TETRA TECH, INC.
3200 Chapel Hill - Nelson Highway
Cape Fear Building - Suite 105
P.O. Box 14409
Research Triangle Park, NC 27709
Telephone: (919) 485-8278 Telefax: (919) 485-8280

MEMORANDUM

Date: 7/26/04

To: J. Todd Kennedy (NC DWQ)

From: Jon Butcher

Subject: Jordan Lake Nutrient Response Model Uncertainty

On July 1, 2004, CH2M HILL produced a document for the Jordan Lake Project Partners entitled "Jordan Lake TMDL Implementation Issues." This report emphasizes model error and uncertainty, and includes as Attachment 1 a memorandum entitled "Statistical Analysis of the Jordan Lake Model." The Project Partners propose using the characterization of model uncertainty as a justification for a phased approach to the TMDL and NSW Strategy.

ANALYTICAL UNCERTAINTY

Evaluation of uncertainty in the model is complicated by the presence of considerable analytical uncertainty in chlorophyll *a* measurements. The CH2M HILL report cites the *Jordan Lake Nutrient Response Model* report (Tetra Tech, 2002) indicating that the chlorophyll *a* fluorometric analytical methodology results in 95 percent confidence intervals of plus or minus 40-50 percent about the observed value. This statement is based on a DWQ memo dated February 26, 200, which states the following referring to a multi-laboratory test of the analytical methods for chl *a*: "Using the HPLC method values as a true value, the percent relative standard deviation (%RSD) was most frequently in the 20-25 percent range." Unfortunately, this statement is not entirely accurate. A discussion of this issue follows.

EPA commissioned a multilaboratory validation and comparison study in 1996 (discussed in the EPA Methods 445 and 447, September 1997). The primary goals of the study were to determine estimated detection limits, assess precision as %RSD, and assess bias or accuracy as percent recovery. The study compared spectrophotometric, fluorometric, and HPLC methods. The average %RSD for the fluorometric methods was approximately 23 percent, ranging from 15 to 33 percent. Median percent recoveries for the fluorometric methods ranged from 104 to 296 percent over all concentrations and species tested. HPLC values ranged from 80 to 252 percent. There was no significant trend by method across test species.

Percent RSD in the EPA multilaboratory study is a measure of *precision, not accuracy.* Precision is the ability of a measurement to be reproduced. Accuracy is the ability of a measurement to match the actual or real value. No comparison to known "true" values was made by EPA. A comparison to HPLC-generated values was made in the study's percent recovery calculation, but the HPLC method itself is uncertain, with %RSD values ranging from 15 to 57 percent. Therefore, the follow on calculation of confidence intervals using the %RSD values is not entirely appropriate. However, the precision of the fluorometric method does provide a lower bound on the expected variability in environmental analyses related to the true chlorophyll concentration.

In sum, there is considerable uncertainty associated with all test methods for chlorophyll *a*. Because of this uncertainty, it is expected that the model and individual observations will frequently differ, even in the absence of model uncertainty. However, the model should be able to capture spatial and temporal trends in observed concentrations.

CHARACTERIZATION OF MODEL UNCERTAINTY

The discussion provided byCH2M HILL is largely based on an evaluation of Relative Absolute Error (RAE), the Nash-Sutcliffe coefficient, and the Root Mean Squared Error (RMSE). Each of these measures is based on the difference between paired observations and model predictions.

We agree that there is considerable uncertainty (in data, modeling tools, and scientific understanding) in the linkage analysis connecting specific levels of nutrient loads to predicted frequency of chlorophyll *a* concentrations greater than 40 µg/L. We feel, however, that focusing exclusively on uncertainty in model predictions of point-in-time/point-in-space chlorophyll *a* observations is not appropriate.

Evaluations based on deviations between paired observations and model output do reflect uncertainty and imprecision in model fit. However, they also reflect data uncertainty and temporal and spatial mismatches between the model and observations. Because the impaired segments of Jordan Lake are highly dynamic environments, we expect significant spatial and temporal variability in algal concentrations. A grab sample obtained at a point in space and time may not be representative of conditions in the segment as a whole, and significant variability may occur from day to day. The fine-scale temporal resolution in the model is also intrinsically limited by the sparseness of the data available on tributary loads and light penetration in the water column. Thus, the model could provide an accurate representation of segment-average conditions in week 25, but still yield a result very different from a single observation collected at a single point during the week.

In general, the calibration strategy for the model was to capture broad spatial trends and fit multiple parameters simultaneously. The relationships between concentrations of multiple parameters at multiple stations are more significant than the fit to individual points at individual stations.

Use of the RAE appears particularly inappropriate, for the following reasons:

- A high value of RAE may reflect a small error if based on a small observed concentration. For instance, the point cited as having an RAE of 300 percent is associated with the lowest observation on record (9 µg/L) in lumped segments 14-15.
- Observed data are subject to a high level of uncertainty, due both to issues of sample representativeness of the volume and averaging period of a modeling segment and analytical issues. Because the measurements are imprecise, it is less meaningful to look at relative error (dividing one uncertain number by another) than to examine the magnitude of the discrepancy between observations and predictions.
- RAE is sensitive to the presence of a few large values, although this can be mitigated to some extent by summarizing RAE by the median, rather than mean, as recommended by Thomann (1982)[1]. The median RAE for Segments 14-15 (excluding 2000 results for comparability to CH2M HILL's calculation) is 55 percent, compared to the average RAE of 66 percent, while the median RAE for TN is 14 percent, half the average RAE of 28 percent.

[1] Thomann, R.V. 1982. Verification of water quality models. *J. Envir. Eng. Div., ASCE*, 108 (EE5): 923-939.

The summary of model error and uncertainty on pages 5-6 of the CH2M HILL report contains a number of misleading statements, which should be corrected:

- The second paragraph of the section states, "If data error and uncertainty are high, it is difficult to make decisions concerning the model calibration parameters. Modifying inputs to match data cannot be justified in these circumstances." This misrepresents the calibration strategy. The aim in calibration was to achieve a simultaneous fit to multiple parameters (nutrients, chlorophyll *a*) at multiple stations, which constrains the calibration process. Moreover, the fit is aimed at reproducing the central tendency of trends in time and the approximate frequency distribution, rather than replicating individual observations of chlorophyll *a*. As such, calibration of the model is appropriate.
- The third paragraph states "model output should be viewed as a range of potential values based on their probability density functions rather than as a precise single output number." We concur fully with this statement, and contend that the model provides a reasonable representation of the expected distribution of concentrations, but not necessarily individual point-in-time concentrations. Unfortunately, the error statistics developed by CH2M HILL are all focused on evaluation of deviations between observations and predictions for individual point-in-time/point-in-space. This statement thus contradicts the uncertainty analysis presented in Attachment 1 of the memorandum.
- Paragraph five admits, "that the model predicts chlorophyll *a* fairly well on average", but then notes that "DWQ staff has proposed that average predictions not be used in the development of the TMDL and nutrient targets." The second statement does not follow logically from the first. The fact that the model predicts average concentrations well is one component in the prediction of the distribution of concentrations. In fact, the model also does a good job of predicting frequencies of concentrations around the criterion value, as explained further below. What DWQ has stated is that the TMDL must be based on the frequency of concentrations greater than 40 μg/L, rather than the long-term average concentration.

COMPARISON TO USGS OBSERVATIONS

The CH2M HILL report discusses USGS water quality observations collected as part of the Triangle Area Water Supply Monitoring Project as another potential source of information for evaluation of the model. The discussion is correct in stating that these data were not directly used during model calibration. There were two reasons for this: First, there were concerns about method comparability, and second the USGS data are quite sparse, with only two summer measurements in most years.

Long runs of USGS monitoring data are available at three locations: Buoy 12 (0209687310, just above SR 1308 in model segment 4), above Highway 64 at Wilsonville (0209799150, in segment 8), and near the dam (0209719700, in segment 14). Data through August 2003 were downloaded from the USGS web site.

On page 11 of the CH2M HILL report, the statement is made "Like the DWQ data, the majority of the USGS data were collected during the growing season. Despite this limitation, the only station that exceeded the threshold of 10 percent violations for evaluating use impairment is at Buoy 12, in the New Hope Creek arm of the lake. Chlorophyll *a* levels at that station exceeded the State standard for less than 15 percent of the samples."

This statement is misleading on a number of grounds. First, the USGS data are not strongly focused toward the growing season (May-September). For instance, at Buoy 12, only 58 percent of the samples are from the growing season. Second, it is not surprising - and fully consistent with the model – that the 10 percent threshold is exceeded only at Buoy 12 over the period of record. The other USGS stations are within portions of the lake where lower concentrations are both predicted and observed.

Over the period of record (1992-2003), 15.7 percent of the total USGS observations at Buoy 12 were greater than 40 µg/L, while 23.6 percent of the growing season observations were greater than 40 µg/L (using interpolated estimates with the Excel PERCENTRANK function). At the station above the dam over the period of record for the HPLC method (1991-2003), 6.6 percent of the total observations were greater than 40 µg/L, while 4.3 percent of the summer observations were greater than 40 µg/L. For 1997-2001 only, 11.9 percent of the USGS observations at this station were greater than 40 µg/L.

Page 12 of the CH2M HILL report states that the USGS "data contrast very sharply with data and modeling results used for the model calibration and Jordan Lake analysis." This is simply not true, and is apparently contradicted in the next paragraph, where it is stated, "The modeling results are more similar to the USGS chlorophyll *a* data."

Based on the CH2M HILL comments, model results were compared to the USGS monitoring. The graphical comparison is shown in Figure 1. This figure uses the same format as the model calibration report, and corrects *observed* values for depth support as in the calibration work. Visually, the USGS observations are in general agreement with both the DWQ observations and model predictions.

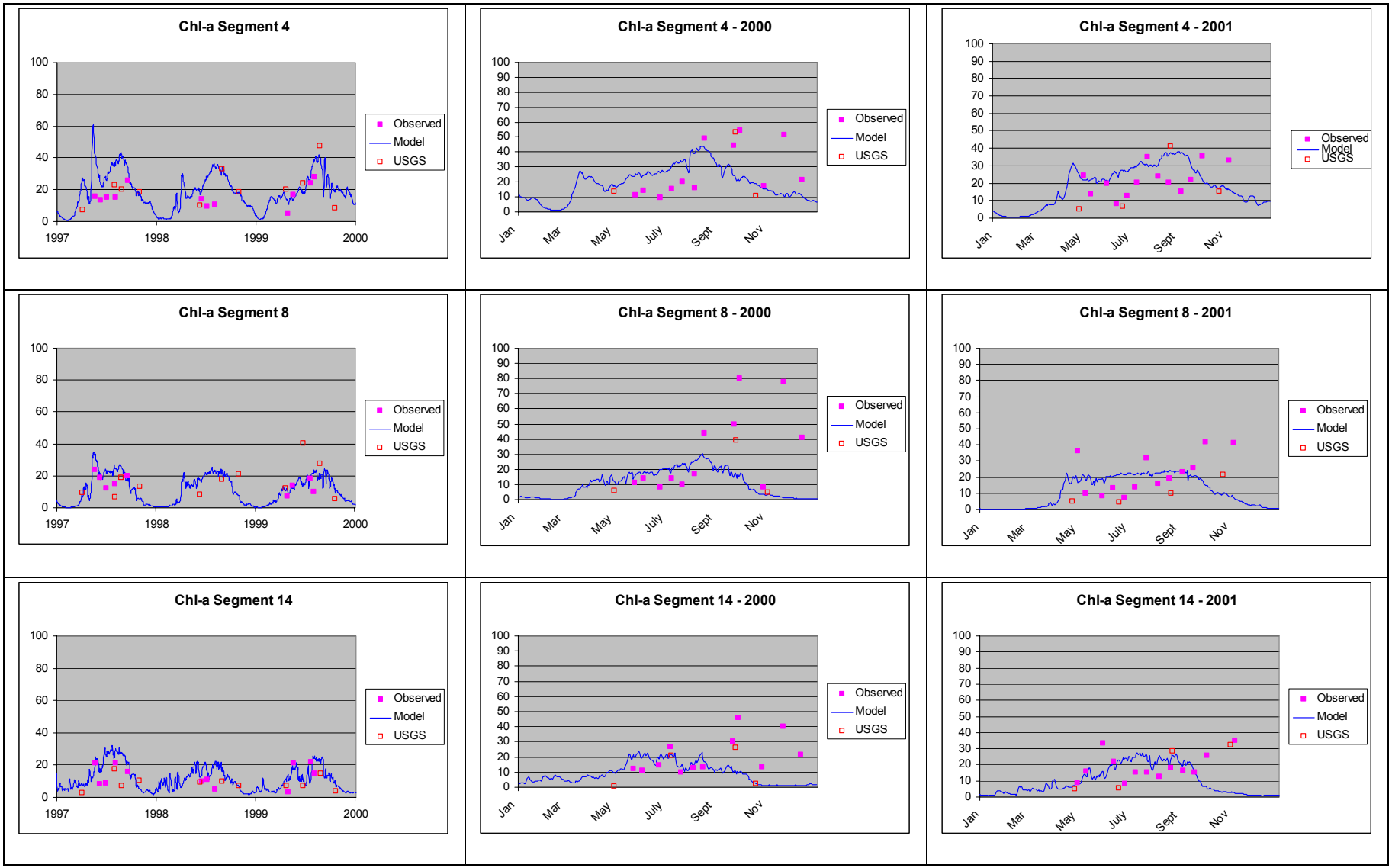

Figure 1. Comparison of Model Output to DWQ Observed and USGS Chlorophyll *a* (µg/L)

Observed values scaled to reflect model depth support in Segments 4 and 14, as in model calibration report.

Error statistics can be calculated for paired model predictions versus observations for the USGS data (which were collected on different days than the DWQ data), as shown in Table 1. These data are presented as measured - that is, the model results are corrected as needed for depth support and the observed data are not transformed. Average absolute error and root mean square errors are similar for paired observations regardless of whether the model is evaluated versus DWQ or USGS data. The USGS observations can thus be construed as providing further validation of the model.

Table 1. Chlorophyll *a* Model Fit Statistics for Paired Observations

	Seg 4 DWQ	Seg 4 USGS	Seg 8 DWQ	Seg 8 USGS	Seg 14 DWQ	Seg 14 USGS
AvSim	25.4	29.8	18.0	15.1	17.1	21.3
AvObs	21.8	24.9	19.4	14.4	16.2	17.9
GMSim	21.3	28.2	12.7	13.3	14.2	17.0
GMObs	17.3	20.1	16.9	11.1	14.2	12.5
AvErr	8.4	4.8	0.2	0.1	2.2	3.1
AvAbsErr	11.7	12.2	9.5	10.5	9.1	12.3
RMSE	13.9	15.9	12.8	12.6	11.9	17.4

Note: DWQ data used for 1997-1999 plus 2001; USGS data used for 1997-2001.

PREDICTION OF EXCURSIONS

The current focus of model application is on reducing the frequency of excursions of the 40 µg/L chlorophyll *a* criterion. Regardless of other strengths and weaknesses of the model, its ability to replicate the observed frequency of excursions of the criterion is of obvious practical importance.

Both the DWQ and USGS data can be used to examine model performance in this regard. Results for full year simulations are shown in Table 2. As noted above, the frequency greater than 40 µg/L is an interpolated estimate using the Excel PERCENTRANK function. Note that the volume-weighted results have been computed only for the DWQ and paired model data; these cannot be calculated from the USGS data as USGS has not monitored segments 1-3 or 15.

On an annual basis, the table shows that the model appears to ***underestimate*** the DWQ observed frequency of excursions of the criterion. This is primarily due to high concentration observations in the fall period, where the model may not perform well and/or the data may be suspect, so that apparent underestimation is not of great concern. However, its presence does contradict the statement made by CH2M HILL that the model is biased high. USGS and paired model predictions provide results that are similar to one another.

Table 2. Observed and Predicted Frequency of Chlorophyll *a* Concentrations Greater than 40 µg/L (Full Year)

Location	DWQ Observed	Model – Paired to DWQ Data	Complete Model Results (1997-2001)	USGS (1997-2001)	Model – Paired to USGS Data
Seg 1-4 (weighted)	41.7 %	31.1 %			
Seg 14-15 (weighted)	20.0 %	6.5 %			
Seg 1	73.2 %	68.2 %	36.5 %		
Seg 2	45.5 %	42.5 %	21.0 %		
Seg 3	39.8 %	33.4 %	14.8 %		
Seg 4	31.5 %	21.1 %	12.5 %	17.4 %	21.1 %
Seg 14	16.7 %	7.2 %	5.0 %	11.9 %	11.7 %
Seg 15	23.7 %	9.3 %	6.9 %		

Table 3 shows the frequency comparison for the growing season (May through September) observations. For segments 1-4, the frequency predicted by the model is slightly higher than observed, but for segments 14-15 the model frequency is again lower than the observed frequency, indicating no evidence of over-estimation of criterion excursions. In contrast, the model frequency is higher than the observed frequency in the USGS data for Segment 14 – but, due to the small sample size, this results from a difference of only one observation greater than the criterion.

Table 3. Observed and Predicted Frequency of Chlorophyll *a* Concentrations Greater than 40 µg/L (May - September)

Location	DWQ Observed	Model – Paired to DWQ Data	Full Model (1997-2001)	USGS (1997-2001)	Model – Paired to USGS Data
Seg 1-4 (weighted)	35.7 %	42.1 %			
Seg 14-15 (weighted)	14.9 %	8.8 %			
Seg 1	71.3 %	87.6 %	74.5 %		
Seg 2	35.8 %	57.6 %	48.4 %		
Seg 3	30.4 %	45.3 %	33.7 %		
Seg 4	25.3 %	28.6 %	29.6 %	36.9 %	39.9 %
Seg 14	11.3 %	4.3 %	11.8 %	12.6 %	23.3 %
Seg 15	15.6 %	12.7 %	16.3 %		

COMMENTS ON MODEL PERFORMANCE

Performance of mechanistic eutrophication models varies widely, based on the nature and stability of the system being modeled and the temporal detail and accuracy of the forcing functions. Some eutrophication model applications clearly perform much better in terms of individual point predictions for chlorophyll *a* than the Jordan Lake model. For instance, the EPA WASP model of the Neuse estuary[2], supported by highly detailed monitoring, achieved root mean square errors on chlorophyll *a* ranging from 7 to 16 percent of the mean predicted value. In contrast, the Jordan model, supported by much sparser data, has root mean square errors on individual chlorophyll *a* observations that range from 42 to 100 percent of the predicted mean. Yet, even in the Neuse model, the R^2 of the correlation between observations and predictions for chlorophyll *a* was relatively low (around 20 percent), and relative absolute errors of individual predictions range up to about 250 percent.

The Jordan model is, in many respects, constrained at the monthly scale, as information on tributary influent concentrations and water clarity, as well as observed data for calibration, is available at approximately this time scale. Håkanson[3] concluded that the natural coefficient of variation (CV) for monthly chlorophyll *a* observations in lakes was on the order of 0.25, while Håkanson et al.[4] proposed that the CV at the monthly scale for large rivers was on the order of 0.8. The CV will increase further with less precise analytical methods. Characteristics of the inflow segments of Jordan Lake intuitively should fall between those of large rivers and lakes. For the volume-weighted analysis based on 1997-99 and 2001 data, the error CV for the Jordan Lake model is 0.46 for segments 1-4 and 0.74 for segments 14-15. The uncertainty in the model relative to individual observations thus appears consistent with the natural variability associated with the temporal resolution of much of the forcing data. This in turn suggests that better model performance cannot be attained without greatly enhanced monitoring data.

In sum, the Jordan Lake model as currently implemented is not a particularly good predictor of individual point measurements of chlorophyll *a* – and cannot be without much better knowledge of external forcing functions. The dynamic, riverine nature of the influent segments of the lake likely means that significant natural variability would still be present even if these forcing functions were known precisely. But, this is not the appropriate test of the model. Instead, the model should be judged on its ability to replicate longer-term spatial and temporal trends and the frequency distribution of chlorophyll *a* concentrations greater than the criterion. For these purposes the model appears to perform well. The significant uncertainty that is present does provide a compelling rationale for use of adaptive management to achieve goals – but is not an excuse for inaction.

---

[2] Wool, T.A., S.R. Davie, and H.N. Rodriguez. 2003. Development of three-dimensional hydrodynamic and water quality models to support Total Maximum Daily Load decision process for the Neuse River Estuary, North Carolina. *Journal of Water Resources Planning and Management*, 129(4): 295-306.

[3] Håkanson, L. 1999. On the principles and factors determining the predictive success of ecosystem models, with a focus on lake eutrophication models. *Ecological Modelling*, 121: 139-160.

[4] Håkanson, L., J.M. Malmaeus, U. Bodemer, and V. Gerhardt. 2003. Coefficients of variation for chlorophyll, green algae, diatoms, cryptophytes and blue-greens in rivers as a basis for predictive modeling and management. *Ecological Modelling*, 169: 179-196.