UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NC 27711

OFFICE OF
AIR QUALITY PLANNING
AND STANDARDS

MEMORANDUM

SUBJECT: Approach for Applying the Upper Prediction Limit to Limited Datasets

DATE: September 3, 2014

FROM: Barrett Parker and Brian Shrager
Office of Air Quality Planning and Standards

TO: Docket ID No. EPA-HQ-OAR-2010-0895

I. What is Our Approach for Applying the Upper Prediction Limit to Limited Datasets?

The UPL approach addresses variability of emissions data from the best performing source or sources in setting MACT standards. The UPL also accounts for uncertainty associated with emission values in a dataset, which can be influenced by components such as the number of samples available for developing MACT standards and the number of samples that will be collected to assess compliance with the emission limit. The UPL approach has been used in many environmental science applications.[1,2,3,4,5,6] As explained in more detail in memorandum entitled "Use of the Upper Prediction Limit for Calculating MACT Floors", which is located in the docket for this action, the EPA uses the UPL approach to reasonably estimate the emissions performance of the best performing source or sources to establish MACT floor standards.

Sample size (the number of values in a particular dataset) is an important component of the UPL approach. First, the use of the UPL approach requires us to identify the distribution of the data, that is, whether a particular dataset is normally or non-normally distributed (see the memorandum entitled "Use of the Upper Prediction Limit for Calculating MACT Floors", which

[1] Gibbons, R. D. (1987), *Statistical Prediction Intervals for the Evaluation of Ground-Water Quality*. Groundwater, 25: 455–465 and Hart, Barbara F. and Janet Chaseling, *Optimizing Landfill Ground Water Analytes*—New South Wales, Australia, Groundwater Monitoring & Remediation, 2003, 23, 2.

[2] Wan, Can; Xu, Zhao; Pinson, Pierre; Dong, Zhao Yang; Wong, Kit Po. Optimal Prediction Intervals of Wind Power Generation. 2014. IEEE Transactions on Power Systems, ISSN 0885-8950, 29(3): pp. 1166 – 1174.

[3] Khosravi, Abbas; Mazloumi, Ehsan; Nahavandi, Saeid; Creighton, Doug; van Lint, J. W. C. Prediction Intervals to Account for Uncertainties in Travel Time Prediction. 2011. IEEE Transactions on Intelligent Transportation Systems, ISSN 1524-9050, 12(2):537 – 547.

[4] Ashkan Zarnani; Petr Musilek; Jana Heckenbergerova. 2014. Clustering numerical weather forecasts to obtain statistical prediction intervals. Meteorological Applications, ISSN 1350-4827. 21(3): 605.

[5] Rayer, Stefan; Smith, Stanley K; Tayman, Jeff. 2009. Empirical Prediction Intervals for County Population Forecasts. Population Research and Policy Review, 28(6): 773 – 793.

[6] Nicholas A Som; Nicolas P Zegre; Lisa M Ganio; Arne E Skaugset. 2012. Corrected prediction intervals for change detection in paired watershed studies. Hydrological Sciences Journal, ISSN 0262-6667, 57(1): 134 – 143

is located in the docket for this action). To determine the distribution of the data, we use well-established tests (kurtosis and skewness tests), and these tests require at least three data points. In prior rulemakings, we used the kurtosis equation that is a built-in function in Excel software, but that equation requires at least 4 values in order to avoid dividing by zero, which precluded its use for datasets of 3. However, we recently further reviewed the application of the UPL where data are limited, and concluded that it is appropriate to use another kurtosis estimator[7] that provides a meaningful result with just 3 values. The use of this estimator is important, as many, if not most, of our new source emissions limits are based on 3 samples collected during an emissions test.

Once we determine the data distribution, the appropriate equation to be used in the UPL approach is selected based on the data distribution. We recognize that the use of the UPL approach for limited datasets introduces some amount of uncertainty in the calculation of MACT standards, and, therefore, we are taking additional steps, discussed in this document, to ensure that the level of the MACT standards is reasonable. We also note that after MACT standards are promulgated, we are required to review those standards periodically, and for such reviews, we typically have significant additional HAP emissions data from the intervening years of compliance with which to further assess the actual performance of the various emission sources.

Regardless of the distribution of the data, UPL equations have three well-defined components: an average, the t-score and a measure of variability that includes the actual variability of the data, the sample size, and the number of data points that are averaged together to determine compliance with a particular emission limit. A t-score is a value that estimates the uncertainty and variability for a certain confidence level associated with a specific number of data points. For a constant confidence level, t-scores decrease as the number of samples increase. This means that, if the mean and variance remain constant, UPL values for a particular confidence level decrease as the number of samples increase. Figure 1 and Table 1 show the t-scores for various sample sizes, and demonstrate that the t-score is highly variable at the smallest sample sizes and becomes relatively constant once the sample size is larger than a few data points. Consequently, we recognize that we need to take special care when smaller, limited sample sizes are used to establish emission limits, as the t-scores can have a disproportionate effect, overwhelming the other components of the emission limit calculation. In addition, we recognize that for a sample size of fewer than three data points, which has a very large t-score and precludes the appropriate selection of a distribution, we should not develop emission limits using the UPL. In other words, if fewer than 3 data points are available for use in determining an emission limit for a particular source, and no other data from sources in the subcategory are available, we would have to establish a different procedure for establishing the MACT floor that does not rely on the UPL.

[7] Doric D, Nikolic-Doric E, Jevremovic V and Malisic J. 2009. On measuring skewness and kurtosis. Qual Quant. 43:481-493. DOI 10.1007/s11135-007-9128-9.

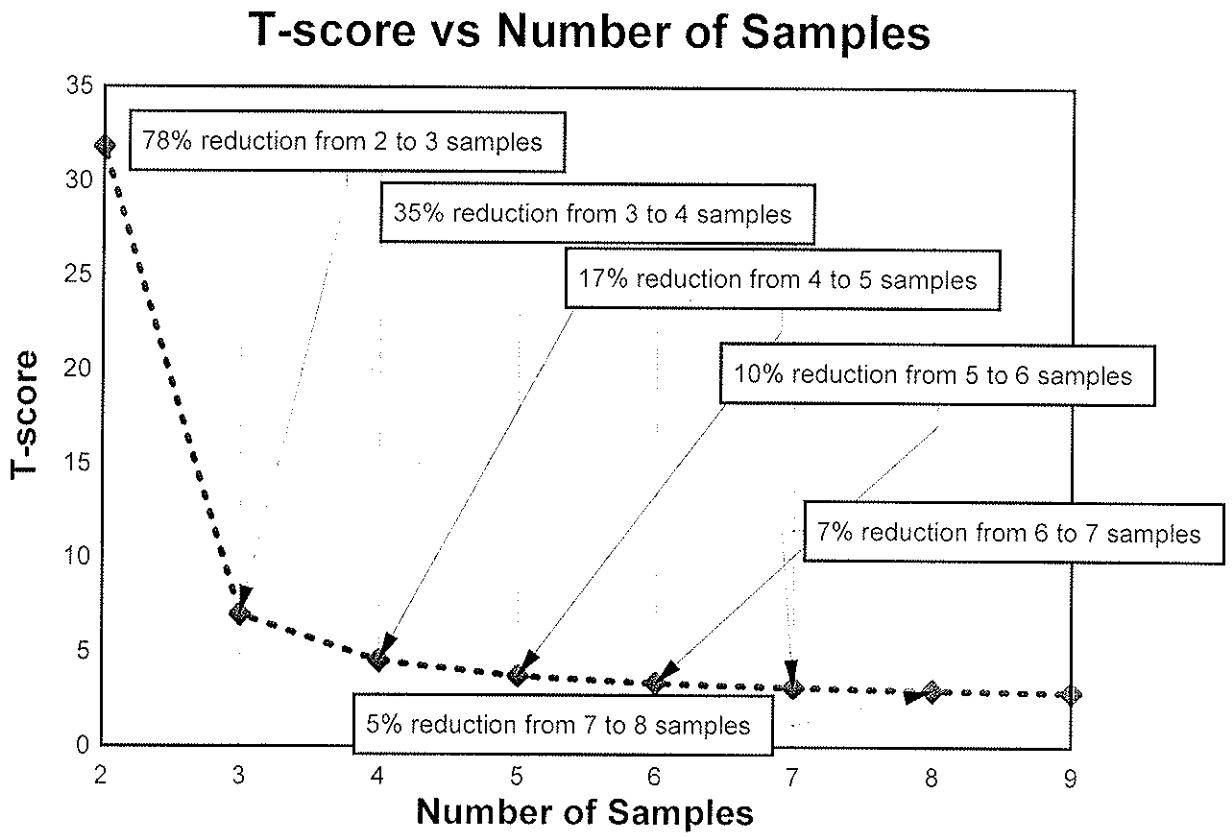

Figure 1. Sample sizes versus one-sided, 99 percent confidence interval t-scores for normal distributions

Table 1. Degrees of Freedom, T-Score for a One-Tailed, 99 Percent Confidence Interval, and Percent Difference in T-Score for Each Sample Size

Sample Size (n)	Degrees of Freedom (n-1)	t-score (99 percent)	Relative Percent Difference[a]
2	1	31.82	
3	2	6.96	78
4	3	4.54	35
5	4	3.75	17
6	5	3.36	10
7	6	3.14	7
8	7	3.00	5
9	8	2.90	3

[a]The relative percent difference between the t-score and the t-score shown on the preceding row

In the case of a dataset with a normal distribution, the equation to calculate the UPL is as follows:

$$UPL = \bar{x} + t_{[(n-1),(1-\alpha)]}\sqrt{s^2\left(\frac{1}{n}+\frac{1}{m}\right)}$$

where $\bar{x}$ = average or mean of test run data;

$t_{[(n-1),(1-\alpha)]}$ = t score, the one-tailed t value of the Student's t distribution for a specific degree of freedom (n-1) and confidence level (1- α);

α = level of significance expressed as a decimal (e.g., 1% significance = 0.01), note that confidence level = $100 - (\alpha \times 100)$;

s^2 = variance of the dataset (test run);

n = number of values (test runs) in the dataset; and

m = number of values used to calculate the test average (generally 3).

As shown in the equation, the sample size (n) directly affects the UPL calculations in two ways. First, as mentioned earlier, the t-score is selected based on the sample size. Second, the sample size is a separate variable in the equation, and as with the t-score, as the sample size gets larger (all other variables being held constant), the UPL gets smaller (i.e., approaches the mean value). The sample size also influences both the mean ($\bar{x}$) and variance (s^2) in that larger sample sizes provide better estimates of the population mean and variance. The confidence level also influences the t-score, and is represented in the UPL equation by the subscript "1- α." As noted in the memorandum entitled "Use of the Upper Prediction Limit for Calculating MACT Floors", which is located in the docket for this action, we typically choose the 99 percent UPL, which is another way of saying that α equals 0.01.

Note that the UPL equation for a lognormal distribution uses a z-score, instead of a t-score.[8] As with the normal UPL equation, the lognormal UPL equation is affected by sample size and the variance of the data, and the basic concepts outlined in this discussion apply to datasets regardless of the distribution.

In summary, sample size is important for UPL calculations, as it is a key component of the t-score value and the calculation of the mean and variance of the sample. For the reasons presented above, for datasets below a certain size, further evaluation is warranted to ensure that the results are a valid representation of the performance of the emission sources whose data are included in the dataset. The remainder of this document provides our assessment of what constitutes a

[8] Bhaumik, D.K. and Gibbons, R.D.``An Upper Prediction Limit for the Arithmetic Mean of a Lognormal Random Variable", Technometrics, 46, 2004, 239-248.

limited dataset and what additional analyses are appropriate for ensuring that the UPL calculations are valid representations of the performance of units as reflected in the available data.

II. What is a limited dataset?

While there are different options for choosing the number of samples that represents a limited dataset (or small sample size), we can use distinctions based on our observations in Figure 1 and Table 1 as an aid in defining what constitutes a limited dataset. As discussed earlier, the t-score is an influential parameter that can have a particularly large influence on an emission limit developed from a limited dataset. As such, the t-score is a key parameter for identifying what constitutes a limited dataset for purposes of MACT analyses. As shown in Figure 1 and Table 1, above, at the 99 percent confidence level there is a clear distinction between t-score values for sample sizes of 3 and fewer when compared to t-score values for sample sizes of more than 6. As Figure 1 and Table 1 show, the t-score changes drastically from sample size equal to 2 to sample size equal to 3. The changes in the t-score are considerably less dramatic as the sample size approaches 9 data points and larger.

As noted earlier, we will not use the UPL approach for datasets with fewer than 3 values because of the high degree of uncertainty that results from the size of the dataset and the related t-score. In fact, the t-score (see Figure 1 and table 1) for a dataset of 3 is 78 percent lower than a t-score for a dataset of 2, meaning that given an equal variance, the component of the UPL that includes the t-score and variance would be more than 78 percent less when evaluating a 3-point dataset compared to a 2-point dataset. The table and figure also show that the relative change of the t-score for each additional data point decreases as the sample size gets larger. As sample size increases, uncertainty decreases, but there is always some amount of uncertainty. The discussion that follows presents our examination of the t-score at various sample sizes. We rely on this examination to help identify a sample size below which we would further evaluate the data and the application of the UPL to ensure that the amount of uncertainty is within a reasonable range. Based on Table 1, the lower bound of the sample size where such further evaluation is needed is 3 values, and a reasonable value (where the size of the t-statistic has much less influence on the eventual emission limit) lies between 4 and 9 values.

An additional factor to consider is that, in the context of MACT analyses, emission tests typically include 3 test runs (independent data points), and, therefore, our MACT floor dataset size typically is a multiple of 3. For this reason, we first considered sample sizes that are multiples of 3 when identifying the size below which further scrutiny is needed. As can be calculated from the t-scores shown in Table 1, the relative difference between the t-score for 6 data points and the t-score for 9 data points is about 14 percent ((3.36-2.90)/3.36). The relative difference between the t-scores for 7 and 9 data points is about 8 percent ((3.14-2.90)/3.14). Given that the test methods that are used to generate the data typically are accurate to within 10 to 20 percent[9] , this difference in t-score (between datasets with 7 and 9 runs) is less than the uncertainty in the test methods and would not have a large influence on the magnitude of an emission limit. Conversely, comparing the t-score for 3 runs to that for 7 runs, the relative

[9] (ReMAP): PHASE 1, Precision of Manual Stack Emission Measurements; American Society of Mechanical Engineers, Research Committee on Industrial and Municipal Waste, February 2001.

difference in the t-score is 55 percent ((6.96-3.14)/6.96), indicating that there is a possibility, especially if the variance is high, that the t-score for 3 test runs may introduce a large amount of uncertainty in an emission limit, and, therefore, additional analysis of a dataset of this size is warranted. We also conducted the same type of evaluation on datasets between 4 and 6 points, and while the change in t-score for datasets of these sizes could be considered to be sufficiently small, to be conservative, we conclude that a limited dataset, for purposes of MACT analyses, is a dataset that includes at least 3 test runs and fewer than 7 test runs. Because of the uncertainty associated with datasets at or below this size, further evaluation is warranted, and the types of additional analyses that we may choose to conduct are explained below.

III. What approach will the EPA follow for MACT floors based on limited datasets?

When a MACT floor for either existing or new sources is based on fewer than 7 data points, we will further evaluate each individual dataset in order to ensure that the uncertainty associated with a limited dataset does not cause the calculated emission limit to be so high that it does not reflect the average performance of the units upon which the limit is based after accounting for variability in the emissions of those units. The evaluation will include one or more of the following, depending on the specific dataset: confirming that the data distribution was selected correctly; after confirming the data distribution, ensuring that we use the most appropriate UPL equation[10]; and, as necessary, comparing UPL equation components for the individual unit upon which a new source floor is based with those of the units in the existing source floor to determine if our identification of the best unit is reasonable. Each of the additional evaluations are discussed below.

Confirming the data distribution is important because UPL equations and the emission limits derived from those UPL equations depend on the distribution. In prior rulemakings, we had not identified a way of ascertaining the data distribution for units with three samples, so we either assumed a normal or lognormal distribution. This particularly affected new source emissions limits, many of which were based on three samples. However, we have since identified and adopted an established technique that compares the ratio of skewness to the standard error of the skewness for both the raw and log transformed data. The lower of the two ratios identifies the data distribution that best represents the sample set. This ability to more precisely ascertain the data distribution enables us to select the best option for the UPL equation given the small sample size.

After confirming that we selected the best distribution based on the available data and ensuring that we used the correct equation to calculate a MACT floor value, we then examine key variables that factor into the calculated floor values. The variables that are factored into the emission limit include the following: average, t-score, confidence level, variance, number of test runs (i.e., sample size), and number of runs in each test (i.e., the number of data points that are averaged together to determine compliance with a particular emission limit). The considerations related to each variable are discussed below.

[10] For datasets with lognormal distributions, we recently determined that the most appropriate UPL equation, which is especially important for the smallest datasets that we use, is based on an approach described in Bhaumik, 2004, op. cit.

The average, which typically is calculated as the mean value of a dataset, does not require additional consideration. Once it is calculated, it is used to rank the performance of units, and also is the key value upon which each emission limit is based. As such, average values have already been compared within a particular dataset, and additional assessment is not necessary. However, in the overall assessment of variables, when multiple best performing units have emission averages that are similar, we may look to other variables like the variance to help to inform our decision as to which unit is the single best performer.

As discussed earlier, the t-score is dependent on the sample size, and, therefore, we discuss these two variables together. As shown earlier in Figure 1, as the sample size decreases, the t-statistic increases. This is important because the t-statistic is multiplied by the variance and a factor that involves the sample size and the number of test runs used for compliance (i.e., the "m" term in the UPL equation). For the smallest datasets (3 data points) that we consider under the UPL approach, a large variance coupled with a t-score that is significantly greater than the t-score for a dataset of 7 or more can create a situation where the emission limit includes a large amount of uncertainty. In such cases, we use our technical expertise to assess what level of emissions could realistically be expected from the type of unit, controls, and other relevant factors, and would ensure that the emission limit is reasonable.

The variance is another key variable that we would evaluate. Generally speaking, if our evaluation showed a very small amount of variability (e.g., a small variance), the emission limit would not contain an unacceptable amount of uncertainty and variability because a small variance would result in an emission limit that would be reasonably close to the mean of the data. With larger datasets, it is more likely that the demonstrated variance is truly representative of the variation in the source's emissions simply because there is more evidence of the source's emissions over time. On the other hand, when a limited dataset includes a large amount of variability (e.g., a large variance), we would need to carefully evaluate the data. For instance, consider the case where the pool of best performing units all have similar averages and the unit identified as the single best performer (based on average emissions only) has a limited dataset and a variance that greatly exceeds the variance of the other similarly-controlled best-performing units. In such a case, the MACT floor analysis may yield an emission limit for that unit (i.e., the new source MACT floor) that is higher than the existing source MACT floor, which is an indicator that further analysis is warranted. Careful consideration of variance will allow us to better ascertain which unit should be considered the best performer.

Another tool that could be used to adjust unreasonable emission limits is the selection of the confidence level. In certain instances, we may determine that a limited dataset includes such a high degree of variability that the 99 percent confidence level results in an emission limit that our experience suggests is higher than the average emissions limitation achieved by the best performing source or sources after accounting for variability. In such cases, we may choose to select a confidence level of 95 percent or 90 percent. As an example, consider a unit identified as the best performing new source with a specific process or control device that has been demonstrated to operate far more efficiently (thus having the potential to lower emissions) in similar units, processes, or control devices. Under these conditions, we may choose to acknowledge the better operation by lowering the confidence level (which lowers the emissions limit).

Establishing and using this approach on a case-by-case basis for limited datasets will ensure consistent application of emissions limit development procedures, which will mitigate the additional uncertainty that could otherwise result from calculating MACT floor standards with limited data.[11]

IV. How did we apply the approach for limited datasets to limited datasets in the ferroalloys source category?

For the ferroalloys source category, we have limited datasets for the following pollutants and subcategories: PAHs for existing and new furnaces producing ferromanganese (FeMn); PAHs for new furnaces producing silicon manganese (SiMn); mercury for new furnaces producing SiMn; mercury for existing and new furnaces producing FeMn; and HCl for new furnaces producing FeMn or SiMn. Therefore, we evaluated these specific datasets to determine whether it is appropriate to make any modifications to the approach used to calculate MACT floors for each of these datasets.

For each dataset, we performed the steps outlined above, including: ensuring that we selected the data distribution that best represents each dataset; ensuring that the correct equation for the distribution was then applied to the data; and comparing individual components of each limited dataset to determine if the standards based on limited datasets reasonably represent the performance of the units included in the dataset. The results of each analysis are presented below.

The MACT floor dataset for PAHs from existing furnaces producing FeMn includes 6 test runs from 2 furnaces. This subcategory includes only two units, and the CAA specifies that the existing source MACT floor for subcategories with fewer than 30 sources shall not be less stringent than "the average emission limitation achieved by the best performing 5 sources." However, since there are only 2 units in the subcategory and we have data for both units, the data from both units serve as the basis for the MACT floor. After determining that the dataset is best represented by a normal distribution and ensuring that we used the correct equation for the distribution, we considered the selection of a lower confidence level for determining the emission limit by evaluating whether the calculated limit reasonably represents the performance of the units upon which it is based. In this case, where two units make up the pool of best performers, the calculated emission limit is about twice the short-term average emissions from the best performing sources, indicating that the emission limit is not unreasonable compared to the actual performance of the units upon which the limit is based and is within the range that we see when we evaluate larger data sets using our MACT floor calculation procedures. Therefore, we determined that no changes to our standard floor calculation procedure are warranted for this

[11] We note that in the recent D.C. Circuit Court of Appeals decision in National Assn. of Clean Water Agencies v. EPA (NACWA), which involved challenges to EPA's MACT standards for sewage sludge incinerators, the court identified two instances where we replaced a new source floor value with the existing source value because the new source value was less stringent than the existing source value (the "anomalous result"). See 734 F.3d 1115. The procedures outlined here would eliminate the anomalous result in both instances in the rule at issue in NACWA. The first instance in which the court in the NACWA decision identified the anomalous result is where we established an emission limit using the UPL on a dataset that included only 2 data points. The other instance in which the court identified the anomalous result is where we did not apply the most appropriate lognormal UPL equation.

pollutant and subcategory, and we are proposing that the MACT floor is 1,400 µg/dscm for PAHs from existing furnaces producing FeMn.

The MACT floor dataset for PAHs from new furnaces producing FeMn includes 3 test runs from a single furnace (furnace #12 at Eramet) that we identified as the best performing unit based on average emissions performance. After determining that the dataset is best represented by a normal distribution and ensuring that we used the correct equation for the distribution, we evaluated the variance of the best performing unit. Our analysis showed that this unit, which was identified as the best unit based on average emissions, also had the lowest variance. Therefore, we determined that the emission limit would reasonably account for variability and that no changes to the standard floor calculation procedure were warranted for this pollutant and subcategory, and we are proposing that the MACT floor is 880 µg/dscm for PAHs from new furnaces producing FeMn.

The MACT floor dataset for PAHs initially identified for new furnaces producing SiMn includes 6 test runs from a single furnace (furnace #2 at Felman) that we identified as the best performing unit based on average emissions. After determining that the dataset is best represented by a normal distribution and ensuring that we used the correct equation for the distribution, we evaluated the variance of this unit (furnace #2 at Felman) and concluded that further consideration of the variance was warranted. In particular, we noted that the variance of the dataset for this unit was almost twice as large as the variance of the dataset for the pool of best performing units that was used to calculate the existing source MACT floor. The high degree of variance in the dataset for the unit with the lowest average prompted us to question whether this unit was, in fact, the best performing unit and to evaluate the dataset for the unit with the next lowest average (furnace #7 at Felman). The dataset for furnace #7 includes 3 test runs, the furnaces are controlled with the same type of add-on control technology, and the average emissions from furnace #2 are only about 22 percent lower than the average emissions from furnace #7. While we find the average performance of these 2 units to be similar, the unit with the higher average has a variance more than 2 orders of magnitude lower than that of the unit with the lower average, thus indicating that the unit with the higher average has a far more consistent level of performance. The combination of components from the unit with the higher average (furnace #7) yields an emissions limit that is lower than that calculated from the dataset of the unit (furnace #2) with the lowest average (71.7 versus 132.8 µg/dscm). For these reasons, we determined that the unit with the lowest average (furnace #2) is not the best performing source for this pollutant and we are instead selecting furnace #7 as the best performing source. After selecting the source upon which the new source limit would be based, we next considered whether the selection of a different confidence level would be appropriate. In this case, we determined that a lower confidence level was not warranted given the small amount of variability in the data for the unit that we identified as the best performer. Based on the factors outlined above, we are proposing that the MACT floor is 72 µg/dscm for PAHs from new furnaces producing SiMn.

The MACT floor dataset for mercury from existing and new furnaces producing FeMn includes 6 test runs from a single furnace. We first determined that the dataset is best represented by a normal distribution and ensured that we used the correct equation for the distribution. Because the floor for both existing and new furnaces is based on the performance of a single unit, our evaluation of the data was limited to ensuring that the emission limit is a reasonable estimate of the performance of the unit based on our knowledge about the process and controls. Accordingly, we compared the calculated emission limit to the highest measured value and the average short-term emissions from the unit, and found that the calculated emission limit is about 2.5 times the short-term average from the unit, which is within the range that we see when we evaluate larger data sets using our MACT floor calculation procedures. The fairly wide range in mercury emissions shown by the available data for this best performing unit indicate that variability is significant, and we determined that the emission limit is representative of the actual performance of the unit upon which the limit is based, considering variability. Therefore, we determined that no changes to our standard floor calculation procedure were warranted for this pollutant and subcategory, and we are proposing that the MACT floor is 170 µg/dscm for mercury from existing furnaces producing FeMn. We also note that while we calculated the same MACT floor value for new sources, we are proposing a beyond-the-floor standard for new sources, which is discussed later in this section of this preamble.

The MACT floor dataset for mercury from new furnaces producing SiMn includes 3 test runs from a single furnace (furnace #7 at Felman) that we identified as the best performing unit based on average emissions. After determining that the dataset is best represented by a normal distribution and ensuring that we used the correct equation for the distribution, we evaluated the variance of this unit. Our analysis showed that this unit, identified as the best unit based on average emissions, also had the lowest variance, indicating consistent performance. Therefore, we determined that the emission limit reasonably accounts for variability and that no changes to the standard floor calculation procedure were warranted for this pollutant and subcategory, and we are proposing that the MACT floor is 4.0 µg/dscm for Hg from new furnaces producing SiMn.

The MACT floor dataset for HCl from new furnaces producing FeMn or SiMn includes 6 test runs from a single furnace (furnace #5 at Felman) that we identified as the best performing unit based on average emissions. After determining that the dataset is best represented by a non-normal distribution and ensuring that we used the correct equation for the distribution, we evaluated the variance of this best performing unit. Our analysis showed that this unit, identified as the best unit based on average emission, also had the lowest variance, indicating consistent performance. Therefore, we determined that the emission limit reasonably accounts for variability and that no changes to the standard floor calculation procedure were warranted for this pollutant and subcategory. We also note that for this standard, the calculated new source floor level was below the level that can be accurately measured (the level that we refer to as "3 times the representative detection level" or 3xRDL). Therefore, we are proposing a new source MACT emission limit of 180 ppm, which is the 3xRDL value for HCl.