Appendix: Estimation of the factor for converting dust mass from revised inlets to equivalent dust mass from original inlets

Purpose of analysis

Regression analysis on pooled data was used to determine the coefficient by which coal dust mass measured with revised inlets should be multiplied to yield the equivalent mass measured with original inlets.

Ordinary least squares regression model

Standard regression analysis, in which values for the slope and intercept are found by the method of ordinary least squares (OLS), was performed first. The ordinary least squares solution is one in which the sum of squared residuals is minimized. Important statistics for the OLS regression model are shown in Table A1.

Table A1. Statistics for OLS regression model

R^2	SEE	Slope	Slope 95% LCL	Slope 95% UCL	Intercept	Intercept 95% LCL	Intercept 95% UCL
0.991	0.090	1.018	1.007	1.028	-0.008	-0.026	0.010

Checking assumptions of regression analysis for OLS model

Evidence was found that two assumptions of OLS regression, namely, normality of residuals and homogeneity of variance of residuals, were violated. Results of both the Kolmogorov-Smirnov and Shapiro-Wilk tests [NIST/SEMATECH, 2012, Sec. 1.3.5.16, 7.2.1.3] [Stephens, 1974] for normality were significant at $p < 0.001$, leading to rejection of the null hypothesis of normality (see Table A2). Visual evidence of non-normality was also found. Figure A1 displays a histogram of the residuals with a normal curve superimposed and Figure A2 displays a normal Q-Q plot. It can be seen in Figure A1 that the distribution is more peaked than the normal distribution and that values more extreme than those expected in a normal distribution are found in the lower and upper tails. These extreme values are also reflected in Figure A2, in which one negative standardized residual has a value less than -4 and one positive standardized residual has a value close to +4. In a normal distribution, less than 0.3% of the values are expected to fall beyond 3 standard deviations away from the mean [Zar, 1984].

Table A2. Tests of normality for OLS model

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	0.081	324	< 0.001	0.970	324	< 0.001

Results of the modified Levene test [Gastwirth, et al., 2009] [NIST/SEMATECH, 2012, Sec. 1.3.5.10] for constant variance of residuals were also significant at $p < 0.001$, leading to rejection of the null

hypothesis of homogeneity of variance. As was the case for non-normality, visual evidence of variance heterogeneity was observed. The plot of standardized residuals against standardized predicted values, shown in Figure A3, exhibits the well-known “megaphone” shape, suggesting that variance of residuals increases as predicted values increase. (The tendency for variance to increase proportionally as collected dust mass loadings increase is also supported by longstanding experience with coal dust sampling.)

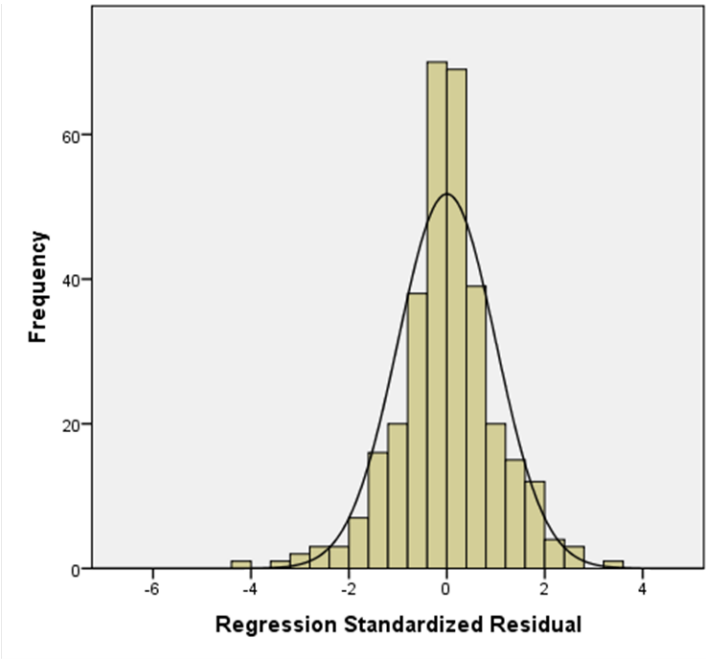

Figure A1. OLS regression: Histogram of standardized residuals.

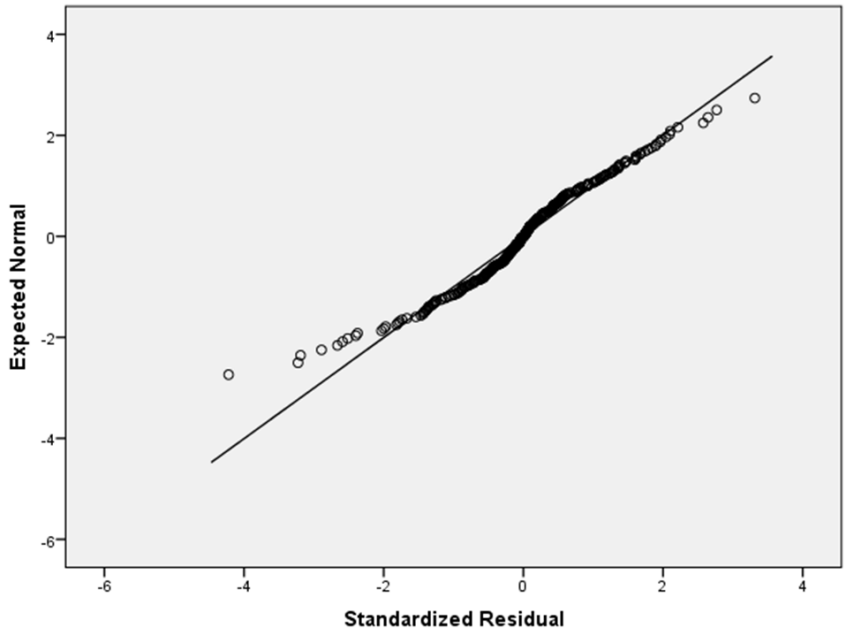

Figure A2. OLS regression: Normal Q-Q plot of standardized residuals.

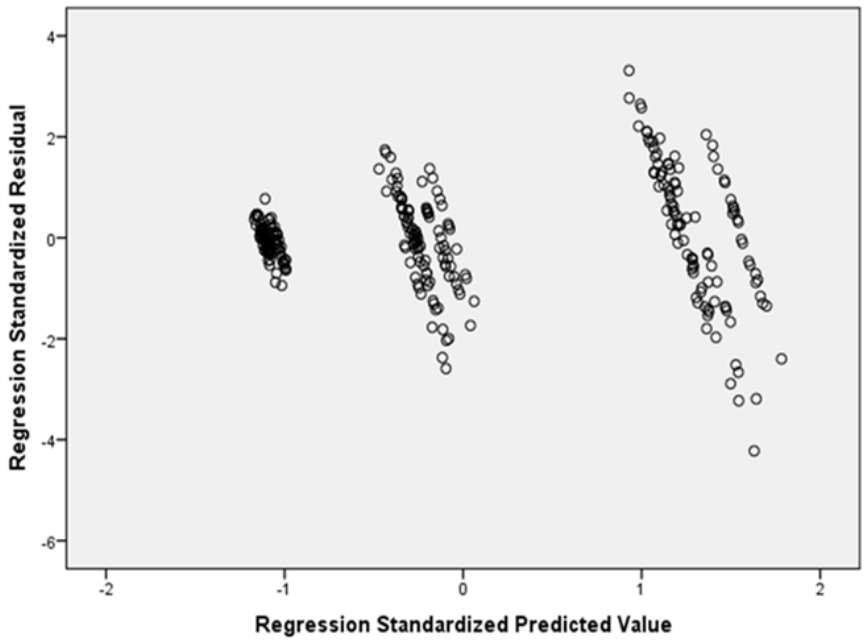

Figure A3. OLS regression: Plot of standardized residuals against predicted values.

The statistics literature indicates that violation of the normality assumption does not have serious consequences, provided that the homogeneity of variance assumption is met and the departure from normality is not extreme. However, violation of the homogeneity of variance assumption, either alone or in combination with violation of the normality assumption, can result in a non-optimal solution for values of the slope or intercept [Kleinbaum, et al., 2008] [Neter, et al., 1996].

Remediation of assumption violations

The two usual approaches to stabilizing variance are data transformation, or the use of weighted least squares (WLS) regression [NIST/SEMATECH, 2012, Sec. 4.1.4.3]. Both of these methods often remediate non-normality, as well as stabilize variance. The data transformation approach involves applying an arithmetic operation to observed values of either the dependent variable alone, or to observed values of both the independent and dependent variables. Taking the logarithm or the square root of values of one or both variables are commonly used transformations. Regression analysis is then carried out using the transformed values.

Whereas in OLS regression, the values of the slope and intercept are found through minimizing the sum of equally weighted squared residuals, in WLS these quantities are found through minimizing the sum of unequally weighted squared residuals. Weighting serves to give more influence to small residuals and less influence to large residuals. The optimal weighting can be approximated through an iterative process, but experience has shown that reciprocals of powers of X or Y often work well. The weighting is applied to the squared residuals and the regression model is then fit so that the sum of weighted squared residuals is minimized.

Comparison of models

The outcomes of regression models using logarithm and square root transformations of both variables, as well as the outcomes of weighed least squares regression using several weighting factors, were reviewed and compared. On the basis of the size of R^2, the size of the standard error of estimate, the range of 95% confidence intervals for the slope, the degree to which the assumptions of normality of residuals and homogeneity of variance appeared to be met, and the presence or absence of outliers, a weighted least squares model using the inverse of the original inlet mass loading as the weighting was selected as being the best model. An advantage of WLS over data transformation is that the need for reversing the transformation in order to get back to the original metric is eliminated.

Final weighted least squares regression model

Important statistics for the final WLS model are shown in Table A3, and the fitted regression line is also shown in Figure A4. In comparing Tables A1 and A3, it is observed that the extra care of employing WLS techniques had little impact on the final derived regression. Still, the next section of this report provides evidence that the WLS method remediated assumption violations and therefore resulted in more reliable estimates of the slope and intercept than those obtained by OLS.

Table A3. Statistics for WLS regression model weighted by inverse of original inlet loading

R^2	SEE	Slope	Slope 95% LCL	Slope 95% UCL	Intercept	Intercept 95% LCL	Intercept 95% UCL
0.993	0.068	1.018	1.008	1.027	-0.013*	-0.022	-0.003

*$p < 0.05$

It can be seen in Table A3 that R^2 was greater than 0.99, indicating that more than 99% of the variation in original inlet mass loadings can be explained by revised inlet loadings. The small, non-zero intercept is of no practical consequence, non-zero values being common in dust research, and -0.013 mg being less than 1% of the dust collected at the exposure limit of 1.5 mg/m^3 over a work shift. Based on the value of the slope, 1.018, it was determined that 1.02 would be used as the conversion (equivalency) factor. (Note that a 1.02 slope is within the 95% confidence interval; a 1.00 slope, inferring uncorrected inlet equivalence, is not.)

Testing assumptions of regression analysis for WLS model

Results of the Kolmogorov-Smirnov and Shapiro-Wilk tests for normality were mixed. Results of the Kolmogorov-Smirnov test were non-significant (see Table A4), whereas results of the Shapiro-Wilk test were significant at $p < 0.05$. (Recall that for the OLS model, results of both of the above tests were significant at $p < 0.001$.) The lack of agreement between the two tests was not surprising, given that the Shapiro-Wilk is known to be sensitive to even slight departures from normality when sample size is large. It is for this reason that the Shapiro-Wilk test is considered to be most appropriate for sample sizes of less than 50 [Kleinbaum, et al., 2008, p. 301]. Results of both normality tests for the WLS model provided evidence of closer approximation to normality. Visual evidence of closer approximation to normality was also observed. Figure A5 displays a histogram of standardized weighted residuals with a normal curve superimposed, Figure A6 displays the related normal Q-Q plot, and Figure A7 displays a plot of standardized weighted residuals against predicted values. When compared to the histogram

shown in Figure A1, the histogram shown in Figure A5 fits the normal curve more closely. Similarly, when compared to the plot shown in Figure A2, the plot shown in Figure A6 shows greater congruence between standardized weighted residuals and values expected in a normal distribution, especially at the middle and upper parts of the range.

Table A4. Tests of normality for WLS model weighted by inverse of original inlet loading

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Std. Weighted Residual	0.041	324	0.200*	0.989	324	0.013

*This is a lower bound of the true significance.

Results of the modified Levene test for constant variance of residuals were non-significant ($p > 0.05$). When considering visual evidence, the plot shown in Figure A7 demonstrates improved equality of variances when compared to the plot shown in Figure A3. Taken together, Table A4, and Figures A5, A6, and A7 suggest that it is reasonable and defensible to conclude that the assumptions of normality and homogeneity of variance were met for the WLS model.

Linearity of the relationship between the predictor and criterion variables is an additional assumption of regression analysis. The most common methods of assessing linearity are visual examination of a scatterplot, and visual examination of the plot of residuals against predicted values. No evidence of a curvilinear pattern was found in either of these plots (Figures A4 and A7). Furthermore, the fact that the value of R^2 is greater than 0.99 suggests that the addition of higher order terms, e.g., X^2 or X^3, would not significantly improve the fit of the model, nor could an assumption of a curvilinear relationship between the variables be justified.

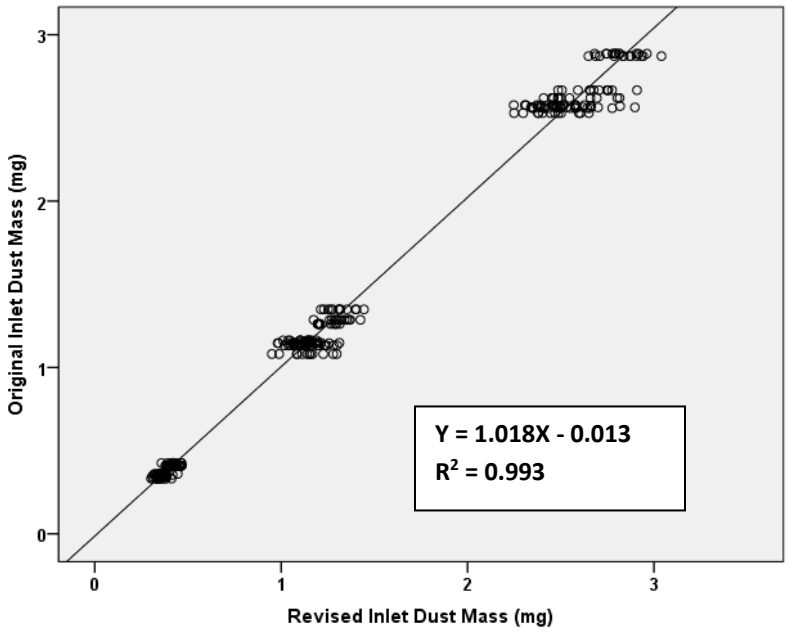

Figure A4. WLS regression (weighted by inverse of original inlet loading): Fitted regression line.

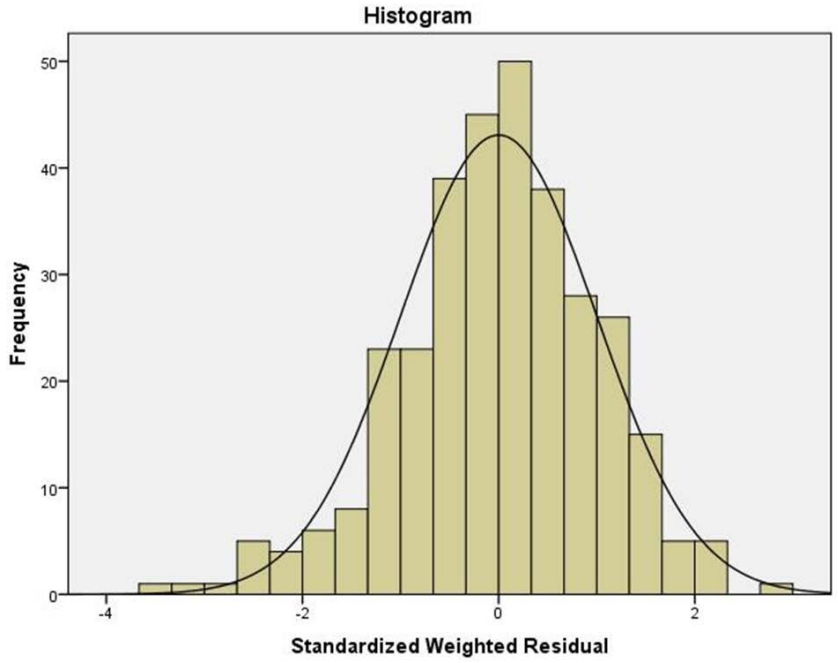

Figure A5. WLS regression (weighed by inverse of original inlet loading): Histogram of standardized weighted residuals.

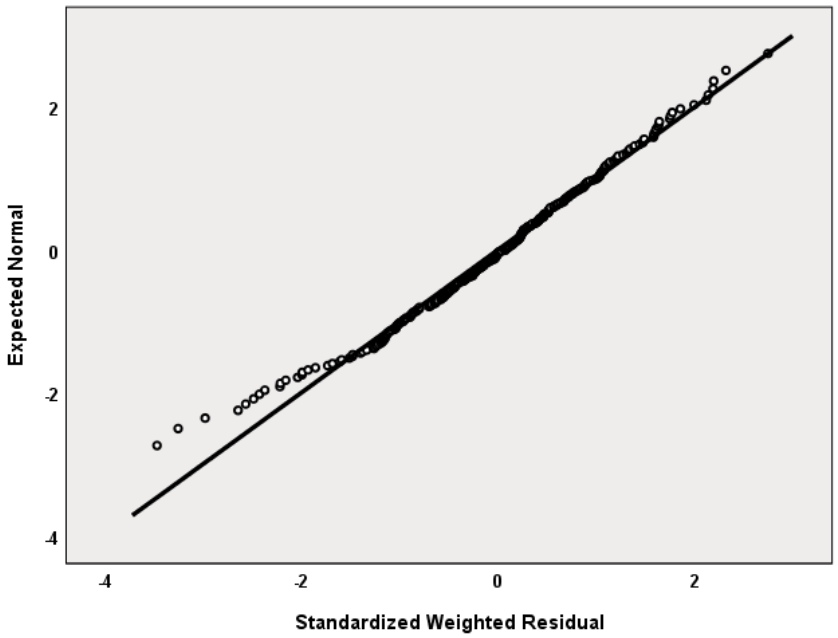

Figure A6. WLS regression (weighted by inverse of original inlet loading): Normal Q-Q plot of standardized weighted residuals.

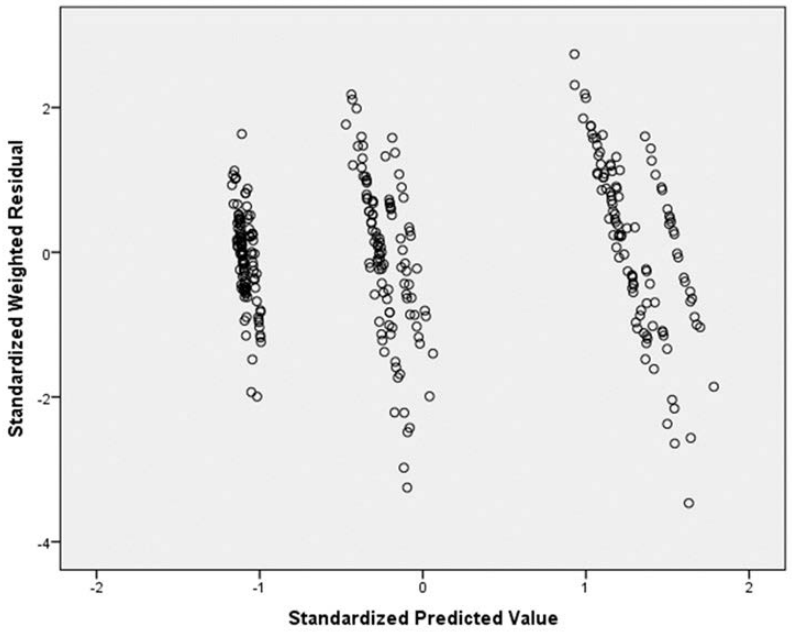

Figure A7. WLS regression (weighted by inverse of original inlet loading): Plot of standardized weighted residuals against predicted values.

Conclusion

As the focus of this appendix, the best statistical evidence from applying WLS regression methods supports a 1.02 factor in converting coal dust masses obtained from revised inlets to equivalent masses obtained from original inlets.

Endnotes

• All statistical calculations were performed with IBM SPSS Statistics for Windows, Version 22.0, Released 2013, IBM Corporation, Armonk, NY.

• The 1.5 mg/m^3 exposure limit mentioned above was lowered from 2.0 mg/m^3 on August 1, 2016.

• References, and the pooled data set for coal dust masses, are recorded on the following pages.

References

Gastwirth, JL, Gel, YR, and Miao, W [2009]. The impact of Levene's test of equality of variances on statistical theory and practice. Statistical Science, 24(3): 343-360.

Kleinbaum, DG, Kupper, LL, Nizam, A, Muller, KE [2008]. Applied regression analysis and other multivariable methods, 4th ed. Belmont, CA: Thomson, pp. 48, 299-305.

Neter, J, Kutner, M, Nachtsheim, C, Wasserman, W [1996]. Applied linear statistical models, 4th ed. Chicago, IL: Irwin, pp. 98-110, 112-114, 400-410.

NIST/SEMATECH [2012]. e-Handbook of statistical methods, Sec. 1.3.5.10 (Levene test), Sec. 1.3.5.16 (Kolmogorov-Smirnov test), Sec. 4.1.4.3 (WLS regression), Sec. 7.2.1.3 (Shapiro-Wilk test), [http://www.itl.nist.gov/div898/handbook//index.htm], accessed Oct. 25, 2016.

Stephens, MA [1974]. EDF Statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69(347): 730-737.

Zar, JH [1984]. Biostatistical analysis, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, p. 483.

Page 1, Pooled Data: Collected Coal Dust Masses (mg)

Revised Inlet Coal Dust Mass	Original Inlet Coal Dust Mean Mass	Revised Inlet Coal Dust Mass	Original Inlet Coal Dust Mean Mass	Revised Inlet Coal Dust Mass	Original Inlet Coal Dust Mean Mass
0.355	0.332	0.374	0.345	0.455	0.410
0.335	0.332	0.351	0.345	0.458	0.410
0.357	0.332	0.337	0.345	0.429	0.410
0.347	0.332	0.343	0.345	0.394	0.410
0.382	0.332	0.308	0.345	0.379	0.410
0.374	0.332	0.371	0.345	0.430	0.410
0.412	0.332	0.385	0.345	0.464	0.410
0.334	0.332	0.334	0.345	0.415	0.410
0.331	0.332	0.368	0.345	0.400	0.410
0.332	0.332	0.370	0.345	0.404	0.410
0.312	0.332	0.363	0.345	0.392	0.410
0.302	0.332	0.350	0.345	0.452	0.410
0.362	0.355	0.380	0.354	0.454	0.410
0.340	0.355	0.351	0.354	0.467	0.410
0.343	0.355	0.356	0.354	0.443	0.410
0.354	0.355	0.361	0.354	0.435	0.410
0.373	0.355	0.381	0.354	0.422	0.410
0.351	0.355	0.366	0.354	0.405	0.410
0.369	0.355	0.418	0.354	0.463	0.410
0.320	0.355	0.360	0.354	0.379	0.410
0.342	0.355	0.351	0.354	0.420	0.410
0.334	0.355	0.382	0.354	0.387	0.410
0.350	0.355	0.355	0.354	0.414	0.410
0.366	0.355	0.354	0.354	0.425	0.410
0.346	0.360	0.340	0.356	0.418	0.425
0.350	0.360	0.372	0.356	0.464	0.425
0.390	0.360	0.380	0.356	0.422	0.425
0.371	0.360	0.349	0.356	0.465	0.425
0.361	0.360	0.377	0.356	0.407	0.425
0.359	0.360	0.360	0.356	0.419	0.425
0.445	0.360	0.378	0.356	0.442	0.425
0.324	0.360	0.366	0.356	0.358	0.425
0.347	0.360	0.365	0.356	0.418	0.425
0.324	0.360	0.316	0.356	0.418	0.425
0.349	0.360	0.371	0.356	0.391	0.425
0.365	0.360	0.380	0.356	0.430	0.425

Data continues on next page.

Page 2, Continued Pooled Data: Collected Coal Dust Masses (mg)

Revised Inlet Coal Dust Mass	Original Inlet Coal Dust Mean Mass	Revised Inlet Coal Dust Mass	Original Inlet Coal Dust Mean Mass	Revised Inlet Coal Dust Mass	Original Inlet Coal Dust Mean Mass
1.298	1.081	1.235	1.132	1.282	1.261
1.087	1.081	1.101	1.132	1.204	1.261
1.226	1.081	1.139	1.132	1.297	1.261
0.989	1.081	1.039	1.132	1.195	1.261
1.139	1.081	1.082	1.132	1.265	1.261
1.083	1.081	1.131	1.132	1.211	1.261
1.157	1.081	1.245	1.132	1.314	1.261
0.950	1.081	1.082	1.132	1.199	1.261
1.151	1.081	1.129	1.132	1.198	1.261
1.113	1.081	1.018	1.132	1.205	1.261
1.168	1.081	1.070	1.132	1.202	1.261
1.279	1.081	1.112	1.132	1.293	1.261
1.231	1.133	1.213	1.148	1.308	1.287
1.195	1.133	1.067	1.148	1.352	1.287
1.300	1.133	1.198	1.148	1.322	1.287
1.163	1.133	1.101	1.148	1.363	1.287
1.094	1.133	1.152	1.148	1.340	1.287
1.204	1.133	1.102	1.148	1.294	1.287
1.281	1.133	1.312	1.148	1.425	1.287
1.095	1.133	1.123	1.148	1.174	1.287
1.067	1.133	1.119	1.148	1.286	1.287
1.070	1.133	1.068	1.148	1.272	1.287
1.049	1.133	1.141	1.148	1.260	1.287
1.154	1.133	1.198	1.148	1.370	1.287
1.153	1.146	1.104	1.163	1.398	1.349
1.062	1.146	1.190	1.163	1.309	1.349
1.183	1.146	1.153	1.163	1.404	1.349
1.068	1.146	1.143	1.163	1.213	1.349
1.143	1.146	1.102	1.163	1.277	1.349
1.129	1.146	1.102	1.163	1.229	1.349
1.257	1.146	1.148	1.163	1.444	1.349
0.986	1.146	1.038	1.163	1.252	1.349
1.138	1.146	1.165	1.163	1.313	1.349
0.981	1.146	1.010	1.163	1.318	1.349
1.131	1.146	1.047	1.163	1.266	1.349
1.044	1.146	1.139	1.163	1.353	1.349

Data continues on next page.

Page 3, Continued Pooled Data: Collected Coal Dust Masses (mg)

Revised Inlet Coal Dust Mass	Original Inlet Coal Dust Mean Mass	Revised Inlet Coal Dust Mass	Original Inlet Coal Dust Mean Mass	Revised Inlet Coal Dust Mass	Original Inlet Coal Dust Mean Mass
2.606	2.529	2.576	2.559	2.705	2.667
2.502	2.529	2.551	2.559	2.755	2.667
2.597	2.529	2.642	2.559	2.775	2.667
2.445	2.529	2.412	2.559	2.655	2.667
2.380	2.529	2.657	2.559	2.748	2.667
2.487	2.529	2.582	2.559	2.658	2.667
2.651	2.529	2.776	2.559	2.909	2.667
2.298	2.529	2.355	2.559	2.486	2.667
2.469	2.529	2.349	2.559	2.750	2.667
2.249	2.529	2.429	2.559	2.506	2.667
2.403	2.529	2.413	2.559	2.592	2.667
2.378	2.529	2.499	2.559	2.677	2.667
2.700	2.563	2.818	2.571	2.786	2.871
2.479	2.563	2.581	2.571	2.943	2.871
2.398	2.563	2.584	2.571	3.040	2.871
2.467	2.563	2.659	2.571	2.687	2.871
2.342	2.563	2.512	2.571	2.876	2.871
2.502	2.563	2.377	2.571	2.832	2.871
2.898	2.563	2.663	2.571	2.870	2.871
2.483	2.563	2.464	2.571	2.709	2.871
2.531	2.563	2.419	2.571	2.931	2.871
2.341	2.563	2.479	2.571	2.649	2.871
2.621	2.563	2.459	2.571	2.838	2.871
2.385	2.563	2.495	2.571	2.907	2.871
2.693	2.620	2.578	2.577	2.812	2.886
2.547	2.620	2.575	2.577	2.906	2.886
2.454	2.620	2.627	2.577	2.917	2.886
2.452	2.620	2.516	2.577	2.792	2.886
2.803	2.620	2.313	2.577	2.788	2.886
2.500	2.620	2.458	2.577	2.777	2.886
2.816	2.620	2.452	2.577	2.962	2.886
2.485	2.620	2.307	2.577	2.682	2.886
2.462	2.620	2.393	2.577	2.797	2.886
2.408	2.620	2.577	2.577	2.743	2.886
2.488	2.620	2.248	2.577	2.747	2.886
2.659	2.620	2.372	2.577	2.816	2.886

End of data set.