pennsylvania
DEPARTMENT OF ENVIRONMENTAL PROTECTION

OFFICE OF WATER PROGRAMS
BUREAU OF CLEAN WATER

WATER USE ASSESSMENT DECISION-MAKING BASED ON PHYSICOCHEMICAL AND BACTERIOLOGICAL SAMPLING

FEBRUARY 2021

Prepared by:

Brian Chalfant
Pennsylvania Department of Environmental Protection
Office of Water Programs
Bureau of Clean Water
11th Floor: Rachel Carson State Office Building
Harrisburg, PA 17105

2013

Acknowledgments
Big thanks to Roy Irwin, Eric Money, Doug McLaughlin, and Karen Blocksom for sharing their gracious meditations, and to Rod Kime and Gary Walters for insightful reviews.

Edited by:

Rebecca Whiteash
Pennsylvania Department of Environmental Protection
Office of Water Programs
Bureau of Clean Water
11th Floor: Rachel Carson State Office Building
Harrisburg, PA 17105

2021

TABLE OF CONTENTS

INTENTION AND AIMS

This document discusses the technical, contextual, and conceptual aspects of procedures applied by the Pennsylvania Department of Environmental Protection (DEP) to inform Clean Water Act (CWA) Sections 303(d) and 305(b) water use support decisions based on physicochemical and bacteriological sampling.

This document contains relatively little discussion of the planning and execution phases of physicochemical and bacteriological water quality sampling projects, such as outlining study objectives, choosing sampling plan designs, and setting data quality objectives. These aspects are described in more detail within the data collection protocols of the *Water Quality Monitoring Protocols for Streams and Rivers* (Shull and Lookenbill 2018).

DEP strongly recommends that anyone planning to collect physicochemical or bacteriological sampling data for consideration in the water use assessment process fully familiarize themselves with this document as well as the following guidance published by the United States Environmental Protection Agency (USEPA) and DEP before initiating sampling:

- Guidance on Choosing a Sampling Design for Environmental Data Collection for Use in Developing a Quality Assurance Project Plan (USEPA 2002a)

- Guidance on Systemic Planning Using the Data Quality Objectives Process (USEPA 2006)

- Consolidated Assessment and Listing Methodology – Toward a Compendium of Best Practices (USEPA 2002b)

- Designing Your Monitoring Program – A Technical Handbook for Community-Based Monitoring in Pennsylvania (DEP 2001)

- Water Quality Monitoring Protocols for Streams and Rivers (Shull and Lookenbill 2018)

REGULATIONS AND GUIDANCE

The following review presents some key regulatory and policy considerations associated with use support decisions based on physicochemical and bacteriological water quality sampling in Pennsylvania. This review does not exhaustively relate all germane regulation and guidance. Rather, this review highlights regulatory and guidance language most relevant to making water use support decisions based on physicochemical and bacteriological sampling with a focus on numeric water quality criteria; application and interpretation of general water quality criteria is not discussed here. The full text of Pennsylvania's environmental protection regulations can be accessed online.

Federal law – specifically section 303(d) of the Clean Water Act – requires that states identify waterbodies not meeting water quality standards (WQS). Pennsylvania regulations at 25 Pa. Code Chapters 93 and 96, and the statement of policy at 25 Pa. Code Chapter 16, present WQS and associated implementation requirements applicable to waterbodies in Pennsylvania. The following excerpted sub-sections of § 96.3 (relating to water quality protection requirements) are particularly relevant to the use support decision process based on physicochemical and bacteriological sampling.

> (c) To protect existing and designated surface water uses, the water quality criteria described in Chapter 93 (relating to water quality standards), including the criteria in §§ 93.7 and 93.8a(b) (relating to specific water quality criteria; and toxic substances) shall be achieved in all surface waters at least 99% of the time, unless otherwise specified in this title. The general water quality criteria in § 93.6 (relating to general water quality criteria) shall be achieved in surface waters at all times at design conditions.
>
> (d) As an exception to subsection (c), the water quality criteria for total dissolved solids, nitrite-nitrate nitrogen, phenolics, chloride, sulfate and fluoride established for the protection of potable water supply shall be met at least 99% of the time at the point of all existing or planned surface potable water supply withdrawals unless otherwise specified in this title.
>
> (e) When a water quality criterion described in Chapter 93, including the criteria in §§ 93.7 and 93.8a(b), cannot be attained at least 99% of the time due to natural quality, as determined by the Department under § 93.7(d) based on water quality observations in that waterbody or at one or more reference stations of similar physical characteristics to the surface water, the natural quality that is achieved at least 99% of the time shall be the applicable water quality criterion for protection of fish and aquatic life.
>
> (f) When the minimum flow of a stream segment is determined or estimated to be zero, applicable water quality criteria shall be achieved at least 99% of the time at the first downstream point where the stream is capable of supporting existing or designated uses.

All of these excerpted sub-sections require that the specific, numeric water quality criteria detailed in Chapter 93 be met at a certain frequency. Note that sub-section 96.3(c) stipulates that *general water quality criteria* contained in § 93.6 (relating to general water quality criteria) shall be achieved in surface waters "*at all times at design conditions*." In 25 Pa. Code Chapter 16 (relating to water quality toxics management strategy – statement of policy), § 16.21 states that aquatic life criteria for toxic substances are developed such that the frequency of occurrence is accounted for through the specification of factors appropriate to the criteria in Chapter 96 (relating to water quality standards implementation), but also, that the basis for the magnitude, duration, and frequency is described in criteria development rationale or other appropriate supporting documentation. Section 16.22 states that DEP looks to National guidelines (USEPA 1985) in establishing aquatic life criteria for toxic substances. In 25 Pa. Code Chapter 96 (relating to water quality standards implementation), § 96.3 (relating to water quality protection requirements) provides that – to protect existing and designated surface water uses – the water quality criteria described in Chapter 93 shall be achieved in all surface waters at least 99% of the time. The National guidelines most often state that the frequency of excursions is not to occur more than once in three years on average (USEPA 1985). The import of these regulatory clauses (e.g., "at least 99% of the time", or "once in three years on average") to water use assessment decisions based on physicochemical and bacteriological sampling are discussed in more detail below.

The specific WQS in § 93.7 (relating to specific water quality criteria) incorporate a variety of nuanced clauses, which are critical to recognize when designing and implementing procedures to determine if waterbodies meet those standards. The standards in § 93.7 can be grouped into a few broad categories based on how each criterion is expressed.

- **Instantaneous standards**: Many standards in § 93.7 are presented as minimum or maximum numeric values without an associated averaging period. Such standards can be considered to have an "instantaneous" duration, applying to every "instant" of time, and can be assessed by any discrete measurement.

    - **Instantaneous maximum standards**: Standards presented as instantaneous numeric maxima include the Potable Water Supply (PWS) standards for chloride, color, dissolved iron, total manganese, nitrite plus nitrate, phenolics, sulfate, and part of the standard for total dissolved solids (TDS).

      Aquatic Life Use (ALU) instantaneous maximum standards include part of the ammonia nitrogen, osmotic pressure, pH, and temperature, with temperature maxima varying seasonally.

- **Instantaneous minimum standards**: The ALU standard for alkalinity and parts of the standards for dissolved oxygen and pH are presented as instantaneous minimal values.

- **Time-averaged standards**: The ALU standards presented as time-averaged maximum values include total ammonia (which includes the pH-dependent and temperature-dependent equations) as a 30-day rolling and four-day average maximum, parts of the dissolved oxygen standards as a seven-day average minimum, iron as a 30-day average maximum, and total residual chlorine as four-day and one-hour average maximums.

  Parts of the Water Contact Sport (WC) standard for bacteria – *Escherichia coli* during the swimming season and fecal coliform during the non-swimming season – are written as maximum values calculated as geometric means, and during the non-swimming season, the geometric mean must be based on a minimum of five consecutive samples with each sample collected on different days during a 30-day period.

  The PWS standards written as time-averaged values include fluoride as a daily average maximum, and part of the TDS standard as a monthly average maximum.

  In § 93.9x (relating to drainage list X, Lake Erie), specific to Lake Erie (Outer Erie Harbor and Presque Isle Bay), the PWS standard for coliform bacteria is written as a monthly average value.

  Section 93.1 (relating to definitions) defines the durations of “one-hour average”, “daily average”, “four-day average”, “seven-day average”, “monthly average”, and “thirty-day average” as arithmetic means, in contrast to the geometric mean specified by the fecal coliform WC standard.

  Section 93.1 further defines the durations “daily average”, as any continuous 24-hour period, not necessarily a single midnight-to-midnight calendar day, “four-day average” and “seven-day average” as any consecutive four-day period and seven-day period (respectively), “monthly average” as calendar months (e.g., August 1 to August 31), and “thirty-day average” as any consecutive 30-day period (e.g., August 15 to September 13).

- **Standards based on proportions of samples**: In § 93.7, part of the bacteria standard for WC is written as a maximum value that no more than 10% of the samples taken during a 30-day period may exceed; this

> incorporates a proportion of samples (i.e., 10%) and stipulates a duration (i.e., 30-day period).
>
> In § 93.9x, part of the bacteria standard for PWS is written as a maximum value that no more than 20% of the samples taken during a monthly period may exceed; this incorporates a proportion of samples (e.g., 20%) and stipulates a duration (e.g., one-month period).

In addition to the specific WQS expressed in § 93.7, water quality criteria for toxic substances are presented in §§ 93.8a (relating to toxic substances), 93.8b (relating to metals criteria), and 93.8c (relating to human health and aquatic life criteria for toxic substances). Regarding ALU criteria for toxic substances, § 16.21 (relating to acute and chronic protection) states:

> To provide for protection of aquatic life, it is necessary to consider both chronic, that is, long-term (reproduction, growth, survival) and acute, that is, short-term (survival) endpoints. Aquatic life can generally survive excursions of elevated concentrations of a pollutant as long as the excursion is of relatively short duration and does not frequently recur. However, to provide protection over a lifetime, a lower concentration shall be maintained. Thus, each aquatic life criterion consists of two magnitudes. The EPA defines these as a criterion maximum concentration (CMC) for acute protection and a criterion continuous concentration (CCC) for chronic protection. Each criterion is defined in terms of magnitude (a scientifically derived number), duration (the period of time over which the number must be achieved), and the maximum desired frequency (the number of repetitions per unit time) of occurrence…The frequency of occurrence is accounted for through the specification of factors appropriate to the criteria in Chapter 96 (relating to water quality standards implementation). Basis for the magnitude, duration, and frequency is described in criteria development rationale or other appropriate supporting documentation.

Regarding ALU criteria development for toxic substances, § 16.22 (relating to criteria development) states:

> The Department will establish criteria for toxic substances to provide for protection of aquatic life in accordance with the following guidelines:
>
> (1) For those toxics for which the EPA has developed criteria in accordance with the National guidelines as set forth in ‘‘Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and Their Uses’’ (1985), as amended and updated, the Department will review and evaluate the criteria. If the Department determines that the criteria are adequate to protect indigenous aquatic communities in the State’s waters, these criteria will serve as the basis for establishing total maximum daily loads (TMDLs) under Chapter 96 (relating to water quality standards implementation) or NPDES effluent limitations under Chapter 92a (relating to

> National Pollutant Discharge Elimination System permitting, monitoring and compliance). If the Department determines that the EPA National criteria are inappropriate, the Department will adjust these criteria in accordance with National guidelines to reflect the levels required for protection of aquatic life in this Commonwealth's waters.
>
> (2) For those toxics identified or expected in a discharge for which the EPA has not developed criteria, the Department will develop criteria using EPA-approved National guidelines.

Guidance from USEPA provides definitions for terminology, including for *digression*, *excursion*, *exceedance*, and *CMC* and *CCC* exposure periods. Following the USEPA guidance, a *digression* occurs when a single, discrete sample, or a set of spatially-composited samples, result in a measured concentration that is outside of a criterion magnitude (USEPA 2005). An *excursion* occurs when a measured concentration that is averaged over the criterion duration falls outside of a criterion magnitude (USEPA 2005). An *exceedance*, as the term is used in the context of assessment determinations, occurs when water quality does not meet the relevant components of a water quality criterion (WQC), which include magnitude, duration, and frequency. It is worth noting that the term *exceedance* had previously been defined by USEPA to be synonymous with *digression* (USEPA 1991) but has since been widely defined to mean not meeting WQS. Using the current definition, "any exceedance of a WQC is grounds for determining impairment" (USEPA 2003). For example:

> The PWS criterion for TDS in § 93.7 specifies 500 mg/L as a monthly average value and a maximum 750 mg/L.
>
> If ten samples were collected during a one-month period and the average concentration of the ten samples were greater than 500 mg/L, this would be an *excursion* from the criterion magnitude.
>
> If one of the ten samples measured greater than 750 mg/L, this would be a *digression* of the criterion magnitude.
>
> If two or more of the ten samples (i.e., > 1%) measured greater than 750 mg/L, this would be an *exceedance* of the criterion and the PWS use of the water would be assessed as impaired.

The USEPA guidance, *Deriving Numerical Water Quality Criteria for the Protection of Aquatic Organisms and Their Uses* (USEPA 1985), typically defines the exposure period associated with CMC as one hour and with CCC as four days, but this may differ per criteria and will be noted within specific DEP assessment protocols when defined otherwise. These national guidelines most often specify that an excursion from the acute (CMC) or chronic (CCC) criteria are not to occur more than once in three years on average for any waterbody, otherwise there is an exceedance of the criteria. The aquatic life criteria for toxic substances – magnitudes (e.g., concentrations), durations

(i.e., averaging periods), and frequencies (i.e., the number of digressions or excursions per unit of time) – are based on biological, ecological, and toxicological modeling studies and have been designed to protect aquatic organisms and ecosystems from unacceptable effects and are based on the ability of aquatic ecosystems to recover within that timeframe (USEPA 1985).

Section 93.8a presents two types of water quality criteria for toxic substances designed to protect human health: threshold effect criteria and cancer risk criteria. Regarding threshold effect criteria, § 16.32 states:

> (a) A threshold effect is defined as an adverse impact that occurs in the exposed individual only after a physiological reserve is depleted. For these effects there exists a dose below which no adverse response will occur. Threshold toxic effects include most systemic effects and developmental toxicity, including teratogenicity. Developmental toxicity includes all adverse effects in developing offspring resulting from prenatal exposure to a causative agent.

And regarding cancer risk, or non-threshold effect, criteria, § 16.33 states:

> (a) A nonthreshold effect is defined as an adverse impact, including cancer, for which no exposure greater than zero assures protection to the exposed individual. Thus, in contrast to the threshold concept discussed in § 16.32 (relating to threshold level toxic effects), the nonthreshold approach to toxics control is based upon the premise that there is no safe concentration of the toxic.

The magnitude of all criteria for toxic substances – aquatic life acute criteria, aquatic life chronic criteria, human health threshold criteria, and human health non-threshold criteria alike – are expressed as maximum values.

The human health criteria for toxic substances located at § 93.8c, Table 5, are instantaneous maximums not to be exceeded at any time. While the National recommended toxics criteria for human health (USEPA 2020) is derived by assuming a lifetime exposure of 70-years (USEPA 2007), this 70-year period is not to be confused with the duration of the criteria.

Chapter 93 (relating to water quality standards) and Chapter 16 (relating to water quality toxics management strategy – statement of policy) also contain water quality criteria and policy statements concerning toxic substances that are specific to Lake Erie and other sites in Pennsylvania. In addition, § 93.8(e) (relating to special criteria for the Great Lakes system) contains toxic substance criteria designed to protect wildlife from bioaccumulation impacts, a type of criteria unique to Lake Erie.

The water quality criteria, associated implementation requirements, and statements of policy outlined in Chapters 16, 93, and 96 apply to the surface waters (e.g., creeks,

streams, rivers, ponds, lakes) of Pennsylvania. These regulations do not apply to finished, treated drinking water.

As described above, an array of state regulations, state policies, and federal recommendations guide the protection of water uses through the establishment and assessment of water quality criteria. These criteria are expressed in a variety of nuanced ways, which must be considered when deciding if surface waters meet certain criteria.

SAMPLING AND INFERENCE

Within the regulatory and policy framework outlined above, DEP must determine if waterbodies meet WQS, allowing for infrequent excursions from some criteria (e.g., once in three years for aquatic life criteria for many toxic substances; less than 1% of the time for many specific water quality criteria). A number of interrelated considerations – outlined in this document – must be addressed when assessing if waterbodies meet WQS based on physicochemical or bacteriological samples, allowing for some infrequent excursions.

Sampling Space-Time

Water quality conditions – like many perceptible phenomena – change in space and in time. Interpreting water quality sample results often requires considering how far in space and for how long in time observed conditions can reasonably be considered representative of unobserved conditions that were not sampled.

This inferential process of using discrete, spatiotemporally-limited observations (i.e., samples) to estimate a larger set of unobserved, continuously dynamic conditions can introduce uncertainty into the use support decision process. Such uncertainty is called sampling error. Uncertainty can also enter the use support decision process through variability attributable to analytical measurement techniques, or measurement error. Sampling error and measurement error subject decisions based on sampling to decision error by introducing the potential for inaccuracy and imprecision into the observational process via sampling and analytic quantification. The aim of much of the rest of this document is to describe how the inherent variation and uncertainty introduced by sampling (i.e., sampling error) is addressed by DEP in the use assessment decision process for physicochemical and bacteriological water quality sampling data. Variability attributable to analytical and laboratory techniques and equipment (i.e., measurement error) is discussed elsewhere (DEP 2010). Both forms of error can be minimized by applying quality assurance procedures during sample collection, processing, and analysis.

Here's a simple scenario to illustrate sampling error. Imagine someone dutifully collects a one-liter water sample from a set location in a creek each month for two years; that's 24 one-liter samples. For the sake of discussion, let's imagine that 600 billion liters of water flow past that sampling location each year. Many considerations arise when we evaluate how representative those 24 liters of sampled water are of all 1.2 trillion liters

of water that flowed past that location in those two years. Imagine that none of the 24 samples show excursion from a relevant water quality criterion. When we look to that sampling data to inform our decision if the creek is meeting standards “at least 99% of the time,” we must ask ourselves how representative we think the *observed* conditions are of the *unobserved* conditions during the monitoring period. We may not observe water quality criteria excursions if they occur at times when – or in areas where – samples were not collected. How we deal with these sampling error considerations is the primary focus of much of the rest of this document.

Sampling Error Implications

Unless we continuously observe – or census – every quantum of water in a stream, and as long as we rely on limited observations derived from sampling, we have to acknowledge the possibility of sampling error.

Ideally, use assessment decisions for surface waters in Pennsylvania based on physicochemical and bacteriological data will be informed by sampling conducted frequently enough to accurately characterize the conditions for each parameter of concern over a long enough time frame to account for variations attributable to changes in all relevant factors. This may be possible for some parameter in some locations through deployment of automated, continuous instream monitoring devices or through extremely intensive monitoring efforts. However, many water quality sampling efforts require human beings to visit sites with chemical test kits or hand-held probes to measure water quality conditions or to collect grab samples of the water for later analysis at laboratory facilities. Such monitoring efforts require personnel, funding, and site accessibility among other considerations. As a result, chemical and bacteriological water quality sampling often provides limited windows into the dynamic continuum of water quality conditions at any given location at any given time.

Continuous instream monitoring (CIM) devices can measure conditions on a relatively frequent basis (e.g., every 15 minutes, every hour). Monitoring water quality conditions at such high frequency minimizes how long sample results have to be extrapolated into unobserved time, thereby minimizing the potential for sampling error. With less temporally-dense monitoring approaches (e.g., periodic grab samples collected by a person and analyzed in a laboratory) the amount of temporal extrapolation will likely need to extend further in time, and the potential for sampling error may increase. Some CIM devices can be deployed in remote locations and set up to report observations via telemetry or through occasional retrievals and downloads. While CIM devices can provide extremely detailed, temporally-dense observational records, funding and staffing considerations limit the number of locations at which these devices can practically be deployed and maintained. Furthermore, many CIM devices can only measure a few water quality parameters (e.g., dissolved oxygen, temperature, conductivity, pH) for which WQS exist.

In the absence of temporally-dense observations, if we understand and have information on enough relevant variables (e.g., stream flow, precipitation, water temperature) as related to the parameter of interest, we may be able to confidently infer

or extrapolate unobserved conditions from observed conditions based on empirical understanding of variability, and thereby reduce uncertainty attributable to sampling error. A wide variety of interrelated factors can contribute to spatial and temporal variation in the concentrations of water quality parameters of concern, including but certainly not limited to: precipitation rates, durations, and locations; thawing of ice and snow; stream flow; geologic and soil characteristics; annual and diurnal cycles of solar input; atmospheric conditions (e.g., cloud cover); discharges from permitted facilities; chemical spills; watershed drainage patterns; watershed land use; and hydrologic alterations (e.g., dams). Different water quality parameters often vary in unique ways relating to these and other factors. For example, dissolved oxygen concentrations in streams often exhibit strong annual and daily patterns attributable to interrelated patterns of solar flux, stream temperature, and photorespiratory activity of green plants. Meanwhile, TDS concentrations often vary much less with diurnal and annual patterns of solar flux, and more often vary primarily with stream flow and related patterns of surface runoff, geologies, and groundwater flow patterns. Knowledge and understanding of such patterns can strengthen inferences about unobserved conditions. USEPA (2005) recommends,

> "… states should decide how far out in time to extrapolate from the time at which a particular single grab was collected. EPA recommends that such decisions be based on contextual information regarding conditions when and where the grab was taken. For example, such information might include: 1) precipitation, 2) streamflow, 3) location of point source discharges in relation to the monitoring site, 4) land use patterns in the vicinity, 5) expected patterns of pollutant loading from the different kinds of sources present in the watershed, 6) occurrence of a chemical spill or other unusual event, and 7) historic patterns of pollutant concentrations in the monitoring segment and/or waterbodies similar to it."

In some situations, it may be possible to extrapolate patterns observed at nearby, physiographically-similar, or hydrologically-similar locations to unobserved locations for certain parameters. For example, data from past monitoring show that TDS concentrations of in many lotic systems in Pennsylvania often exhibit an inverse power type response to stream flow, so it may be possible to predict – or inferentially estimate – TDS concentrations with some confidence if we know stream flow and have reason to believe the often-observed relationship between stream flow and TDS concentrations holds for the particular situation at hand.

Even if we confidently observed or inferred the condition of every possible quantum of flow at a particular location, and even if we quantified the amount or concentration of the parameter of interest with immaculate accuracy and precision, the frequency component of many water quality criteria allow for some infrequent criteria excursions, as discussed previously, which introduces another set of considerations to the use assessment process aside from sampling error.

Sampling and Criteria Frequency

Along with considerations of uncertainty introduced by spatiotemporally-limited sampling of continuously-dynamic conditions, the fact that many criteria allow for some infrequent criteria excursions introduces to the use support decision process the additional consideration that some temporally rare criteria excursions are acceptable (e.g., ≤ 1% of the time for some criteria or once in three years on average for other criteria). In Chapter 96 (relating to water quality standards implementation), § 96.3(c) provides that the narrative general water quality criteria contained in § 93.6 shall be achieved in surface waters "at all times at design conditions." Section 96.3(c) also provides that – to protect existing and designated surface water uses – the water quality criteria described in Chapter 93 shall be achieved in all surface waters "at least 99% of the time" unless otherwise specified. Some of the specific water quality criteria in Chapter 93 define averaging periods and minimal sampling requirements, but many of these criteria do not explicitly specify associated sampling requirements or temporal aspects (e.g., frequency and duration components).

The vast majority of physicochemical and bacteriological water quality criteria – as expressed in Chapter 93 – are written as numeric values which are the amounts or concentrations of various parameters, which comprise the magnitude component of the criteria. The amounts or concentrations of these parameters are measured and often averaged over a specified duration of time for determining whether an excursion or exceedance of the criteria has occurred.

Digressions or excursions from criteria are not to occur more than a specified frequency. For aquatic life criteria for toxic substances, USEPA recommends certain acceptable digression or excursion frequencies, most often once in three years on average. The frequency of "at least 99% of the time" addresses the temporal aspect of criteria for which this consideration is not otherwise specified in Chapter 93, such as some specific water quality criteria listed in § 93.7. Additional discussions on criteria written as minima and maxima can be found in the *Clarifications* section.

The frequency of "at least 99% of the time" needs further consideration and definition as it does not specify a time period to which it applies (e.g., 99% of one year, 99% of one month, 99% of one day, 99% of one minute). Contrast this with the comparatively exact specificity of the WC use criterion for fecal coliform in § 93.7 (i.e., a maximum geometric mean based on a minimum of five consecutive samples collected on different days during a 30-day period). Therefore, interpreting the frequency of "at least 99% of the time" requires context-specific considerations that take into account the particular standard(s) being evaluated, as well as site-specific evaluation of expected patterns of variability in the parameters of interest. These context-specific considerations are also relevant to assessing criteria for which excursions are allowable no more than once in three years on average (e.g., many aquatic life criteria for toxic substances).

Considering these temporal aspects of water quality criteria, the underlying concept in the phrase "at least 99% of the time" and the allowance for one excursion in three years on average is straightforward: there is some acceptable frequency – albeit relatively

low– at which digressions or excursions from the water quality criteria concentrations presented in Chapter 93 are allowed (i.e., the digressions or excursions do not parameter an exceedance).

Sampling Design

Sampling error is influenced by how, when, where, why, and by whom samples are collected. As such, these considerations play a critical role in the use support decision process. While sampling plan design is not the focus of this document, some considerations on this topic are discussed in this section because sampling plan design largely determines what analytical procedures can tenably be used to assess the sampling data. The following discussion of sampling plan design is not intended to be exhaustive or complete by any means, rather to highlight some issues that are particularly relevant to making water use assessment decisions in the regulatory and policy context of Pennsylvania. As stated above, DEP strongly suggests that anyone planning to collect physicochemical or bacteriological sampling data for consideration in the water use assessment process familiarize themselves with the guidance documents listed in the introduction of this document for more thorough discussions of sampling plan design, data quality objectives, and other study planning considerations.

Thoughtful study design and execution are critical to assuring water quality sampling efforts provide the information necessary to address the study questions. USEPA (2006) details step-by-step considerations of study design and data assessment. USEPA (2002a) provides further specific details on designing a sampling plan. While all these sampling plan design considerations are not repeated here, it is important to address a few interrelated implications of sampling plan design in light of the fact that many criteria allow for infrequent criteria excursions or digressions.

The fact that many water quality criteria allow for infrequent criteria digressions or excursions is a critical consideration in designing a sampling plan. For these criteria, digressions or excursions are acceptable for a very small proportion of time. As such, any study or investigation aiming to assess criteria must aim to observe an extreme end – or ends – of water quality parameter distributions.

Due to interrelated considerations of decision error rates, sample sizes, and extreme percentiles of water quality parameter distributions, it will very often be impractical to employ a probability-based sample design to assess against criteria that allow for infrequent digressions or excursions without collecting large numbers of samples, at least at any reasonable decision error rates. It will often be the most resource-effective approach – especially when accounting for monitoring costs – to focus monitoring at times when excursions are most likely to occur based on understanding of the factors affecting the parameter of interest in the particular monitoring situation being assessed. In the rest of this document, DEP refers to these times when criteria excursions are most likely to occur as **critical sampling periods**. Sampling focused on these critical sampling periods will necessarily be based on human understandings of the variables at play. In the terminology used by USEPA (2002a), these critical sampling periods can be thought of as temporal “hot spots,” and sampling targeted to observe these “hot spots”

based on understandings of context-specific variations is referred to as “judgment-based sampling” (as contrasted with probability-based sampling). Since this targeted, judgment-based sampling is not as suited to some forms of quantitative statistical analyses as probability-based sampling, assessment processes based on judgement-based sampling may involve a different analytical toolset than assessment processes based on probability-based sampling designs.

Of the various sampling plan designs discussed by USEPA (2002a), DEP believes the so-called “judgment-based” sampling design is the most suited method to assess extreme, infrequent ends of water quality parameter distributions. Other sampling plan designs (e.g., simple random sampling, systematic sampling) presented by USEPA (2002a) are unlikely to provide accurate, precise estimates of such extreme ends of distributions while maintaining reasonable decision error rates without requiring large numbers of samples. For example, systematic sampling (i.e., sampling at regular temporal intervals) may be useful for certain applications (e.g., determining temporal trends) and can be attractive in terms of scheduling personnel and logistics, but is unlikely to directly observe infrequent, extreme events (e.g., heavy storm flows or conditions that occur 1% of the time or less) unless many samples are collected at relatively short intervals. Such systematic sampling will usually require a very large number of samples to accurately and precisely estimate extremely infrequent conditions.

Regarding systematic sampling, USEPA (2002a) notes,

> “… if the scale of the pattern or feature of interest is smaller than the spacing between sampling locations [or times], then the systematic pattern of sampling is not an efficient design unless the spacing between sampling locations [or times] is reduced or some other procedure such as composite sampling is introduced into the design.
>
> Systematic sampling would be inappropriate if a known pattern of contamination coincides with the regularity of the grid design. Such a coincidence would result in an overestimation or underestimation of a particular trait in the target population of interest.”
>
> “Systematic/grid sampling may not be as efficient as other designs if prior information is available about the population. Such prior information could be used as a basis for stratification or identifying areas of higher likelihood of finding population properties of interest.”
>
> “… if nothing is known about the spatial characteristics of the target population, grid sampling is efficient in finding patterns or locating rare events unless the patterns or events occur on a much finer scale than the grid spacing. If there is a known pattern or spatial or temporal characteristic of interest, grid sampling may have advantages over other sampling designs

depending on what is known of the target population and what questions are being addressed by sampling."

For example, a systematic sampling plan for dissolved oxygen where samples are collected the 15th day of every month at noon would be likely to sample the highest dissolved oxygen concentrations because photosynthetic activity usually peaks around midday.

Regarding judgment-based sampling designs, USEPA (2002a) states that,

> "In judgmental sampling, the selection of sampling units (i.e., the number and location and/or timing of collecting samples) is based on knowledge of the feature or condition under investigation and on professional judgment. Judgmental sampling is distinguished from probability-based sampling in that inferences are based on professional judgment, not statistical scientific theory. Therefore, conclusions about the target population are limited and depend entirely on the validity and accuracy of professional judgment; probabilistic statements about parameters are not possible. As described in subsequent chapters, expert judgment may also be used in conjunction with other sampling designs to produce effective sampling for defensible decisions."

As noted by USEPA (2002a), many commonly-used statistical analysis methods assume either implicitly or explicitly that data were obtained using a probability-based – often simple random – sampling design. Probability-based sampling designs allow for application of certain statistical tools, which can facilitate quantification and control of decision error rates. In short, probability-based sampling designs offer the benefit of being amenable to certain statistical analyses but can have the drawback of requiring a lot of sampling to sufficiently characterize the full range of water quality conditions. Judgment-based sampling designs may not be as conducive to some standard inferential statistical analyses – primarily due to sample selection bias – but offer the benefit of more resource-efficient sampling (i.e., needing fewer observations to achieve a given level of precision) by incorporating existing understandings of the site and systems being sampled. USEPA (2002a) notes that,

> "Judgmental sampling is useful when there is reliable historical and physical knowledge about a relatively small feature or condition …."
>
> "… whether to employ a judgmental or statistical (probability-based) sampling design is the main sampling design decision …."
>
> "An important distinction between the two types of designs is that statistical sampling designs are usually needed when the level of confidence needs to be quantified, and judgmental sampling designs are often needed to meet schedule and budgetary constraints."

> "Data obtained from convenience or judgment sampling cannot be used to make formal statistical inferences unless one is willing to assume that they have the same desirable properties as probability samples, an assumption that usually cannot be justified …."

> "Although statistical methods for developing the data collection design … are strongly encouraged, not every problem can be resolved with probability-based sampling designs. On such studies … the planning team is encouraged to seek expert advice on how to develop a non-statistical data collection design and how to evaluate the results of the data collection."

When designing a sampling plan, USEPA (2002a) recommends considering tradeoffs among considerations of desired data quality (e.g., characteristics of the parameters of interest, applicable analytical approaches, decision error estimates) and practical constraints (e.g., budgets, personnel, time, site accessibility) for a given parameter in a given situation.

USEPA (2002a) also suggests that some sampling plan designs – like stratified random sampling – draw on aspects of both probability-based designs and judgment-based designs. Stratified random sampling can be used to more efficiently focus sampling resources to critical sampling periods at a given location based on existing understanding about variability of the parameters of concern at the particular study location (i.e., where and when criteria excursions are most likely to occur). For example, an understanding – or "conceptual model" – of dissolved oxygen concentrations could be used to define three temporal strata based on likely concentration ranges: likely low-level (i.e., pre-dawn, summer), likely mid-range (i.e., autumn and spring mornings and evenings), and likely high-level (i.e., mid-day, winter). Such a stratified approach may also incorporate spatial aspects with backwater, less-turbulent areas being more likely to have lower concentrations of dissolved oxygen than faster-flowing, more-turbulent areas mid-channel. In stratified random sampling designs, each member of the target population has a known – although perhaps unequal – probability of selection into the sample. Therefore, techniques of statistical inference can be applied to data resulting from stratified random sampling designs. Regarding stratified random sampling, USEPA (2002a) notes,

> "When stratification is based on correlation with an auxiliary variable which is adequately correlated with the variable of interest, stratification can produce estimates with increased precision compared with simple random sampling or, equivalently, achieve the same precision with fewer observations. For increased precision, the auxiliary variable used to define the strata should be highly correlated with the outcomes being measured. The amount of increase in precision over simple random sampling depends on the strength of the correlation between the auxiliary variable and the outcome variable being measured."

> "Stratified sampling needs reliable prior knowledge of the population in order to effectively define the strata and allocate the sample sizes. The gains in the precision, or the reductions in cost, depend on the quality of the information used to set up the stratified sampling design. Any possible increases in precision are particularly dependent on strength of the correlation of the auxiliary, stratification variable with the variable being observed in the study."

USEPA (2002a) also acknowledges that no sampling plan design is completely objective, noting (emphasis original),

> "Implementation of a judgmental sampling design should not be confused with the application of professional judgment (or the use of professional knowledge of the study site or process). Professional judgment should *always* be used to develop an efficient sampling design, whether that design is judgmental or probability-based. In particular, when stratifying a population or site, exercising good professional judgment is essential so that the sampling design established for each stratum is efficient and meaningful."

MAKING DECISIONS

In the past, DEP adopted an approach that stipulated minimum data requirements for chemical and bacteriological use assessment datasets that were applied across all criteria. These requirements stipulated a minimum number of samples (i.e., eight), a minimum sampling frequency (i.e., at least quarterly), and minimum sampling time period (i.e., at least one year) needed to assess sampling data against any criteria. While this approach attempted to direct sampling so that a variety of conditions would be observed (e.g., different flow conditions, different times of year), this approach did not address the idea of critical sampling periods, discussed above. Regarding data quantity, USEPA (2005) states,

> "EPA encourages the collection of adequate data to make well-grounded attainment determinations. EPA has not established, required, nor encouraged the establishment of rigid minimum sample set size requirements in the WQS attainment status determination process. EPA is particularly concerned with application of such thresholds state-wide, without regard to key factors like the manner in which applicable WQC are expressed, variability in segment-specific conditions, and fluctuations in rates of pollutant loading. Rather if employed, target sample set sizes should not be applied in an assessment methodology as absolute exclusionary rules, and even the smallest data sets should be evaluated and, in appropriate circumstances, used. While it may be appropriate to identify target sample sizes as a methodology is developed, states should not exclude from further consideration data sets that do so solely because they do not meet a target sample size. A methodology may provide for an initial sample size screen but should also provide for a further assessment of sample sets that do not meet

the target sample size. (EPA suggests that states avoid setting target sample set sizes higher than the amount of data available at most sites.)"

Presently

Presently, DEP recommends context-specific and site-specific approaches to evaluate various water quality criteria, accounting for the fact that the criteria in Chapter 93 are presented in different ways, and because some parameters vary in different ways with changing natural conditions (e.g., diurnal and annual cycles of solar radiation, changes in stream flow) and may exhibit variable responses to these factors at different locations. DEP believes it inappropriate to develop data requirement guidelines applicable to all criteria across the board since different monitoring efforts may utilize different means and may have different goals, and because different parameters, criteria, and situations call for different monitoring approaches. The present approach is consistent with recommendations from USEPA (2005) that,

> "Any target sample set size thresholds must be consistent with the state's EPA-approved water quality standards. Hence, when making a determination based on comparison of ambient data and other information to a numeric WQC expressed as an "average" concentration over a specified period of time, a statement of a desired number of samples may be appropriate. Still, the methodology should provide decision rules for concluding nonattainment in cases where the target data quantity expectations are not met, but the available data and information indicate a reasonable likelihood of a WQC exceedance (e.g., available samples with major digressions from the criterion concentration, corroborating evidence from independent lines of evidence such as biosurveys or incidence of waterborne disease, indications that conditions in the waterbody and loadings of the pollutant into the waterbody have remained fairly stable over the period in question)."

All relevant data will be considered in DEP's use support assessment process regardless of sample size, but – because waterbody assessments are made on a continual basis in an effort to document current conditions – more recent data take precedence over older data, especially in situations where conditions have recently changed (e.g., installation of pollution remediation projects, alteration of permit limits in the watershed, changing land use patterns, discontinuation of combined sewer overflows). In some instances, older and newer data may be considered in concert to document temporal trends.

DEP makes every effort to verify the accuracy of all data used in the use support decision process. DEP strongly encourages anyone submitting data to familiarize themselves with DEP Bureau of Laboratories quality assurance and quality control procedures (DEP 2010) regarding record keeping, methods documentation, sampling techniques, selection of analytic laboratories, chain of custody concerns, and so forth. DEP will not drop extreme values (outliers) from a dataset unless there is reason to believe the extreme value is invalid. For example, a dissolved oxygen concentration of 100 mg/L is physically impossible at tropospheric temperatures and pressures – it is

likely that such a record is a typographical error meant to really be 1 mg/L or 10 mg/L. Similarly, in a water temperature dataset submitted in degrees Celsius where one value is recorded at 72, it is highly unlikely this is a valid reading and may be recorded in degrees Fahrenheit. DEP does not want to discount any data or information from consideration outright, so no strict guidelines are set forth with regard to what sampling designs are acceptable because different sampling approaches may be necessary to answer different questions in different situations depending on the particular parameter and waterbody in question.

DEP strongly recommends that any physicochemical or bacteriological water quality sampling datasets intended for consideration in the use assessment decision process be collected using a "judgment-based sampling" design – as discussed above – with sampling targeted to critical sampling periods when water quality excursions (or human exposure in the case of bacteriological samples) are most likely to occur based on knowledge of the conditions affecting the parameter(s) of interest. Some considerations are common to sampling design decisions for many parameters.

Diurnal cycles of solar radiation – The rising and setting of the sun drives daily cycles of photosynthesis and respiration across much of the earth's surface, including in surface waters. During peak influx of solar radiation, photosynthetic activity tends to peak, resulting in higher instream dissolved oxygen concentrations and pH levels. Incident sunlight also increases instream temperatures, which affects levels of dissolved oxygen in the water and photosynthetic rates. Other parameters (e.g., TDS, alkalinity) may also exhibit some diurnal cycling.

Annual cycles of solar radiation – Northern hemisphere locations, such as Pennsylvania, receive their most intense and prolonged sunlight in the summer months of June and July, with less intense and shorter exposure to sunlight in winter months of December and January. Stream systems reflect these cycles in annual cycles of water temperature and dissolved oxygen. Fluctuations in pH are often less dramatic in winter months as well, likely due to reduced photosynthetic activity with the colder temperatures. Some nutrient parameters may also exhibit variation with annual seasons.

Annual cycles of precipitation, evapotranspiration and stream flow – Across Pennsylvania, stream flow patterns vary annually with the lowest stream flows typically observed from July through September, gradually increasing through autumn and winter with peak flows often observed January through April and tailing off again May through June to summer base flows, although hurricane and tropical storm remnants occasionally dump heavy rain on Pennsylvania in early autumn. These stream flow patterns reflect annual cycles of precipitation, snow and ice melt, and evapotranspiration. Such patterns vary locally with a variety of factors (e.g., soil types, hill slopes) and fluctuate year to year. Some parameters of concern often exhibit predictable patterns of response with precipitation and stream

flow. For example, TDS concentrations tend to decrease consistently with increasing stream flow; in many situations this has to do with surface runoff containing lower dissolved ion concentrations than groundwater inputs to a stream, which tend to be higher in dissolved minerals. Likewise, some parameters, like various forms of phosphorus, usually enter stream systems attached to particulate matter washed in during periods of high surface runoff. Such patterns may also vary with surrounding land use, like fertilizer application, impervious surfaces and so on.

Conservative and non-conservative substances – Some water quality parameters – like sulfate – are considered conservative in that their concentrations are not directly affected by biological processes. These conservative substances do not decay, are not selectively incorporated by living organisms, and concentrations are affected mostly by sedimentation and dilution. Non-conservative substances – such as phosphorous – are removed from the water column by biological processes.

Decision Framework

DEP will implement the following framework when evaluating physicochemical or bacteriological water quality monitoring data in the use assessment decision process. The details of this appraisal process may vary from application to application based on the unique characteristics and contexts of each situation. However, DEP will follow this process as much as possible in order to maintain consistency in the use support decision process and so that interested stakeholders can clearly see how DEP evaluates physicochemical and bacteriological sample results. This process will be documented for each use support decision. The decision framework aims to document and communicate each step of the decision process in a clear, consistent manner addressing the study designs, data quality, data analysis, assumptions, uncertainties, and consequences associated with each use assessment decision. DEP attempts to be as concise as possible within this framework while not compromising adequate discussion of critical issues influencing the decisions.

(1) Describe monitoring effort. Describe the waterbody and the watershed, including basin size, land uses, geologies, and other characteristics. Discuss any germane history and context pertaining to the monitoring effort. To the extent possible, describe the motivations and intentions of the monitoring effort, including the individuals and organizations involved as well as the intended use of the information collected. Clearly state study goals. Describe and map monitoring locations. Include any photographs.

(2) Check data quality. Evaluate any study plans and objectives, including sampling plan design details such as record keeping, data management, training, sampling techniques, and analytical methods. Check data for typos and other anomalies. Document non-detects and censored data.

(3) **Gather information on likely sources of variation.** At a minimum, this information will typically include characterization – and quantification where possible – of tributary locations, upstream discharges, geologies, and land uses. Potential sources of this information include stream gages, climatological records, and discharge monitoring reports. Include maps, figures, and diagrams as needed. Discuss relevant physical, chemical, and biological processes and other potential sources of variation for the parameter(s) of concern. Address context-specific considerations (e.g., dams).

(4) **Explore data.** Perform various graphical analyses (e.g., histograms, probability distribution functions, boxplots, time-series plots, scatterplots with likely sources of variation, LOWESS) to visually explore and illustrate data characteristics. Document summary statistics (e.g., minimum, maximum, mean, median, standard deviation).

(5) **Evaluate data representativeness.** Evaluate how representative samples are of unmonitored conditions, mindful of the sampling plan design (e.g., sample collection frequency, locations, timing, targeting) and the likely sources of variation with special attention to any critical sampling times and locations. Consider if the system is likely to be spatially well-mixed at monitoring location(s) and how quickly conditions are likely to change in time.

(6) **Describe the relevant standards.** Identify which criteria are being evaluated and the uses to which they apply. Describe how the parameters of concern impact the protected use (i.e., exposure pathways, detrimental effects). Review the associated regulatory language including any relevant criterion rationale documentation.

(7) **Apply appropriate analytical procedures.** Select and apply appropriate analytical techniques, mindful of the sampling plan design, monitoring objectives, and the relevant criteria, parameters, and context. State and verify any assumptions associated with each analytical technique. Evaluate decision error rates, if applicable. For hypothesis tests, evaluate null hypothesis choice. Discuss the frequency, duration, and magnitude of any criteria digressions or excursions.

(8) **Consider other sources of relevant use support information.** Additional sources of information may include: previous or concurrent monitoring efforts; data from water supply intakes; biological surveys; and discharge monitoring reports.

(9) **Evaluate all relevant lines of evidence.** Bring together the previous steps into a narrative that addresses contextual data interpretations, possible counter arguments, alternative decision choices, and decision

consequences, including evaluation of decision error consequences. Explicitly address any policy ramifications if applicable.

(10) Decide. Decide what to do with the dataset and waterbody in question. At a minimum, each decision will include placing the waterbody in one of the Integrated Report categories.

Clarifications

When compared to physicochemical or bacteriological sampling datasets used to make listing decisions, datasets used to inform **delisting decisions** (i.e., decisions to remove a waterbody from Category 5 of the 303(d) list) must: (1) have been collected more recently; (2) have been collected as frequent or more frequently; and (3) contain more samples.

Criteria written as time-average criteria

DEP encourages multiple sampling events within the specified averaging periods. For example, the ALU criterion for total iron is written as a 30-day average, so DEP encourages sampling multiple times in a given 30-day period to compare existing conditions to the criterion. Like any sampling, this sampling should represent the most likely excursion times and spaces as discussed above. As a general guideline, DEP encourages at least three sampling events in the averaging period expressed in various time-averaged criteria.

For assessment decisions based on aquatic life criteria for toxic substances, DEP generally follows relevant Federal guidelines (USEPA 1985) as discussed above in the *Regulations and Policy* section above (i.e., a toxic substance concentration, averaged over the respective durations for the CMC or CCC, above the criterion magnitude more than once in three years on average parameters an exceedance). To be clear on how the three-year period provision for toxic pollutants is applied, if a site is dutifully monitored for a year and a single excursion is observed at the end of that year, another excursion must not be observed at that site for the next two years for the water to be considered attaining the relevant protected use for that toxic substance. Similarly, if the first sample at a previously unmonitored site shows a toxic substance criteria excursion, another excursion must not be observed at that site for the next three years for the water to be considered attaining the relevant protected use for that parameter.

DEP generally considers discrete samples (a.k.a., grab samples) to be representative of one day unless convincing evidence exists to suggest otherwise (e.g., a documented spill, influence of a known biological process, supporting high-frequency monitoring data). Under this presumption, for criteria expressed without any time-averaging period (a.k.a., instantaneous criteria) for which no frequency component is specified in the criteria or incorporated via federal recommendations, four days with observed criterion excursions in a year parameter an exceedance, and thus an impairment of the relevant protected use because the criteria excursions occur more than 1% of the time (i.e., 4 days / 365 days ≈ 1.1%, which means the criteria are being met less than 99% of the time). For criteria expressed as 30-day or monthly averages to which the “99% of the

time” provision applies as the frequency component of the criteria, any one month or 30-day period showing a criterion excursion will be considered an exceedance based on the reasoning that the water is not meeting standards 99% of the time (i.e., 1 month / 12 months ≈ 8.3%).

LITERATURE CITED

DEP. 2001. Designing Your Monitoring Program – a Technical Handbook for Community-based Monitoring in Pennsylvania. Harrisburg, Pennsylvania.

DEP. 2010. Quality Assurance Manual for the Pennsylvania Department of Environmental Protection Bureau of Laboratories. Harrisburg, Pennsylvania.

Shull, D. R., and M. J. Lookenbill (editors). 2018. Water Quality Monitoring Protocols for Streams and Rivers. Pennsylvania Department of Environmental Protection, Harrisburg, Pennsylvania.

USEPA. 1985. Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and Their Uses. PB85-227049.

USEPA. 1991. Technical Support Document for Water Quality-Based Toxics Control. EPA/505/2-90-001.

USEPA. 2002a. Guidance on Choosing a Sampling Design for Environmental Data Collection for Use in Developing a Quality Assurance Project Plan. EPA QA/G-5S.

USEPA. 2002b. Consolidated Assessment and Listing Methodology – Toward a Compendium of Best Practices. First edition.

USEPA. 2003. Guidance for 2004 Assessment, Listing and Reporting Requirements Pursuant to Sections 303(d), 305(b) and 314 of the Clean Water Act; TMDL-01-03.

USEPA. 2005. Guidance for 2006 Assessment, Listing and Reporting Requirements Pursuant to Sections 303(d), 305(b) and 314 of the Clean Water Act.

USEPA. 2006. Guidance on Systematic Planning Using the Data Quality Objectives Process. EPA QA/G-4.

USEPA. 2007. Water Quality Standards Handbook. Second edition. EPA-823-B-94-005.

USEPA. 2020. National Recommended Water Quality Criteria – Human Health Criteria Table. U.S. Environmental Protection Agency, Office of Water, Washington, D.C., 2020 Feb 20 [accessed 2021 Jan 8].