amsa

Association of
Metropolitan
Sewerage Agencies

March 31, 2005

Water Docket
EPA Docket Center, Mail Code 4101T,
1200 Pennsylvania Ave., NW.
Washington, DC 20460
Attention: Docket ID No. OW-2004-0037
Via E-mail: OW-Docket@epa.gov

Dear Sir or Madam:

The Association of Metropolitan Sewerage Agencies (AMSA) appreciates the opportunity to comment on the U.S. Environmental Protection Agency's (EPA or Agency) draft *National Whole Effluent Toxicity (WET) Implementation Guidance Under the NPDES Program*, November 2004 (*Draft Guidance*). AMSA and its members have requested additional guidance on WET for years as a way to address some of the implementation issues that arise when WET monitoring requirements and numeric limits are incorporated into the National Pollutant Discharge Elimination System (NPDES) permit program.

AMSA's members have reviewed the *Draft Guidance* and welcome several of EPA's recommendations. Many of the Association's long-standing issues, however, are not addressed, and new concerns have been raised regarding the effect the guidance may have on current implementation approaches. If the *Draft Guidance* is finalized as written, AMSA believes it will hinder current efforts to use improved WET approaches in many states and will increase the likelihood that dischargers will receive a numeric WET limit simply due to the statistics of EPA's reasonable potential process.

Over the past 10 years AMSA has supported the use of WET testing as a useful tool for assessing uncertainties in effluent quality and impacts on receiving water biota. AMSA, however, continues to be concerned about implementing numeric WET limitations on a pass/fail basis for determining NPDES permit compliance. While many of AMSA's concerns apply to the entire suite of WET methods and endpoints, the Association's primary focus continues to be on the use of chronic, sub-lethal endpoints, where permit compliance or reasonable potential may be more a function of the method itself, instead of effluent quality.

AMSA's recent advocacy on WET issues has focused on a substantive legal challenge of the methods themselves and on parallel efforts to address WET implementation issues. On the legal front, the U.S. Court of Appeals for the D.C. Circuit in December 2004 (*Edison Electric Institute, et al. v. EPA*), upheld the WET methods but did note that the WET tests are "not without their flaws" and emphasized the ability of permitting authorities to account for the methods' limitations at the local level. EPA's *Draft Guidance*, therefore, takes on added significance as it must now guide implementation of the flawed and variable methods. Several groups have sought a rehearing of the WET decision before the D.C. Circuit. AMSA, however, determined it would not participate in the rehearing process.

Rather, AMSA has chosen to continue a sustained effort to improve current NPDES permit program implementation approaches for WET. In April 2003, AMSA met with key Office of Wastewater Management officials, including Linda Boornazian, Director, Water Permits Division, to discuss some of the Association's concerns with respect to no and low-dilution environments, where the flaws of the WET tests are exacerbated. It was in the context of that meeting that AMSA first discussed with EPA managers its notion of a tiered or step-wise approach to WET implementation, which is discussed in more detail below.

The April 2003 meeting was intended to be the first of many meetings between EPA and AMSA on the issue of WET implementation. The WET Coalition had also expressed an interest in meeting with EPA regularly to discuss various implementation issues concurrent with EPA's efforts to develop its *Draft Guidance*. Unfortunately this series of collaborative meetings to discuss implementation issues never materialized and the *Draft Guidance* fails to address critical municipal concerns.

Accordingly, AMSA offers the following comments on the *Draft Guidance* and supporting documents.

General Comments

National Consistency vs. State/Regional Flexibility

One of the goals of the *Draft Guidance* is to promote national consistency. AMSA understands the importance of national consistency in the implementation of NPDES permit requirements, but is concerned that technically reasonable and defensible approaches now being used by states may be abandoned in an effort to ensure consistency with the *Draft Guidance*. The language used in the *Draft Guidance* that EPA Regions are expected to be consistent with the *Draft Guidance* and that states are "strongly encouraged" to follow the *Draft Guidance* and may need to "revise their current procedures to fully implement the national recommendations" is too restrictive and does not seem to allow for alternative, defensible approaches. This directive to the states is inconsistent with the D.C. Circuit's emphasis on the fact that states have the discretion to set toxicity thresholds to compensate for local conditions at the permitting stage. This helps to mitigate the fact that the correlation between laboratory toxicity and instream impacts is weaker at low levels of toxicity.

AMSA suggests revising the *Draft Guidance* to further clarify, consistent with the Notice and Disclaimer (p. iii), that it is just that, guidance, and that Regions and especially states can use other approaches that are technically defensible. The national WET program can consistently apply the same principles and concepts (e.g., weight of evidence, accounting for instream exposure, addressing test uncertainty, etc.) across all programs without going to the level of specificity used in this document.

For example, on page 18 of the *Draft Guidance* EPA notes that use of the statistical procedures in the *Draft Guidance* with few data points can result in "conservative projections about possible effluent toxicity and thus may result in unnecessary permit limits." AMSA agrees and believes that the Agency needs to be clear that States may use alternatives to those procedures. Many states have been addressing WET in their permitting programs for years, using a variety of regulatory approaches. Some apply different statistical techniques, while others use procedures that rely more on professional judgment, after evaluation of information about the discharges at issue. EPA has neither shown any basis for believing that those state programs have failed, nor provided any other reason why those state approaches should all be replaced with a single, uniform statistical procedure. EPA should clarify in the *Draft Guidance* that states can use alternate approaches as long as they effectively address toxic effects on aquatic life in the state's waterbodies.

A True "Step-wise" Approach Needed

When AMSA met with EPA in April 2003, the attendees of the meeting discussed the concept of a step-wise or tiered approach to implementing WET limits, specifically chronic, sub-lethal limits. Rather than establishing a numeric pass/fail limit and then relying on EPA's enforcement discretion policy[1], EPA and AMSA discussed the possibility of establishing a non-numeric limit consisting of accelerated follow-up tests and potential toxicity reduction evaluation (TRE) steps that would be triggered by a test failure. Failure to conduct the additional testing and possible TRE, not the initial failure, would be considered the violation.

Participants at the April 2003 meeting discussed the feasibility of establishing such a 'narrative limit' consisting of confirmatory testing and toxicity investigation in lieu of a numeric limit. Although it was agreed that such a scenario would be harder to address in NPDES permits, everyone, including the EPA staff present, agreed that for most cases, with exceptions for obvious (extreme) toxic events, such an approach could work and would be protective of the environment.

At a subsequent meeting between the WET Coalition (including AMSA) and EPA, where enforcement personnel were present, it was made clear that enforcement officials sought a clear limit that would be easy to enforce. An enforcement official present made it clear that the step-wise approach AMSA and the WET Coalition had been advocating for would make assessing compliance too difficult and would not result in a well-defined violation (i.e., a test failure).

The D.C. Circuit's opinion supports this step-wise approach. Concurring with AMSA and others that WET tests "will be wrong some of the time," the court highlights EPA's own warning "against using a single test result to institute an action for a civil penalty." EPA's proposed 'step-wise' approach as outlined in the *Draft Guidance*, *however,* misses the point of AMSA and the Court's concerns and

[1] *National Policy Regarding Whole Effluent Toxicity*, August 14, 1995. The policy states that EPA "does not recommend that the initial response to a single exceedance of a WET limit, causing no known harm, be a formal enforcement action with a civil penalty."

proposals. While EPA's approach requires additional testing to determine whether a TRE should be initiated following an initial test failure, the initial test failure is still a violation. It was not AMSA's intent to simply add more testing. Without defining the requirement to conduct the additional testing and toxicity evaluation as the actual limit, the EPA approach simply increases the chances of another test failure. AMSA understands that EPA policy states that single WET test failures are subject to enforcement discretion (i.e., no civil penalty), but a violation is a violation and AMSA's members do not take violations lightly, regardless of whether a civil penalty is assessed. As such, AMSA can not support EPA's approach as currently drafted.

Instead, where chronic limits are deemed necessary, AMSA suggests that permits require the following:

1. Exceedance of WET trigger values results in increased testing (failures taken together with other evidence that indicate environmental toxicity would be considered a violation) to determine the presence of persistent toxicity;
2. Finding of persistent toxicity requires the permittee to begin a TRE within a certain period of time (to be specified in the permit);
3. Monthly progress reports would be required with each discharge monitoring report to outline identification and solution of toxicity problem (certain provisions would need to be made for instances where the source of toxicity can not be determine after a certain level of effort has been expended).

Failure to follow the permit requirements or 'narrative limit' as outlined above would be the NPDES permit violation. This type of approach is more desirable as it:

1. Results in the actions EPA deems necessary to address the toxicity;
2. Is faster and more efficient than current approaches which usually require interaction between the permittee and regulatory agency before actions are taken; and
3. Does not trivialize permit violations.

(Though AMSA's approach above differs somewhat from that of the WET Coalition, a more detailed description of the concept was outlined in a Coalition white paper provided to EPA in 2003)

AMSA acknowledges that EPA has also applied its step-wise thinking to the development of permit requirements by allowing a delayed WET limit so that permittees have additional time to collect more information. Though a step in the right direction, this approach still means that a permittee will have a limit based on few or no tests and yet be responsible for conducting additional tests in the hope that the regulatory agency will later remove the limit through a permit modification.

In addition, 18 months is not enough time to gather the additional data (in some cases, 6 data points, probably collected on a quarterly basis, to provide an estimate of year-round performance), then review the data and submit a request for a permit modification. The state would need to review the request, prepare a draft permit modification, issue that draft for public comment, review the comments, and then issue a final permit modification. It is simply not possible to do all of that in 18 months, when states would have to review multiple modification requests for their various dischargers, and would also have to continue with their existing permitting efforts. If the permit modifications have not been finalized within the 18 months, antibacksliding restrictions would apply, and the dischargers could be unable to have the WET limits removed, even though they have made adequate demonstrations that the limits are not needed. To avoid that problem, longer compliance schedules should be provided and allowances for state-specific procedures that permit longer compliance schedules should be made.

Issues Associated with Technical Support Document
One of EPA's stated goals for the *Draft Guidance* is to restate and clarify, where necessary, elements of previous Agency guidance, policy, and regulations concerning WET, including the Technical Support Document for Water Quality-based Toxics Control, 1991 (TSD). The procedures for reasonable potential determinations and limit derivation calculations in EPA's TSD are referred to throughout the *Draft Guidance*. However, EPA has yet to address any of AMSA's long-standing comments and concerns regarding elements of the TSD. Since the *Draft Guidance* relies heavily upon the TSD, AMSA is reiterating some of these concerns here.

If EPA chooses to release this document without adequately addressing these long-standing issues, AMSA recommends that EPA provide broad statements in the *Draft Guidance* acknowledging these issues and their importance in making defensible permitting decisions, and recommending that states/regions address them on a site-specific basis.

1. *Test and Instream Exposures*

Differences in exposure assumptions between lab and receiving waters and how these assumptions should impact monitoring frequency and triggers for other actions have not been addressed in the *Draft Guidance* despite the fact that stakeholder groups, which included states and EPA staff, have concluded that excessive margins of safety exist in the WET program. The proceedings of the WET Pellston Workshop of 1995 state that:

> *(t)he workgroup agreed that most major concerns about the application of WET test data arise… from the application of toxicity test data to conditions that often do not accurately reflect test exposure.*

The proceedings go on to state that:

> *(t)he participants at this Workshop identified exposure in the receiving system as one of the more important issues limiting the predictive relationship between WET test results and receiving system impacts. One of the fundamental reasons for this is the large margins of safety that are built into discharge permits. The larger the margins of safety, the less the likelihood that an exceedance of a WET limit would result in an adverse exposure in the receiving system. Permit requirements that contribute to large margins of safety and to overestimates of exposure use conservative allowances for dilution, such as basing available*

> *dilution on the 7Q10 flow, and for the low probability that maximum toxicity will coincide with such infrequent flows.*

Other than reference to the use of dynamic models versus static assumptions for instream mixing, EPA has avoided this issue in its *Draft Guidance* even though EPA staff have agreed that considerations of exposure in the current program (basically unchanged since 1995) are very important. Even previous EPA memos and the 1997 WET Stakeholder Implementation meeting state that nothing precludes the testing of instream conditions dependent on seasonal variations, but this *Draft Guidance* document does not reference these observations. The Pellston proceedings conclude that "WET testing is an effective tool for predicting receiving system impacts when appropriate considerations of exposure are considered," that "[r]egulatory agencies need to reevaluate the safety margins used in establishing WET requirements," and that "[t]he significance of an exceedance of WET limits depends on receiving water conditions, especially dilution at the time the exceedance occurs. All exceedances of the same magnitude do not necessarily have the same receiving system impacts."

EPA also held a WET Stakeholder Implementation meeting in 1996. This meeting consisted of permittees, consultants, state and federal representatives, and labs broken into several groups to address several topics. One group was assigned to address exposure issues and AMSA members participated in this group. This group, contrary to EPA's report on the meeting, reached 100% consensus that:

1. WET guidance needs to allow evaluation of discharge-specific, time exposure-related effects;
2. Duration of exposure in tests must equal that instream; and
3. A two-tiered test interpretation approach should be adopted where exceedance of a trigger or limit based on conservative assumptions of exposure is followed by an investigation of the actual exposure at the time of testing to see if the trigger or limit was actually exceeded. If the trigger or limit was not exceeded at the time of testing the discharger resumes normal testing. This approach can be balanced by testing more frequently to reduce uncertainty in representing actual instream exposures. However, all other points of uncertainty (test variability, species representativeness, duration and frequency of exposure, etc.) must be addressed before accurate reasonable potential or compliance determinations can be conducted.

EPA's draft WET implementation strategy of 1997 states that EPA proposes to "[b]egin to reevaluate the safety margins (e.g., critical low flows) used for establishing WET requirements to increase the confidence in the limits developed to predict receiving stream impacts" and "[p]rovide guidance to stakeholders on the appropriate examination of effluent flow and monitoring data in order to establish a realistic link of exposure assumptions to permit development."

Despite numerous recommendations and input from its own staff and other stakeholders, EPA has offered nothing in its *Draft Guidance* to help determine if exposure has been appropriately addressed in the WET program. For example, a discharger's outfall achieves acute dilution in less than one minute following discharge. Since planktonic organisms are used in these tests, the organisms are only exposed to the effluent at this concentration for seconds, yet the duration of the tests are 48 hours. The basics of toxicology hold that organism response will increase until it is maximized (in the case of survival, complete mortality) as duration of exposure increases. Therefore the response expected instream from the most sensitive organisms, in this example, will likely be much less than that predicted by these tests. Another issue tied to test duration is frequency of exposure. The tests assume that organisms are continuously exposed to the same concentration of effluent for the duration of the test, but, again, planktonic organisms used in these tests do not have the ability to maintain their position in the water column (horizontally and vertically) or in relation to an outfall. The location of these organisms relative to outfalls is a function of hydrodynamic forces like tides and currents. The tests do not address this; therefore the implementation program must address it.

2. Use of Toxicity Units (TUs)

The D.C. Circuit ruled that the use of TUs will result in a "grossly inflated result" when used to calculate the coefficient of variation (CV) of data. The CV is critical to making reasonable potential determinations and deriving limits. If TUs significantly overestimate the CV they will result in an overestimation of reasonable potential to exceed or contribute to an exceedance of standards as well as limits which are more stringent than necessary. The court also ruled that the use of TUs to characterize the distribution of WET data is a "mistake". EPA assumes that WET data is distributed lognormally, an assumption also critical to reasonable potential determinations and limit derivation, but this is based on an analysis of data in the TSD using TUs. Additionally, the use of TUs is not supported in the 40 CFR Part 136 methods nor are TUs included in federal NPDES regulation. There are also technical concerns regarding TUs including:

1. They are not additive unless the dose response curves used to develop each TU are parallel and linear at the level in question (D.R. Ownby and M.C. Newman, 2000);
2. They are unitless and encourage the pooling of data from different tests with different exposures (TU for a 2 day acute test is the same as a TU for a 4 day acute test, for example), different biological endpoints (a TU based on reproduction is not the same as a TU based on survival) and different statistical endpoints (a TU based on a LC50 is not the same as a TU based on the NOAEC);
3. The effect level for each TU can be very different when based on hypothesis test endpoints;
4. A different TU can be found for the same test depending on which statistical endpoint is used (IC25 or NOEC, NOAEC or LC50); and
5. A lab must test 100% effluent to use the TU approach even though dilution instream may be significant.

The 1996 WET Stakeholder Implementation meeting held by EPA also concluded unanimously that TUs should not be used in favor of percent effluent. Therefore, the use of TUs in the WET permit program is fatally flawed and unsubstantiated, and will lead to erroneous conclusions regarding reasonable potential, limit derivation and compliance. AMSA has recommended that EPA consider the use of the "Percent Effect" (PE) approach to interpret WET data, in response to concerns over TUs, but EPA made no mention of this approach in the *Draft Guidance*. This approach interprets test results relative to effects at the receiving water concentration (RWC), rather than the concentration at which certain effects are measured. The PE approach offers numerous advantages to that of TUs, including:

1. It is intuitive in nature without artificial manipulation (higher PE means more toxicity);
2. It renders conversion factors like the LC1/LC50 ratio obsolete; and
3. It more directly translates to narrative water quality standards.

AMSA recommends that EPA reconsider use of the PE approach and, at a minimum, include a provision in the *Draft Guidance* that recognizes that endpoints other than those mentioned in the WET methods can be used in regulatory programs.

3. *Reasonable Potential Multipliers*

EPA is aware of a report drafted by the Water Environment Research Foundation (WERF) (Project 00-ECO-1, Whole Effluent Toxicity: Improving Reliability in Regulatory Programs) characterizing conservatism built into its reasonable potential calculations. Specifically, a report provided by Dr. Robert L. Wolpert of Duke University (see Attachment 1) found EPA's TSD approach to be extremely conservative. The report by Dr. Wolpert finds that the TSD uses either the upper 95% or 99% confidence bound on the upper 95th or 99th percentile of toxicity, respectively. EPA appears to have ignored the existence of this report in its *Draft Guidance* and provides no evidence that the conservatism built into the TSD approach is necessary to protect the environment. AMSA has always believed that the TSD approach was conservative, but work has now been brought forward better quantifying this fact. Dr. Wolpert also recommended that reasonable potential be based on the average response rather than the highest value due to the certainty in basing conclusions on single values versus those calculated from multiple observations. This is also not addressed in the *Draft Guidance*.

4. *Data Distribution Assumptions*

The TSD assumes that WET data follows a lognormal distribution, but provides little data to make this point and even this data is based on TUs which should not be used based on the recent court ruling. The *Draft Guidance* provides some text addressing the condition where data follows another distribution, but the *Draft Guidance* falls far short of enough detail. As said in other parts of these comments, states and regions will default to the TSD approach assuming a lognormal distribution because they usually do not have the expertise or resources to test this question properly. EPA must provide more detail on how to address data sets with different characteristics than those assumed in the TSD.

5. LC50 to LC1 Conversion

The TSD recommends that LC50s be converted to LC1s using a factor of 0.3 multiplied by the LC50. However, even EPA's data in the TSD shows that the LC1/LC50 ratio is greater than 0.3 ninety percent of the time. This means that over 90% of the time the TSD approach will be biased towards findings of reasonable potential and more stringent limits. EPA should allow the use of calculated LC1s if an approach is used requiring such a conversion.

6. Species Sensitivity

Federal NPDES regulation requires that species sensitivity be addressed when determining the reasonable potential to exceed or contribute to an exceedance of water quality standards. However the TSD and the *Draft Guidance* document fail to properly address this variable. Species sensitivity will be a function of the test design, test statistics as well as the species.

7. Variability Assumption

The TSD assumption that the CV of any toxics data set is 0.6 was set 15 years ago and has little relevance to the performance of POTWs today. Dischargers have become more aware of the importance of maintaining their systems and producing effluent consistent in quality. EPA should have more than enough data at this point in its program to determine a more accurate default value for WET variability given today's standards of POTW performance.

8. Uncertainty at Low Dilutions

The *Draft Guidance* attempts to address concerns for dischargers when dilution is not available and limits are to be included in a permit. This is presumably to address uncertainty associated with permit decisions under these circumstances. However, EPA failed to address this same concern during the reasonable potential determination. Given its very nature, the TSD approach is biased to findings of potential to exceed or contribute to exceedances of water quality standards. Examples of this bias include the fact that none of the multipliers in EPA's TSD are less than 1.1, and the approach uses the most toxic test in combination with the multiplier.

Translations of Narrative Criteria

Except for EPA's recommendation to use 0.3 TU_a and 1.0 TU_c, the *Draft Guidance* does not provide for translations of narrative criteria into numeric triggers or benchmarks for monitoring and permit actions. EPA proposes in its 1997 draft WET Implementation strategy to: "develop methodologies to support the development of site-specific toxicity criteria" and "develop guidance and/or regulatory language to address the appropriate and necessary elements that comprise a narrative WET standard so that it may be easily translated and implemented into the NPDES program."

AMSA understands that the TU "criteria" from the TSD only represent recommendations, but when EPA does not provide alternatives or even procedures to develop alternatives, states and regions are left without choices and will blindly adopt EPA's recommendations. Further, water quality criteria must have a frequency, magnitude and duration component, but the recommendations do not define duration of the criteria. States and EPA regions are allowed to develop their own criteria, but EPA has repeatedly warned them that criteria that are not equally protective will be rejected. Procedures to translate narrative criteria must be provided as part of this guidance.

Weight of Evidence vs. Independent Applicability

AMSA continues to believe that a weight-of-evidence approach to determining reasonable potential (RP) is appropriate for WET. Instream biological survey data demonstrating the presence or absence of adverse effect from an effluent on aquatic life use attainment – a direct measure of the environment – is superior to solely considering past, variable WET test results when determining the need for WET limitations. It is unfortunate that the *Draft Guidance* fails to offer the weight-of-evidence approach as an alternative to its RP approach rigidly based on statistics with safety factors compensating for uncertainty. We believe that the weight-of-evidence approach would provide the very kind of WET implementation flexibility to which the D.C. Circuit referred in its December 2004 decision:

> *The role of state permitting authorities … should allay the concern, which petitioners express, that the correlation between laboratory toxicity and instream impacts grows weaker at lower levels of toxicity…Individual dischargers* [should] *remain free to challenge their permits, on a case-by-case basis, if they believe that local authorities are regulating at a level that poses only a minimal risk to aquatic life.*

Furthermore, EPA's 1997 draft WET Implementation Strategy proposed to "continue evaluating the feasibility of a more integrated bioassessment program, including the use of biological assessments, WET test results, and chemical analyses in a weight-of-evidence decision-making process to assess receiving system impacts caused by effluents."

Establishment of water quality standards and determination of compliance with these standards should be based on an integration of chemical, toxicity and biological monitoring, rather than on single types of monitoring. Each monitoring type has its own advantages and uses, and different capacities to predict or indicate beneficial use impacts. Participants in the Pellston WET Workshop (1995) overwhelmingly supported that "biological assessments, WET test results and chemical analyses be used in concert for integrated decision-making."

EPA's *Draft Guidance* ignores this overwhelming support for a weight-of-evidence approach.

Hypothesis Test versus Point Estimate Endpoints

The *Draft Guidance* does not adequately address the limitations associated with the use of hypothesis test endpoints and should further promote the use of point estimates. The Society of Environmental

Toxicology and Chemistry (SETAC), using funds provided by EPA, developed WET training materials that emphasize the problems associated with the use of No-Observed-Effect-Concentrations (NOECs) and No-Observed-Adverse-Effect-Concentrations (NOAECs) in the program, particularly because these endpoints :

1. Are a function of the concentrations tested;
2. By definition can never result in a "greater than" value and therefore infer more toxicity than is present;
3. Do not reflect the level of effect measured; and
4. Cannot be mathematically manipulated as required in researching the need for limits, limit calculations and limit compliance.

EPA has stated in other documents that it prefers the use of point estimates to hypothesis test endpoints. That said, EPA must also address technical limitations of the point estimates it uses in the program. EPA proposed in its 1997 draft WET Implementation Strategy to "initiate studies to evaluate improvements for the statistical analysis of toxicity test data…[and] enhance the current statistical approaches for both hypothesis testing and point estimate models and approaches to quantify the confidence around test endpoints to improve interpretation of WET test results."

Specifically, the IC25 has a number of issues associated with it that EPA recognizes but failed to address in its method rulemaking. This includes the IC program's failure to use all data available to calculate an endpoint, its data smoothing procedure which artificially increases the control response and decreases the IC, and its inappropriate use of dose response curves that are irregular leading to erroneous IC calculations. EPA should revisit its draft implementation strategy to determine what needs to be added to the *Draft Guidance*.

In many cases the IC calculation is misapplied to data that do not fit the assumptions of the model (monotonicity; piecewise linear response function; and random, independent, and representative sample). If data from a test violate these assumptions, the point estimates may be invalid. Furthermore, the issue of hormesis should be addressed. The model pools control and test concentrations that show stimulation and calculates an elevated response, increasing the chances that a WET test will fail. EPA should allow new curve fitting software to be used to generate does-response curves and resulting IC25 values that do not penalize permittiees by artificially increasing the response of the controls to make the statistics work.

Test Endpoint Uncertainty

The *Draft Guidance* does not address uncertainty in test endpoints within the context of program activities such as reasonable potential and compliance determinations. At the 1995 Pellston Workshop it was stated that "confidence limits should be considered if point estimates are used" and that regarding WET variability, "(t)his uncertainty must therefore be considered in the test result interpretation and incorporated into the regulatory decision-making process." An advantage of using point estimates is that "ECp values and their associated confidence intervals can be compared among multiple experiments or among laboratories. This capability facilitates quantification of intertest or interlaboratory performance, intratest QA/QC control, and evaluation of multiple-sample toxicity test results."

EPA staff attending the Pellston Workshop stated that "we have not been particularly good at incorporating statistical variability of test data into the decision-making process for effluent toxicity" and that "it is important that we incorporate an understanding of all sources of statistical error into our decision-making process."

Clearly the attendees of the Workshop, which included EPA staff and members of several state agencies, recognized the importance of adequately accounting for the uncertainty in test endpoints. Even the recent court ruling stated that EPA has to address uncertainty in test results in its implementation program and EPA itself states that uncertainty in test endpoints is +/- 100% (EPA WET methods). However, EPA's proposal to address this uncertainty (i.e., more testing) assumes that the tests as designed and conducted accurately reflect conditions instream. The comments provided herein should provide sufficient evidence that this is not the case. More testing under these circumstances does not address test uncertainty, it only allows one to reproduce erroneous conclusions. This *Draft Guidance* should take the recommendations made by the WET Pellston Workshop and its own staff and provide guidance that directly addresses uncertainty in test endpoints.

WET Data Validity and Representativeness

In Section 4.1.2 EPA states that "[v]alid and representative data should not be ignored." The converse of that principle is that invalid data should not be used in making decisions. However, the *Draft Guidance* seems to contradict that principle, when it states that "[t]he permitting authority may require additional information, for example, results of tests determined to be invalid for any reason (e.g., too many control organisms in the test die)." If too many control organisms die, that test is obviously not one that should be considered in making WET permitting decisions; it says nothing about whether the discharge is toxic. Test results that are determined to be invalid should simply not be considered.

In Section 4.14, the *Draft Guidance* states, "If an effluent is known to contain residual chlorine at levels which may result in unacceptable toxicity instream, *in situ* testing is recommended." We believe that the *Draft Guidance* should clarify that *in situ* WET testing would be redundant and is not necessary wherever chemical monitoring is performed to determine the adequacy of effluent dechlorination.

Program Quality Assurance (QA)

Quality assurance (QA) for the program is still lacking or problematic, particularly with lab accreditation, analyst certification, methods to address data representativeness and comparability, and the WET discharge monitoring report quality assurance (DMR QA) program. QA will affect data variability that in turn impacts whether a discharger receives a limit, the magnitude of the limit, and whether the discharger can comply with the permit limit.

Lab accreditation was determined to be critical during the 1996 WET Implementation Stakeholder meeting. Efforts such as the National Environmental Laboratory Accreditation Conference (NELAC) barely go beyond the requirements of the methods, effecting little to no improvements for the current program. The *Draft Guidance* does attempt to address data representativeness by discussing the importance of sampling but it fails to recognize that even the best efforts to collect representative data by POTWs can be negated by factors that cannot be controlled by POTWs (system failures, inclement weather, illegal discharges noted after the test is completed, etc.). There should be a procedure to reject outliers in a dataset, particularly since the TSD reasonable potential approach uses the test exhibiting the most toxicity to drive decisions.

EPA attempted to address intra-test precision issues by including MSD limits for some chronic test endpoints, but the criteria used to develop these limits were arbitrary at best. EPA is aware of WERF's studies (Project 95-PQL, Whole Effluent Toxicity Testing Methods: Accounting for Variance, 1999, and Project 00-ECO-1, referenced above) of variables that impact test results and in some cases test results are highly dependent on which lab conducts the test or when the test was conducted, rather than a function of the test concentrations and effluent quality. Therefore data comparability cannot be assumed.

AMSA has previously provided comments emphasizing the need for a way to compare data across labs through the WET test method rulemaking process. EPA's interlaboratory study showed that the same results are not attained in labs across the country, but despite EPA's own QA guidance documents there is no way to satisfy this data quality objective within the current program. Perhaps the best way to do this is for EPA to select a number of reference toxicants for its program, based on a number of variables (reproducibility, safety in use, type and pathway for effect, etc.), and develop acceptance limits for each toxicant, test type, duration, temperature, species and biological endpoint using measures other than confidence intervals for LC50s, NOECs or IC25s. The spread of data using traditional endpoints, as can be seen by reviewing typical DMR QA results, is too wide to gauge lab performance as acceptable or not. EPA must consider using other measures for the DMR QA program or similar efforts for lab qualification to be meaningful. Examples of variables that could be measured are the MSD and the presence of a dose-concentration relationship. The EPA interlab study showed that labs make frequent and sometimes significant errors when calculating results. The DMR QA program could also test labs' abilities to correctly calculate test results by providing data sets to the labs and requesting that they provide the results. These results could then be compared to the correct answers.

Regardless of EPA's response to this concern the Agency must recognize that the current DMR QA program does not go far enough. The width of acceptance limits for each test and endpoint almost guarantees that labs will comply with the program, regardless of the quality of the lab. This unfairly equates the best and worst labs. The DMR QA program must be overhauled or abandoned. AMSA recommends that EPA review its own data quality objective (DQO) guidance documents and resolve which data quality indicators (DQIs) and measurement quality objectives (MQOs) are necessary to bring the WET implementation program in line with that used for chemical-specific parameters.

Magnitude of Toxicity and Implementation

The *Draft Guidance* does not address magnitude of toxicity as it relates to compliance and triggered actions like accelerated testing and TIEs. TIEs have limitations (for example, need LC50 or NOEC/IC25 < 80%), but EPA has not acknowledged or incorporated these limitations into their guidance. Minor exceedances of triggers or limits cannot be successfully followed with a TIE to identify and remove toxicants. The Draft Guidance should be modified to alert regulatory agencies to the technical limitations of TIEs and TREs and protect permittees from unnecessarily spending public funds.

Specific Comments

Limits without Data

The *Draft Guidance* continues to support the concept of limits based purely on assumptions and without data. This is totally unacceptable and should be deleted from the Draft Guidance. Perhaps 15 years ago when the TSD was written there may not have been enough data to start the program, but this is not the case today. The program's assumptions are far too conservative, leading to unnecessary limits or unnecessarily stringent limits. Limits without data will even be worse. Further, it is hard to understand how EPA can hold in one part of the document that more data is necessary to make decisions and in another part state that it is acceptable to make the same decisions without any data.

Median Limit of 1.0 TU

AMSA appreciates the fact that the *Draft Guidance* has provided an option of expressing the monthly chronic WET limit as a median rather than as an average in low-flow dilution situations. We believe this reflects a correct recognition by the Agency that, as a numeric WET limitation approaches 1.0 chronic toxicity units (TU_c), the probability of the WET limitation being exceeded by a single WET test result greatly increases independent of any real effluent toxicity. Because this probability is a function of the numeric magnitude of the effluent limitation rather than receiving water conditions, and because WET limitations can be established this low also for policy or other reasons, we urge the Agency to expand this option's availability beyond low-flow dilution situations to include all situations where chronic numeric WET limitations are as low. For example, there are many other dischargers who have limited dilution but not enough to address the uncertainty of making regulatory decisions. A discharger with 3:1 acute dilution will still receive a 1.0 TU acute limit. EPA should consider those dischargers that fall in between the scenarios where dilution is unavailable or low and those that have enough dilution to compensate for the conservatism and uncertainty in the permitting process. One way to do this is to base the approach on the uncertainty in test results, which EPA claims is +/- 100% (EPA WET methods). If this is the case then perhaps the median limit should be applied to those whose wasteload allocations are less than 2.0 TUs (a measurement of 2.0 TUs could actually be as low as 1.0 TUs given uncertainty in test results).

For the median option to be beneficial as intended, at least three chronic WET tests would need to be performed within the averaging period. Conducting this many chronic WET tests within one month can pose a virtually insurmountable challenge to some laboratories, considering the difficulties associated with scheduling additional, consecutive or even overlapping tests while maintaining sufficient, viable populations of test organisms. To ease these laboratories' burden and make the justified median option more practicable, we urge the Agency to provide that the averaging period for compliance with a low chronic WET limitation may be extended to beyond one month's duration. AMSA recommends that EPA extend the averaging period for tests to 3-month periods to allow enough time for dischargers and labs to schedule tests when necessary, and consider the use of a running average concept (the evaluation would look back over the last three months, present month included). Also helpful in this regard would be allowing, when applying either an average or median as the basis for compliance, the use of different numbers of tests for different species (e.g., three tests for *C. dubia* and one test for *P. promelas*) within the averaging period. Third, the Agency should recognize that alleviating the burden associated with performing multiple WET tests by using more than one qualified laboratory may be necessary under such circumstances and is appropriate.

Furthermore, AMSA questions the *Draft Guidance*'s prerequisite condition that the permittee must have conducted three-species screening to be eligible for use of the median option. We note that the *Draft Guidance* has recommended three-species screening for WET testing more generally to ensure that the most sensitive species are used for routine testing. We are aware of no reason for an additional, permittee-specific requirement for three-species screening where the permitting authority has already determined that routine testing of the two species selected is protective. AMSA is not aware of any consideration specific to using the median for determining compliance that would necessitate such an additional requirement.

Maximum Limit of 1.6 TU

EPA is proposing a maximum limit of 1.6 TUs for dischargers with zero instream dilution. However this limit ignores the uncertainty in test results recognized by EPA to be +/- 100%. Therefore the maximum limit should 2.0 TUs since the confidence limits around such a result include a limit of 1.0 TUs. Additionally, dischargers with limited dilution should also be allowed to use this option because even limited dilution can result in a daily limit of 1.0 TU which the Draft Guidance was intending to address.

Accelerated Testing

The *Draft Guidance* is requiring that 6 additional tests be conducted following exceedance of a trigger or limit to determine if a TIE/TRE should be initiated. This approach might be defensible if magnitude of exceedance and other sources of conservatism were addressed in developing those triggers and limits, but this is not the case. This approach assumes that an exceedance of a trigger or limit actually represents a significant potential for impact instream. Based on quotes provided elsewhere in these comments from the Pellston Workshop of 1995 and conclusions of the 1997 WET Implementation Stakeholder meeting this assumption is invalid in many cases. The *Draft Guidance* continues to support an approach that concludes impact without addressing certainty in that conclusion.

Frequency of Testing

The *Draft Guidance* is recommending that more tests be conducted to determine reasonable potential and the need for a TIE/TRE. Superficially this seems like a good idea; more data provides more information and addresses uncertainty in some assumptions. However more data used improperly will only lead to inappropriate permit decisions and actions. As stated throughout these comments the current implementation program has significant flaws, and most have not been addressed by EPA in the TSD or in this *Draft Guidance*. EPA must address the conservatism in their program and determine if it is justified by its uncertainty. Where uncertainty is low, less testing is required; where uncertainty is high, higher testing frequencies are necessary.

The net effect of this *Draft Guidance* is clear – the flaws in the reasonable potential determination will result in virtually all dischargers ending up with WET limits in their permits. From there, every single test failure (lethal and sublethal) will be a permit violation subject to EPA enforcement discretion. AMSA had hoped that some of its long-standing issues with WET implementation would be addressed in this *Draft Guidance* or that EPA would consider an alternative approach to implementing chronic WET limits in NPDES permits. Unfortunately, AMSA now believes that this Draft Guidance will only result in states and regions more consistently using the flawed RP approach outlined in the TSD and this *Draft Guidance* and further complicate WET implementation.

Again, AMSA is committed to meeting with the Agency to discuss how these implementation issues can be resolved. We hope that the Agency will seek out stakeholders, including AMSA, following the close of the comment period to gather additional insight into the types of approaches that are currently being used and currently working and how the *Draft Guidance* should embrace those approaches.

Sincerely,

Chris Hornback
Director, Regulatory Affairs

ATTACHMENT

ATTACHMENT 1

Concentration Limits

Robert L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Durham, NC 27708-0251, USA

Revised November 11, 2002

Summary

A statistical methodology is presented in EPA Technical Support Document (1991) to "project an estimated maximum concentration" for an effluent, based on a small number of concentration measurements. This method is shown here to be less efficient and more variable than proposed alternative methods, leading to unnecessarily high rates of "false positives" (false declarations that concentrations exceed acceptable limits).

Simple and efficient alternatives are proposed here, and the performance of both the current EPA method and the proposed alternatives are illustrated with synthetic data based on the probability distribution the EPA assumes for actual effluents.

The current EPA method is "doubly conservative" in that it attempts to offer a 95%-confident upper bound for C_{99}, the 99%-percentile of concentration measurements. The red vertical bar in Fig. 1 represents C_{99} (with a numerical value of 10.24 in our illustration); the area to its left under the black curve represents the 99% probability that each individual concentration measurement falls below C_{99}, while the areas to its right under the red, blue, and green curves represent the 95% probability that the estimate $\hat{C}_{99}^{\text{EPA}}$ will exceed C_{99} with samples of size 5, 10, and 100, respectively. This double conservatism leads to extraordinarily conservative upper bounds that exceed actual concentrations by orders of magnitude.

The EPA's attempt to be doubly-conservative in fact fails, due to their inefficient use of sample data to estimate concentration variability and a failure to reflect this variability. We offer two alternatives to the EPA approach: one that, like the EPA method, would be doubly-conservative if the

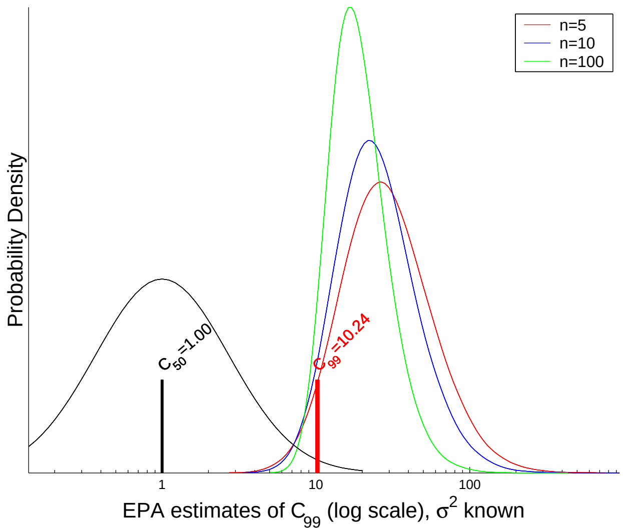

Figure 1: Concentration measurements (black) and EPA bounds $\hat{C}_{99}^{\mathrm{EPA}}$ for C_{99} (vertical red bar) with sample-sizes of $n = 5, 10, 100$ (red,blue,green, respectively).

concentration variability were known without error, but which makes more efficient use of sample data to get more realistic upper bounds; and another, that attains the goal of double conservatism even when the variability is unknown. We would argue however that the intended double conservatism is unwarranted and leads to gross overestimates of the concentration quantiles without offering a commensurate improvement in public safety, and that it would be more appropriate simply to estimate C_{99} as accurately as possible.

1 Introduction

In EPA Technical Support Document (1991, *p.* 52 and Box 3-2) a statistical methodology is presented to "project an estimated maximum concentration" for an effluent, based on a small number n of concentration measurements $\{X_1, ..., X_n\}$. The "maximum concentration" is defined to be the 99%-percentile of the concentration distribution, *i.e.*, the number C_{99} with the property that $\Pr[X_i > C_{99}] = 0.01$ for each concentration measurement X_i. The EPA method is intended to be *conservative* in the sense that, with high probability, it offers a upper bound for C_{99}— the "projected estimate"

$\hat{C}_{99}^{\text{EPA}} \equiv \hat{C}_{99}^{\text{EPA}}(X_1, ..., X_n)$ has only a small probability α of understating the true value of C_{99} (both $\alpha = 1\%$ and $\alpha = 5\%$ are considered), so that $\Pr[\hat{C}_{99}^{\text{EPA}} < C_{99}] \leq \alpha$. The method assumes that the concentration measurements X_i have independent lognormal distributions, an assumption justified by empirical evidence presented in (EPA Technical Support Document, 1991, Appendix E).

We present here a critique of this methodology and an alternative estimator $\hat{C}_{99}^{\text{ALT}} \equiv \hat{C}_{99}^{\text{ALT}}(X_1, ..., X_n)$ which is less variable, more robust, and still satisfies the doubly-conservative requirement $\Pr[\hat{C}_{99}^{\text{ALT}} < C_{99}] \leq \alpha$.

The random variables X_i are independent lognormal $X_i \sim \mathsf{LN}(\mu, \sigma)$ if their natural logarithms have independent normal distributions with mean μ and variance σ^2, $\ln X_i \sim \mathsf{No}(\mu, \sigma^2)$. Their quantiles are related to those of the standard normal distribution by the relation

$$\begin{aligned} F(x) &\equiv \Pr[X \leq x] \\ &= \Pr\left[\frac{\ln X - \mu}{\sigma} \leq \frac{\ln x - \mu}{\sigma}\right] \\ &= \Phi\big((\ln x - \mu)/\sigma\big), \end{aligned}$$

where $\Phi(z)$ is the Cumulative Distribution Function (CDF) for the standard Normal distribution and, in particular, the "maximum concentration" satisfying $F(C_{99}) = 0.99 = \Phi(2.326)$ can be expressed in the form

$$C_{99} = \exp(\mu + 2.326\,\sigma)$$

as a function of the (in general, unknown) parameters μ and σ.

2 The Box 3-2 Approach

The approach of (EPA Technical Support Document, 1991, Box 3-2) (henceforth, "the Box 3-2 approach") to identifying $C_{99} = \exp(\mu + 2.326\,\sigma)$ depends on just two features of the data set:

- The maximum observed concentration $X_n^* \equiv \max\{X_1, ..., X_n\}$, and
- The empirical coefficient of variation, $\widehat{\mathsf{CV}} \equiv S/\bar{X}$.[1]

[1] Here $\bar{X} = \frac{1}{n}\sum X_i$ and $S = \sqrt{\frac{1}{n-1}\sum(X_i - \bar{X})^2}$ are data-based estimates of the mean and standard deviation of a very long string of concentration measurements at a site.

These statistics X_n^* and $\widehat{\mathsf{CV}}$ do not reflect all the evidence in the data about $C_{99} = \exp(\mu + 2.326\,\sigma)$ (the technical term is that they are not *sufficient statistics* for the lognormal distribution), so any estimator of C_{99} based on them will be inefficient (more variable and less precise than necessary); see Section 3 for more details. The EPA bases its estimates of C_{99} on percentiles of the maximum concentration measurement $X_n^* \equiv \max\{X_1, ..., X_n\}$, which are related to those of the individual concentrations X_i as follows.

For the lognormal distribution, the population Coefficient of Variation $\mathsf{CV}[X] \equiv \mathsf{V}[X]^{1/2}/\mathsf{E}[X]$ (the standard deviation divided by the mean) is given by $\mathsf{CV}[X]^2 = \exp(\sigma^2) - 1$, so the parameter σ^2 may be estimated from the empirical coefficient of variation $\widehat{\mathsf{CV}} \equiv S/\bar{X}$ as

$$\sigma^2 \approx \ln\left(1 + \widehat{\mathsf{CV}}^2\right) = \ln\left(1 + S^2/\bar{X}^2\right).$$

Under the assumed independence of concentration measurements the maximum observed concentration $X_n^* \equiv \max\{X_1, ..., X_n\}$ has CDF

$$\begin{aligned} F_n^*(x^*) &\equiv \Pr[X_n^* \le x^*] \\ &= \Pr[X_1 \le x^*, ..., X_n \le x^*] \\ &= \Pr[X_1 \le x^*]^n \\ &= F(x^*)^n \\ &= \Phi\big((\ln x^* - \mu)/\sigma\big)^n. \end{aligned}$$

There is a 5% chance that the observed maximum concentration X_n^* will fall below its 5%-percentile, the point $x_{.05}$ with $F_n^*(x_{.05}) = .05$. Since $F_n^*(x) = F(x)^n$ for every x and n, there must also be a 5% chance that $\Phi\big((\ln X_n^* - \mu)/\sigma\big) < .05^{1/n}$; for example, with $n = 5$, there is a 5% chance that $\Phi\big((\ln X_n^* - \mu)/\sigma\big) < .05^{1/5} = .5493 = \Phi(0.1238)$, so $\Pr[X_n^* < \exp(\mu + 0.1238\,\sigma)] = 0.05$. Since $C_{99} = \exp(\mu + 2.3263\,\sigma)$ is also the exponential of $\mu + x\sigma$ for some number x, we can multiply both sides by an appropriate constant to achieve an estimated upper bound of

$$\begin{aligned} 0.05 &= \Pr[X_n^* < \exp(\mu + 0.1238\,\sigma)] \\ &= \Pr[X_n^* \exp(2.2025\,\sigma) < \exp\left(\mu + (0.1238 + 2.2025)\,\sigma\right)] \\ &= \Pr[X_n^* \exp(2.2025\,\sigma) < \exp\left(\mu + 2.3263\,\sigma\right)] \\ &= \Pr[\hat{C}_{99}^{\text{EPA}} < C_{99}], \quad \text{with} \quad \hat{C}_{99}^{\text{EPA}} \equiv X_n^* \exp(2.2025\,\sigma), \end{aligned}$$

giving a conservative upper bound for the maximum concentration. For values of n from one to ten, the same approach with $\alpha = .05$ and $\alpha = .01$ leads to the similar estimates given in Table 1 below.

$\alpha = .05$	$\alpha = .01$
$\Pr[\exp(3.9712\,\sigma)X_1^* < C_{99}] = \alpha$	$\Pr[\exp(4.6527\,\sigma)X_1^* < C_{99}] = \alpha$
$\Pr[\exp(3.0864\,\sigma)X_2^* < C_{99}] = \alpha$	$\Pr[\exp(3.6079\,\sigma)X_2^* < C_{99}] = \alpha$
$\Pr[\exp(2.6624\,\sigma)X_3^* < C_{99}] = \alpha$	$\Pr[\exp(3.1140\,\sigma)X_3^* < C_{99}] = \alpha$
$\Pr[\exp(2.3944\,\sigma)X_4^* < C_{99}] = \alpha$	$\Pr[\exp(2.8046\,\sigma)X_4^* < C_{99}] = \alpha$
$\Pr[\exp(2.2025\,\sigma)X_5^* < C_{99}] = \alpha$	$\Pr[\exp(2.5846\,\sigma)X_5^* < C_{99}] = \alpha$
$\Pr[\exp(2.0549\,\sigma)X_6^* < C_{99}] = \alpha$	$\Pr[\exp(2.4163\,\sigma)X_6^* < C_{99}] = \alpha$
$\Pr[\exp(1.9361\,\sigma)X_7^* < C_{99}] = \alpha$	$\Pr[\exp(2.2813\,\sigma)X_7^* < C_{99}] = \alpha$
$\Pr[\exp(1.8371\,\sigma)X_8^* < C_{99}] = \alpha$	$\Pr[\exp(2.1694\,\sigma)X_8^* < C_{99}] = \alpha$
$\Pr[\exp(1.7528\,\sigma)X_9^* < C_{99}] = \alpha$	$\Pr[\exp(2.0743\,\sigma)X_9^* < C_{99}] = \alpha$
$\Pr[\exp(1.6795\,\sigma)X_{10}^* < C_{99}] = \alpha$	$\Pr[\exp(1.9920\,\sigma)X_{10}^* < C_{99}] = \alpha$

Table 1: Formulas to implement Box 3-2 method for $n =$ 1–10.

The "Box 3-2 approach" is to approximate σ^2 from the empirical coefficient of variation, as above, and use that with Table 1 the table above to achieve a 95% (or 99%) conservative upper bound for the maximum concentration C_{99}.

3 A Simple and More Efficient Approach

In this section we develop an alternative upper bound $\hat{C}_{99}^{\text{ALT}}$ for C_{99} that, like the Box 3-2 approach, is "doubly conservative" in providing a 95%-confident upper bound for the 99%-percentile of the concentration distribution, so that $\Pr[\hat{C}_{99}^{\text{ALT}} < C_{99}] \leq \alpha = 0.05$.

The Maximum Likelihood Estimators (MLE's) for μ and σ^2 are simply the sample mean and variance of the log concentration measurements,

$$\hat{\mu} \equiv \frac{1}{n}\sum \ln X_i \qquad\qquad \hat{\sigma}^2 \equiv \frac{1}{n}\sum(\ln X_i - \hat{\mu})^2.$$

These statistics are *sufficient* for the lognormal distribution (see Bickel and Doksum, 2001, *p.* 41), so by the Raô-Blackwell result (*op cit.*, *p.* 45) any estimator of any feature of the distribution (such as the Box 3-2 projected upper bound $\hat{C}_{99}^{\text{EPA}}$ for C_{99}) that does not depend on the data only through $\hat{\mu}$ and $\hat{\sigma}^2$ can be improved by replacing it with its conditional expectation, given $\hat{\mu}$ and $\hat{\sigma}^2$. Note that the Box 3-2 approach does *not* depend only on these sufficient statistics— instead, it summarizes the data by the sample maximum X_n^* and the empirical coefficient of variation $\widehat{\mathsf{CV}} \equiv S/\bar{X}$, and so does not take full advantage of the information in the data. This makes it more variable and less reliable than other conservative upper bounds on C_{99}.

If σ^2 were known exactly (the conditions under which the Box 3-2 approach is justified as a 95% upper bound), then MLE for the Maximum Concentration C_{99} would be

$$\widehat{C_{99}} \equiv \exp\big(\hat{\mu} + 2.326\,\sigma\big).$$

The average log-concentration $\hat{\mu}$ has a normal distribution $\hat{\mu} \sim \mathsf{No}(\mu, \sigma^2/n)$, so a conservative 95% upper bound $\Pr[C_{99} \le \hat{C}_{99}^{\text{ALT}}] = 0.95$ is given by

$$\hat{C}_{99}^{\text{ALT}} \quad \equiv \quad \exp\big(\hat{\mu} + (2.326 + 1.645/\sqrt{n})\sigma\big)$$

For large enough n the variance σ^2 is well approximated by its MLE $\hat{\sigma}^2$ (better than the less-reliable Box 3-2 estimate $\ln(1 + \widehat{\mathsf{CV}}^2)$) so replacing σ with its estimate $\hat{\sigma}$ in this formula still gives an approximate bound; for smaller n we may use the known probability distribution of the estimator $\hat{\sigma}^2$ to construct a valid 95% upper bound for C_{99} (see Section 5). If prior experience or data from related sources are available, Bayesian methods can be employed to construct even better solutions.

4 Numerical Illustration

Below we summarize results from 100,000 replications of drawing samples of size 5, 10 and 100 from a $X_i \sim \mathsf{LN}(\mu = 0,\ \sigma^2 = 1)$ distribution, with exact median ("typical"), mean and maximum concentration measurements

$$C_{50} = e^{\mu} = 1.0000 \quad \mathsf{E}[X_i] = e^{\mu+\sigma^2/2} = 1.6487 \quad C_{99} = e^{\mu+2.326\sigma} = 10.2405.$$

The statistical inefficiency of the Box 3-2 method is revealed in Table 2 and Fig. 2 by its high probability of gross overestimates, while the MLE method is fairly efficient, with moderate overshoot for small trials (100% with $n = 5$, 50% with $n = 10$) and negligible overshoot for trials with $n = 100$ concentration measurements.

Even with $n = 100$ measurements the EPA method's estimate is more than double the true C_{99} about 40% of the time and, with $n = 5$ concentration measurements, is nearly four times the true 99%-percentile on average.

Box 3-2 Method

n	$\mathsf{P}[\hat{C}_{99}^{\mathrm{EPA}} > 2 \cdot C_{99}]$ (Gross Overshoot)	$\mathrm{Md}[\hat{C}_{99}^{\mathrm{EPA}}]$ (Median)	$\mathsf{E}[\hat{C}_{99}^{\mathrm{EPA}}]$ (Mean)	$\mathsf{P}[\hat{C}_{99}^{\mathrm{EPA}} < C_{99}]$ (Should be 0.05)
5	0.684	27.90	36.71	0.050
10	0.608	23.89	30.02	0.052
100	0.390	18.17	21.09	0.049
Exact	0.000	10.24	10.24	0.050

Alternate MLE Method

n	$\mathsf{P}[\hat{C}_{99}^{\mathrm{ALT}} > 2 \cdot C_{99}]$ (Gross Overshoot)	$\mathrm{Md}[\hat{C}_{99}^{\mathrm{ALT}}]$ (Median)	$\mathsf{E}[\hat{C}_{99}^{\mathrm{ALT}}]$ (Mean)	$\mathsf{P}[\hat{C}_{99}^{\mathrm{ALT}} < C_{99}]$ (Should be 0.05)
5	0.535	21.29	23.60	0.050
10	0.290	17.18	18.06	0.051
100	0.000	12.07	12.13	0.050
Exact	0.000	10.24	10.24	0.050

Table 2: Performance of EPA and MLE methods on synthetic data with σ^2 known.

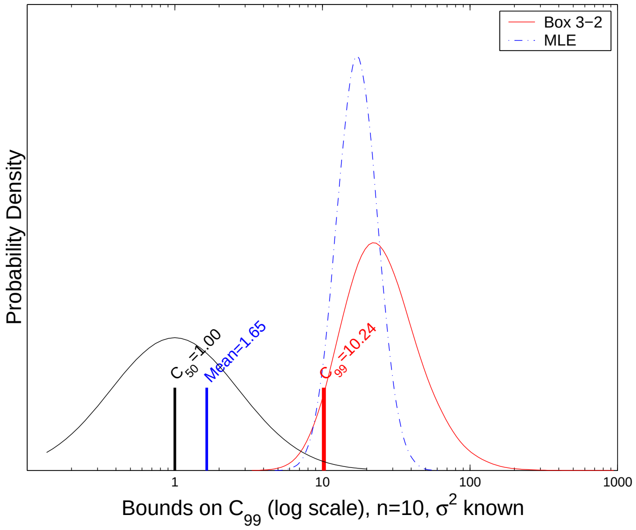

Figure 2: Bounds for C_{99} for sample-size $n = 10$.

The mean is much higher than the median for the Box 3-2 method, indicating highly skewed values with high probability of dramatically exceeding C_{99}, while the mean and median are closer together for the MLE, indicating less of a problem with skewness. The last column shows that both methods achieve the intended 95% conservatism, giving a upper bound for $C_{99} = 10.24$ in approximately 95% of the 100,000 simulated trials, under the assumption of known variance σ^2.

5 Failure of EPA Method with Estimated σ^2

Table 4 shows similar (and more realistic) simulation results with σ^2 treated as unknown, and estimated from the data using $\sigma^2 \approx \ln\left(1 + \widehat{\mathsf{CV}}^2\right)$ as the Box 3-2 method specifies. The right-most column shows that the Box 3-2 method failed to be even close to 95% conservative, leading to an upper bound for C_{99} for only about 60–70% of the simulated trials for sample sizes $n = 5$–10, rather than the promised 95%. Table 4 also shows exceedence probabilities for an extension (described below) of the Alternate MLE approach introduced above to accommodate estimated variance; note that this method does achieve double conservatism (*i.e.*, the probabilities of failing to bound C_{99} given in the right-most column are all below 5.0%).

The MLE approach can be extended to the case of unknown σ^2, as follows. Recall that the sufficient statistics $\hat{\mu} = (1/n)\sum \ln X_i$ and $\hat{\sigma}^2 = (1/n)\sum(\ln X_i - \hat{\mu})^2$ are independent, with the $\mathsf{No}(\mu, \sigma^2/n)$ and $\frac{\sigma^2}{n}\chi^2_{n-1}$ distributions, respectively, so $Z \equiv \sqrt{n}(\mu - \hat{\mu})/\sigma$ and $Y \equiv n\hat{\sigma}^2/\sigma^2$ have independent $\mathsf{No}(0,1)$ and χ^2_{n-1} distributions (see Bickel and Doksum, 2001, *p.* 495). It follows that $C_{99} = \exp(\mu + 2.326\sigma)$ satisfies

$$\begin{aligned}
0.95 &= \Pr[C_{99} \le \exp(\hat{\mu} + r\hat{\sigma})] \\
&= \Pr[\mu + 2.326\sigma \le \hat{\mu} + r\hat{\sigma}] \\
&= \Pr\left[\frac{\mu - \hat{\mu}}{\sigma/\sqrt{n}} + 2.326\sqrt{n} \le r\sqrt{n}(\hat{\sigma}/\sigma)\right] \\
&= \Pr\left[\frac{Z + 2.326\sqrt{n}}{\sqrt{Y/(n-1)}} \le r\sqrt{n-1}\right] \\
&= \mathrm{nct}(r\sqrt{n-1},\, n-1,\, 2.326\sqrt{n}),
\end{aligned}$$

the CDF of the noncentral t distribution with $n-1$ degrees of freedom and noncentrality parameter $2.326\sqrt{n}$. Thus even when σ^2 is unknown we have conservative upper bounds of the form

$$0.05 = \Pr[\exp(\hat{\mu} + r_{95}\hat{\sigma}) < C_{99}], \qquad 0.01 = \Pr[\exp(\hat{\mu} + r_{99}\hat{\sigma}) < C_{99}]$$

where $r_{95} \equiv \text{nctinv}(0.95,\ n-1,\ 2.326\sqrt{n})/\sqrt{n-1}$ and $r_{99} \equiv \text{nctinv}(0.99,\ n-1,\ 2.326\sqrt{n})/\sqrt{n-1}$ are based on the 95%-percentile and 99%-percentiles of the noncentral t distribution, respectively.

n	$\alpha = .05$	$\alpha = .01$
2	$\Pr[\exp(\hat{\mu} + 37.09\hat{\sigma}) < C_{99}] = \alpha$	$\Pr[\exp(\hat{\mu} + 185.59\hat{\sigma}) < C_{99}] = \alpha$
3	$\Pr[\exp(\hat{\mu} + 10.55\hat{\sigma}) < C_{99}] = \alpha$	$\Pr[\exp(\hat{\mu} + 23.89\hat{\sigma}) < C_{99}] = \alpha$
4	$\Pr[\exp(\hat{\mu} + 7.04\hat{\sigma}) < C_{99}] = \alpha$	$\Pr[\exp(\hat{\mu} + 12.39\hat{\sigma}) < C_{99}] = \alpha$
5	$\Pr[\exp(\hat{\mu} + 5.74\hat{\sigma}) < C_{99}] = \alpha$	$\Pr[\exp(\hat{\mu} + 8.94\hat{\sigma}) < C_{99}] = \alpha$
6	$\Pr[\exp(\hat{\mu} + 5.06\hat{\sigma}) < C_{99}] = \alpha$	$\Pr[\exp(\hat{\mu} + 7.33\hat{\sigma}) < C_{99}] = \alpha$
7	$\Pr[\exp(\hat{\mu} + 4.64\hat{\sigma}) < C_{99}] = \alpha$	$\Pr[\exp(\hat{\mu} + 6.41\hat{\sigma}) < C_{99}] = \alpha$
8	$\Pr[\exp(\hat{\mu} + 4.35\hat{\sigma}) < C_{99}] = \alpha$	$\Pr[\exp(\hat{\mu} + 5.81\hat{\sigma}) < C_{99}] = \alpha$
9	$\Pr[\exp(\hat{\mu} + 4.14\hat{\sigma}) < C_{99}] = \alpha$	$\Pr[\exp(\hat{\mu} + 5.39\hat{\sigma}) < C_{99}] = \alpha$
10	$\Pr[\exp(\hat{\mu} + 3.98\hat{\sigma}) < C_{99}] = \alpha$	$\Pr[\exp(\hat{\mu} + 5.07\hat{\sigma}) < C_{99}] = \alpha$

Table 3: Formulas to implement MLE method with unknown σ^2 for $n = 2$–10.

Table 4 and Figs. 3 and 4 show the results of a numerical simulation. Even for large sample sizes the Box 3-2 method fails to meet its conservative goal of a 95% bound when (as in usual practice) σ^2 is unknown, despite exceeding $2 \cdot C_{99}$ nearly 40% of the time— the area to the left of C_{99} (given in the right-most column of Table 4) exceeds 0.05 by far. The alternate method introduced here does meet its 95% goal, with minimal overshoot for large samples, but its faithful reflection of the uncertainty about σ^2 in small samples leads it to overstate the true value of C_{99} for sample sizes as small as $n = 5$ or $n = 10$.

The EPA Technical Support Document (1991, Box 3-2) recommends ignoring sample data about σ^2 when $n < 10$ and, instead, using the nominal values of 0.6 for $\widehat{CV}$ (this implies a value of $\ln(1+0.6^2) = 0.3075$ for σ^2). If that nominal value is at least as large as the true value of σ^2 then the "known σ^2" results of Table 2 apply, with minimal overshoot for $\hat{C}_{99}^{\text{ALT}}$; if not, as in our numerical example (where $\sigma^2 = 1.0 > 0.3075$), then $\hat{C}_{99}^{\text{EPA}}$ will again fail to be conservative and the EPA method will not deliver honestly its stated goal of double conservatism.

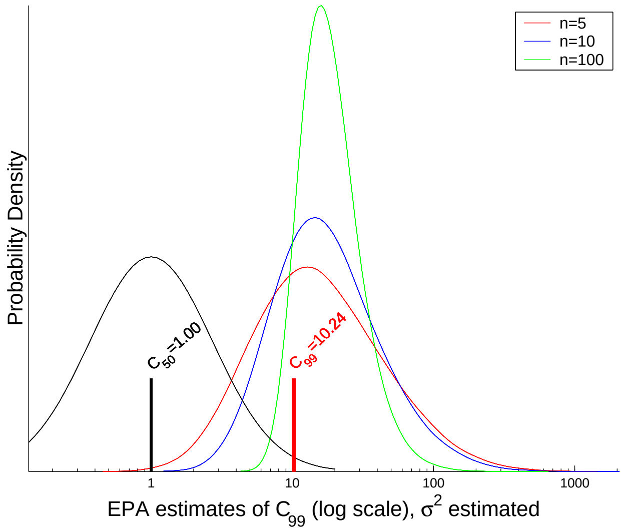

Figure 3: Failure of EPA method's conservatism for estimated σ^2.

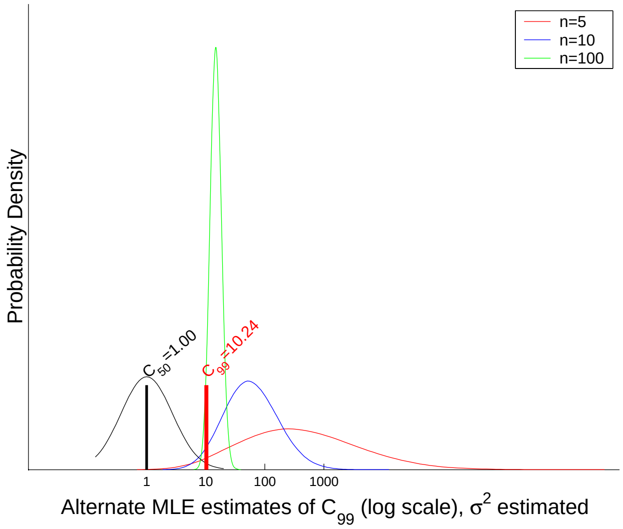

Figure 4: Success of Alternate method's conservatism for estimated σ^2.

Box 3-2 Method, σ^2 estimated

n	$\mathsf{P}[\hat{C}_{99}^{\text{EPA}} > 2 \cdot C_{99}]$	$\text{Md}[\hat{C}_{99}^{\text{EPA}}]$	$\mathsf{E}[\hat{C}_{99}^{\text{EPA}}]$	$\mathsf{P}[\hat{C}_{99}^{\text{EPA}} < C_{99}]$
	(Gross Overshoot)	(Median)	(Mean)	(Should be 0.05)
5	0.372	14.45	28.04	0.368
10	0.397	16.34	26.90	0.279
100	0.375	17.58	21.18	0.084
Exact	0.000	10.24	10.24	0.050

Alternate MLE Method, extended for estimated σ^2

n	$\mathsf{P}[\hat{C}_{99}^{\text{ALT}} > 2 \cdot C_{99}]$	$\text{Md}[\hat{C}_{99}^{\text{ALT}}]$	$\mathsf{E}[\hat{C}_{99}^{\text{ALT}}]$	$\mathsf{P}[\hat{C}_{99}^{\text{ALT}} < C_{99}]$
	(Gross Overshoot)	(Median)	(Mean)	(Should be 0.05)
5	0.920	349.86	9640.10	0.034
10	0.847	56.48	103.53	0.037
100	0.065	14.73	15.10	0.045
Exact	0.000	10.24	10.24	0.050

Table 4: Performance of EPA and MLE methods on synthetic data with σ^2 unknown. Note failure of double conservatism for the Box 3-2 method.

6 Robustness

Both the EPA method and the alternate MLE method are based on an assumed lognormal distribution for individual concentration measurements, and both methods may give misleading results if this assumption fails. The lognormal distributional assumption will fail, for example, if there is even a small chance for data to be misrecorded, mistranscribed, misread, or entered incorrectly. It will fail if there is even a small chance of laboratory analytical error or mishandling of concentration samples.

The sample maximum X_n^* is particularly vulnerable to such departures. A single misplaced decimal point can distort its value by an order of magnitude, invalidating it. The sample mean $\bar{X}_n$ (a component of the sample coefficient of variation), though less sensitive than X_n^* by a factor of n, is still strongly affected by departures from log-normality; fortunately robust alternatives (trimmed means, sample median, *etc.*) are available.

The log-scale sample standard deviation $\hat{\sigma}^2$ used in the MLE method is relatively robust, but the log-scale sample mean $\hat{\mu}$ is not; again, the same robust alternatives as before are available on the log scale, leading to far greater robustness for the MLE method (or a small variation on it, using the trimmed mean or sample median instead of $\hat{\sigma}^2$). No such simple improvements are possible for the EPA's Box 3-2 method, due to its fundamental dependence on the inherently-nonrobust sample maximum X_n^*.

7 Alternatives to Double Conservatism

Both the EPA method and the alternate MLE method are "doubly conservative" in offering a 95%-confident upper bound for the 99%-percentile of the concentration distribution. This doubly-conservative approach leads to estimated bounds that are much higher than typical concentration measurements. Is this necessary or helpful in protecting public health?

n	$\mathsf{P}[X \le \hat{C}_{99}^{\mathrm{EPA}}]$	$\mathsf{P}[X \le \hat{C}_{99}^{\mathrm{ALT}}]$	$\mathsf{P}[X \le \widehat{C_{99}}]$	$\mathsf{P}[X \le \widehat{C_{995}}]$
5	0.9978	0.9974	0.9829	0.991
10	0.9976	0.9967	0.9867	0.993
100	0.9969	0.9934	0.9897	0.995

Table 5: Performance of EPA and MLE methods on synthetic data with σ^2 known.

On average the doubly-conservative methods give much higher than a 99%-percentile for concentration measurements, closer to a 99.7% or 99.8%-percentile— at the high cost of grossly overstating the concentrations. Less extreme bounds, such as $\widehat{C_{995}} = \exp(\hat{\mu} + 2.576\hat{\sigma})$ (which gives more than a 99%-percentile on average for all sample sizes), would still meet the conservative concern of ensuring that the "projected maximum concentration" would be at an acceptable level, without unnecessary gross overestimates.

8 Conclusions

The method used by the EPA to "project" a "maximum concentration" is flawed in two ways, each of which leads to grossly overstating concentrations. First, the method is *doubly conservative* in that it tries to find a projected maximum concentration that has a 95% chance of being greater than a bound that itself has a 99% chance of exceeding each concentration measurement. Second, the method makes inefficient use of the available evidence about effluent concentrations— it is based on ad-hoc summaries (the observed maximum and empirical coefficient of variation) that are more widely variable and capture less information than the best available statistical summaries (the sample mean and variance of the log concentrations), which are just as easy to compute. This unnecessarily variable doubly-conservative estimate exceeds actual measured concentrations by factors of 20–30, approximately double the exceedence of more efficient doubly-conservative estimates and approximately ten times the exceedence of available methods that still give 99% upper bounds to effluent concentrations. It would be difficult to justify continued use by the EPA of the Box 3-2 method.

We also note that uncertainty about the variability of concentration measurements necessarily leads to large overestimates of maximum concentrations for any method that reflects the uncertainty honestly. Sample sizes as small as $n \leq 10$ are particularly vulnerable to this problem. The added expense of larger sample sizes may well offset the cost associated with the risk of unnecessary remediation due to the "false positive" declarations that concentrations exceed acceptable limits when in fact they do not.

References

Bickel, P. J. and Doksum, K. A. (2001) *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. 1. 2nd edn. Upper Saddle River, NJ: Prentice Hall.

EPA Technical Support Document (1991) Technical support document for water quality-based toxics control. Document EPA/505/2-90-001 PB91-127415, U.S. Environmental Protection Agency.