Comparison of Methods for Tracking Marsh Phosphorus Concentrations in Loxahatchee National Wildlife Refuge Under the Consent Decree

prepared for

U.S. Department of the Interior

By

William W. Walker, Jr., Ph.D.
Environmental Engineer

December 20, 2006

The Consent Decree requires the Technical Oversight Committee to compare the Consent Decree's long-term levels/limits with the State of Florida's Class III numeric criterion and determine, for the Refuge, which is "lower," *see* Consent decree ¶¶ 8.C, 8.D, 10.B, App. B, and for the Park as to which is "more stringent." *See* Consent decree App. A. Specific numerical values for the long-term levels and limits are prescribed in the Consent Decree. However, at the time the Consent Decree was executed, no analogous numerical values had yet been established for the numeric criterion. Such values have now been established by the State of Florida. Consequently, it is now possible to make the comparison of numerical values for the levels/limits and criterion called for under the Consent Decree, and the determination of which is lower. That determination can be made based upon a straight-forward comparison of the concentration regimes that each test was designed to provide (< 10 ppb for the 4-Part Test and < 7 ppb for the long-term levels, expressed as long-term geometric means (LTGMs)). Additional comments and numerical analyses supporting the above straightforward comparison are presented below.

This report evaluates the statistical power of each test based upon the best available information. Power is defined as the probability of non-compliance (test failure) when the actual LTGM marsh concentration is above the target value (7 ppb or 10 ppb). The tests were developed with different objectives, target concentration regimes, databases, assumptions, monitoring program designs, and data-reduction procedures (Table 1). From a numerical perspective, the power of each test depends on the difference between the specified numerical value (the term "limit" is used here for simplicity) and the marsh data screened and processed using different protocols. A limited numerical comparison can be made with respect to the limits and statistical power of the tests based upon historical data from the 14 interior marsh stations that are included in both monitoring plans, as described below.

One factor that influences the relative statistical power of the tests is the fact that different "confidence levels" were used to set the numeric limits. The 4-Part test set the limits at the 95th percentile for the one-year single-station GM (15 ppb) and one-year spatial average GM (11 ppb), based upon the measured variability in the yearly geometric mean at marsh stations in WCA-2A and the Refuge (FDEP, 2003). The CD marsh levels were set at the 90th percentile of 1978-1979 measurements at three sites considered to be least impacted, with concentrations at the upper end of the range measured at other least-impacted marsh sites in the Everglades Water Conservation Areas and National Park (4 - 7 ppb, Walker, 2005, p. 11). Using the 95th percentile has the effect of raising the compliance limits for the 4PT relative to those that would occur if the limits were set at the 90th percentile, consistent with the CD specification. FDEP (2003, Table 5-4) computed one-year single-station limits ranging from 15.1 (WCA2A data) to 14.6 ppb (WCA-2A + WCA-1 data) using the 95th percentile. These values were apparently rounded in specifying the 15 ppb annual limit under the 4PT. Had the 90th percentile been used, the computed limits would have been 13.7 ppb and 13.4 ppb, respectively, and the power of the 4PT would have increased.

Another quantitative factor that influences statistical power is the fact that different round-off conventions are used for the measured marsh geometric means. The FDEP protocol (2006) rounds off the measured geometric means to the nearest 1 ppb before comparing them with the numeric limits. This has the effect of raising the compliance limit by approximately 0.5 ppb and decreasing the failure rate with a given distribution of measurements. For example, Table 2 shows that for concentrations between 9.5 and 11.0 ppb in 0.1 ppb increments, the failure rate (> 10 ppb) would be 63% with the measured geometric mean rounded to the nearest 0.1 ppb (consistent with the Consent Decree procedure), as compared with 38% with the measured geometric mean rounded to the nearest 1 ppb. Furthermore, the round-off convention for the 4PT adopted by FDEP in 2006 (1 ppb) is different from that employed by FDEP in developing and demonstrating the test, (0.01 ppb, Tables 5-1, 5-2, 5-3, 5-4, & 5-5, FDEP 2003).

A previous analyses (Goforth, 2006) focused only on the numeric limits without considering actual marsh data. Problems and limitations of that analysis are described in a document submitted to TOC by federal representatives (November 2006). Dr. Goforth's analysis was based upon application of the 4-Part test to a hypothetical time series of marsh P concentrations that were exactly equal to the Long-Term Levels in each month. The likelihood that such a time series would be encountered in the future is nil. The analysis involved numerous assumptions and results indicated very small differences in the numeric limits when applied to 1999-2006 time series. Results are invalid because the averaging method used to compute the annual geometric mean (arithmetic mean of 12 monthly values) was inconsistent with the protocol established by the FDEP (geometric mean of

12 monthly values). The comparison also ignored the fact that the FDEP protocol excludes data from stations with less than 6 samples per year, which tend to have lower water depths and higher phosphorus concentrations. As a consequence of these errors, that analysis over-states the power of the 4-Part test relative to the Long-Term Levels, as applied to the 14 interior stations.

A limited, but more realistic numerical comparison of the tests can be made using historical data from the 14 interior marsh stations that are included in both monitoring plans. The marsh data "passed" the 4-Part test (Nearhoof, 2005) over the 1999-2005 monitoring period, when numerous excursions and exceedances of the Long-Term Levels would have occurred had they been in effect (SFWMD TOC Report, 2006). That alone suggests that the CD levels are "lower" in the sense that there would be a greater risk of failure when applied to historical data from the interior marsh. The comparison can be amplified by rescaling the historical data to different long-term geometric mean concentrations and re-computing the failure rates for each test using the following procedure:

1. Compile measurements from each of the 14 sites over the May 1995-April 2006 period. Exclude data from May and June of 2005, considered by TOC to be influenced by "error or extraordinary natural phenomena".

2. Screen and summarize the data and compute limits using the procedures specified for each test.

3. Compute failure rates for each component of the 4-Part test and for the combined result, expressed as a percent of Water Years when the measured value exceeded the specified limit. The failure rates are computed using May 1999 - April 2006 data, when sufficient data are available for applying all 4 components of the test. The May 1995-April 1999 values are used only for computing the 5-year geometric means used in Part 1 of the test.

4. Compute the failure rates for the Long-Term Levels, expressed as percent of 12 monthly sampling event intervals when an "excursion" (monthly marsh GM > LTL) or an "exceedance" (2 or more excursions in any 12 consecutive sampling events) occurred.

5. To evaluate sensitivity to changes relative to historical conditions, rescale the historical data by fixed ratios to generate a hypothetical series of datasets with long-term geometric means (LTGM) ranging from 4 to 15 ppb (vs. 1999-2006 value of 8.7 ppb). Compute the LTGM of each series using the FDEP protocol (arithmetic mean of yearly geometric mean across all sites and years). This rescaling assumes that spatial and

temporal variability around the LTGM (expressed on a percentage basis) is independent of the LTGM.

6. Repeat steps 2-4 for each rescaled time series and plot the failure rate for each test as a function of the LTGM.

Results are shown in Figure 1.

For the CD test, the risk of an "excursion" increases from ~0% to ~100% as the LTGM increases from 6 ppb to 15 ppb. The risk of an "exceedance" (2 or more excursions in any 12-sample interval) increases from ~0% to ~100% as the LTGM increases from 6 to 10 ppb. The risk of failing the 4-Part test increases from ~0% to ~100% as the LTGM increases from 9 ppb to 12 ppb.

When evaluated based upon expected failure rate, the two tests are essentially identical at LTGM values < 6 ppb (failure rate ~0%) and > 12 ppb (failure rate ~100%). With LTGM's between 6 and 12 ppb, the risk of failing the CD test (exceedance) is consistently greater than the risk of failing the 4-Part test. For example, when the LTGM is 10 ppb, the expected failure rate is ~ 100% for the LTL vs. ~25% for the 4-Part Test.

The LTGM value corresponding to a 10% failure rate is ~7.0 ppb for the CD LT Levels and ~ 9.5 ppb for the 4-Part test. While this comparison does not consider all dimensions of the tests (Table 1), results indicate that the LT Levels are "lower" relative to the metrics used to express the measured marsh concentrations. This is manifested in the result that the expected failure rate (hence, statistical power) is higher when the measured interior marsh LTGM's are between 6 and 12 ppb (Figure 1).

As discussed in the Consent Decree (Appendix B), achieving compliance with the Long-Term Levels is expected to provide a LTGM less than 7 ppb, as observed at the 3 least-impacted sites in 1978-1979 and at the upper end of the range of values measured at other least-impacted sites in the Everglades (4 - 7 ppb, Walker, 2005, p 11). If that objective were achieved, Figure 1 shows that the risk of failure would be less than ~10%. That risk level is consistent with the decision to set the marsh levels at the 90th percentile of historical values in deriving the CD compliance test.

References

FDEP, "Calculation of Annual and 5-Year Geometric Mean Total Phosphorus Concentrations to Assess Achievement of the Phosphorus Criteria for the Everglades Protection Area", submitted to TOC, August 2006.

FDEP, "Development of Numeric P Criterion", Everglades Consolidated Report, Chapter 5, 2003.

Federal technical response to June 2006 Goforth memo to TOC.
November 7, 2006

Goforth, G., "Comparison of Class III Phosphorus Concentrations with Long-Term Levels for the Refuge", submitted to TOC, June 2006, revised November 2006.

Nearhoof. F., "Report on Water Quality Sampling in the Everglades", Expert Testimony submitted to Special Master, August 2005.

SFWMD, "Settlement Agreement Report", July - Sept 2006, November 2006.

Walker, W.W., "Expert Report Concerning Exceedances of the Interim Phosphorus Levels at the A.R.M. Loxahatchee National Wildlife Refuge in July 2002 and August 2004", submitted to Special Master, August 2005.

Table 1

Comparison of 4-Part Test with Long-Term Levels Based upon Various Factors

Factor	4PT	LTL
Basis	P Criterion	1978-1979 Data
Long-Term Geo Mean	<= 10 ppb	<= 7 ppb
Moderating Provision	Yes	No
Interior Sites	> 14	14
Exterior Sites	> 0	0
Stage-Dependent	No	Yes
Limit Percentile	95%	90%
Test Interval	Water Year	12 Consecutive Sampling Events
Minimum Duration	1-5 years	~ 1 year
Water Depth	> 10 cm	> 10 cm
Marsh Stage		> 15.4 ft
Sample Frequency	> 6 samples/yr	
Round-off Convention	1 ppb	0.1 ppb
Historical Failure Rate (14 Interior Stations, 1999-2006)	0%	~22% excursions
LTGM at < 10% Failure Rate	< 9.5 ppb	< 7.0 ppb

Table 2

Effect of Round-Off Convention on Test Failure Rates

Data Rounded to Nearest 0.1 ppb	Freq > 10 ppb	Data Rounded to Nearest 1.0 ppb	Freq > 10 ppb
9.5	Pass	10	Pass
9.6	Pass	10	Pass
9.7	Pass	10	Pass
9.8	Pass	10	Pass
9.9	Pass	10	Pass
10.0	Pass	10	Pass
10.1	Fail	10	Pass
10.2	Fail	10	Pass
10.3	Fail	10	Pass
10.4	Fail	10	Pass
10.5	Fail	11	Fail
10.6	Fail	11	Fail
10.7	Fail	11	Fail
10.8	Fail	11	Fail
10.9	Fail	11	Fail
11.0	Fail	11	Fail
Failure Rate	63%		38%

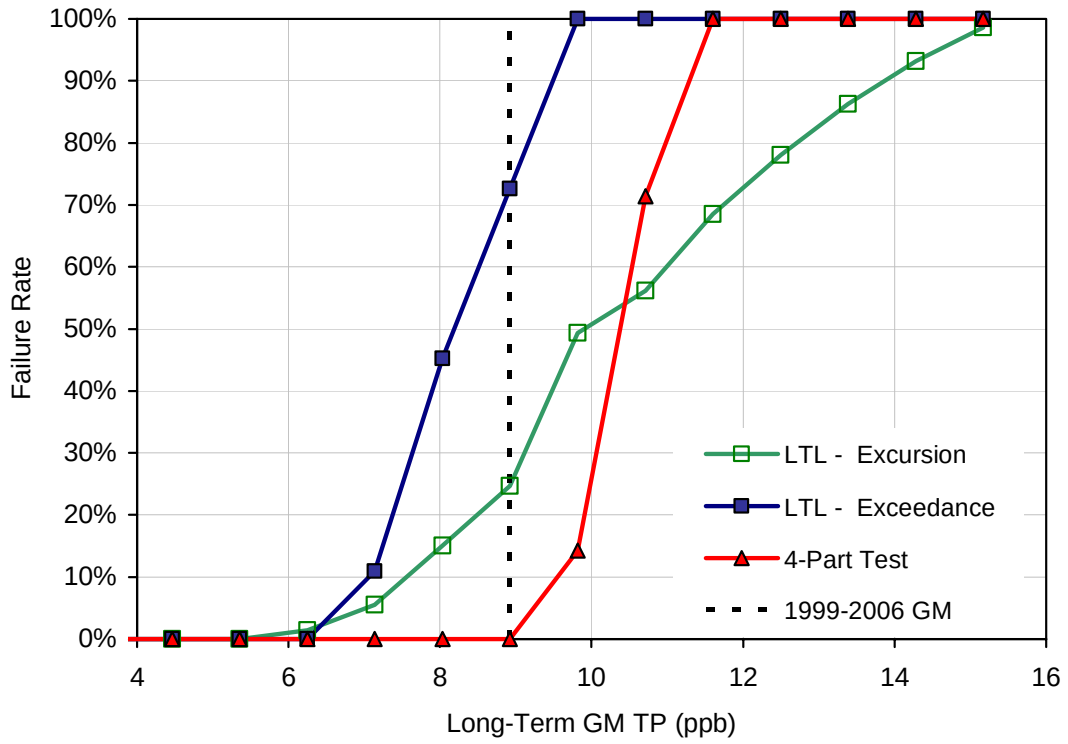

Figure 1

Power Curves for Long-Term Levels and 4-Part Test

LTL - Excursion	monthly excursion from CD Long-Term Levels
LTL - Exceedance	>= 2 excursions in any 12 consecutive monthly sampling events
4-Part Test	fails any part of the test in any water year

"Power" is defined as the failure rate when the actual LTGM is greater than the expected value or target for the LTGM. Curves developed by rescaling the May 1995 - April 2006 data from 14 interior marsh stations. Frequencies computed for May 1999 - April 2006 (6 Water Years). Computation of limits, data screening, and data reduction procedures are consistent with the protocols established for each compliance test.