TETRA TECH, INC.
4401 Building, Suite 200
79 T.W. Alexander Drive
P.O. Box 14409
Research Triangle Park, NC 27709
Telephone: (919) 485-8278 Telefax: (919) 485-8280

MEMORANDUM

To:	Claire Hunt Ed Garvey (TAMS/NJ)	**Date:** January 19, 1998
From:	J. B. Butcher	**Project:** Hudson
Subject:	Low Res vs High Res MDPR/DMW stats	**Pjn:** 1182-06

NUMBER OF PAGES: 8 Original will follow as e-mail attachment

Purpose

This memo is based on a reanalysis of electronic data and analyses provided by Claire Hunt, 1/13/98. The general purpose of these analyses is to compare molar dechlorination product rations (MDPR) and change in molecular weight (DMW) between the low resolution and high resolution sediment data sets, and look for consistency, or lack of consistency in the relationships. There are two related types of questions we can ask: (1) Are the low resolution results consistent with the high resolution results, and do the low resolution results confirm the statistical models derived from high resolution results?; and (2) Are the two sets of data drawn from the same population?

The analyses were completed using the spreadsheets supplied by you. It was, however, necessary to make several corrections. First, in the files PCBDMW and PCBMDPR the predicted value use the wrong independent variable column. Second, prediction confidence limits are not seem correctly calculated. Finally, the spreadsheet for Theil's U uses rounded off values of the coefficients, instead of linking them directly to the regression. The result is somewhat sensitive to use of exact values, and these should be linked. I will send electronic copies of the spreadsheets back to you.

Three relationships were tested: DMW as a function of MDPR, MDPR as a function of PCBs, and DMW as a function of PCBs. For each of these relationships, five different statistical tests were applied in a weight-of-evidence approach.

DMW Predicted from MDPR

DMW is predicted as a linear model of MDPR. It should be noted that there is likely a correlation between disturbance terms for DMW and errors in calculating MDPR, as both are derived from analysis of underlying data and may be strongly affected by changes in quantitation of a few congeners. In this type of stochastic regressors problem it is likely that standard statistics will over-estimate the strength of the relationship, and model coefficients are likely to be biased.

Coefficient estimates are given below

	High Resolution	Low Resolution	Pooled Data
Intercept	-0.052	-0.027	-0.044
Slope	0.282	0.235	0.263

1. *Calculate an R^2 predicting Low Res results with High Res model and compare to High Res R^2*

This test applies the High Resolution fitted model coefficients to predict Low Resolution results, i.e., DMW from concentration. Quality of fit is summarized by the R^2 value, which represents the percent of total variation explained by the regression model. To test Low Resolution results with the High Resolution coefficients we manually computes an R^2 for the Low Resolution application as 1 - SSE/SST. If this R^2 is not significantly below the R^2 obtained in the original High Resolution regression, the specification is said to be satisfactory.

The R^2 from the High Resolution regression is 93.6, indicating a very tight fit, while the R^2 from a direct analysis of the Low Resolution data is 94.9. When the High Resolution coefficients are applied to the Low Resolution data, a calculated R^2 of 91.0 is obtained. Thus, in this case, the High Resolution and Low Resolution interpretations appear to be consistent in terms of ability to predict the data (whether or not the coefficients are significantly different).

2. *Proportion within confidence bounds*

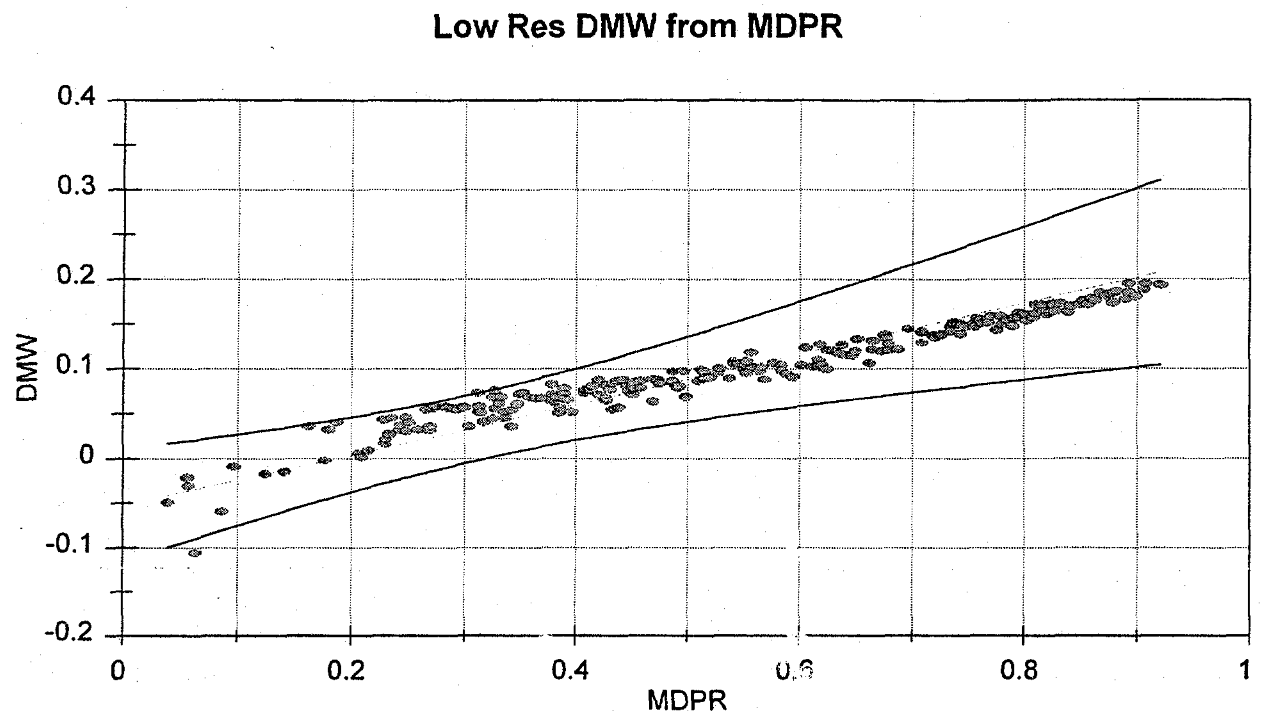
This is an informal, qualitative test, in which we examine the proportion of Low Resolution data which fall within the 95% confidence bounds on the High Resolution model. The confidence bounds we want here are those for the prediction interval, the error in predicting specific realizations, from the High Resolution regression. The confidence interval for an individual point prediction, y_0, is given by

$$y_0 \pm \frac{t_{n-2}}{\sqrt{n-2}} s_{yx} \sqrt{n + 1 + n(x_0 - \bar{x})^2 / s_x^2}$$

where t_{n-2} is approximately 1.96 for 95% confidence intervals and large sample size (Normal approximation), and $s_{y.x}$, n, x, and s_x are all derived from the High Resolution regression, with

$$s_{yx} = \sqrt{\sum (y_{obs} - y_{est})^2 / n}$$

The position of the low resolution data points within the high resolution prediction limits is shown below. 240 out of 242 points (99.2%) fall within the prediction limits.


Low Res DMW from MDPR

3. Regress predicted on observed Low Resolution values and test coefficients

For this test we want Low Resolution values predicted with the High Resolution regression coefficients. We can then test whether the intercept is significantly different from 0 and the slope significantly different from 1. The graph presented above suggests there is a slight bias in the relationship of the Low Resolution data to the High Resolution regression line, with most points falling above the regression line in the region of MDPR of 0.2 to 0.5. A regression of Low Resolution predictions (with the High Resolution model) on Low Resolution observations provides a good fit, with R^2 of 94.9; however, the intercept of -0.0137 is statistically different from 0 at the 95% confidence level (confidence interval -0.0175- -0.0098) and the slope of 1.140 is statistically different from 1 (confidence interval 1.106–1.173). These slight deviations may reflect the fact that the original High Resolution regression is slightly biased because of stochastic regressors.

4. Theil's U Statistic

Theil's U statistic gives a measure of the consistency between forecasts (e.g., Low Resolution predictions using the High Resolution model) and the data used to develop the forecasts. It ranges from 0 to 1, with 0 indicating perfect prediction, and it's variance can be approximated (for U less than 0.3) as U^2/T, where T is the number of samples in the "forecast".

The U statistic may also be decomposed into portions attributed to bias (U^m), variance (U^s), and covariance (U^c). When U is non-zero, we would ideally like the decomposition to show that

U	0.07
Var (U)	1.9 E-5
Lower 95% Confidence	0.06
U^m	0
U^s	0.32
U^c	0.68

the difference is entirely attributable to the covariance component, which represents non-controllable random variability. Weight on the bias component indicates that the linear relationship differs between the two data sets. Weight on the variance component indicates that the difference is attributable primarily to differing variances between the two data sets.

For the DMW on MDPR regression, results are shown in the box. The U statistic is close to, but significantly different from zero, indicating the presence of some difference between the data sets. The decomposition shows that there is minimal bias. A significant portion (one third) of the difference, however, is attributable to the variance component. This indicates that the decline in fit is partly due to differing variance between the Low Resolution and High Resolution data.

5. Chow's F Test

Chow's F test addresses the hypotheses that the parameters have or have not changed between the two data sets. It is developed by calculating SSEs for regressions on each of the data sets individually and an SSE for a regression on the pooled data. The comparison is made by forming an *F* statistic with *k* and t_1+t_2-2k degrees of freedom, formed as

$$F = \frac{[SSE(constrained) - SSE(unconstrained)]/k}{SSE(unconstrained)/(t_1 + t_2 - 2k)}$$

in which *SSE(unconstrained)* is the sum of the *SSE*s from the two separate regressions, *SSE(constrained)* is the *SSE* from the regression on the pooled data, t_1 is the number of observations in the first sample set, t_2 is the number of observations inn the second sample set, and *k* is the number of parameters. The resulting statistic can then be compared to a tabulation of the *F* distribution.

In this case, $k = 2$, because both a slope and an intercept are calculated. The resulting *F* statistic (calculated after correcting the spreadsheet), is 28.1, which has a probability value under the null hypothesis of no change in parameters of 3.33×10^{-12}. In other words, the *F* test indicates that the change in parameter values between the High Resolution and Low Resolution regressions is statistically significant, matching the results seen in item 3.

MDPR Predicted from PCB Concentration

Change in molecular dechlorination product ratio (MDPR) is predicted via a linear model from log of total PCBs.

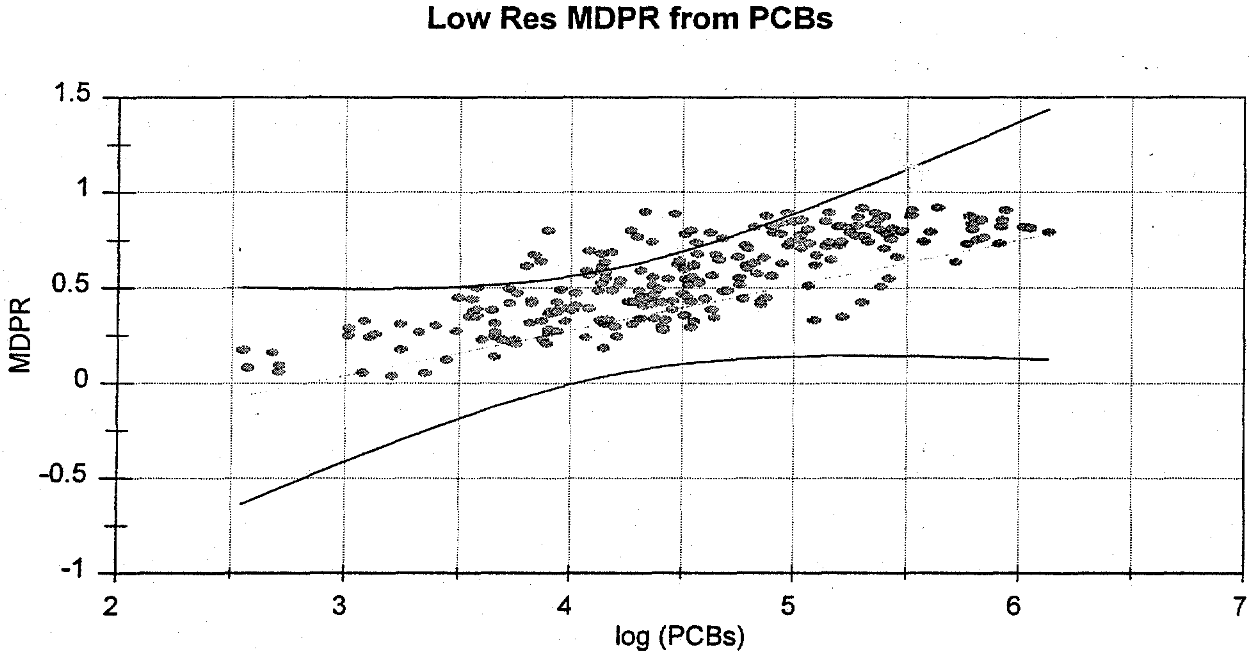
	High Resolution	Low Resolution	Pooled Data
Intercept	-0.670	-0.515	-0.661
Slope	0.237	0.236	0.254

1. Calculate an R^2 predicting Low Res results with High Res model and compare to High Res R^2

The R^2 from the High Resolution model is 70.0, while a direct regression on Low Resolution results provides an R^2 of 62.8. Applying the High Resolution coefficients to Low Resolution data, however, results in a calculated R^2 of 16.6, showing a dramatic reduction in explanatory power. This indicates that the High Resolution coefficients constitute a substantially biased model when applied to the Low Resolution data.

2. Proportion within confidence bounds

219 out of 242 Low Resolution Points fall within the High Resolution 95% prediction confidence intervals (90.5%). The percentage is relatively large primarily because the confidence bounds are wide, as shown in the following figure. Note that the High Resolution and Low Resolution data sets have approximately the same slope against log PCBs, but the Low Resolution dataset is displaced upward (has a lower intercept value), such that most points fall well above the High Resolution regression line.



3. Regress predicted on observed Low Resolution values and test coefficients

U	0.20
Var (U)	1.6 E-4
Lower 95% Confidence	0.17
U^m	0.55
U^s	0.05
U^c	0.40

The R^2 for a regression of Low Resolution predictions (using the High Resolution coefficients) on Low Resolution observations has a relatively low value of 75.4%. Both the intercept (0.0469; 95% confidence interval 0.006-0.088) and slope (0.642; 95% confidence interval 0.563-0.721) are significantly different from target values of 0 and 1.

4. Theil's U Statistic

The U statistic of 0.20 is much higher than for the DMW-MDPR regression, and is significantly different from zero. More importantly, the decomposition of the U statistic indicates more than 50% of its value is attributable to bias, which reflects the fact that the High Resolution and Low Resolution models appear to have similar slopes, but different intercepts.. This suggests there may be some systematic difference in the data sets.

5. *Chow's F Test*

The *F* statistic is 48.7, with a probability value of 8 x 10^{-20}. This result also supports the conclusion of a significant difference between the two regression models.

DMW Predicted from PCB Concentration

DMW is predicted as a linear model of the log of total PCB concentration. For this model, like the previous, the slopes are similar but the intercepts differ between the High Resolution and Low Resolution regressions.

	High Resolution	Low Resolution	Pooled Data
Intercept	-0.242	-0.156	-0.221
Slope	0.067	0.057	0.067

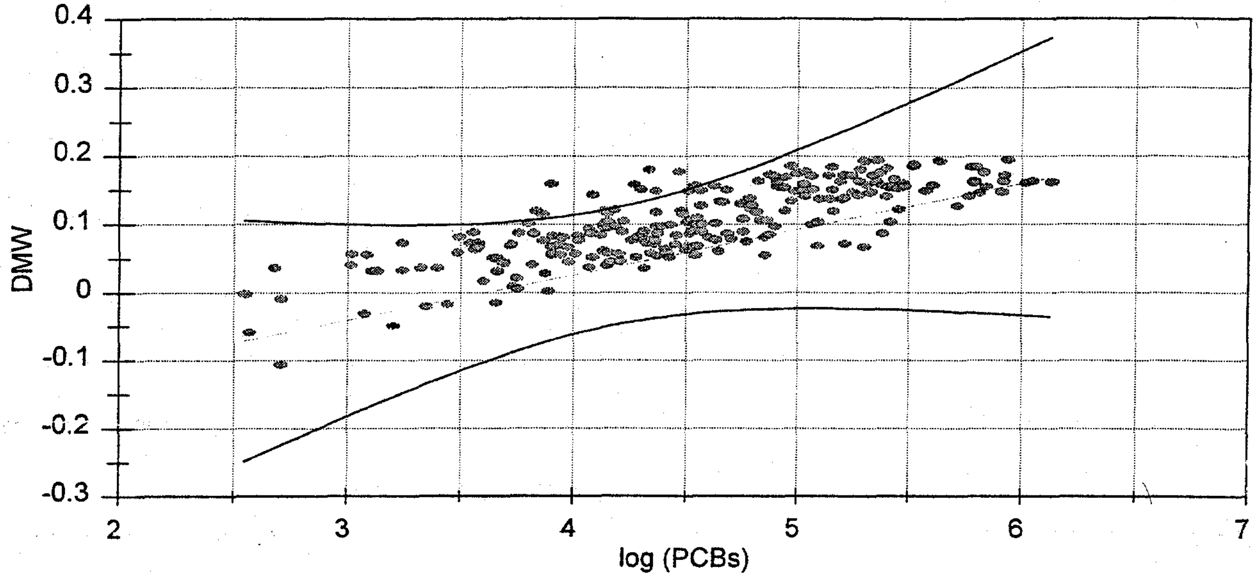
1. *Calculate an R^2 predicting Low Res results with High Res model and compare to High Res R^2*

For the High Resolution regression the R^2 is 65.8, while a direct regression on Low Resolution results provides an R^2 of 63.3. Applying the High Resolution coefficients to Low Resolution data, however, results in a calculated R^2 of 23.9, representing a substantial fall off and suggesting that there are inconsistencies between the High Resolution and Low Resolution results.

2. *Proportion within confidence bounds*

232 of 243 individual Low Resolution observations (95.5%) fall within predictive confidence intervals of the High Resolution regression. As with the previous model, the confidence bands are broad because the predictive power of the model is not particularly strong (see figure next page).

Low Res DMW from PCBs



3. Regress predicted on observed Low Resolution values and test coefficients

The R^2 for a regression of Low Resolution predictions (using the High Resolution coefficients) on Low Resolution observations has a relatively low value of 63.3%. Both the intercept (-0.016; 95% confidence interval -0.024- -0.007) and slope (0.741; 95% confidence interval 0.670-0.813) are significantly different from target values of 0 and 1.

4. Theil's U Statistic

The U statistic is 0.28, suggesting a substantial inconsistency between the two data sets. Decomposition of the U statistic suggests that it is predominantly attributable to bias. In other words, the High Resolution regression does not provide a very good representation of the Low Resolution data.

U	0.28
Var (U)	3.2 E-4
Lower 95% Confidence	0.24
U^m	0.61
U^s	0
U^c	0.38

5. Chow's F Test

As expected, the Chow *F* test also indicates inconsistency. The *F* statistic is 53.6, with a probability value of 1.5 x 10^{-21}.

Discussion

All three relationships show some discrepancy between the High Resolution and Low Resolution data sets. For predicting DMW from MDPR, the discrepancy is small. The discrepancy which does arise is likely due to slightly biased coefficient estimates from the High Resolution model due to the stochastic regressors problem. I recommend combining the High Resolution and Low Resolution data sets for a model relating DMW to MDPR, which should help reduce the stochastic regressors problem through larger data count.

Predictions of either DMW or MDPR from log PCBs show similar discrepancies between the High Resolution and Low Resolution data sets. In both cases, the slope of the relationship shows little change between High Resolution and Low Resolution data, but the intercept is different. As a result, the High Resolution models are biased low with respect to the Low Resolution data. This suggests the data sets should not be combined without adjustment. Why do the relationships differ in intercept? The most parsimonious explanation would be that less total PCBs relative to key congener components of DMW and MDPR were recovered or reported in the High Resolution data than in the Low Resolution data, or, alternatively, consistently greater amounts of dechlorination products relative to total PCBs were obtained in the Low Resolution data. One less attractive possibility is a change in analytical methods/results between analysis of the High Resolution and Low Resolution data. I wonder, however, if there might be another simple solution. I note that the Low Resolution data were all collected in July and August. The High Resolution cores were collected throughout the year. The sediment temperature, and biological activity, should thus be higher, on average, in the Low Resolution data set. This in turn could lead to increased production, and transiently increased concentration, of the most mobile dechlorination products. The issue could be investigated by looking at the relationship of DMW to log total PCBs in High Resolution cores stratified into summer and winter samples.