VIII. EPA FAILED TO PROPERLY APPLY THE STATUTORY STANDARD IN SETTING ITS LIMITS

A. The Agency Must Use a Consistent Measure of "Actual Emissions" to Set the Floor

The statute requires EPA to establish new source standards that are no less stringent than "the emission control that is achieved in practice by the best controlled similar source," and existing source standards that are no less stringent than "the average emission limitation achieved by the best performing 12 percent of the existing sources." 42 U.S.C. § 7412(d)(3). The floor for both standards must reflect, in other words, "what the best performing sources actually achieve." *Cement Kiln Recycling Coalition v. Environmental Prot. Agency,* 255 F.3d 855, 862 (D.C. Cir. 2001). In setting its utility standards for mercury and HCl (but not PM), EPA has failed to consistently follow that command: it has used one measure of sources' 'actual performance' to select its best performing plants, and an entirely different measure of actual performance to set standards that reflect those plants' emissions.

When *selecting* its "best performing" units, EPA claims that plants' single lowest stack test properly represents those plants' actual performance. Memorandum from Jeffrey Cole to Bill Maxwell dated May 18, 2011 at 8-9 ("Floor Memo") (Docket no. EPA-HQ-OAR-2009-0234-9858). When *calculating floors* based upon those plants' test data, however, the Agency claims that the plants' 'actual emissions' are best represented by the average of all tests, further manipulated for variability. *Id.*

As a result, a handful of outlying test results substantially bias EPA's ultimate standards. For example, there were two mercury tests for the Indiantown Cogeneration plant with averages of 8.5384 E-09 lb/MMBtu and 5.49E-07 lb/MMBtu – the latter 64 times the former. Similarly, there were two tests for Spruance Genco, GEN4 with averages of 1.177E-08 lb/MMBtu and 1.6075E-06 lb/MMBtu (the latter almost 137 times the former). There

were 8 tests for BL England, the average of the lowest being 2.9649E-08 lb/MMBtu and the average of the highest being 2.434E-06 lb/MMBtu (the latter over 82 times the former). There were three tests for Scrubgrass Gen 1, the average of the lowest being 3.53E-08 lb/MMBtu and the average of the highest being 3.0E-06 lb/MMBtu (the latter almost 85 times the former). Lastly, there were also three tests for Cherokee Unit 4, the average of the lowest being 3.1654E-08 lb and the average of the highest being 1.12E-06 lb/MMBtu (the latter over 35 times the former).[1]

To test the sensitivity of this assumption, we set aside the highest average tests of the five units noted above and conducted the same variability adjustment with the remaining (75 rather than 80) tests for the top 40 sources. The resulting value (applying a 99% upper prediction limit) is 0.43 lb/TBtu instead of EPA's 1.18 lb/TBtu. The standard is thereby reduced by two thirds just by setting aside the highest outlying tests for 5 units (each of which was one or more orders of magnitude greater than the lowest tests at the same units). As noted above these outlying (and other older tests) were generally not conducted using Method 30B; the results likely reflect test variability rather than actual source emissions variability, and should be excluded on that basis.

EPA's chosen approach is, at any rate, beyond the agency's authority. The statute provides EPA some discretion in defining the measure of plants "actual performance"; it does not, however, allow the Agency to use two inconsistent definitions of that term. If a plant's actual emissions are best approximated by averaging all of that plants' test data, and then statistically adjusting that data to address possible variability – then the "best performing plants" are those with the lowest emissions according to that

[1] It appears that while the best tests were obtained using EPA Method 30B, others may have been obtained using the older EPA Method 29. Tests obtained using different methods should not be lumped together and assumed to reflect unit emissions variability.

measurement. If, on the other hand, a plant's actual emissions are best approximated by its single lowest test result, then the MACT floor must be defined using the single lowest test results of the best performing plant (or plants).

Any consistent approach would yield lower floors than those chosen by the Agency. To generate its primary mercury standard for coal fired units (burning fuel with a caloric content above 8300 Btu/lb), for example, if EPA used the single best test for each of the top 40 plants to define a plant's actual emissions both to select its best performing sources, and to measure those sources' emissions, the floor would be 0.05 lb/TBtu instead of the proposed 1.2 lb/TBtu – even EPA's variability adjustments and other manipulations.

Alternatively, if EPA used variability-adjusted average emissions to both select the best performing sources, and to measure their emissions, it would similarly generate lower standards. Units EPA identified as "best performing" sources would be replaced by others with lower and less variable historical data.[2] For example, EPA's mercury standard for existing units burning coal of less than 8300 BTU/lb is based upon 40 "best performing" units, selected on the basis of those units' lowest 3-run average mercury data (in lb/MMBtu, typically obtained via Method 30B testing). If EPA had begun by considering *all* of the available test data for the various units (say, by taking the average of all of the available data), determined the "best performing" units on that basis, many of EPA's claimed "best performers" would not qualify as such, including: Scrubgrass Generating Company (Gen 1); Spruance Genco (Gen 4); B L England (Unit 2); Cherokee (Unit 4); Martin Drake (Unit 7); and Indiantown Cogen (Unit 001). Other units, currently not in the best performing list, would become best performers, including:

---

[2] In determining the variability adjusted average emissions, only emissions conducted pursuant to the same test method should be used, so that the variability reflects variability in emissions rather than test methods. The latter is irrelevant for standards, such as these, for which EPA has specified a single test method.

Hopewell (Units 1 and 2); Clover (Unit 2); PSEG Mercer (Unit 1); and possibly Piney Creek Project (Unit BRBR1); and San Juan (Unit 2). These units have far less variance (and thus lower "actual emissions," if actual emissions are best represented by variability-adjusted average emissions) than the units EPA selected as its best performers.

Instead, EPA used the lowest test result to select its best performers, but included all available tests for these "best performers," further adjusted for variability, in order to determine the floor. In so doing, it has failed to observe the statutory standard.

B. EPA's Variability Adjustments Result in Floors that Do Not Reflect the Average Emissions of the Best Performing Unit or Units

EPA set its existing source floors by, first, calculating the mean of all the test data submitted by the plants it selected as the best performing 12%. It then subjected that mean to a statistical manipulation intended to generate an "upper prediction limit" (UPL) at a confidence level of 99 percent. 76 Fed. Reg. at 25041. EPA asserts that this calculation results in a floor that a "best performing source can expect to meet 'every day and under all operating conditions.'" *Id.* (citation omitted).

That statistical adjustment is unwarranted. Section 112(d)(3) allows EPA to account for variability between the best performing units only by setting floors that reflect the "average" emission limitation these sources achieve. 42 U.S.C. § 7412(d)(3)(A). EPA has not supported its assumption that the best-performing sources vary to the maximal degree predicted by a 99% UPL; operating conditions (and other factors affecting emissions) are not randomly variable – much less randomly variable to the extreme predicted a 99% UPL. The UPL, by design, provides a number which exceeds the expected performance of all the "best-performing" unit or units. In using the UPL to set its existing floor limits, EPA has not, as a result, set its limits at the "average" emissions of the best performing 12% of existing sources. 42 U.S.C. § 7412(d)(3)(A). By definition, such an average would be below the

worst-case performance of approximately half of the units in EPA's "best-performing" category (assuming, as EPA has, a normal distribution). EPA has, moreover, impermissibly inflated its standards in a manner reflecting variation between sources, rather than variation in the performance of individual units, as well as variation in test methods that is unrelated to plants' actual emissions.

The practical effect of that manipulation is to heavily inflate floors. EPA's variability adjustment projects (allegedly worst-case) emissions well in excess of any actual emissions reported by the best-performing units. The most startling example is EPA's calculated mercury floor for coal-fired units designed to burn coal of less than 8300 Btu/lb – a floor that exceeds the actual emissions of all thirty sources in the category. Beyond the Floor Memorandum dated March 14, 2011, at 4 (Docket no. EPA-HQ-OAR-2009-0234-3004). While EPA prescribed a lower mercury limit for that category, it has otherwise – by ignoring Section 112's "beyond the floor" standards – refused to examine methods of improving upon the worst-case (or beyond) performance resulting from its variability analysis.[3] *See* Section VIII.D, below. In combination, those two choices eviscerate the statutory standard. According to the data provided by EPA, over 60% of the units for which it has information already meet its mercury floor of 1.2 lb/TBtu. Its hydrogen chloride and particulate matter limits are both set at levels exceeding the reported emissions of over 70% of the units in EPA's dataset.

1. The Agency Has Not Justified Its Assumption of a Normal Underlying Distribution or Reliance on the Central Limit Theorem

The formulas for the Upper Prediction Limit (UPL) used in EPA's analysis are based upon an assumption of normality of the underlying distribution, or the application of the Central Limit Theorem (CLT) when sample size exceeds N = 15. Floor Memo at p.4 (Docket no. EPA-HQ-OAR-

[3] EPA should, at a minimum, examine whether there are available means of reducing the variability of HAP emissions from utility power plants.

2009-0234-9858). Statistical experience with contaminant concentration data suggest that such data are rarely from a single normal distribution. Moreover, environmental data are typically highly skewed and much larger samples are needed before the CLT can be applied on the distribution of the sample mean.

The CLT assumes that the sample is from a single population with mean μ and standard deviation σ. Neither is the case here. For example, for the coal mercury (Hg) data set (for coal units in EPA's >8300 BTU/lb subcategory), has 80 data points from 40 of the "best performing" units and each result is the mean of 3 underlying test runs. The combined sample, therefore, is from 40 (or more) different sub- populations, each with its own mean and standard deviation (see Figure 1). The 80 data points for the 40 best performing units includes data available for the past 5 years at particular units. As discussed elsewhere in these comments, there are clear issues with including this historic data for the "best performing" sources since the older results in many cases, were collected using different methods and under different conditions, and are therefore unlikely to belong to the same population.

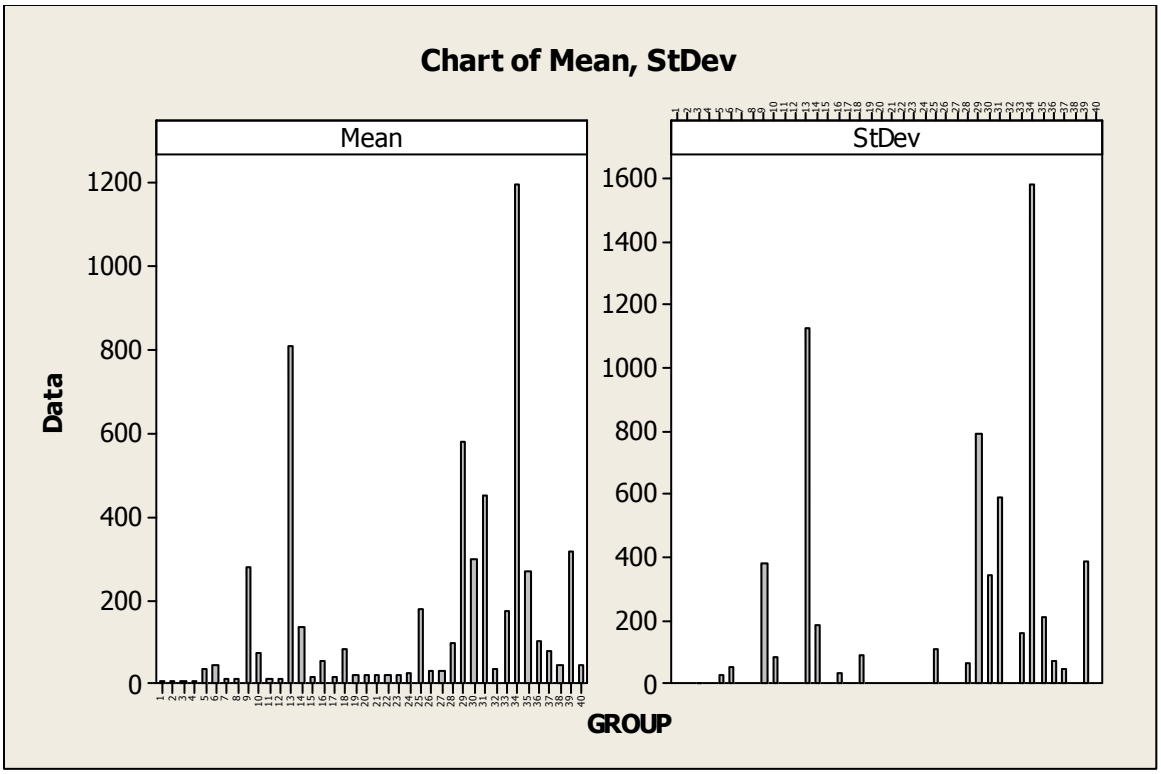

Figure 1: Bar chart of sample mean and sample sd of the 40 sub-populations in Coal_Hg data

We next ran the Kolmogorov-Smirnov test of normality on the Coal_Hg sample (see Figure 2). The P-value of the K-S test is less than .01, and the probability plot of the data clearly shows that Hg data is far from being normally distributed.

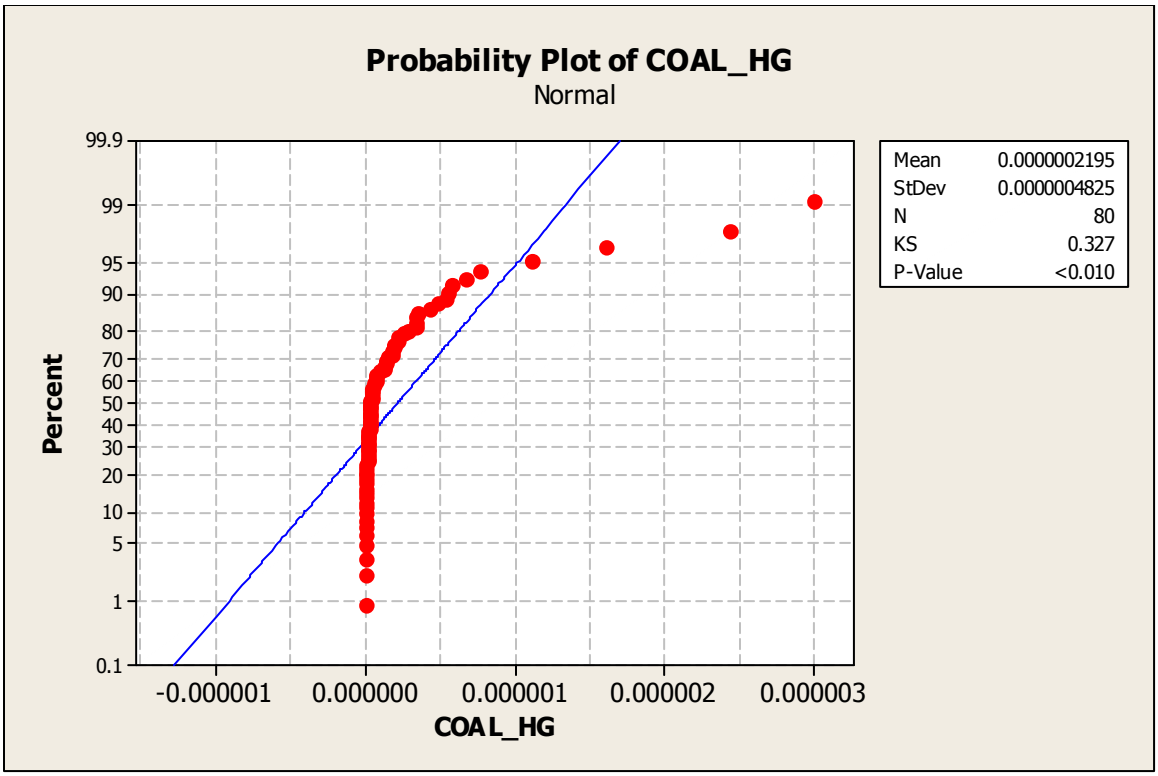

Figure 2: Probability plot of the Coal_Hg data

We also used the bootstrap method to see if the CLT applies to the Coal_Hg sample. The CLT verification method has the following steps: a) Generate a sample of size N with replacement from the Coal_Hg sample of size 80; b) Compute the mean of the bootstrap sample in Step 1. By repeating the steps (a) and (b) above 1000 times, we obtained sample means of 1000 bootstrap samples, each of size N. Figure 3 is the histogram and Figure 4 is normal Q-Q plot for bootstrap sample size of 15. Figure 5 is the histogram and Figure 6 is normal Q-Q plot for bootstrap sample size of 25.

Figures 3-6 demonstrate that the sample mean is not normally distributed for samples of sizes 15 and 25. Since each observation is the mean of 3 measurements, this shows that samples of sizes 45 to 75 are not large enough to guarantee CLT.

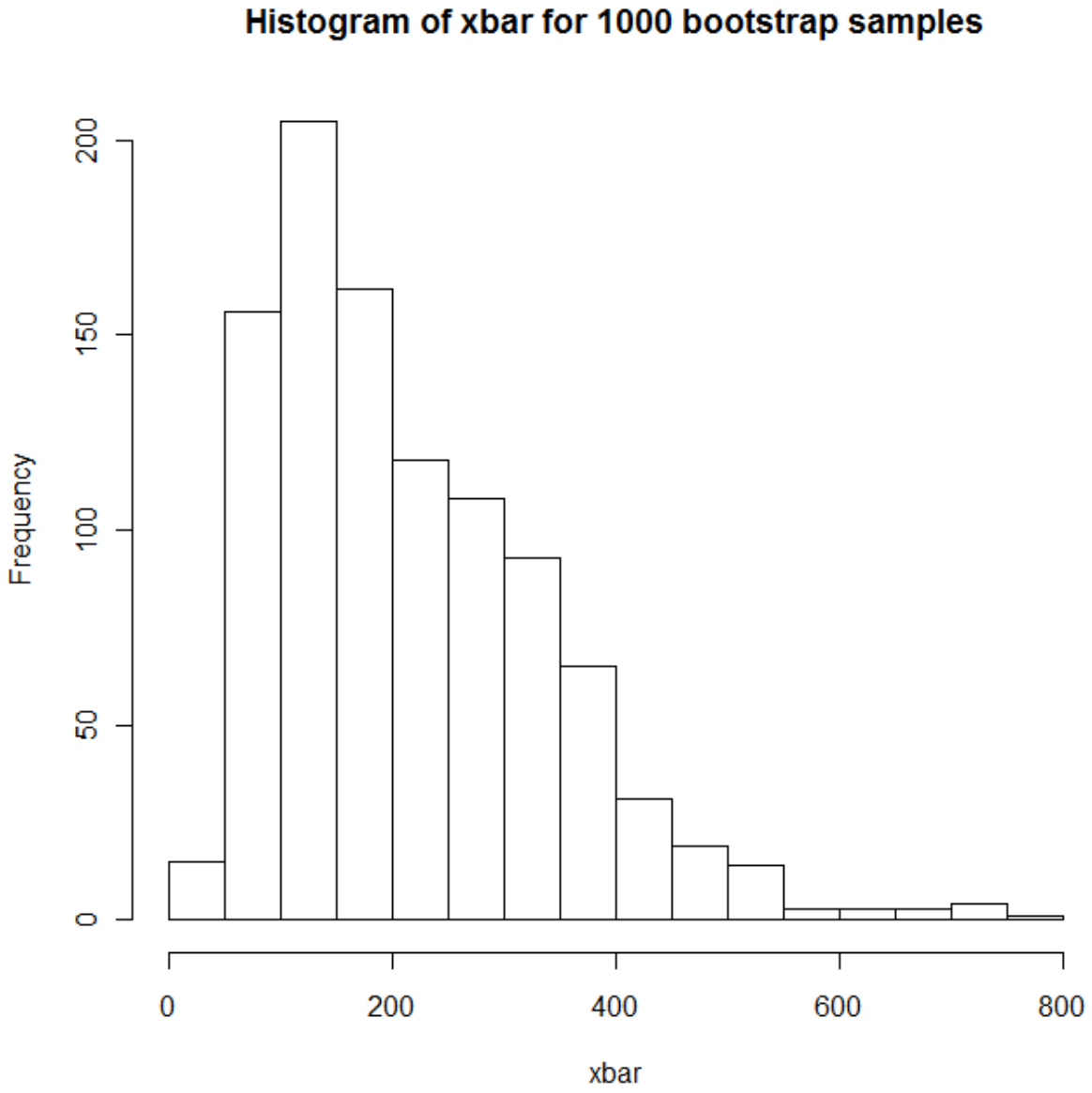

Figure 3: Histogram of xbar for 1000 Bootstrap Samples of size 15

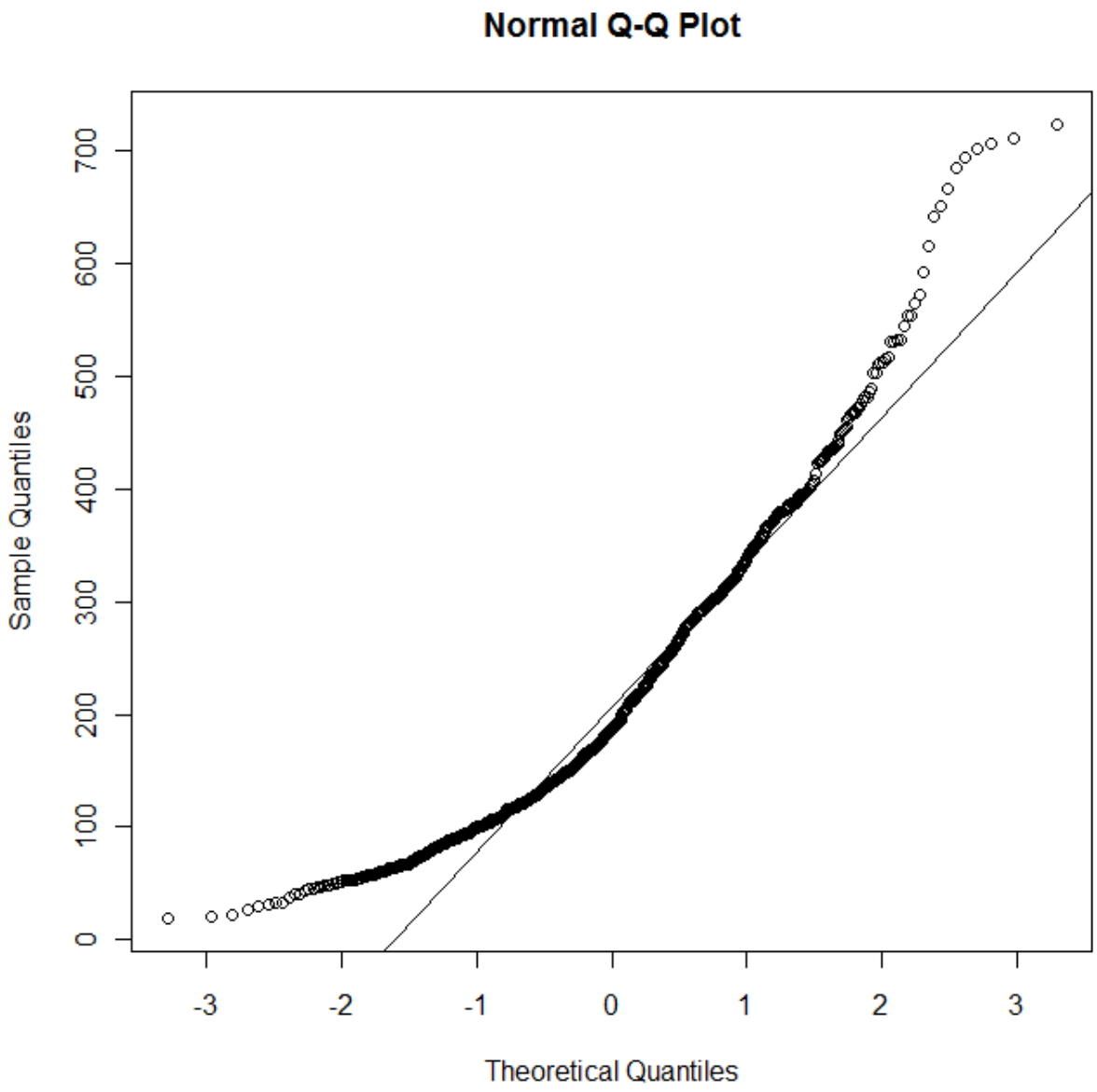

Figure 4: Q-Q Plot of xbar for assessing normality of sample mean

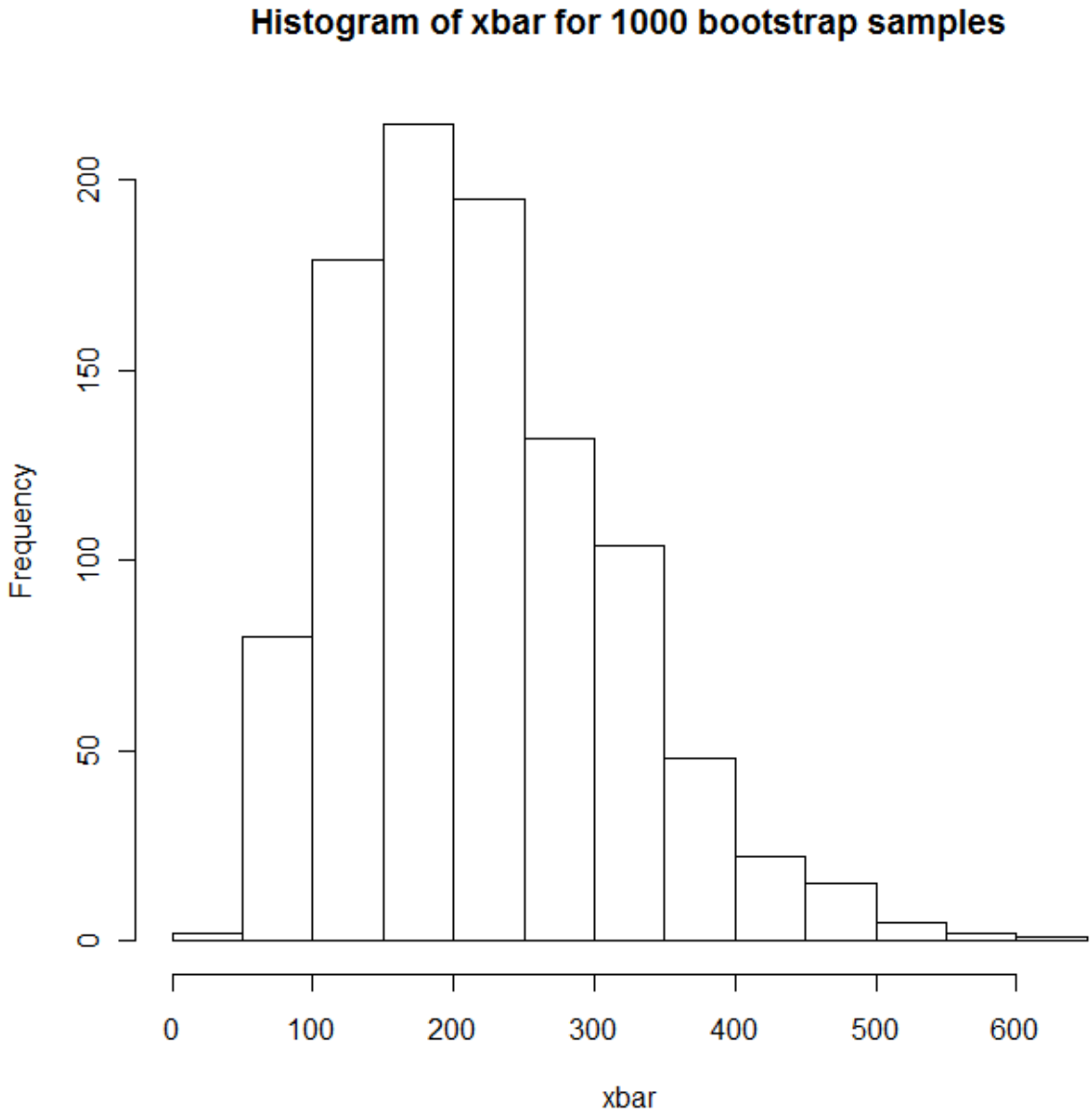

Figure 5: Histogram of xbar for 1000 Bootstrap Samples of size 25

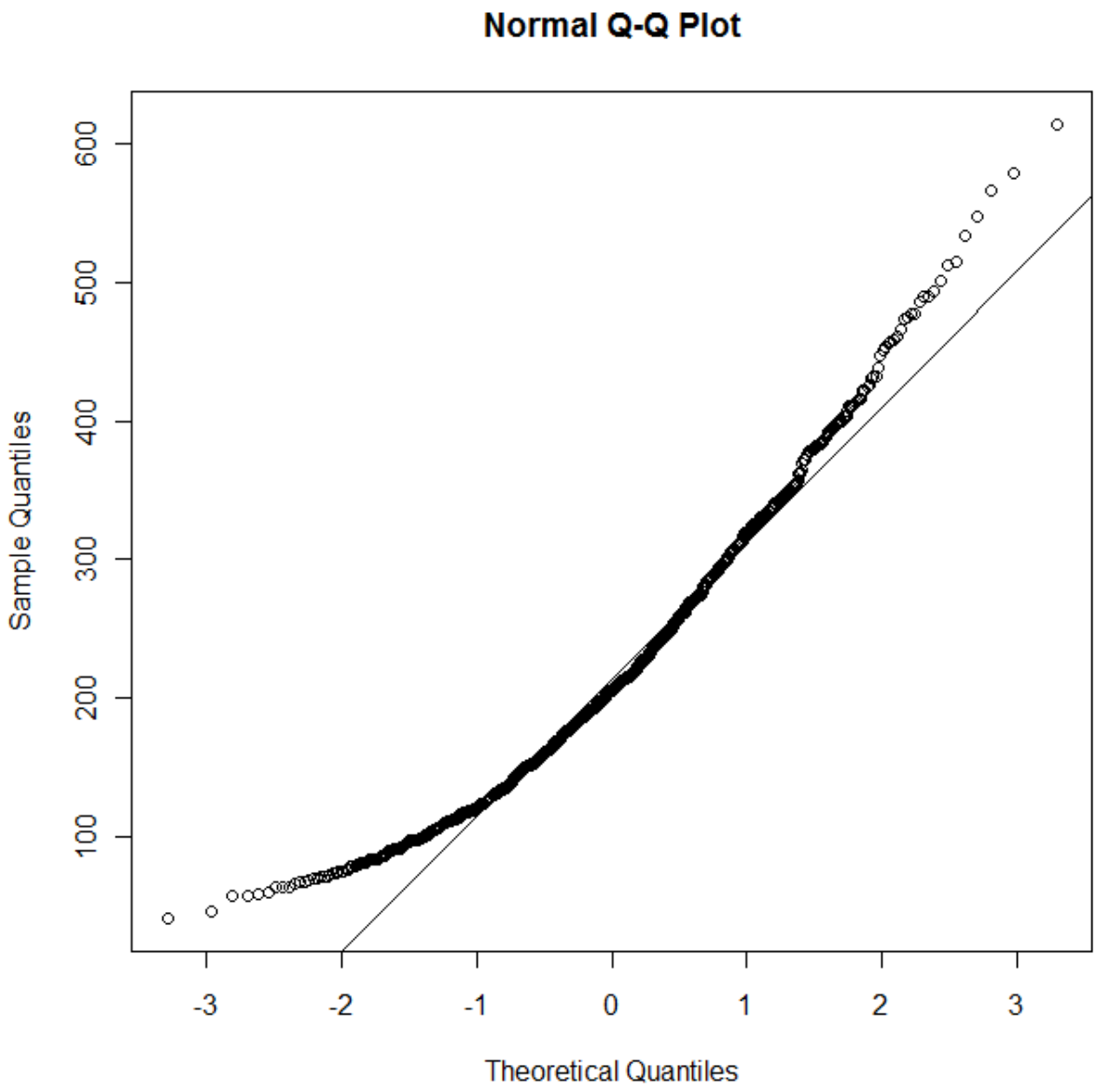

Figure 6: Q-Q Plot of xbar for assessing normality of sample mean

Based on the above, it is unlikely that the mechanical computation of the 99 UPL, even if were otherwise permissible, has any real meaning or significance. EPA should first determine the properties of the underlying distributions before computing such UPLs. For example, if we consider the best performing 40 units that belong in the existing-plant mercury floor calculation (for the coal >8300 BtU/lb subcatgory), and include only those results that were used to determine that these were the best performing units (i.e., the most recent ICR data – which were overwhelmingly collected using the same EPA Method 30B test method), we can calculate the various UPL values for this data set. We used EPA's approved statistical tool (ProUCL, available at http://www.epa.gov/esd/tsc/software.htm) to do the analysis. The results are as follows, for this data set: the best fit distribution for this data set was a gamma distribution and the calculated 99% UPL value was 7.51E-08 lb/MMBtu or 0.075 lb/TBtu. EPA's standard of 1.2 lb/TBtu is 16 times greater.

2. The Agency's Variability Assessment Is Distorted By Its Treatment of Results Below Detection Limits

In calculating the emission rates used to determine its floors, EPA assumed detection limit values were equal to the detection limit. When a unit's emissions are at or below the detection level, its emissions could be just below the detection level – or much lower. To account for such test results in developing emission factors, EPA has typically assumed emissions were at half the detection level, when there are other test results for the unit that measured above detection limits.[4] That approach would yield significantly

[4] *See* US EPA, Procedures for Preparing Emissions Factor Documents, EPA 454/R-95-015, November 1997, at 4-21, available at http://www.epa.gov/ttn/chief/efdocs/procedur.pdf.

lower limits.[5] Maximum likelihood estimation (MLE) methods, imputation methods, or the Kaplan-Meier methods would be more accurately account for below-detection-limit results.[6] Instead, EPA has assumed that emissions are equal to the detection limit of the test at that unit, thereby has overestimated the average emission rates of the best performing units.

C. The Agency Should Not Set Limits at Triple the Method Detection Limit

EPA seeks comment on an approach by which it would set standards at triple a "representative" Method Detection Limit (MDL), when-ever its calculations would otherwise yield a MACT floor below that triple-detection-limit threshold. 76 Fed. Reg. at 25,044. EPA's Floor Memorandum indicates that EPA has adjusted its new source standards for HCl and HF for coal, HCl for IGCC, and HCl, HF, Beryllium, and Cadmium standards for oil categories in this manner. Memorandum from Jeffrey Cole to Bill Maxwell dated May 18, 2011 ("Floor Memorandum") at 8-9 (Docket no. EPA-HQ-OAR-2009-0234-9858). EPA claims that this adjustment is necessary to address "measurement imprecision" associated with test results near the method detection limit. EPA should not adopt such an approach.

First, the record does not support the "measurement imprecision" suggested by the Agency. In its proposal, EPA notes that "[t]he expected measurement imprecision for an emissions value occurring at or near the MDL is about 40 to 50 percent. Pollutant measurement imprecision decreases to a consistent relative 10 to 15 percent for values measured at a level about three times the MDL." 76 Fed. Reg. 25,044. As support EPA cites the

---

[5] *See* Helsel, D.R., "Fabricating Data: How Substituting Values for Non-Detects Can Ruin Results and What Can Be Done About It," Chemosphere Vol. 65, 2006. Ex. 9-33.

[6] See Helsel, D.R., "More Than Obvious: Better Methods for Interpreting Non-Detect Data," Environmental Science and Technology, October 15, 2005, p. 419. Ex. 9-34 (detailing flaws in substitution of detection limits or half-detection limits for non-detect results)

American Society of Mechanical Engineers, *Reference Method Accuracy and Precision (ReMAP): Phase 1, Precision of Manual Stack Emission Measurements,* CRTD Vol. 60, February 2001 (Docket no. EPA-HQ-OAR-2009-0234-2937). That document does not support EPA's statements. ReMAP notes that its purpose was to "determine the precision of pollutant emission measurements based on analysis of available simultaneous sample test data" for various pollutants, including those analyzed by EPA Method 26 for HCl.[7] The ASME ReMAP analyzed data using Method 26 and not Method 26A, which EPA used for HCl analysis in the ICR data collection. ReMAP uses the standard deviation of the distribution as a basis for estimating precision and selects two other parameters: (i) the anticipated range for 99 out of 100 future measurements; and (ii) the anticipated range for 99 out of 100 future triplicate measurements.[8] However, ReMAP includes a critical assumption: that in its analysis "the measurement method is being applied to a hypothetical stack with an average pollutant concentration that does not vary with time." It acknowledges that "[i]n real world applications, single and triplicate measurement results will indicate variation from both random errors and systematic errors (bias) in the measurement process as well as temporal variation in facility operation."[9] This assumption does not hold in the case of the current data set – we do indeed have both random and systematic errors as well as temporal variations.

With the caveats and assumptions discussed above, ReMAP found that while the "[relative standard deviation (RSD)] does increase when the method is applied to stacks with very low concentration," the "RSD is typically in the range of 5% to 10%." The RSD is the standard deviation divided by the average. In fact, ReMAP data for 99 out of 100 triplicate measurements of HCl over a wide range of concentrations show the following:

---

[7] ReMAP, p. ES-1 (Docket no. EPA-HQ-OAR-2009-0234-2937).
[8] ReMAP, p. ES-1 and ES-2 (Docket no. EPA-HQ-OAR-2009-0234-2937).
[9] ReMAP, p. ES-2 (Docket no. EPA-HQ-OAR-2009-0234-2937).

True Stack HCl Concentration (mg/dscm)	99 out of 100 Triplicate Measurements	
	Lower Limit	Upper Limit
1	0.80	1.20
5	4.26	5.74
10	8.71	11.3
20	17.7	22.3
50	45.3	54.7
100	91.8	108.2

When the true concentration is 1 mg/dscm in the stack, the precision in that data set (i.e., its standard deviation) would predict the 99 percentile interval for triplicate future measurements to be 0.80-1.20 mg/dscm – i.e., around +/- 20% of the true value. When the true concentration at the stack is 100 mg/dscm, i.e., 100 times the previous case, the 99 percentile interval is 91.8-108.2, or around 8% or so. Over a very wide range of true concentrations, the variability is such that the prediction range varies by 8% to 20%. This does not support EPA's statement regarding imprecision of "40-50% . . . at or near the MDL."

Second, even if the claimed imprecision existed, Section 112 does not permit the Agency to substitute triple the detection limit for standards based on the actual performance of the best performing similar source (or sources). The Act requires new source standards to be no higher than the emissions of the "best performing similar source." 42 U.S.C. § 7412(d)(3). The Agency does not suggest – nor does the record – that the actual emissions of units reporting emissions below the detection limit are, in fact, triple the selected detection limit. Even if, as the Agency suggests, its emissions data from sources at or near a detection limit are imprecise by 40 to 50 percent, at most that source's actual emissions would be 50 percent higher than the reported detection limit for that source (it is equally likely that the actual value would be 50 percent lower). The Agency has, instead, substituted a number that is *triple* the detection limit obtained by a laboratory for (in most cases) a

completely different source. Likewise, where the Act requires existing source standards to be at least as stringent as the average emissions achieved in practice by the best performing 12 percent of existing sources, 42 U.S.C. § 7412(d)(3), the Agency has substituted a figure that does not reflect the actual emissions of *any* source within the category.[10]

D. The Agency's Beyond-the-Floor Analysis Does Not Meet Its Statutory Obligations

The Clean Air Act requires EPA to establish standards that require "the maximum degree of reduction in emissions . . . that the Administrator, taking into consideration the cost of achieving such emission reduction, and any non-air quality health and environmental impacts and energy requirements, determines is achievable for new or existing sources . . . ." 42 U.S.C. § 7412(d)(2). EPA has, for the most part, refused to apply the substantive requirements of that provision, instead setting its standards (with only two exceptions) according to the maximum "floors" set out in sections 112(d)(3). *See* 76 Fed. Reg. at 25,027 (proposing beyond-the-floor mercury standard for existing coal-fired units designed for coal less than 8,300 Btu/lb, and for new IGCC units).

In justification, EPA primarily asserts that it "could not identify better HAP emissions reduction approaches," aside from duplicate controls (an option EPA discards as too costly). 76 Fed. Reg. at 25,046. In its supporting materials, EPA provides some additional consideration of pollution reductions achievable through use of natural gas as a fuel, but dismisses such reductions as overly expensive. *See also* Beyond the Floor Memoranda dated March 14, 2011 (Docket no. EPA-HQ-OAR-2009-0234-3004), and March 15, 2011 (Docket no. EPA-HQ-OAR-2009-0234-2925),

[10] Nor can the Agency justify its floor adjustment by asserting an inability to measure emissions below its triple-maximum-detection limit floor. *See* Chapter IX.E, below.

That perfunctory analysis fails to justify EPA's standards according to the statutory criteria. Section 112's command to prescribe the "maximum degree of reductions," requires, at a minimum, that EPA identify and assess all available methods of reducing plants' pollution. *Cf.* 42 U.S.C. 7479(3) (defining "best available control technology limits" in similar terms). EPA instead noted only that the best performing sources, for mercury, utilized activated carbon injection (ACI) in conjunction with a fabric filer (FF), and that no superior removal methods existed. ACI is not, however, a "single" technology, but rather a generic approach for mercury reduction in coal-plant exhaust gas streams. Different injection systems use different types and amounts of carbon (per unit of flue gas volume). For example, utilities have used carbons of varying surface areas, and with different impregnated agents (such as bromine) to enhance mercury removal. Injection rates also affect mercury capture. Typical rates range from 1 lb/mmscf to over 10 lb/mmscf, with removal efficiency increasing accordingly. EPA should have considered, at a minimum, such alternatives within the range of applicable ACI systems.

Moreover, in requiring the "maximum achievable" limits, the statute requires EPA to identify the *maximum* reductions achievable using a particular technology. *See Alaska Dep't of Envtl. Conservation v. EPA*, 540 U.S. 461, 485-86 (2004) (noting, in interpreting other portions of Act, that "permitting authorities' discretion is constrained by the [Act's] strong, normative terms 'maximum' and 'achievable'"). That is especially important in light of EPA's extraordinarily aggressive adjustments in the name of variability. As a result of those adjustments, plants can and do achieve reductions well in excess of the reductions prescribed (as set forth in further detail below). At a minimum, the statutory standard requires EPA to investigate whether methods or techniques are available to improve upon that statistically extreme worst-case performance. 42 U.S.C. § 7412(d)(2) (requiring consideration of all "measures, processes, methods, systems or techniques" to reduce emissions).

Finally, the statute requires EPA to justify its standard by "taking into consideration" costs, the reductions achieved, non-air quality environmental impacts, and energy requirements. 42 U.S.C. § 7412(d)(2). EPA has provided no substantive discussion of those criteria to justify its standards. *See, e.g.,* 76 Fed. Reg. at 25,046 (noting only that switching to natural gas is more expensive that other control alternatives, without assessing reductions achieved, or other impacts).

The Agency's failure to provide an analytic basis tethering its limits to the statutory criteria is reflected in its inconsistent, and arbitrary, rationales for its decisions. EPA refused to enquire as to whether lower mercury emissions can be achieved by most coal-fired plants, offering as its rationale the absence of "other technology" than that used to establish the MACT floor. Yet it simultaneously prescribes a beyond-the-floor limit for units designed for coal of less than 8300 Btu/lb, because, *inter alia,* "the technology installed to achieve the MACT floor would be the *same* technology used to achieve the beyond-the-floor MACT limit." 76 Fed. Reg. at 25,046-47. It thereby applies the same rationale to essentially identical facts – to justify divergent decisions. Similarly, EPA's beyond-the-floor analysis for units designed for coal of less than 8300 Btu/lb concludes its mercury standard is cost-effective at a cost of $17,275/lb. Beyond the Floor Memo dated March 14, 2011 at 5 (Docket no. EPA-HQ-OAR-2009-0234-3004). Elsewhere, it refuses to examine available methods of reducing emissions without any detailed comparison of costs and reductions. *See, e.g,* 76 Fed. Reg. at 25,047-8 (refusing to consider combinations of HCl control systems or fuel switching based only on "significant additional cost" or "high cost" for reductions).

In its refusal to prescribe reductions beyond those of the best performing unit(s), EPA has failed to properly apply the statutory standard.