



September 1, 2006

Ms. Song Her, Clerk to the Board  
State Water Resources Control Board  
P.O. Box 100  
Sacramento CA 95812-0100

Re: Comments on statistical aspects of Panel of Experts report "**The Feasibility of Numeric Effluent Limits Applicable to Storm Water Discharges**"

Dear Ms. Her,

I was retained by Flow Science in connection with their work for the Western States Petroleum Association and was asked to evaluate and comment upon statistical issues associated with the formulation of numerical limits for constituent concentrations in stormwater effluents.

**My qualifications.**

My professional background as a statistician began with my research specialization in the field for my Ph.D. in mathematics from Cornell in 1966. Since that time, I have been continuously engaged in research and teaching of statistics at Northwestern University, UC Berkeley, and Caltech, where I have been Professor of Mathematics since 1977 and was department chair from 2003 to 2006. I am a fellow of the Institute of Mathematical Statistics, and have been active for the last thirty-five years as a statistical consultant for Caltech colleagues and for various governmental agencies and private companies. I have also served as a statistical expert witness in a variety of legal and regulatory matters, including statistical issues of water quality.

**Variability of pollutant concentrations in stormwater flows.**

It is clear from stormwater datasets I have examined in my statistical consulting that the pollutant concentrations associated with stormwater flows are highly variable, even over short time scales. Within a given hour the measured concentrations of pollutants in grab samples can be expected to vary substantially in relation to the mean for that hour. Therefore the probability that a single-grab-sample-per-storm monitoring system will accurately reflect the true impact of effluents on receiving waters is low. Moreover, the application of numerical effluent limits to grab-sample data is inherently less effective from a statistical point of view than the use of composite samples, for example. Because CTR criteria are specified as one-hour (or longer) average concentrations, it is essential to consider the additional variability of effluent concentrations that occur on a sub-hour

basis. In particular, it is important to recognize the fact that any numerical limit applied to grab samples inherently imposes a smaller numerical limit on hourly averages.

### **Sources of Data Variability.**

It is critical that numerical limits be derived with adequate consideration given to *all significant sources of variability of constituent concentrations in stormwater discharges.*

The sources of variability can be described in three categories:

- 1) Input Variables. These include influent characteristics, storm characteristics (e.g. rainfall intensity, and the rate and volume of flow), site-specific hydrologic features, and receiving water characteristics, such as those affecting dilution.
- 2) Treatment Characteristics. As discussed in the panel's report, the treatment capabilities of different facilities under different inputs vary widely, as do limits on their capacity for handling different flow rates and volumes.
- 3) Output Variables. Constituents in discharges vary in concentration in different parts of the effluent flow, the results of laboratory analysis are subject to measurement errors, and sampling techniques like one-grab-sample-per-storm introduce additional variability compared with hourly averages or storm composite concentrations.

Because of the systematic and widespread differences in these characteristics from facility to facility, storm to storm, and sample to sample, it is necessary to carry out an extensive and well-designed data collection effort at a representative set of facilities over a period of years. One or two years of data cannot represent the range of variability in number and severity of storms from year to year.

### **Unusual events and exceedance probabilities.**

The report of the panel of experts contains repeatedly acknowledges the need to consider that storm water flows are dramatically affected by "unusual events". Here are a few examples:

- "...there is wide variation in stormwater quality from place to place, facility to facility, and storm to storm." (p.6)
- "Since the storm-to-storm variation at any outfall can be high, it may be unreasonable to expect all events to be below a numeric value." (p.6)
- "...several to more times each year, the runoff volume or flow rate from a storm will exceed the design volume or rate capacity of the BMP. Stormwater agencies should not be held accountable for pollutant removal from storms beyond the size for which a BMP is designed." (p.10)

- “The Panel recommends that Numeric Limits and Action Levels not apply to storms of unusual event size and/or pattern (e.g. flood events).” (p.18)

“Unusual events” aside, the statistical approach used in the State Implementation Policy (SIP) relies heavily on two features:

- the assumption that pollutant measurements in stormwater flows follow a lognormal distribution, and
- the idea of setting numeric limits for a facility by considering “exceedance frequencies” based upon calculations using the lognormal distribution.

The latter idea is revealed clearly on page 10 in Step 5, which discusses “a factor (multiplier) that adjusts for the averaging periods and exceedance frequencies of the criteria/objectives ...”.

Clearly the use of “never to be exceeded” limits for stormwater effluents in permits needs to be eliminated. In light of the Panel’s discussion and the statistical rationale used in setting limits, provision should be made for two kinds of exceedances—

- exceedances caused by carefully-defined “unusual events”—for example, storms whose severity and/or flow volumes exceed a “design storm”, and
- “random” exceedances—resulting from the unavoidable fact that even ideal data, such as data from the assumed “standard”, lognormal distributions, have some frequency of exceedance of any specified numerical limit.

#### **Is the assumption of Lognormal Distributions valid? Analysis of datasets.**

The most frequently used statistical model for datasets of stormwater constituent measurements is the lognormal distribution. In particular, the SIP relies heavily on the assumption that lognormal distributions adequately describe such data. It is therefore important to evaluate the validity of that assumption— i.e. to test it on actual data.

Flow Science provided me with three datasets containing grab sample measurements of copper at three outfalls of a California facility. I analyzed the datasets A, B, and C to determine whether the *maximum value* in each sample is too large to be reasonably explained by the lognormal distribution best fitting the data as a whole. In my experience with stormwater datasets, the largest value is “too large”, indicating that the so-called *right-hand tail* of the actual data distribution is *heavier* than it would be if the data distribution were lognormal. In other words, the largest values that are found in datasets are not explained by the “general shape” of the lognormal distribution.

The following table gives the results of my analysis.

Summary Statistics for Copper datasets at 3 Outfalls  
(water quality objective= 14 mg/L)

	Outfall A	Outfall B	Outfall C
Sample Size	23	32	20
Sample Median	2.8	2.8	3.2
Sample Maximum	55	12	39
p-level*	.003	.012	.110

\*defined below

The p-level is a quantity used by statisticians to measure the *statistical significance* of a finding obtained from data analysis—in this instance, that the maximum value in each of the three datasets is large in relation to the sample as a whole. It is calculated as the probability that such an extreme value would occur “due to chance” *assuming that the data do come from a lognormal distribution.*

For Outfall A, the p-level is .003, indicating that a value of the sample maximum as extreme as the one observed (in relation to the overall dataset) would occur only about 3 times in 1000 samples. Similarly, for Outfall B, the p-level of .012 indicates that a maximum value as high as 12 (in relation to the other values) would occur only about 12 times in 1000 samples. These two results are what statisticians call “highly statistically significant”, meaning that they provide strong evidence that the hypothesis being tested (in this case, that the true data distribution is lognormal) is not true. The p-level for Outfall C is .110, indicating that a maximum as large as 39 (in relation to the other measurements from Outfall C) would occur in about 1 sample out of every 9. This is what statisticians consider “marginally significant”.

Taken as a whole, these three analyses indicate strongly that the lognormal distribution does not describe the behavior of copper measurements at these outfalls, in that the largest values in samples of size 20 to 30 are “too large to be explained by the lognormal distribution”. Since the setting of numerical limits for constituent concentrations in stormwater effluents, as well as the monitoring of compliance with them, is mainly concerned with the behavior of the largest values in datasets, this apparent failure of the lognormal distribution to describe that behavior seriously undermines the statistical methods used in the State Implementation Policy (and elsewhere) to establish numerical limits.

### How much data is needed to set numeric limits? An example.

As the report of the panel of experts and the State Implementation Policy (SIP) statistical methods make clear, setting numerical limits is critically related to controlling frequencies of exceedance—i.e. satisfying numerical criteria or objectives.

It is very important to recognize that the amount of data needed to determine frequencies of exceedance reasonably accurately and with high confidence is large—larger than would be expected on “common sense” grounds.

For example, suppose one takes a sample of  $n$  measurements designed to “demonstrate with 95% confidence that the value  $L$  is exceeded at most 5% of the time”. (Here  $L$  is arbitrary and the statement is equivalent to saying that “the 95<sup>th</sup> percentile of the true distribution of the data is  $L$  or smaller”.) If  $n=153$  (say), then it turns out that that confidence statement will be true *provided that at most 3 exceedances occur*.

Since 3 is about 2% of 153, we note that in order to *demonstrate* that the true exceedance percentage is at most 5%, we need to have exceedances of 2% or less in the sample. This is because we want to have *95% confidence* in the conclusion that the required 5% rate is truly satisfied.

Moreover, if we ask “How low does the true exceedance percentage have to be so that we can have 90% confidence that there will be at most 3 exceedances in the sample?”, then the answer is “about 1.1%”. If the *true* exceedance rate were 2%, we would get about 3 exceedances *on the average* in a sample of 153, but nearly half the time we would “flunk the demonstration” by getting more than 3 exceedances. Thus, one needs an “extra margin of safety”—in this example, a 1.1% true exceedance rate—to have a 90% probability of a successful *demonstration* that  $L$  is exceeded at most 5% of the time.

This disparity between *having* a low exceedance rate and *demonstrating* a low exceedance rate I call the “Caesar’s wife effect”. To avoid the *suspicion* of “having more than 5% exceedances” of a numerical limit,  $L$ , one must actually have a *true exceedance rate* much lower than 5%.

The sample size  $n=153$  was used for this example. If a smaller sample size were used, the Caesar’s wife effect would be even more pronounced—that is, either one would have to have an *extremely low exceedance rate* (a fraction of 1%) or else would have a probability larger than 10% of getting too many exceedances in the sample to succeed in demonstrating that the 5% objective is attained.

To perform more difficult analyses, such as testing the “fit” of distributions better suited than the lognormal to describe the data, requires even larger sample sizes. Investigation of a large body of data can shed light on the question of whether the lognormal distribution can be replaced by some other shape of distribution that better represents actual data. My expectation is that no *simple* family of distributions can represent adequately the range of behaviors of datasets. Accordingly I expect that the most useful

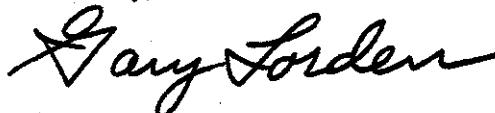
statistical calculations will turn out to be based upon so-called *nonparametric* or *semiparametric* methods, relying more upon estimating from data the actual frequencies of high concentrations rather than upon estimating parameters such as the coefficient of variation.

**Recommendations for developing a sound statistical basis for numerical limits.**

To establish a statistically sound basis for setting numerical limits for storm water effluents, I believe it is necessary to carry out a well-designed data collection effort at a representative set of facilities over a period of years sufficient to incorporate the substantial year-to-year variability in the number and severity of storms.

Investigation of a large body of data can shed light on the question of whether the lognormal distribution can be replaced by some other shape of distribution that better represents actual data. My expectation is that no *simple* family of distributions can represent adequately the range of behaviors of datasets. Accordingly I expect that the most useful statistical calculations will turn out to be based upon so-called *nonparametric* or *semiparametric* methods, relying more on estimating actual frequencies from data than on estimating parameters like the coefficient of variation. It is an inescapable "statistical fact of life" that substantial sample sizes are required to determine appropriate numerical limits and to monitor compliance with them.

Sincerely,

A handwritten signature in cursive script that reads "Gary Lorden". The signature is written in dark ink and is positioned above the printed name.

Gary Lorden