Calculation of MACT Floors in EPA's LMWC Proposal

Bruce C. Buckheit

March 25, 2024

The following analyses use the LMWC emissions data and calculations based on that data to provide examples of the various problems with EPA's methodology to calculate the MACT Floors in the LMWC proposed rule. I am happy to discuss these calculations with EPA staff and provide the underlying calculations upon request.

Section 112 requires that the new source MACT floor be "not less stringent than the emission control that is ***achieved in practice*** by the best controlled similar source" and that the existing source MACT floor be "not less stringent than the average emission limitation ***achieved*** by the best performing 12 percent of existing sources (for which the Administrator has emissions information)." Section 129 is similar in that it also requires that the floor for new sources "not be less stringent than the emission control that is ***achieved in practice*** by the best controlled similar unit" and that the existing source floor "shall not be less stringent than the average emission limitation ***achieved*** by the best performing 12 percent of units." Section 129 differs from section 112 in that section 129 (1) does not contain the limitation "for which the Administrator has emissions information" and (2) excludes recently constructed units that applied LAER level controls from the pool of best performing units.

It is important to recognize that these provisions employ the *past* tense "***achieved***", not "may achieve" or "can be expected to achieve." This language, the legislative history and common sense suggest that Congress intended a simple review of past performance data to determine what emission limitation has been achieved. EPA's "Upper ***Predictive*** Level" does not address what emission limitation has been ***achieved by*** the best performing units. Rather, it attempts to predict what might have happened in the past had sources been tested more frequently or what might happen in the future, and as demonstrated below, fails at doing so.

The use of "average" in the provision suggests that the floor is to be the average of the emission limitations "***achieved***" by the individual units in the pool of the best performers – not the pooled variation of the UPL of those units. There is no reason to believe that Congress intended any exotic statistical test when it used the word "average." Accordingly, the best reading of the word "average" is its common usage, i.e. arithmetic average or mean.

EPA has asserted that the inclusion of the concept of "averaging" emission limitations late in the development of the CAAA introduced an ambiguity that was difficult to resolve. Originally, the Senate version of the legislation called for a cutoff at the emission limit achieved by the top 10 percentile of the best performing sources, the House version called for 15 percent, and the compromise was "the average of the top 12 percent". However, while intended as a compromise, this language is more stringent than either of the prior versions (since it nominally would represent the performance of 6 percent of the best performing sources). At the time EPA OGC advised its Office of Inspector General that it believed the 88th percentile was appropriate, that it had identified a "work around" and that no clarification of legislative intent was necessary.[1] In order to effectuate what it believed to be Congressional intent, EPA's "workaround" should have approximated a floor at the performance of the 88th percentile. But EPA's workaround goes far beyond what would approximate the 88th percentile performance.

One might argue that the test of what has been ***achieved*** in the past might be the arithmetic average of the best (i.e., lowest) emission rate documented in any test of the unit. A more conservative reading is that the floor be no less stringent than the arithmetic average of the "worst" test result of the best performing units.[2] Since the best performers have by now been tested on multiple occasions under what are required to be "representative conditions", the "worst" test result is a reasonable accommodation for "test-to-test" variability when one considers the overall "stack testing" scheme incorporated in the regulations (see below). Additional protection against an "impossible" standard is provided by the statute's choice of the average emission limitation achieved by the top 12 percent performers – or even the emission limitation achieved by the 12th percentile unit.

It is also useful to consider the implications of the term ***"in practice"*** in the requirement for new source floors. "***In practice"*** is

> "used to describe what really happens as opposed to what you think will happen in a particular situation."[3]

EPA asserts that it uses the "UPL to estimate the average emissions performance of the units used to establish floor standards at times other than when the stack tests were conducted." In doing so EPA ignores the term *"in practice"* and substitutes what it thinks will happen (or may have happened) – the UPL – in lieu of the stack tests themselves that document what "really happened". While this phrase does not appear in the existing source floor language it is reasonable to assume that Congress intended a similar approach for those floors as well.

EPA has apparently not examined whether a straightforward reading of the statute's floor language would produce unreasonable results. Table 1, below, sets out the outcome of three alternative interpretations of "achieved in practice" – (1) the average of the best single test result for each of the top 12 percent performers, (2) the average of all tests for each of the top 12 percent performers and (3) the average of the worst test result for each of the top performers.[4] As noted in the table, some results were ignored as outliers.[5]

[1] mactsrep.pdf (epa.gov)
[2] Excluding those test results that were terminated early or otherwise not reported to a permitting authority.
[3] IN PRACTICE | definition in the Cambridge English Dictionary
[4] The top performers are as ranked by the metric being examined.
[5] More low values were considered outliers than high values.

Table 1. EPA Emission Guidelines vs. Alternate Floor Options – Top 12 Percent

Pollutant	EPA 99th percentile UPL	Average of Lowest Test Result	Average Test Result	Average of Highest Test Result
Cd	1.44	0.19	0.59	1.2
Pb	54.74	1.78	7.45/7.1/7.92[6]	17.6/25.9[7]
PM	7.36	0.98[8]	2.9	5.25
Hg	10.291	0.93	3.27	7.89
DF	7.18	0.77/0.40[9]	2.81/1.35[10]	5.42/2.59[11]
HCl	12.49	1.9	5.28	10.46[12]

Similarly, Table 2 examines the outcome of setting new source floors on the basis of emission limitations "achieved in practice."

Table 2. EPA Emission Guidelines vs. Alternate Floor Options – Best Performing Unit[13]

Pollutant	EPA 99th percentile UPL	Lowest Test Result	Average Test Result	Highest Test Result
Cd	0.492/1.1[14]	0.23	0.29	0.38
Pb	12.19	0.87	3.58	6.61
PM	4.81	0.95/0.067[15]	1.28	2.50
Hg	6.07	0.77	2.31	3.83
DF	1.73	0.45	0.79	1.39
HCl	7.8	0.87	2.27	3.65

Which of these interpretations is most reasonable should be evaluated against the notion that Congress intended that the floors should require most units to reduce emissions.

EPA has also stated that it uses its UPL approach to address variation in the emission performance of units. At times it talks about future performance – i.e., what would happen if

[6] Depending on which of the several data sets is used.
[7] Depending on whether outliers are excluded.
[8] Outlier excluded.
[9] Depending on whether outliers are excluded.
[10] Depending on whether outliers are excluded.
[11] Depending on whether outliers are excluded.
[12] Depending on whether outliers are excluded.
[13] Based on the data in Tabs A12-A17.
[14] EPA proposes a limit (1.1) based on 3 times the representative detection limit. The 99th percentile UPL is 0.492.
[15] Depending on whether an outlier is excluded.

the same units are retested at some later time. At other times the agency has argued that it is intended to assess the variation inherent in the past testing that forms the basis of the average. However, the most common – and commonly understood – way to address variation in performance over time is to "average" individual values. Thus, we ordinarily track a baseball player's batting and fielding average, the average of political polls, daily and monthly temperature, rainfall and so on. Here, the variability in performance has been addressed by using the average of three test runs to determine a "test result" and then again for existing sources by the statute's use of the average of the performance of a group of sources.

A source that is not subject to a meaningful constraint on emissions will likely show more variability than one that emits at a level that is close to an applicable limit. One can expect more and greater non-random actions to limit emissions variability in the latter case. There is no reason to believe that on a going-forward basis, these Congressionally directed techniques for addressing variation are insufficient. EPA has not suggested that it would be impossible for sources to meet a limit based on the arithmetic average of the tests of the top performers, merely that, for unjustifiable reasons, it prefers to provide additional measures to address variability – measures that severely reduce the effectiveness of the Congressional mandate.

EPA has and will continue to employ a different test for determining the emission limitation *achieved* in compliance testing than it does for determining the emission limitation *achieved* in setting the floor. If EPA believes that it is appropriate to employ a UPL approach for determining the emission limit (so as to consider variability in the performance of a unit) it should also adopt a UPL approach for determining compliance with that limit. As EPA argues, the CAA requires continuous compliance – not just compliance for 3 hours every year.

Industry may argue that calculating a UPL based on 3 individual runs leads to absurdly high variability – as we have seen with the use of UPL in calculating new source floors. However, operators are free to reduce the effect of test-to test variability by increasing the number of test runs, better operation of control devices or increasing the compliance margin of the control device. The cost of this should not be a concern because if EPA's calculations on the UPL of the "best units" are to be believed, those calculations undercut the agency's assertion that (increasingly more) intermittent test results are sufficient to document continuous compliance with emissions limits. There are 2,920 3-hour periods in a year. The 99th percentile UPL process predicts that a unit that meets a UPL-derived limit during each of several stack tests *will* nonetheless exceed that limit at some times over the year. If applied to a single test, where the average of three runs is only slightly below the limit, adopting a UPL approach to compliance would suggest that the source will be out of compliance almost half of the time. When EPA's "Rounding Policy" is taken into account, and the source "rounds down" a test result to demonstrate compliance, the UPL would assess that the unit will be in noncompliance more than half of the time. Even in instances where the average result is comfortably under the limit, the run-to-run variation may lead to a UPL prediction of substantial periods of noncompliance. To the extent that EPA is authorized to consider variability it should be limited to determining the variability in the mean of any set of tests of the top performing units, not the variability of an individual unit in the set. The difference between this concept and that employed by EPA are quite significant.

> "For similar reasons, prediction intervals to contain the mean of a future sample are also relatively insensitive to deviations from normality, unless either the given sample is very small or the deviation from normality is relatively pronounced. This limitation, however, includes the important case of a prediction interval for a future single observation (i.e., m=1). Thus, the prediction interval given here to contain a single future observation may be seriously misleading when sampling from a non-normal distribution."[16]

EPA cites several scholarly references as authority for its use of the UPL approach. However, these references generally include cautionary language about the use of statistical approaches. Most important among the several cautions regarding the use of these techniques is the assumption that the data from the population is distributed in a certain way (e.g. normally) where one does not have sufficient data to know what that distribution is. Other issues include examining whether the data are changing over time (and so may be expected to continue to change in the future) and whether the data include outliers that are not representative of the assumed distribution.

> "Tolerance intervals to contain a specified proportion of a distribution and prediction intervals to contain all of *m* future observations tend to involve inferences or predictions about the tails of a distribution. Deviations from normality are generally most pronounced in the distribution tails. For this reason, such intervals could be seriously misleading when the underlying distribution is not normal, especially for high confidence levels, and for tolerance levels when the proportion of the distribution to be contained within the interval is close to 1.0."[17]

EPA chooses 0.99, which is quite close to 1.0. Meeker, et al, includes several chapters on sample size requirements for statistical analyses, depending on the precision needed and the existing data and includes a number of case studies to illustrate the analysis needed to reach a given goal. These case studies call for 20 to several hundred measurements, depending on the UPL percentage desired and the underlying data. These analyses and case studies suggest that EPA would need far more than the 5 or 6 tests per unit that it uses to properly evaluate individual unit variability.[18] Here it should be noted that it is likely that the incinerators in this sector have been tested 20 or 30 times since 1990. EPA could use its authority to obtain these data but chooses to rely on a more limited set of data that for most pollutants only includes 5 or 6 tests per unit.

Several facts demonstrate that EPA's UPL approach is wrong:

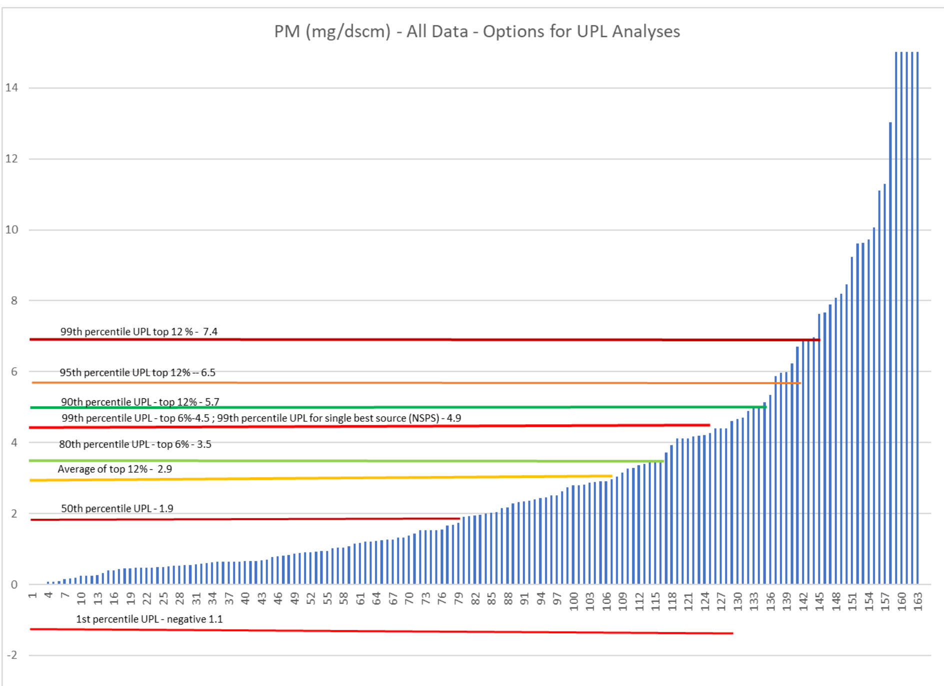
---

[16] Meeker, William Q, Hahn, Gerald J, Escobar, Luis A, *Statistical Intervals – A Guide for Practioners and Researchers Second Edition*, 2017, John Wiley and Sons, at page 59.
[17] *Id.*
[18] *Supra. at Chapters 9-12.*

1. The most obvious demonstration that the EPA's UPL approach is wrong is that it fails to implement the clear Congressional intent that the majority (94%[19]) of existing units in a sector reduce emissions to the level achieved by the average of the top 12 percent. In many instances, including the recent Integrated Iron and Steel proposal, all existing units met the UPL-generated floors for all pollutants.

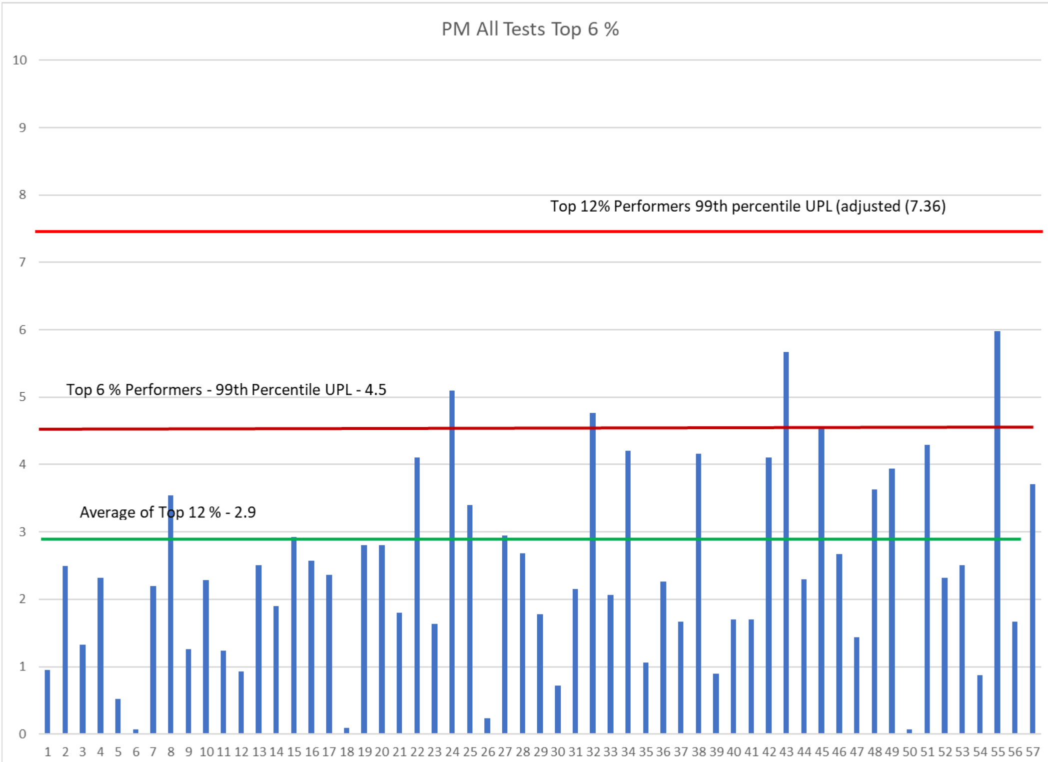
As the charts below show, the result of the UPL approach is that only the very "worst" units will have to reduce emissions from earlier levels. These charts also demonstrate that the UPL approach leads to different outcomes for different pollutants and that for most pollutants reducing the emission limit by selecting more protective UPL percentiles affects only a small subset of the overall fleet. These charts also show that reducing limits in the range under consideration would not impose catastrophic costs on the industry. The reader's "eye" can discern a reasonable cut point for each pollutant where a floor would not only impact relatively clean units.

<u>Figure 1. PM Emissions All Test Data and Options for Alternate Floor Levels</u>



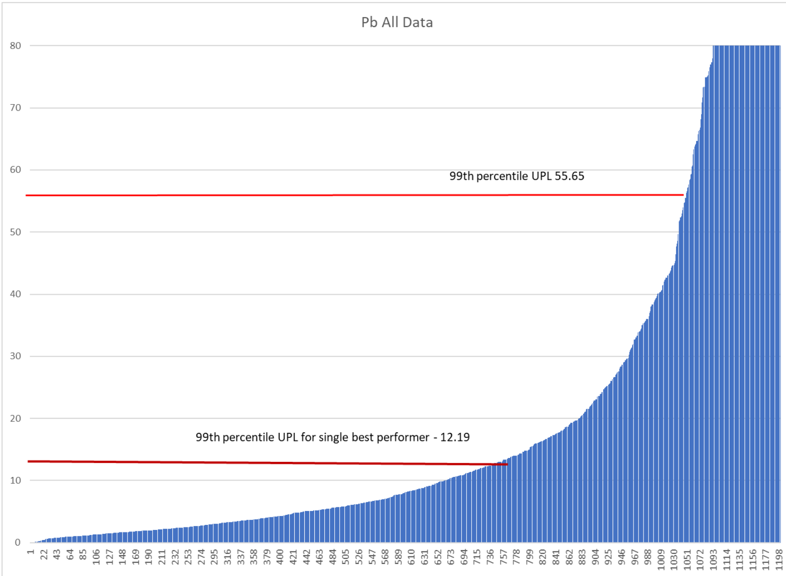
[19] Nominally, half of the top 12 percent would have to reduce emissions to meet the average level of the group. If one considers the earlier House and Senate bills, 85 to 90 percent of units would be expected to reduce emissions.

Figure 2. PM Emission Top 6 Percent and Options for Alternate Floor Levels

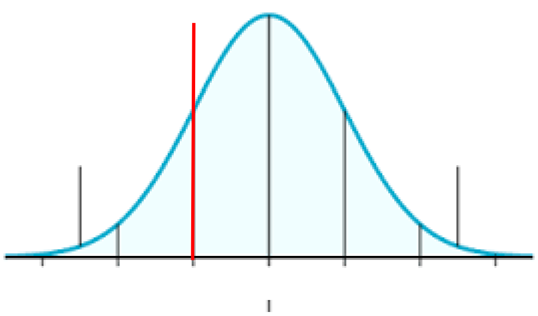


2. A second showing is found in those rulemakings where the UPL-generated floor for new sources (based on the best performing single unit) was less stringent than the UPL-generated floor for existing sources (based on the average of the top 12 percent). EPA has adopted "beyond the floor" limits where this has occurred, but, as shown below, it still accepts new source floors that are more lenient than the 99^{th} percentile UPL for the top 6 percent of sources that can be expected to comply with an existing source floor. See, Figure 1, above. See, also, Figure 3.

Figure 3. Pb Emissions - All Data and UPL



3. A third demonstration of the excessive variation determined by the 99th percentile predictive approach – as applied by EPA - is found if one examines the *lowest* predicted level generated by EPA's formula (the LPL). In many instances if one subtracts EPA's calculated variability from the mean of the group a negative value is predicted, even where the technique establishes that the distribution is "normal"; i.e. has a "bell-shaped" curve as below. The chart below includes a superimposed red "zero" line to illustrate this point. This occurs in several of EPA's calculations of the UPL, including the Pb calculation for new sources where the mean of the measurements is 3.50 and the calculated variance is 8.70. This results in a 99th percentile UPL of 12.2 and a LPL of -9.7. Similar results are found for the new source calculations for PM, HCL and Hg, among others.



We can all agree that zero is a lower bound on emissions and so, any prediction of future emission test results being negative, where the data are reported to be normal, must be wrong.[20] Just as there is a practical limit to the lowest emission rate, there are also real bounds on the expected maximum emissions from a properly operated pollution control device, and,

[20] Where the logarithm of the data are shown to be normal and the log-transformed data are used, the result will never be less than zero.

indeed, control device manufacturers guarantee contracted levels of performance. Operators will ordinarily allow these devices to have some degree of variation in input concentrations, flow and other parameters, but will ordinarily constrain that variation so as to meet applicable limits plus a compliance margin. These facts demonstrate that the inherent assumption behind use of the UPL statistical approach – that variation in emission performance during a test is random and unbounded – is incorrect.

4. EPA's UPL approach is further flawed because it employs a "pooled UPL" approach where, in addition to the test-to-test variability of the individual units it considers the difference in performance among the units in the top 12 percent to contribute to the overall variability of the pool. The adverse effect of this choice is somewhat offset by the fact that pooled data provide a larger value for "n" in the statistics. EPA likely does this to avoid the limited testing "new source" floor problem. However, in many cases there is a substantial difference in performance between the best units and those at the upper end of the top 12 percent. This can and does lead to overly large estimates of *variability.*

This creates the appearance of greater "variability" in emissions testing, when a large part of the difference is the underlying *difference in performance* between units with better or worse pollution controls rather than the *variability in the performance* of a given unit. To illustrate this issue, consider a data set of two idealized units, each of which is tested 10 times. Unit A has a nominal performance of 1. A random number generator is used to permit deviations from 0.9 to 1.1. Unit B's nominal performance is 10. To avoid "random" differences, Unit B's random numbers are scaled to reflect those of Unit A.

<u>Table 3. Evaluation of the Effect of "Pooling" Variation of Individual Units – Illustrative Data.</u>

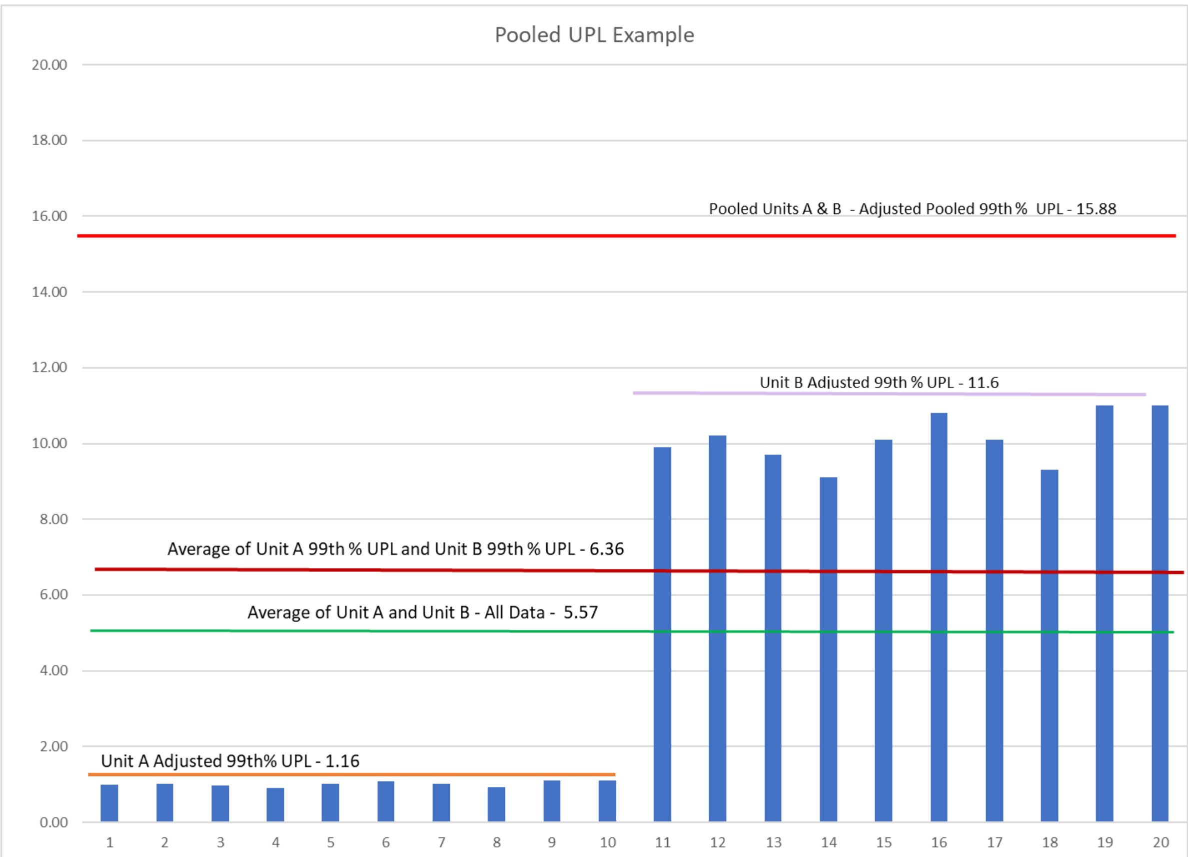
	Unit A	Unit B
Test one	0.99	9.90
Test two	1.02	10.20
Test three	0.97	9.70
Test four	0.91	9.10
Test five	1.01	10.10
Test six	1.08	10.80
Test seven	1.01	10.10
Test eight	0.93	9.30
Test nine	1.10	11.00
Test 10	1.10	11.00
Average (mean)	1.01	10.1
99^{th} percentile UPL	1.21	12.1
Adjusted 99^{th} % UPL	1.16	11.6

The *average* value of Unit A's performance is 1.01 and its "adjusted" 99^{th} percentile UPL is 1.16. Similarly, Unit B has a mean of 10.1 and an adjusted UPL value of 11.6. The *average* performance of this group is 5.57. In this example the average of the adjusted 99th percentile UPL of Unit A and the adjusted 99th percentile UPL of Unit B is (1.16 +11.6)/2, which calculates to 6.36. But if one "pools" all data for Units A and B, the 99th percentile UPL of the "pool" is 15.88.

Table 4. Evaluation of the Effect of "Pooling" Variation of Individual Units - Results

Average Performance Achieved by Units A & B	5.57
Average of 99th % of A & B (1.16 +11.6)/2	6.36
Pooled UPL of A & B	15.88

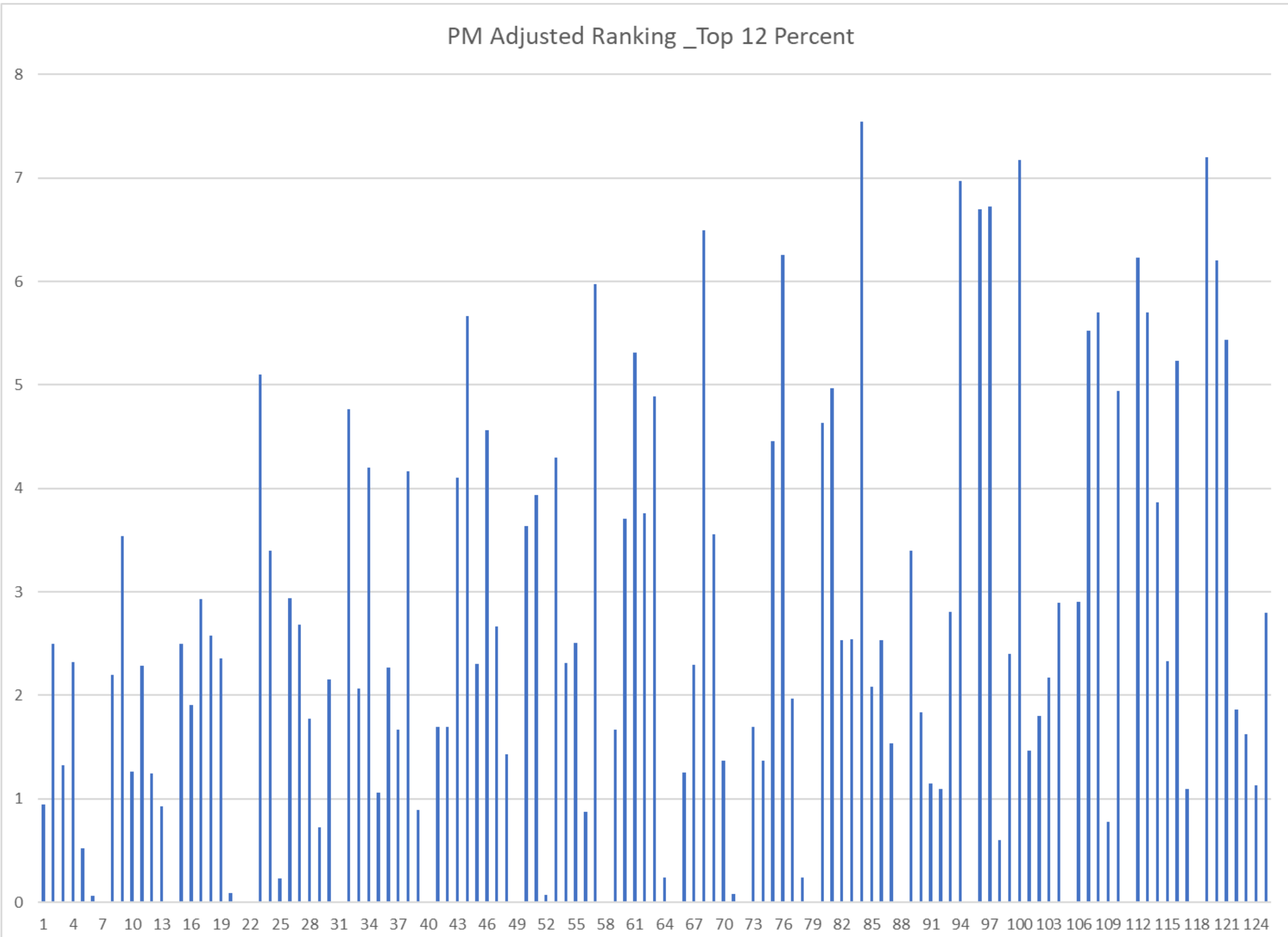
Figure 4. Evaluation of the Effect of "Pooling" Variation of Individual Units - Results



EPA asserts that applying the pooled UPL is needed to properly assure that all units in the top 12 percent will meet a limit. Aside from the fact that the statute explicitly calls for a determination of past performance, EPA fails to contend with the fact that in any normally distributed set of data from a group of sources nominally half of the data will be expected to exceed the *average* of the data. Thus, in our example above, Unit B would not be expected to meet either the average of all data or the average of the 99th percentile UPL of each unit.

5. A fifth significant issue is EPA's failure to follow peer-reviewed standard procedures for identifying and excluding, without further examination, values that are statistical outliers or that may violate existing permit limits. There are a number of test results in EPA's compendium that appear to be outliers.

Figure 5. Individual Test Results for EPA Best PM Performers[21]



These include at least the PM test results for Baltimore Unit 2, Hennepin Unit 2, and Stanislaus (Modesto) 2. EPA has not queried the original test reports or operators/regulators to determine whether those results are in any sense of the term "representative" of the operation of the unit, whether they include "non-standard" operating conditions, operations in violation of existing limits or, indeed, whether the underlying measurements and calculations were properly performed and reported. The PM test data for these three units illustrate the issue.

[21] This figure sets out the test results for each of the 17 Top Performers. A blank value has been inserted along the horizontal axis at 7, 21,40 etc. to distinguish between test results for different units.

Figure 6. Baltimore Unit 2 PM Test Results

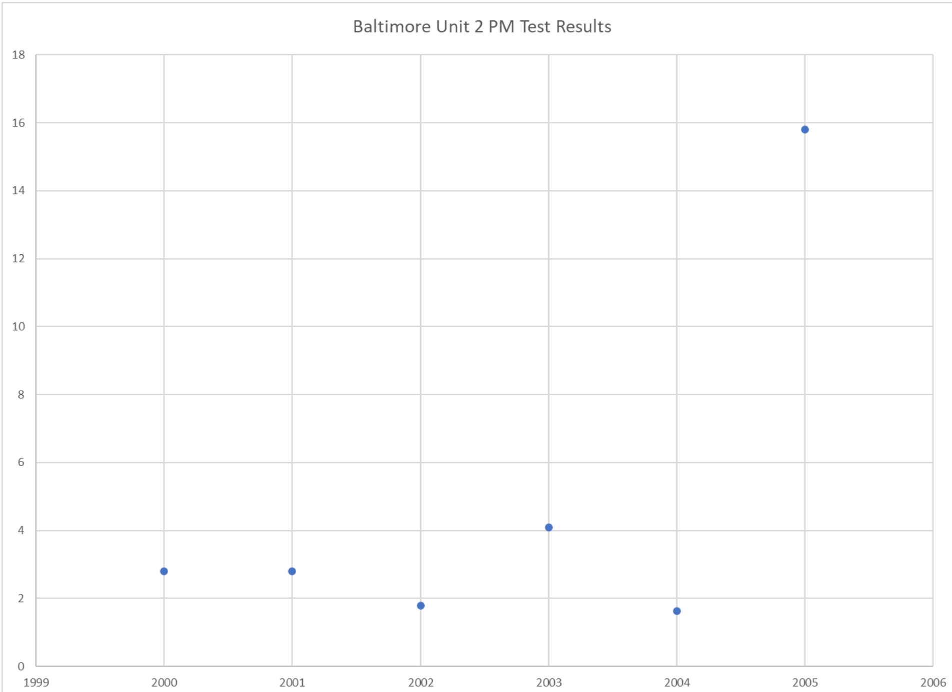


Figure 7. Hennepin Unit 2 PM Test Results

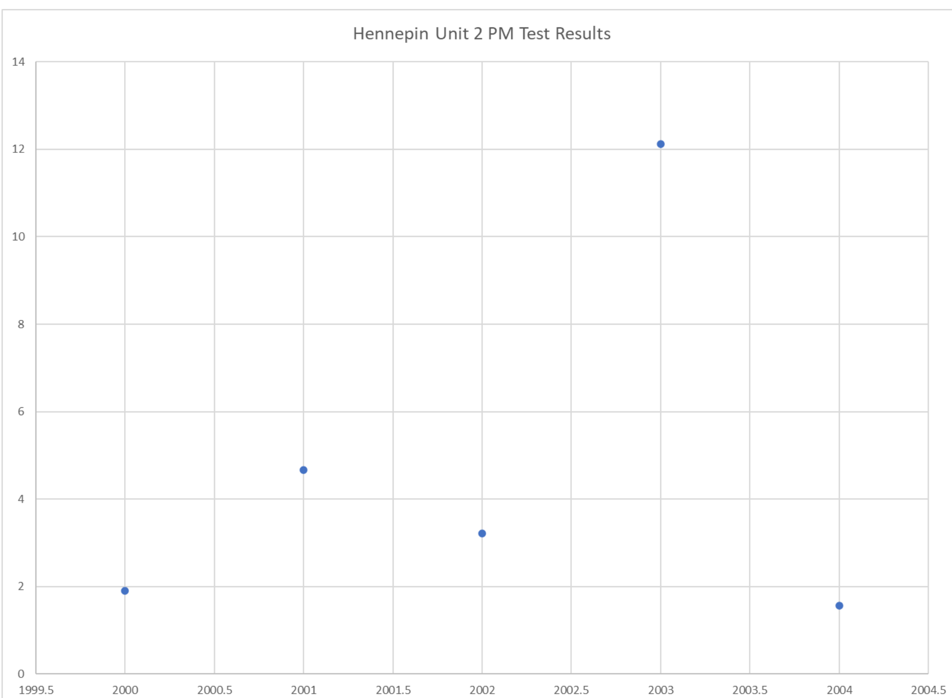
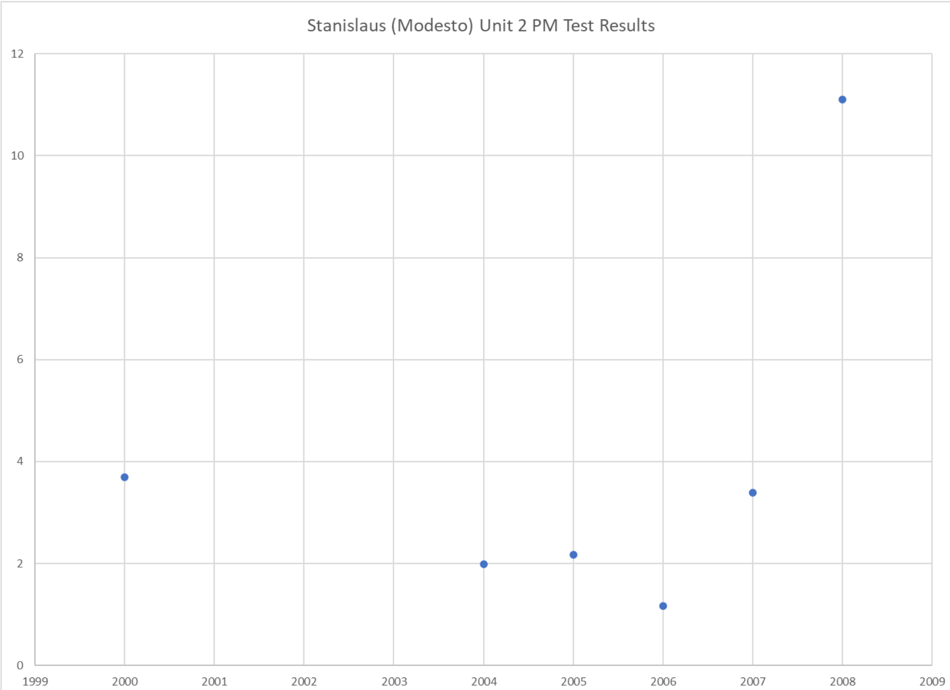


Figure 8. Stanislaus (Modesto 2) PM Test Results



The 2005 test report for Baltimore Unit 2 is approximately 8 times higher than the average of all tests and 4 times higher than the next highest test report. Noting this, EPA should have applied at least one of the standard statistical tests for outliers and either eliminated those results - or at least investigated whether the reported results are correct and reasonably represent emission levels from lawful operation of the unit. The National Institute for Science and Technology among others recommend several possible tests as useful statistical tests for outliers. One commonly recommended test is the Grubbs Test.[22] This test confirms that each of the suspect test results for the three units is a statistical outlier at the one percent significance level.[23]

> "It should be pointed out that almost all criteria for outliers are based on an assumed underlying normal (Gaussian) population or distribution. When the data are not normally or approximately normally distributed, the probabilities associated with these tests will be different. Until such time as criteria not sensitive to the normality assumption are

---

[22] The National Institute of Standards and Technology (NIST) was founded in 1901 and is now part of the U.S. Department of Commerce. NIST is one of the nation's oldest physical science laboratories. Congress established the agency to remove a major challenge to U.S. industrial competitiveness at the time — a second-rate measurement infrastructure that lagged behind the capabilities of the United Kingdom, Germany and other economic rivals. About NIST | NIST; 1.3.5.17.3. Generalized Extreme Studentized Deviate Test for Outliers (nist.gov); Grubbs 1969 - Detecting outlying observations in samples.pdf (ship.edu)

[23] Grubbs 1969 - Detecting outlying observations in samples.pdf (ship.edu) at page 3.

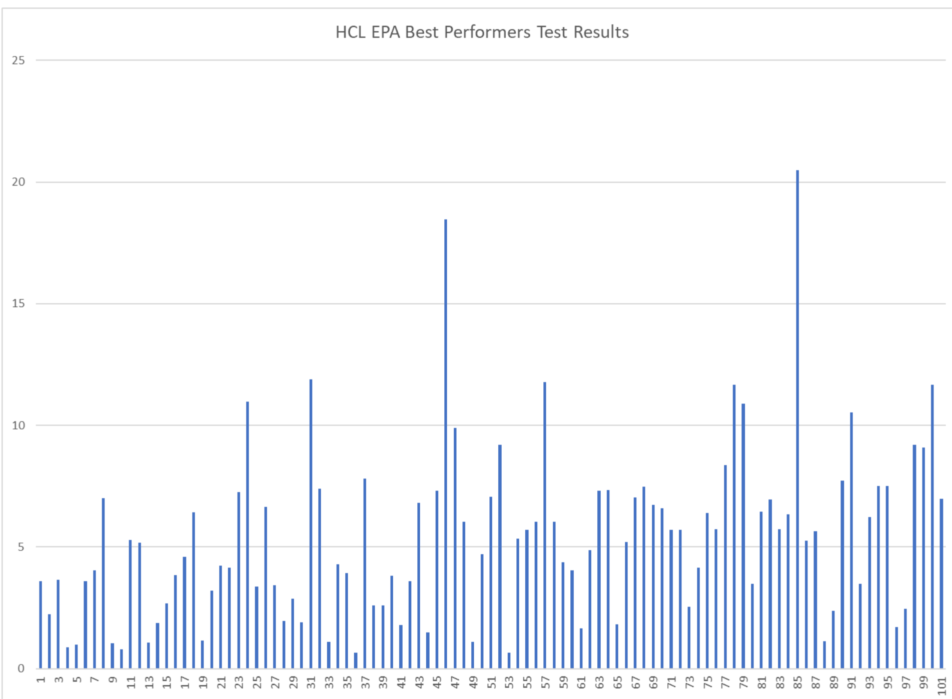
developed, the experimenter is cautioned against interpreting the probabilities too literally when normality of the data is not assured."[24]

In addition to ascertaining whether a particular test result is questionable, such tests also provide information about whether the underlying assumption about the distribution of the overall population is correct.

"Although our primary interest here is that of detecting outlying observations, we remark that the statistical criteria used also test the hypothesis that the random sample taken did indeed come from a normal or Gaussian population. The end result is for all practical purposes the same, i.e., we really want to know once and for all whether we have in hand a sample of homogeneous observations."[25]

EPA's compilation of HCl test results for its best performers reveals that at least two (data points 46 and 85 below), and perhaps five of the data points in the set, are likely to be outliers.

Figure 9. HCL Best Performers, All Tests



If one rejects the two data points at 46 and 85 as "outliers", the UPL-based floors are reduced as set out in Table 5 below.

[24] *Id.* at 3.2.
[25] *Id.* at 3.3.

Table 5. Effect of Including Outlier Data

UPL Percentile	Including outliers	Excluding outliers
99	13.7 (12.52 corrected)	12 (11.2 corrected)
95	11.2	9.91
90	9.92	8.84
80	8.35	7.54

Figure 10. EPA Data Used to Determine the Proposed Pb Emission Guideline

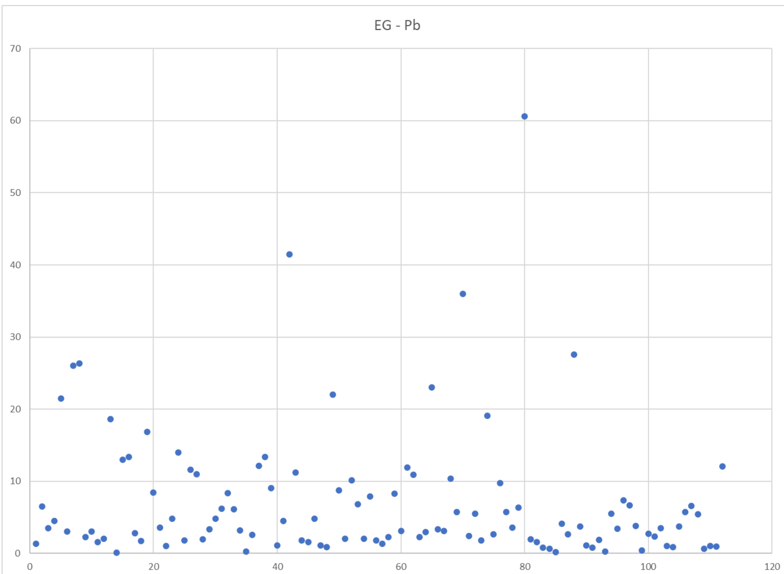


Figure 11. EPA Data Used to Determine the Proposed PM Emission Guideline

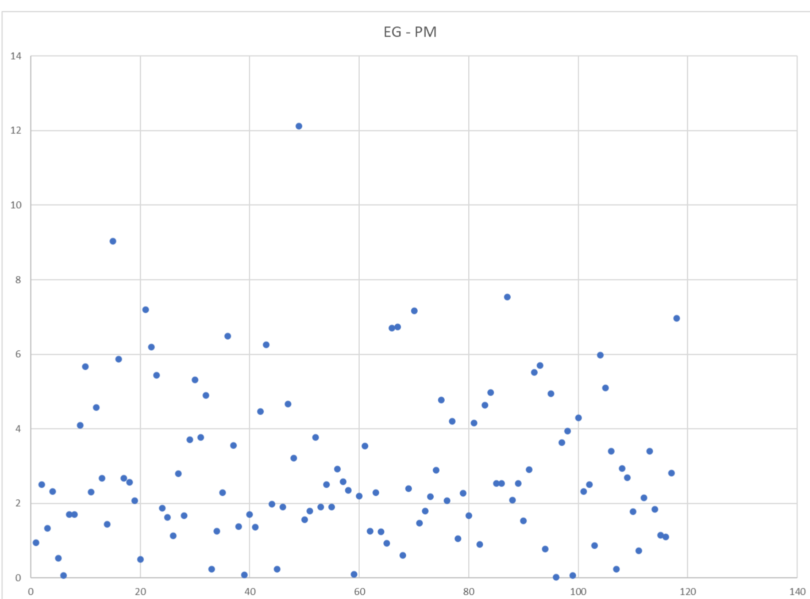


Figure 12. EPA Data Used to Determine the Proposed Hg Emission Guideline

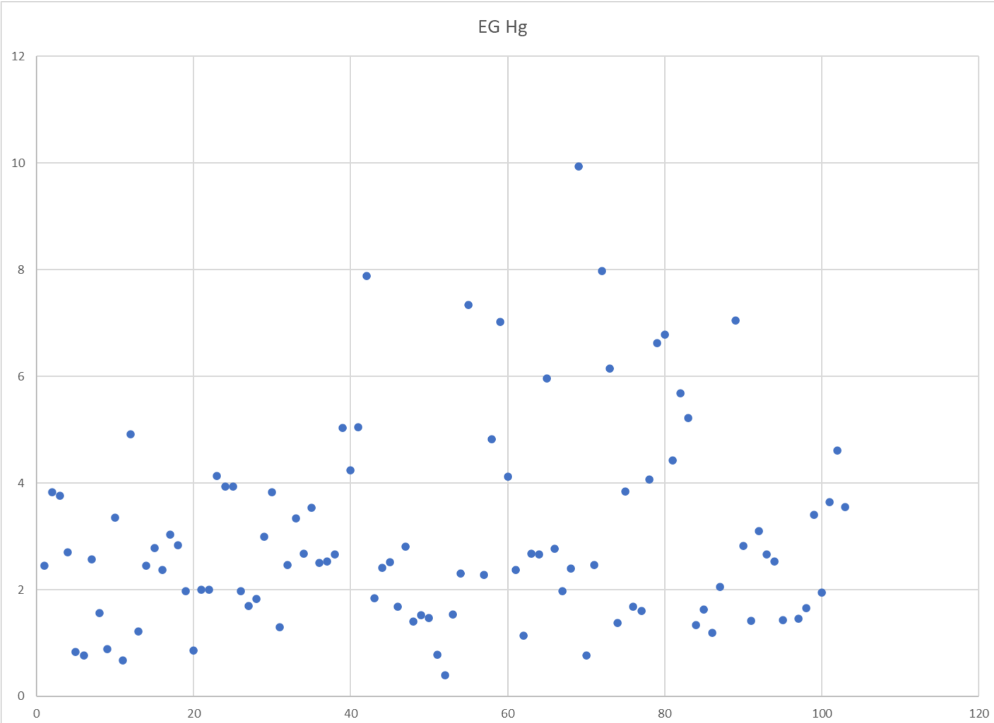


Figure 13. EPA Data Used to Determine the Proposed Dioxin/Furan Emission Guideline

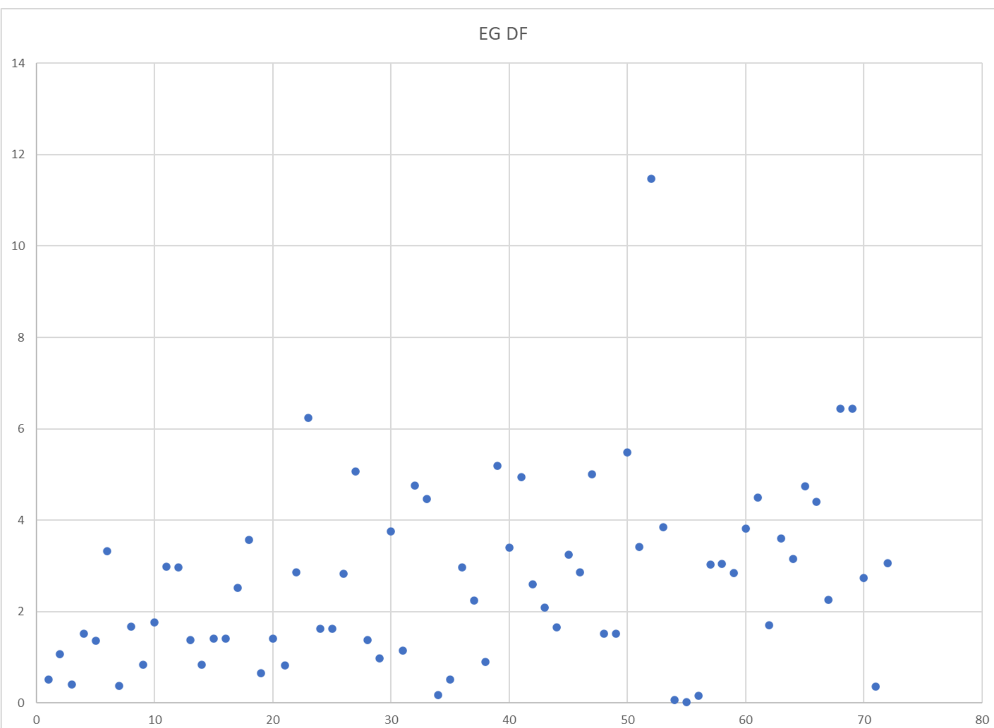
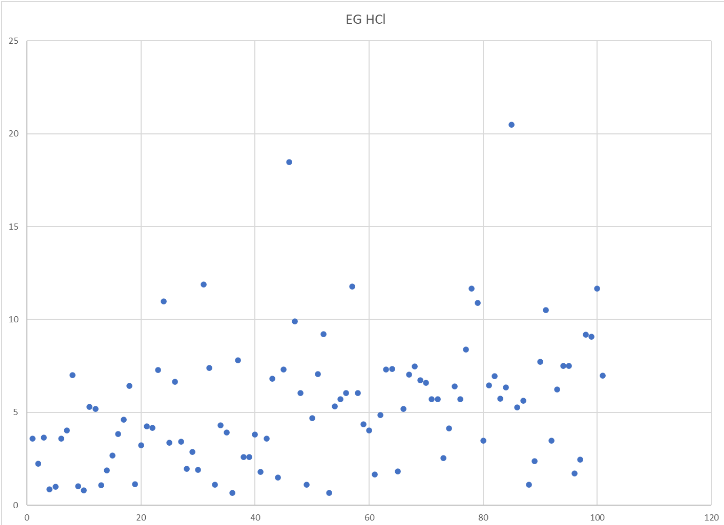


Figure 14. EPA Data Employed to Determine HCl Emission Guideline



7. EPA applies its UPL procedure to the entire data set of the top 12 percent (here 17 units) and seeks to set a floor that every unit in the entire top 12 percent would be projected to meet 99 percent of the time (when tested once every year). However, in any data set nominally half of the values in any set can be expected to be greater than the average. Accordingly, units in the 88^{th}-94^{th} percentile are not expected to be better (lower emitting) than the average of the best performing top 12 percent. Only units that are at or better (lower emitting) than the average should be expected to meet a floor set at the average performance of the group of top 12 percent. Thus, to the extent that EPA seeks assurance that units upon which the floor is based will meet floor limits in the future the agency should limit its variability calculations to the top half of the top 12 percent, i.e., the top 6 percent – top 9 units. As one might expect, this correction reduces floor levels significantly.

Table 7. PM - Options for Alternate Floor Levels

UPL Percentile percent	UPL Top 12 percent (uncorrected)	UPL Top 6 percent (uncorrected)
99	7.95/7.36 corrected	6.31/5.82 corrected
98	7.36	5.85
95	6.47	5.16
90	5.68	4.56
80	4.74	3.84
70	4.06	2.48
50	2.94	1.12
1	-2.07	-1.35

Table 8. HCl – Options for Alternate Floor Levels

UPL Percentile percent	UPL Top 12 percent (uncorrected/corrected)	UPL Top 6 percent (uncorrected/corrected)
99	13.7/12.9	11.1/10.4
98	12.7	10.3
95	11.2	9.08
90	9.91	8.02
80	8.34	6.76
70	5.35	4.37
50	2.37	1.98
1	-2.98	-2.38

8. Finally, we come to the arbitrary choice of the 99th percentile predictive level. EPA never provides an objective rationale for its judgment that the 99th percentile properly predicts the inherent variability in future test results (or why, as above, this is a relevant consideration). The argument that the CAA must be complied with all of the time is misleading and fundamentally incorrect. While sources may monitor some operating parameters continuously, reference method compliance tests are conducted only rarely. In practice this means that units must meet limits based on reference method testing only 0.0342%[26] of the time (three, one-hour test runs every year). Over the next 20 years (which exceeds the remaining useful life of nearly all existing LMWCs) units will be tested only 8 times. It is unlikely that if an individual unit is only tested 8 more times that the result of those tests will exceed a level set at a one in one hundred prediction level.

Current practice allows sources to retest if an exceedance occurs. If formalized by EPA, the 99th percentile UPL the represents a one-in-ten-thousand likelihood of failure.[27] Under either scenario sources may pre-test and ensure optimal performance prior to conducting the testing required by the rule. In my experience enforcement only occurs after sustained and repeated testing demonstrates that the unit cannot routinely comply with applicable limits.

Further, even the 99th percentile does not provide for compliance *all* of the time. Note that there is no particular reason to use the 99.74th percentile (as industry urged), representing less than one day of violation per year, since sources are not allowed to violate the CAA once every year or once every 366 days or once every five years. EPA has rejected industry request for 99.72th percentile (based on one day of violation per year) and has on occasion adopted the 90th percentile. EPA should evaluate the impact of using other metrics – and perhaps different metrics for different pollutants – in consideration of the clear Congressional intent that the floor requires reduced emissions from a larger segment of the existing fleet than under current procedures. A clear candidate for a different metric is PM which, if properly maintained and operated, is not as likely to be variable as e.g. Cd, which depends on what is in the waste.

[26] Not three percent, but three one-hundredth of one percent.
[27] The probability of a 1 percent event occurring twice in a row is 0.01 x 0.01 or 0.0001.

EPA never explains why the 99th percentile would properly predict the inherent variability in future test results for a given unit *after it has become subject to a limitation* or why this would be a relevant consideration for units in the "bottom half" of the top 12 percent.[28]

EPA did not review more recent test data or require that such test reports be submitted to its CEDRI/WebFIRE database. EPA's ECHO database contains information about such tests, including one facility that was tested nine times from 2019 to 2023. It is likely that most LWMCs are also tested regularly. Had EPA collected all of the test results available to it, and calculated the average of some percentile UPL, the floors would have been lower.

<u>NSPS for Stack Test-monitored Pollutants</u>

For NSPS floors EPA applies its UPL approach to the test results in its database for the unit that has the lowest average test result. This limits the data set to only a handful of tests and leads to an extremely high calculation of variability under the extreme (99th percentile) UPL employed. In the past EPA has ignored this result and set "beyond the floor" limits for units where the 99th percentile UPL for the single "best" unit was less stringent than that for the average of the top 12 percent. Here, the NSPS floors are slightly more stringent than the EG floors and so EPA accepts the NSPS floors. As Table 9 below shows the NSPS floor for PM is less stringent than the average of the top 12 percent, less stringent than the average of the top 6 percent and less stringent than the 99th percentile UPL of the top 6 percent that are expected to meet the EG floor.

<u>Table 9. New Source PM</u>

NSPS New Source PM	Emission Rate (ug/dscm)
UPL %	
99	4.81
98	4.17
95	3.39
90	2.83
80	2.25
70	1.87
50	1.28
20	0.32
1	-3.365[29]
Average of Top 17 Units	2.94
Average of Top 9 Units	2.48
99th % UPL Top 17 Units	7.95, corrected to 7.46
99th% UPL Top 9 Units	5.82

[28] That is, for units in the 88th - 94th percentile.
[29] EPA reports that this data set is normally distributed.

The result for HCl is not as extreme, but as Table 10 below demonstrates the new source floor for HCl is less stringent than the average of the top 12 percent and less stringent than the average of the top 6 percent.

Table 10. New Source HCl

NSPS New Sources HCL	Emission Rate (ug/dscm)
UPL %	
99	7.8
98	6.64
95	5.42
90	4.53
80	3.66
70	3.11
50	2.27
20	0.88
1	-3.26[30]
Average of Top 17 Units	5.73
Average of Top 9 Units	4.37 (as re-ranked)
99th percentile UPL Top 17 Units	13.7, corrected to 12.92
99th percentile UPL Top 9 Units	10.4

Floors and limits for CEM-monitored Pollutants

EPA's approach to setting floors for CEM-monitored pollutants is clearly inadequate. The agency proposes to set floors based on the single highest value in a year. EPA specifies either a 4-hour or 24-hour average for CEMS. A unit operating 6000 hours per year will record nominally 1500 four-hour averages or 300 daily averages and so the highest single value in a year corresponds to the 99.7-99.9th percentile of recorded values.

> "Unlike stack test pollutants, there are no individual run data for CEMS pollutants. Instead, data for CO, NO_x and SO_2 are collected continuously, and available data comprise only **peak** annual values which the current rule requires reporting. The annual peak values reported for each unit that was operating in 1990 were adjusted to reflect 1990 control levels and then used to identify the best performing units."

> "The limits were reevaluated simply by averaging annual peak CEMS data corresponding to the top performers for each pollutant and applicable subcategory."

[30] EPA reports that this data set is normally distributed and so this value is just as likely as the 99th percentile UPL.

However, as EPA explicitly acknowledges, it did not exclude from its analysis data during periods that are exempt from its proposed emission limits.

> "CO data are annual highest CEMS reading averages and suspect most of these are reported during operational transition periods so *do not reflect compliance conditions*."[31]

Notwithstanding the fact that EPA quite reasonably suspects that the peak CEMS data *do not reflect compliance conditions* EPA proposes to base the limit for *normal* operation on the highest levels recorded during what are likely startup, shutdown and malfunction (SSM) periods. EPA proposes a separate (higher) emission limit (in the form of an O_2 level correction that effectively triples the applicable limit) for startup and shutdown periods.[32] This approach leads to floor calculations that for NOx and CO in several subcategories are less stringent than the 2005 limits that all sources presumably exceed.[33] This demonstrates that EPA's approach is wrong, but rather than seeking an approach that more meaningfully implements the obligation to set a floor, EPA proposes to retain the existing limit.[34]

EPA's excuse for such a flawed approach is that "available data comprise only peak annual values that the current rule requires reporting." Here the focus is on the term *"available."* Section 114 of the Act specifically authorizes EPA to obtain any information "relating to" or "assisting in the development of any regulation" under section 112 or "any regulation under section 7429 of this title (relating to solid waste combustion)".

The highest single reading of a continuous monitor provides little useful information as to which units are the best performers, much less a basis to set a floor as required by the Act. EPA's attempt to do so is clearly arbitrary, capricious and not in accordance with the Act. There are several options for setting CEM-based standards, but EPA can't rationally adopt any reasonable approach to set the statutorily mandated floors without additional data.

A limit on the highest single 4/24 hr reading in a year – that may contain periods of where the source was operating in a high emitting mode not covered by the limit - is a nearly useless constraint on either short-or long-term operation of the facility and allow much greater emissions than a limit based on the best CEMs-based performance. Moreover, a single limit does not effectively characterize the "best" performer. Consider two units. Unit One emits 100 tpy and has a maximum 4-hour emission rate of 100 lb. Unit Two emits 50 tpy and a maximum 4-hour

[31] EPA-HQ-OAR-2017-0183-0014, Attachment 1 at Cell CP1

[32] "Secondly, CEMS data collected while the large MWC unit is warming up (no waste is introduced to the grate), starting up (warmup period is over and waste is first fed to the grate but not at steady state operation) and shutting down (waste is no longer being fed but is burning down on grate) will be flagged as warmup, startup, or shutdown period data. CEMS data collected during warmup, startup, or shutdown periods will be averaged at stack oxygen content and not corrected to 7 percent oxygen, as are data during normal operations." 89 Fed Reg 4243,4256

[33] See Proposal at Table 2. 89 Fed Reg 4243,4351

[34] Id. at Cell CP7

emission rate of 125 lb. Which is the better performer? The answer depends on the relative risk associated with acute exposures and long-term chronic exposures. EPA does not have the option to address each of these concerns with reference method testing but can do so with CEM-monitored pollutants. EPA can and should set out a limit based on long term average CEMS data plus a short-term limit. This is common in permitting where agencies are attempting to address both short and long-term NAAQS.

Rather than a CEM limit that must be met at all times, except for an unspecified number of times when the alternate startup/ shutdown limit is to be met or a long-term average that is insufficiently protective against short-term spikes, Australia and others require that a given percentage of the data (e.g. 99%) must be less than the relevant limit – with no excuses for SSM.

If EPA develops a CEMS limit for a pollutant for which EPA only has stack test/reference method-based data –the 99th percentile will not provide a sensible outcome. Here it would be most useful for EPA to obtain the full suite of stack tests over the years for the best performing units to better characterize the distribution of the data. In the absence of other data one supposes that EPA could use a modest UPL of top 6 percent performing units. The average of approximately 110 tests over a period of years for a group of performers should adequately address variation in the performance of the group. EPA might also choose to set to a 30-day limits based on the 91.77th percentile or a daily limit based on the 99.77th percentile *where there are a substantial number (i.e. 30) of test results for each of the top unit(s) and one averages the UPL of the individual units (i.e., not pooled).* Using a pooled UPL to set a CEM-based standard has the same issues and deficiencies as using those data to set a reference method-based standard and would likewise lead to numerically lax standards. As suggested above, one could lessen the impact of the limited number of test results by choosing a UPL% less than those stated above (70th?, 80th?) for 30 days.

When calculating a CEMS floor using stack test data the most likely options involve large numbers of test results where the sensible approach is to choose whatever is determined over the appropriate time frame. An annual CEM limit is an *average* of 35,000 15-minute *averages*; a 10-day limit is the *average* of 2880 15-minute *averages*. EPA breaks these down and requires reporting of 4-hour and 24-hour averages. For this reason, there is no need to look for or consider variation. EPA has accepted this in setting CEM floors in other sectors. EPA could reasonably base either an annual or a 30-day CEM-based emission limit on the average of the top performer(s). There is not a lot of value in attempting a 24-hour or shorter standard without actual CEM data, since the UPL would lead to absurdly high limits. One might consider the Australian approach of requiring X number of values to be below Y, but again this approach requires data.

Further complicating this issue is what I call "The Beauty Contest". Source operators know when a reference method test will be conducted and, if they are at risk of exceeding a standard because they have allowed their pollution control device to deteriorate (or otherwise), they can and have taken special measures (including the use of "compliance coal", process unit variations and other clever measures) to attempt to pass the test. Testing at one time was supposed to be conducted under "worst case", then "representative worst cast" and now "representative" conditions. Where a source operates well below a limit, the reference test might provide "representative" performance, but, if a source operates near an applicable limit current reference method testing may only tell you what a source can do “on its best day”, or perhaps “on average”, not what a source will do on a day-to-day basis. EPA’s UPL calculations are a fairly strong indictment of reference method testing. All of this suggests that a CEM-based limit set at a higher number will be more protective than a reference method test-based limit and should be encouraged. I note also that CEMS-based limits incorporate RATA testing that includes reference method testing to confirm the accuracy of the CEM.

<u>Other Issues</u>

EPA assumes general control efficiencies of 99/95% for ESP/FF and uses this assumption to calculate the uncontrolled emission rate of units that added PM controls after 1990. The effect of EPA’s use of 1990 uncontrolled emission rates is to eliminate those units from the top 12 percent pool and replace them with higher emitting units.

An additional issue is the mismatch between EPA’s approach to identifying the top performers and its method for assessing the performance of those units. EPA selects the top performers on the basis of the mean (i.e. average) of the test results for those units. But then EPA applies its UPL test to calculate the performance of those units. This can, and occasionally does, lead to situations where the agency excludes a unit with lower overall variability and includes a unit with higher variability (but a lower average), leading to a higher floor calculation and less stringent limit. In the example below, both units fell within the top 12 percent.

<u>Table 11. Determining “Best Performers”</u>

Unit	Average	Range	Standard Deviation	Mean + 3 x std deviation
Bridgeport 1	3.78	0.5 – 9.03	3.11	13.1
Bridgeport 2	4.67	1.1 – 6.23	2.02	10.75

EPA’s “practice” of “rounding up to 2 significant figures” is contrary to standard rounding conventions and unlawful given the requirement that the floor *be not less stringent* than the average. Here the NOx NSPS MACT Floor calculation was 130.5 ppmv; EPA rounds to 140 ppmv. Note that the stack test data are reported to four or five (or more) significant digits. And then, on the compliance side, EPA allows rounding “down”, so 144.9 complies. EPA policy says not less than two, nor more than 3 significant digits. EPA offers no rationale for the use of 2 rather than 3 significant digits. The use of 3 significant digits would reduce the impact of EPA’s rounding policies, since 130.5 would then round up to 131, not 140.

There appears to be some inconsistency between the “unscreened data” presented in EPA’s Appendix A and the inputs to the UPL calculations. For example, EPA’s “unscreened data” tab contains 107 Cd test results for the top, with an average reading of 0.636; Tab A-6, used to calculate the UPL for Cd lists 115 test results, with an average reading of 0.59.

Recommendations with respect to future standard setting

It is theoretically possible that all units in a sector apply state of the art controls and so all could be considered "above average". EPA should distinguish between those sectors and pollutants where controls are state of the art and universally applied (if any) and other sectors and pollutants. Otherwise, EPA should be guided in structuring 112/129 limits by the clear Congressional intent that many units within a sector should have to reduce emissions of hazardous air pollutants.

EPA's approach to existing source MACT standards may be influenced by a fear of excessive cost, leading to a blowback that might cause repeal of the program. With nominally straightforward implementation one might think the program would impose compliance costs on 94 percent (or 88 percent if one accepts OGC's interpretation) of the facilities in the 174 covered Section 112 source categories. But, as EPA has noted, no one source is in the top 12 percent for all covered pollutants, and so, the program could be anticipated to impose substantial compliance costs, all within a 3-year compliance period for each covered sector. EPA's response to this concern, thus far, has been to use its 99^{th} percentile UPL approach to require emission reductions only from the very worst sources within a category. But, as the charts above show, this approach leads to different outcomes for different pollutants. For some pollutants lowering the standard will only affect a few units, for others the impact might be more substantial.

Perhaps more concerning than EPA's initial round of standards has been its general approach to the 5 or 7-year reviews that are required under the CAAA. EPA generally has attempted to declare "Mission Accomplished" with respect to these standards and only use the mandated reviews to offer small changes or fixes as needed, and states that it has no obligation to revisit MACT floors, even as it concedes that substantial additional performance data are now available and that the stringency of standards developed using the UPL process are highly dependent on the amount of data that is available.

1. EPA should set floors on the arithmetic average of the test results of the top performing unit(s).
2. If it seeks to make an adjustment for "unit variability" EPA should acquire emission testing results sufficient to establish a robust data set for all covered subcategories. Where those data are not available, EPA should require additional testing and reporting. These data should be available to the public in a readily retrievable format.
3. UPL and some arithmetic averages can be significantly affected by single outlier tests (both high and low results). EPA should screen its data to detect and exclude outliers and data for periods when compliance with numerical limits is not required. This effort should be focused on those individual test results that significantly affect the ultimate calculation of a floor.[35]

[35] The data for Pb, in particular, contain several extremely high values that should be examined.

4. If the agency continues to consider statistical approaches to evaluating floors, it should retain the services of an independent expert statistician to conduct a public-facing review of potential options for statistical analysis of the current and newly acquired data.
5. EPA should reject any upper predictive level percentiles where the LPL is less than or near zero.
6. EPA should set floors that distinguish the best performers from the rest of the fleet.
7. If it attempts to set a UPL percentage based on the argument that the UPL reflects actual past or likely future performance of a unit, EPA should set compliance obligations that reflects that argument.
8. Where the result of a floor analysis would lead to an outcome where few, if any, sources must reduce emissions, EPA should conduct a technology review to determine whether that outcome is a consequence of its statistical approach or is the result of homogeneity within the industry in terms of controls, raw materials and other factors that influence emissions.
9. EPA should not set CEM-based limits based on the worst single CEM reading in a year for the unit with the lowest worst reading – especially where it is likely that such reading occurred under conditions that are subject to an alternate limit (i.e., startup, shutdown). EPA is authorized, and here required, to obtain sufficient data to set the Congressionally mandated floors and limits. EPA could and should obtain more comprehensive CEM data from the better performing units. Nominally, EPA should request the CEM data from each source's best quarter for each of the relevant pollutants in the past 5 years.
10. EPA must incorporate at least 3 significant digits in any limit and "round down" so that the emission limit is no less stringent than the calculated floor.
11. EPA could address many of the objections to its floor analysis by a more pro-active approach to setting limits that reflect the *maximum degree of emission reduction that is achievable.*