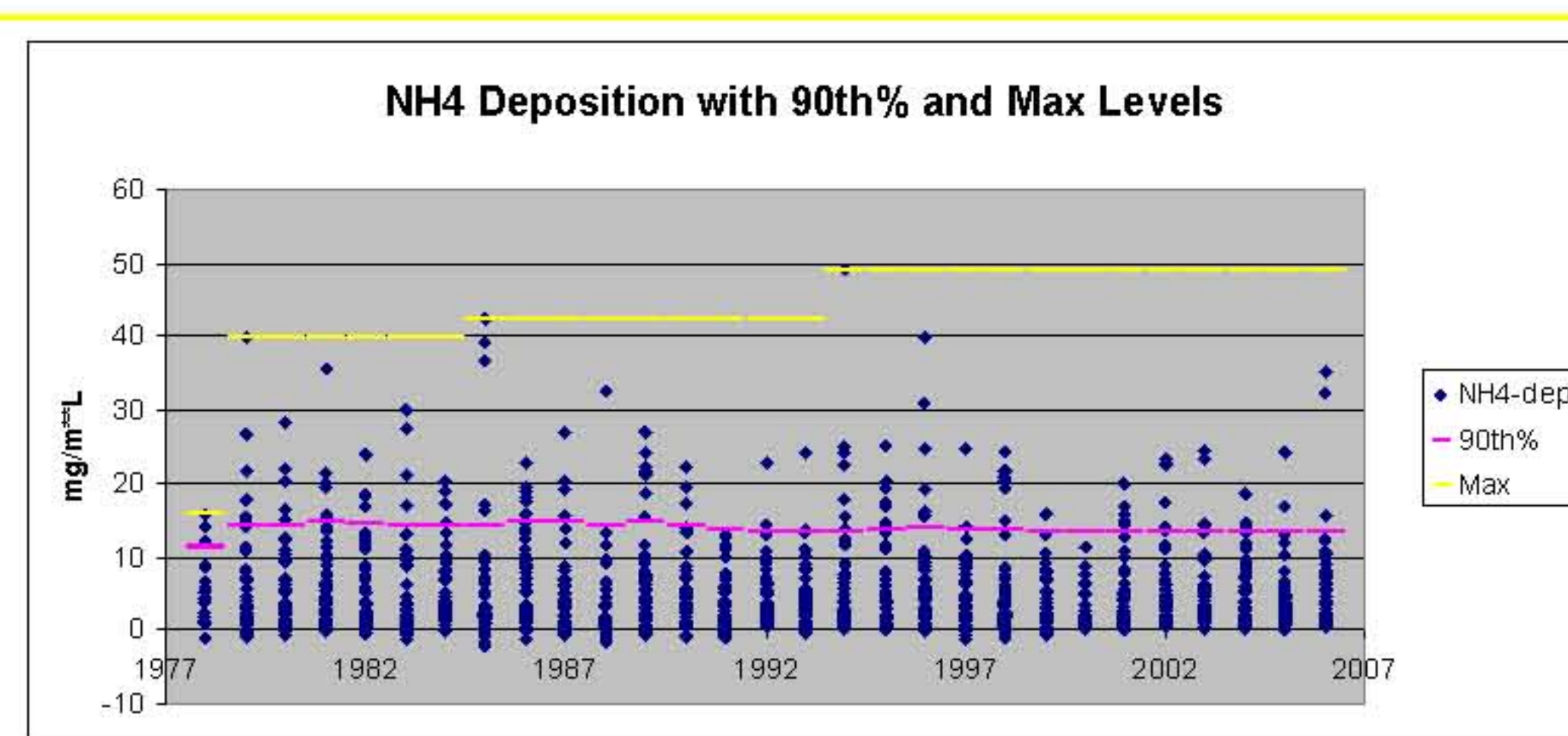
Barbara L. Suever, Jane Rothert, Tom Bergerhouse
National Atmospheric Deposition Program, Central Analytical Laboratory, Illinois State Water Survey

Introduction/Background

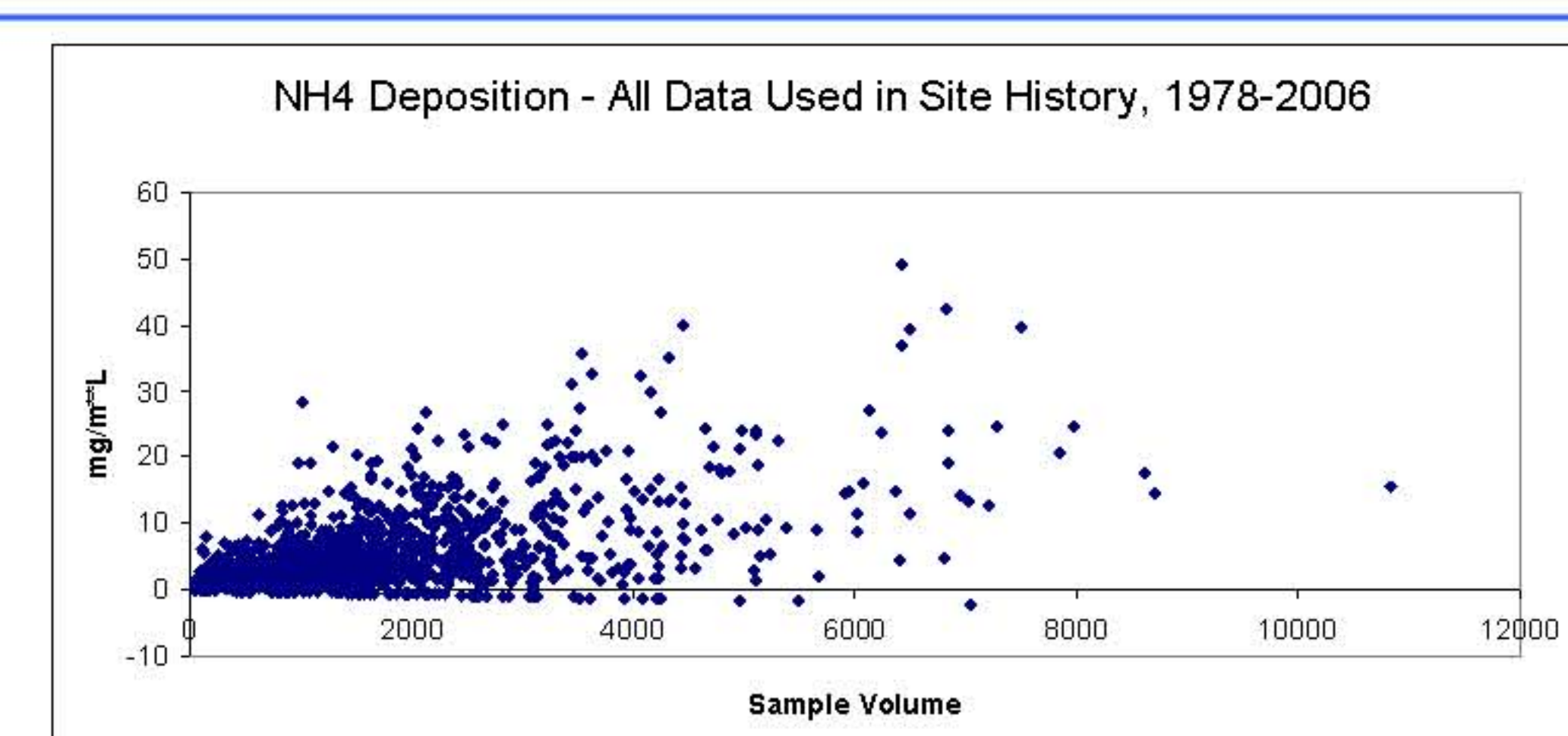
- The NADP/NTN is committed to providing high-quality, consistent data to researchers and other data users. To this end, one of our responsibilities is to ensure that contaminated samples have been thoroughly screened for anomalous chemistry.
- If a sample is found to be grossly contaminated samples are given a Screening Level (SL) code of "C" indicating that this sample should not be used by researchers or other data users.
- Screening for grossly contaminated samples is currently being done by comparing the contaminated sample's *concentration* of analytes to all wet-type samples without noticeable contamination from the start-up of that site.
- We have noticed, over time, an ever-expanding distribution of chemistry data which may lead to a more liberal screening for anomalous chemistry of contaminated samples.
- We examined if utilizing *deposition* values in place of *concentration* (currently being used) would limit the expansion of the distributions.
- We also examined if employing a 99th percentile distribution instead of the maximum values currently used would limit the expansion of the distributions.
(We examined chemistry data for SO₄, NO₃, and NH₄ from WV18, our longest running site.)



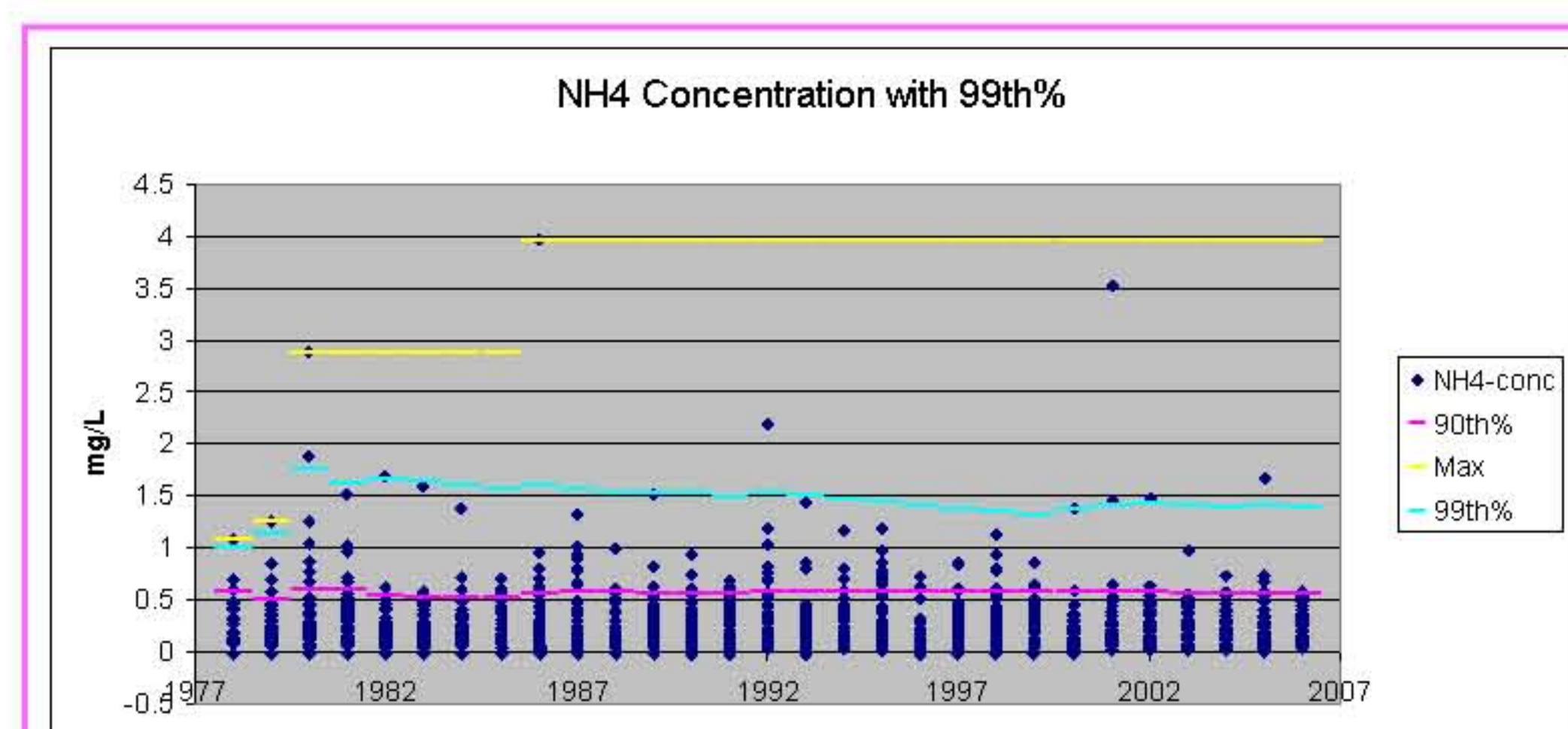
Using a maximum value for site history appears to be too liberal of an exclusion criteria; the maximum values are likely contaminated samples that have not been marked as contaminated and should not be included in site history.



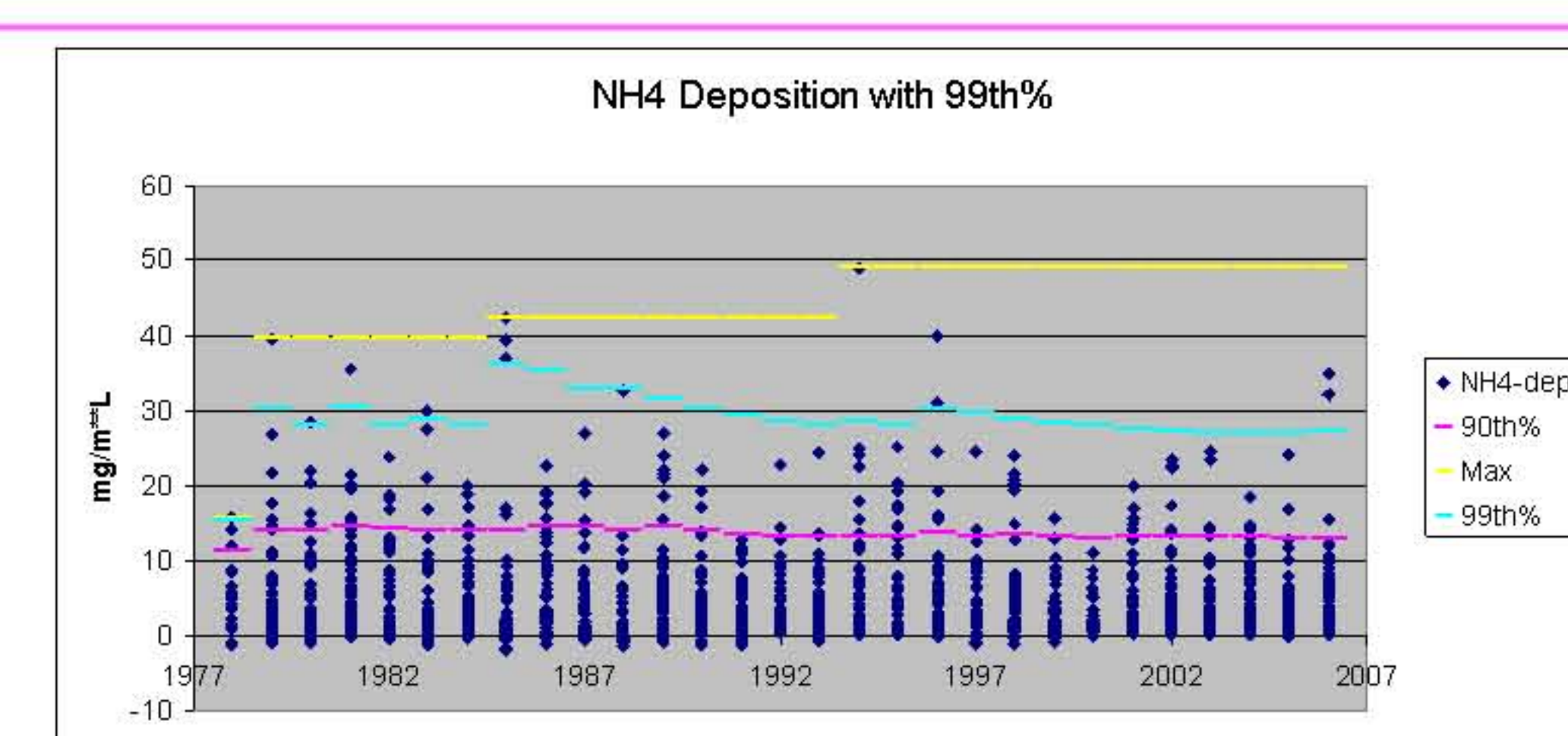
The samples that are outliers in deposition are not typically outliers in concentration and vice versa (see diagram upper right).
Deposition outliers are typically high sample volume; Concentration outliers are typically lower sample volumes.



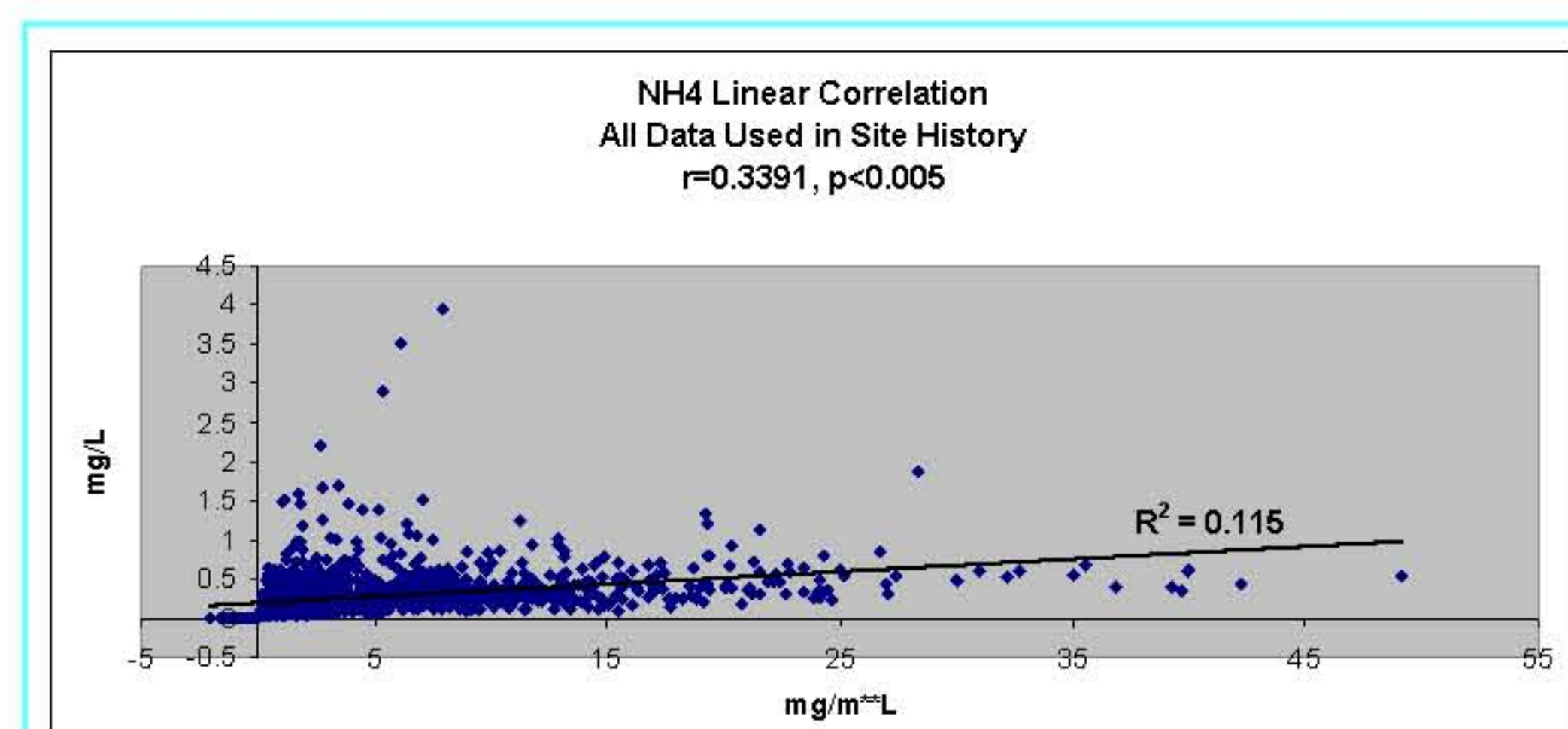
			Z	p
All Data	vs	99th% Outliers for Both	-3.70	0.0001
All Data	vs	99th% Outliers for Conc	-3.46	0.0003
All Data	vs	99th% Outliers for Dep	0.11	0.4562
99th% Outliers for Both	vs	99th% Outliers for Conc	0.24	0.4052
99th% Outliers for Both	vs	99th% Outliers for Dep	3.80	0.0001
99th% Outliers for Conc	vs	99th% Outliers for Dep	3.26	0.0004



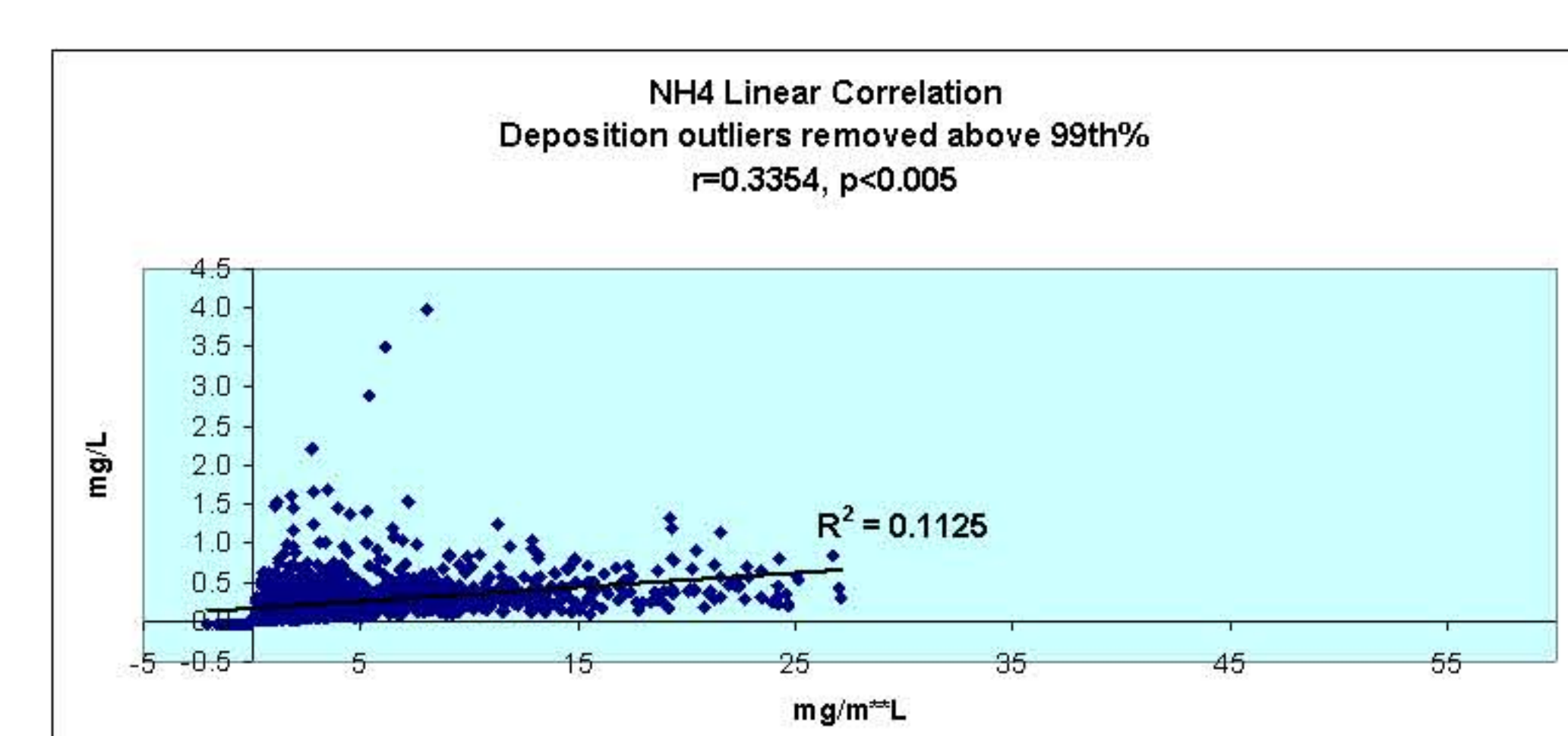
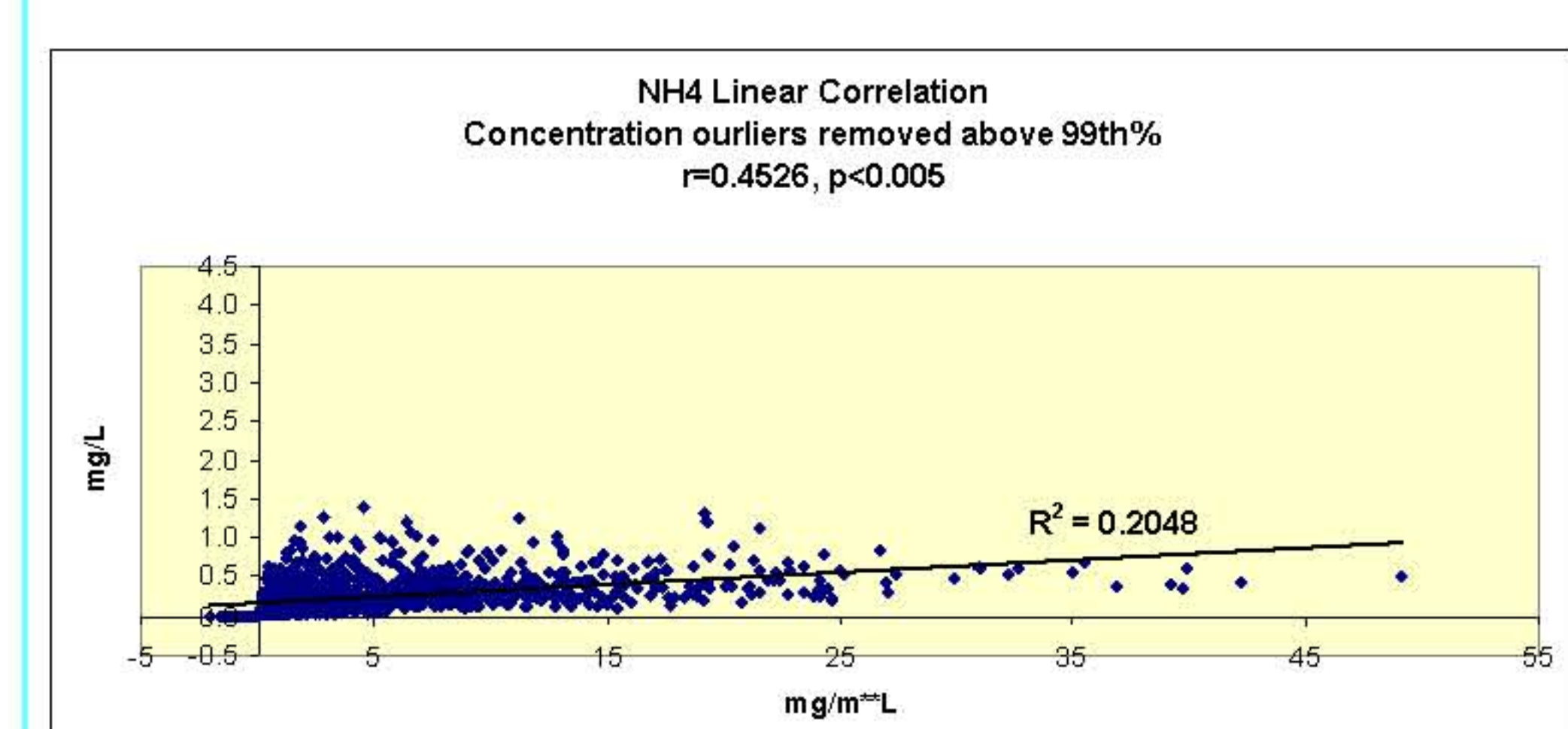
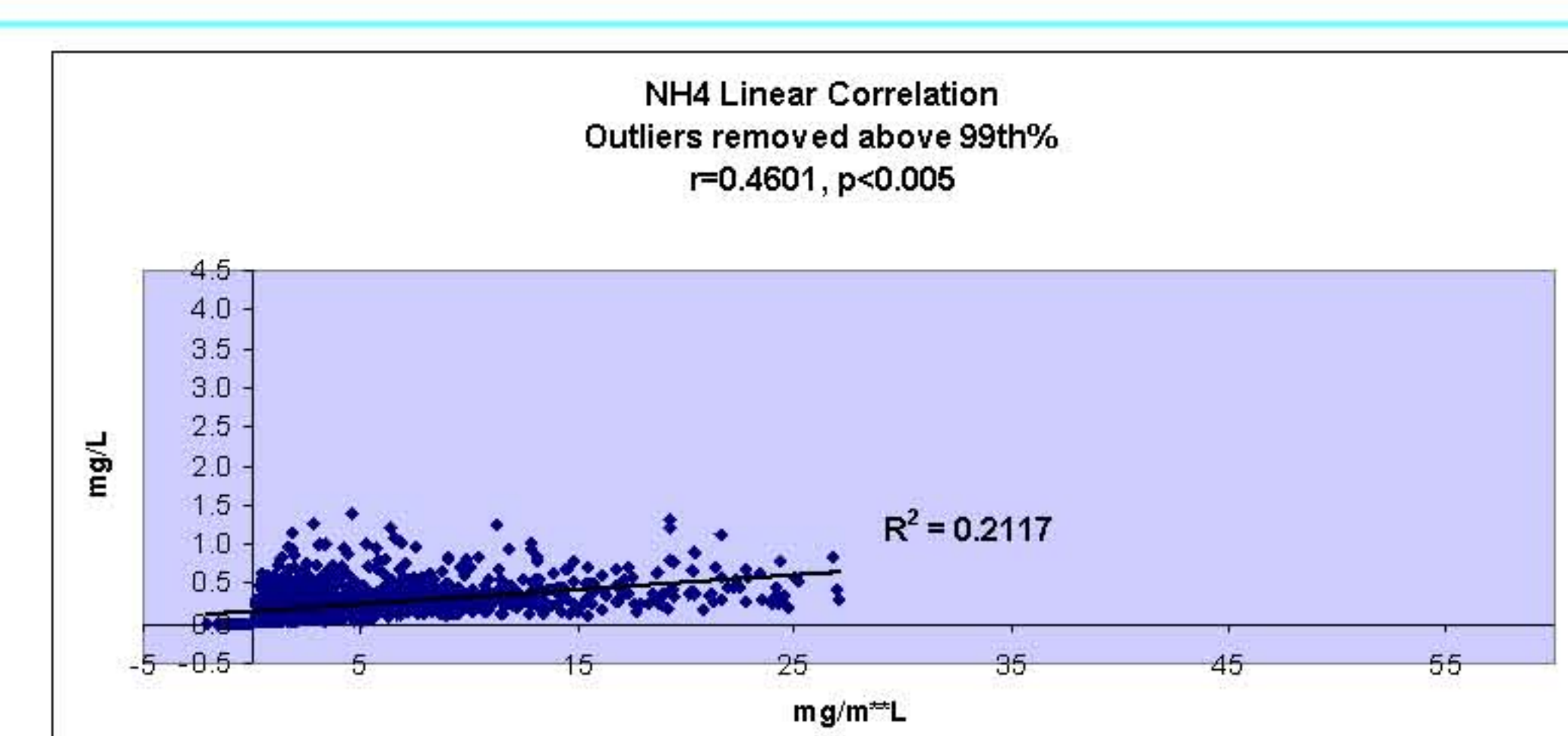
Since contaminated samples may continue to find their way into the site history, a new method for an upper-level exclusion criteria, in place of the maximum value, is needed. We propose using a 99th percentile exclusion criteria in place of the maximum value.



			Z	p
All Data	vs	99th% Outliers for Both	-2.54	0.0055
All Data	vs	99th% Outliers for Conc	-3.09	0.0010
All Data	vs	99th% Outliers for Dep	0.42	0.3372
99th% Outliers for Both	vs	99th% Outliers for Conc	-0.55	0.2912
99th% Outliers for Both	vs	99th% Outliers for Dep	2.95	0.0016
99th% Outliers for Conc	vs	99th% Outliers for Dep	3.50	0.0002



Using the 99th percentile to eliminate outliers significantly improves the linear correlation of the site history. This suggests that removal of outliers using the 99th percentile allows us to more significantly predict deposition from concentration and vice versa.
However, as we look closer, we find that only removing concentration outliers, not deposition outliers, improves the linear correlation of the data.



			Z	p
All Data	vs	99th% Outliers for Both	-1.32	0.0934
All Data	vs	99th% Outliers for Conc	-1.50	0.0668
All Data	vs	99th% Outliers for Dep	0.14	0.4443
99th% Outliers for Both	vs	99th% Outliers for Conc	-0.18	0.4286
99th% Outliers for Both	vs	99th% Outliers for Dep	1.46	0.0721
99th% Outliers for Conc	vs	99th% Outliers for Dep	1.64	0.0505

Conclusions

To make the site history data more linear and to remove potentially contaminated samples (while still keeping the needed variation in the data) we can remove outliers at the 99th percentile level from concentration to significantly improve the overall fit of the data in site history.

Our data does not show that concentration is a "more accurate" or "more valid" measure of determining outliers. Our data simply states that the linear relationship between deposition and concentration is improved when concentration outliers are removed.