Independent Panel Review of the Delta Regional Monitoring Program (Delta RMP) Monitoring Design

Phase I: Initial Review

A report to the

Delta Science Program

Prepared by

Peter Raimondi, Ph.D. (Panel Chair) – University of California, Santa Cruz
Barry Noon, Ph.D. (Lead Author) – University of Colorado
Michael MacWilliams, Ph.D. – Anchor QEA
Allan Stewart-Oaten, Ph.D. – University of California, Santa Barbara (Emeritus)
Laura Valoppi – United States Geological Survey

Septebmer 2016

Delta Stewardship Council
Delta Science Program

Table of Contents

Independent Panel Review of Delta Regional Monitoring Program (Delta RMP) Monitoring Design

Phase I: Initial Review

I. Executive Summary

Overall Comments

The full charge to the Panel is at the end of this report. In brief, it asks us:

A) Is the Monitoring Design adequate to answer the management and assessment questions?
B) To recommend scientific criteria for distributing limited resources towards monitoring.

A short answer to A) is "Probably not." A major reason is that B) has been largely ignored.

The main tasks of a monitoring plan are to define the quantities and summaries needed to address the management and assessment questions, and to specify a sampling scheme that will lead to estimates of these summaries that are reliable enough to be useful. If the ideal plan would cost too much, it would be scaled back based on priorities of concerns and feasibility of useful results.

The Monitoring Design Summary (MDS**)** is silent on most of these tasks. It defines some quantities without saying how they should be summarized and used for management, and specifies a sampling scheme without saying how it can be used to estimate useful summaries. Costs are mentioned but priorities and feasibility are not. The Quality Assurance Program Plan (QAPP**)** addresses reliability in detail, but mainly to describe control of sampling error for estimates at a single site and time, not the errors for summaries over time or space. The core management and assessment questions are in Table 1of the MDS (pp. 3-7). Many refer to "beneficial uses" which are discussed in the QAPP (pp. 15-24). All of them require judgments about quantities which vary over time and space, as well as due to sampling error.

A sample can estimate the level of an indicator (e.g., contaminant) at a particular point in the Delta at a particular moment in time. This estimate is useful for management or assessment only if it can be used to tell us about levels at other points and times that were not observed.

In some cases, the fact that the level varies continuously may be enough. A high pathogen level found at a recreational site or drinking water intake might trigger management action by itself, because levels nearby in time and space are likely to be high too. More often, however, a summary is needed, such as a trend over time at an important site, an average summer level over a sub-region, or a time trend in such averages.

The Panel cannot be certain that the Monitoring Design is inadequate. It is possible that appropriate summaries could be defined, and that models and methods could be developed by which they could be estimated reliably from this sampling design. Some of this work may have been done in the discussions that led to the design. However, none of this supporting information appears in the MDS.

As it stands, the design will lead to a large collection of data, measuring contaminant levels at a discrete set of sites and times which constitute a vanishingly small part of the Delta and the time period of interest. These data will be of little use unless they can be combined and interpreted to form a description of contaminants over larger areas or periods, or of processes that management action might affect. The "Example Data Products" are graphs that display data over space, time or both, but do not extract major messages, uncertainties, or implications for action. The MDS (p. 16) says "Interpretation and reporting methods will be described in a Communications Plan" but they are not.

Panel Recommendations

A list of main recommendations follows. Others more specific to separate studies are in Section III.

1. **We recommend that the monitoring team include one or more environmental statisticians, employed full-time, to refine the sampling design and develop the methods for data analysis.**

2. Monitoring and assessment of the state of the Delta is based on a sample of the study area—that is, not all possible locations are sampled and indicator values measured. Therefore, the ability to use the sample data to draw inferences about unmonitored sites is a key part of sample site selection. This has several components. One is to use models of flow, transport and degradation to help estimate values up- or down-stream of monitored sites. The five pesticide sampling sites may allow crucial areas to be estimated this way (but they are likely to be small and no methods are given). Another approach is statistical. The standard approach has a large literature based on mathematics, simulation and experience. It selects sites partly randomly, using an objective procedure like computer-generated random numbers. (Haphazard, intuitive, or convenience sampling are not substitutes.) While easy in principle, this approach can be hard in practice. Stratification is often needed to ensure that sites in different subareas or of different types are adequately represented. The number of sites may need to be increased by reducing the frequency of sampling or other changes (see Section III). Methods may need to allow for some selected sites turning out to be inaccessible. Some details are in Section II and the Appendix. One motive for the first recommendation is that some of these details require familiarity with statistical methods.

3. Tidal phase and variation in flow need to be taken into account in the sampling plans**.** This may not be relevant for all constituents (such as pathogens at the water intakes or mercury testing in fish), but it is likely to be important for some. The presentation to our review team on August 23, 2016 included a figure from the USGS showing how sparse sampling on a tidally oscillating time series could lead to erroneous conclusions. For example, some variables such as pesticides and nutrients are likely to be associated with local point source introductions to the Delta. If two samples are made at a location near a significant point source but the tidal currents are in opposite directions, one sample could measure the pesticide and the other would not. If these 2 samples are months apart, no significant seasonal difference can be inferred because the difference may be due to tidal flow direction instead. If tidal phase is not considered in the sampling plan, there is no way to isolate the effect. Several recommendations for taking tidal phase into account are included in the detailed discussion of specific constituents (see Mercury and Nutrients under section III).

4. A useful beginning would be **to restate Table 1 of MDS to more specifically address the management questions, monitoring goals, and likelihood of achieving these goals for each constituent.** In some cases, numerical goals (albeit approximate) may be needed for areas (rather than single sites) or periods (rather than single times). How well do the "lower", "midrange" and "higher" sampling levels achieve the monitoring goals? How were the prioritization decisions (shown by stars in Table 4) made? Careful assessment should help ensure that resources are allocated efficiently, are directed towards achievable goals, without being spread too thin. In some cases, the sampling may not be worth doing, because it is not tied to management goals or is too sparse to be useful. For example, the medium cost level for pesticides is over half of the total for the entire recommended program; the higher level is 2.5 times larger, with nothing in between. Yet we don't know that there is a problem or signs of a problem (the initial question). Less costly sampling might tell us. Until then, some of these resources might be better spent on other constituents.

II. Features of an effective monitoring program.

The principle goals of environmental monitoring programs are to inform the management decision making process. Effective programs:

- provide reliable descriptions (usually quantitative estimates) of the state of the resources being managed and of changes over space or time
- use these descriptions to assess the need for possible management actions
- evaluate the implementation and effectiveness of management actions
- update our understanding of how the system operates.

Successful environmental monitoring programs have several characteristics in common. These include:

In the design document, the monitoring program identifies state variables (e.g., indicators) to be measured at sample locations but does not fully explain why these indicators were selected. For example, lab analyses do not assess "pesticides" or "nutrients": they assess particular pesticides and nutrients. Each one added can increase costs, each one ignored can increase risks, and there may be legal requirements. What logic was invoked to justify the selection of the indicators to be measured?

The monitoring objectives are clearly defined quantities that can be observed or estimated from objective measurements. Initial management questions in the documents were usually in words, not numbers: "is there a problem?", "what is the status?", or "is toxicity too high?" These need to be restated in measurable terms, usually as means or trends over time or space (including subregions or tributaries, etc.) or both. Even when a numerical quantity is given, as for some water quality objectives, it may refer to a single observation or to an average over a sample size, area or time period which has not been specified.

The spatial and temporal domain of the population of interest is defined. The spatial domain will usually require a map. Ideally, the sampling scheme should allow the value of any quantity of interest at any place or time in the domain to be estimated, as well as patterns such as means or trends over space or time. It should usually take into account influences originating outside the domain (especially the spatial domain) which contribute to the values inside it.

The desired reliability for important estimates is specified. The "important" estimates are those in the monitoring objectives whose values are key to answering management or assessment questions, and have a clear potential to trigger management actions. (Other estimates might be made in passing because they are cost-free, but need not be in the plan details.) The measure of "reliability" needs to be defined, not only for estimates at a given place and time but for expanded inference in time and space.

The desired reliability for important estimates is justified in detail. We separate this from the previous point because carrying it out requires other steps. The justification can take several forms:

- Some actions may be legally required if a threshold is crossed.
- Other thresholds and actions might be recommended by the monitoring team.
- Some cases might involve several possible management actions, each with a set of possible outcomes whose probabilities can be calculated conditionally on the estimates.

Each of these forms involves an analytical protocol, such as a statistical test or a formal decision analysis, whose effectiveness depends on the reliability of the estimates. Thus the justifications require:

- explaining how each important estimate can lead to management actions, either on its own or as part of a more general assessment of the Delta or a subregion of it;
- describing the protocols that might be used to decide the action;

- explaining why the specified reliability is adequate for these protocols.

The sampling plan is shown to be likely to achieve the desired reliability. The protocols do not pre-empt the role of management, which weighs economic and other concerns as well as protocol results. However, unreliable protocol results may not be worth their cost. Here is an artificial example:

Suppose we test a town's drinking water by giving it to mice and then examining them for cancer. If lab costs limit us to 20 mice, we might issue an alarm if any have cancer. Some mice get cancer anyway: suppose this background rate is 2%. If the town water is in fact safe, the 2% risk gives a 33% chance of getting at least one case and issuing a false alarm. Even if the water doubles the risk of cancer to 4%, the chance of a "true" alarm is only 56%. We can reduce the false alarm rate to 6% by requiring two or more cancers, but then the true alarm rate drops to 19%, so we miss 81% of cases we want to detect. For all its seeming importance, this test is probably too unreliable to use at all.

Determining reliability from a given spatio-temporal sampling plan, or designing a plan to achieve sufficient reliability for a given cost, are not easy tasks. We discuss them in Appendix I not to solve the problem for the Delta program, but to show that solving it requires effort and expertise similar to those needed to choose the most important contaminants and find ways to measure them.

The program allows for ongoing estimates of uncertainty and updates to the sample design as needed. All components of the monitoring program are accompanied by uncertainty. However, as data are collected over time, these uncertainties can be narrowed if the monitoring data are used to update your understanding of how the system works. In some cases, the data might lead to a redirection of effort. For example, the current plan samples mercury in water only because it may suggest ways to control methyl mercury in fish. If, after a reasonable time, the two measures seem to be unrelated, indicating that mercury does not predict methyl mercury even allowing for time lags or flow between sites, it might make sense to drop the mercury sampling and extend the fish sampling.

III. Effectiveness of the current monitoring plan.

This section contains comments from Panel members.

Combining data over time and space

The main weakness of the current plan is given in the Summary: it has little connection to management action or assessment because it does not combine data from different sites and times to form a description of the current state of the Delta, its changes over time, or the processes involved.

The QAPP (p. 12) says the program arose from "shortcomings of existing monitoring efforts to address questions at the scale of the Delta [and] recognition that data from current monitoring programs were inadequate in coverage, could not easily be combined, and were not adequate to support a rigorous analysis of the role of contaminants …" However, there are several statements like the following:

- Communications Plan, p. 9: "The exact methods for data analysis are not prescribed in this plan because doing so would limit the options for the program."
- QAPP (p. 15): "decisions … will be made by the Water Board using its own process. Therefore, the Delta RMP does not have a detailed assessment framework for data interpretation and follow-up."
- Fact Sheet: Ambient Toxicity (p. 3): "Available information allows conclusions about monitored areas and sites but cannot be used to make assumptions about unmonitored areas."
- QAPP (p. 64): "Individual results produced by the Delta RMP are not intended to trigger enforcement actions, even though collectively the data may guide management actions …"

The first statement is misinformed. Statistical methods can change when the data suggest that some assumptions are wrong, but not often and usually only in details such as the covariates to use in prediction or the shape of a probability distribution. The second is an understatement: there is almost no framework at all. The third is confusing: since only sites are monitored, a "monitored area" is an area containing monitored sites, which can be used to "make assumptions" (inferences) about it; it is unclear what an "unmonitored area" is. The fourth seems to regard linking data to action as someone else's job.

This stance contradicts the idea of a monitoring design. Why sample monthly if bi-monthly or annual samples would be nearly as good, and allow more sites? Why are sites for monthly pesticide samples all near the edge of the Delta if these are not informative about interior sites? (Pages 24 and 38 of the MDS lists reasons for site choices but they are vague.) How would one decide whether the proposed design is better than one with half as many times and twice as many sites? The QAPP aims to ensure that results from individual (site, time) samples meet reliability criteria: how are these determined? How would one decide whether to relax some of them so as to add more sites or times, or tighten others due to health risks?

Mercury

Use of Sportfish as a monitoring parameter

What is the goal of the mercury program? If it is to set human fish consumption advisories, then are the sampling locations chosen where people actually fish? Using large sportfish to monitor potential human health impacts from eating those fish makes sense.

However, using sportfish to monitor impacts on MeHg from large restoration projects does not make sense. Large sportfish have fairly large territories/home ranges, so it would be hard to attribute change to a specific restoration action or location. Also, the change would be hard to detect, since large sportfish have higher Hg body burdens that vary more between individual fish. As a result, a small change from a management or restoration action won't stand out. Small, resident fish with small home ranges would reflect such changes more quickly and clearly. Ideally a Before-After-Control-Intervention design could be used.

Sampling schedule

The sportfish are sampled annually. Do we know if mercury varies seasonally in sportfish, as it does in smaller fish? If so, then annual samples are unlikely be adequate unless people catch and consume the fish in only one season, or there is a way to adjust for other seasons (without sampling at those times). If mercury in sportfish varies spatially within a subregion, then sampling one location per subregion is unlikely be adequate. This could be a case where the goal is useful but the effort is far short of what is needed, and thus achieves nothing. How will the data be analyzed to compare trends among sites?

The mercury water samples are monthly. What connects them to the fish tissue samples? Are they at the same sites (including Mokelumne River)? Are they to be compared to the water quality (WQ) criterion of 0.06 ng/L of MeHg in unfiltered water (QAPP, p. 24, Table 3.4)? What will a monthly grab sample at 4 sites in the Delta tell you about MeHg status in the entire Delta? How were the number and locations to be sampled determined? What are the flows at these locations? Will all samples be taken under the same tide/flow conditions?

Additional questions and comments from the panel

- Why is there a low level of fish sampling and a medium level of water sampling? What is the value of the water sampling? How does current fish sampling data relate to previously collected

sampling data? If the primary management question is trends over time, are there existing long term data sets that can be built on. The study plan mentions but does not elaborate on these points (MDS, p. 38).

- How were the bin lengths for the Largemouth Bass determined (QAPP p. 86)? The Central Valley Basin Plan has water quality objectives (WQO) for fish 150-500 mm TL, and for fish <50mm TL, so the proposal's sampling divisions (200-249, 250-304, 305-407 and > 407 mm) are not consistent with this Plan. Fish Hg will often vary by length of fish (surrogate for age). How will the data be compared to WQO? Will bins be analyzed separately? The sampled fish can be assumed random within bins, but not between them; is the plan to fit a regression of fish Hg against length? Note the Basin Plan is specific as to trophic level of fish for the WQO: any alternative predator species should be at the same trophic level.

- Some plan details seem missing or contradictory:
    - Little Potato Slough is not shown on Figure, MDS p. 38.
    - MDS p. 9 says "10 sites" and "350 mm length". QAPP p. 34 shows 6 sites. p. 86, Table 8.3, says 11 fish per site to get 66 individuals, in 4 size bins (200 mm to 500 mm length), so 6 sites.
    - MDS p. 9 says water samples are monthly. QAPP p. 25 says they will be "analyzed" quarterly.
    - QAPP p. 24. Table 3.4. What is the time period for these benchmarks? Are they average concentration? Does the information here agree with Appendix 43 of Central Valley Water Board? Where is the "0.06" level of MeHg in the WQCP? (Is it an average? If so, over what?)
    - MDS p. 39. How will these ancillary parameters be used in analyzing or interpreting the MeHg in water data? Are data relevant to these parameters already being collected by others? e.g., suspended sediment. What are the "conditions" in the footnote?
    - MDS pp. 38, 41. The figure shows 10 fish collection locations, but the table lists 9. <u>It would be helpful to show sampling location numbers in both the figure and the table</u>. Some names in the figure do not match those in table - e.g., "S. Fork of the Mokolumne@ Staten Island" and "Mokelumne River at Benson's Ferry" are in the figure, not in the table, or have different names. It would also be helpful to show the demarcation of the TMDL subareas, since that is the primary rationale for selecting sites.

<u>Pesticides and Toxicity</u>

This is by far the most expensive program and has the potential to become much more so if new or unknown pesticides become an issue. Yet, at present, we do not know the answer to the basic Table 1 question: “What are the spatial and temporal extents of lethal and sub-lethal toxicity?” In fact, we were told at the presentation meeting that so far there has been no observed toxicity, but there was information (not specified) that pesticides may be contributing to toxicity in the Delta.

More detail on toxicity tests is needed. It seems more cost-effective to document the toxicity problem first, by postponing pesticide analyses to pay for toxicity testing over more sites, more widely spread, and during times of year when pesticide use/runoff would be expected to be high. When the sites or areas experiencing toxicity, and the times of the year are known, then samples from these sites and times can be analyzed for the chemicals that might cause that toxicity. This information can then be used to determine source(s), which can then lead to control/management.

Toxicity design

The vague categories used (non-toxic, some, moderate and high: MDS p. 26, 27) are not useful. At present it is proposed to conduct "Pesticide-focused TIEs for samples with > 50% reduction in the organism response compared to the lab control treatment (not to exceed 20% of samples or $40,000)" (MDS p. 21). What criteria led to these numbers? The toxicity tests use "EPA, 2002, Appendix H" (QAPP, p. 61, it should be "2002a"). It is an old t-test (its formal pre-tests are not useful). How the test is to be used (what action it might lead to), and how reliable it should be (a function of sample sizes and variances) are not clearly discussed. (The aims and meaning of the measurement quality objectives column in Table 4.10 is not clear.)

Pesticide sampling design

What samples sizes will be used, and why? In the 2-stage approach above, a decision procedure will be needed to decide which sites and times are candidates for pesticide analysis, and perhaps to choose the pesticides to look for. Thresholds, trigger points, and estimates of reliability will be needed, especially if information from different sites or times is to be combined.

When samples are collected from locations with observed toxicity and analyzed for chemicals, will current use pesticides be the only targets? Is there reason also to consider personal care products, PBDEs (flame retardants), pharmaceuticals, legacy pesticides in sediment (e.g., DDT) or Hg as causes or contributors to observed toxicity?

If protection of human health is a major goal, then sampling is needed in those areas expected to be used as a drinking water sources (e.g., at specific drinking water intake locations). Sampling for pesticides in water not near drinking water intakes (or perhaps recreation areas) does not seem to provide useful information to address this goal.

Sediment sampling design

The plan is not clear about methods for sampling sediments. The QAPP has no information on sediment collection or analysis. Is the Stream Pollutions Trends Monitoring Program (SPoT) collection, toxicity testing and chemistry of sediments considered part of the Delta RMP? Where are those sample locations? A yearly grab sample seems very limited - what is known about the spatial distribution of pesticides in sediment, or their seasonal variation? There are no standards, criteria, or objectives for the prevalence of current use pesticides in sediment, so what would be done with this information? What will the estimated concentrations be compared to in order to evaluate the presence and degree of sediment toxicity? The map on p, 26 of MDS shows there are existing sediment and/or water toxicity test locations in the Delta that have known toxicity (at least within the vague categories). Can these locations be used as negative and positive controls, respectively?

Additional questions and comments from the panel

- Some water samples are scheduled and others triggered by events. If these are to be combined over time, how will they be analyzed? Presumably "event" times have special characteristics, and wet ones are different from dry ones. (This is a question, not a criticism of taking the two types of samples.)
- It seems that monthly samples are not taken when "events" occur. In that case, why are the "event" sites different from the regular sites?
- MDS Table 2, p. 12, does not show toxicity testing.
- QAPP p. 11. There are 3 different entities analyzing water -- does each entity collect its own samples? Can sample collection be consolidated?
- QAPP p. 30. Are the same sites used for both pesticide analyses and toxicity testing?

Nutrients

Monitoring design

One of the initial driving questions (p. 44) is "are there important data gaps associated with particular water bodies within the Delta subregions." It seems appropriate to answer this question before designing the sampling plan and locations for the Delta RMP.

How are tides, flows, and other hydrodynamic conditions considered in choosing where and when to sample?

The MDS (pp. 47-52) shows several ways to display the data, including its variation over time and space. Displays like these are informative, and might help in developing the nutrient monitoring design, or redirect or focus future sampling. However, displays are not a sufficient end point. They do not provide clear criteria for management actions. Such criteria usually need to be numerical estimates, with estimates of reliability. They will arise from comparisons to water quality objectives or other benchmarks of environmental or human health.

We recommend that a PhD-level statistician be added to your team to help develop the nutrient monitoring design.

Synthesis

An allocation of $435,000 seems high for mostly synthesizing the existing data (MDS, pp. 45-52).

Pathogens

Sampling design and data interpretation

It seems too late to make changes in this program. Our concerns over using the data to make inferences about unsampled sites are less here, because many of the sampling sites are important in themselves.

However, it is still unclear what inferences can be drawn about ambient levels elsewhere, which are listed as a goal. How are the sites called "general characterization" (MDS, p. 61) to be used? The Fact Sheet for Pathogens (p. 6-7) says monitoring for ambient levels and sources " should entail representative discharge /effluent locations such as wetlands, urban runoff, POTWs, agricultural/farmland animal areas." It is not apparent that the locations selected for the study are near such areas (see Figure, MDS, p. 62).

Additional questions and comments from the panel

- MDS, p. 14. Pathogens - Cyrptosporidium and Giardia only have narrative WQO - "Waters shall not contain C and G in concentrations that adversely affect ...MUN beneficial uses." What is that level? How do we know what a reasonable detection limit needs to be?
- MDS, p. 60. This involves "triggers". What are they and how are they determined?
- MDS, p. 61. Fate and transport should include a consideration of hydrodynamics. How will sources be identified with this study design?
- QAPP, p. 31.
    - Another program is also collecting pathogens at different sites? Are the analytical methods, quantification limits, etc. similar between the lab that MWQI uses and that which RMP uses?
    - "MWQI … at each of the locations shown in Table A-1…" There is no Table A-1.

QAPP, p. 112. Table 3.5 lists values for Cryptosporidium only - are those values what the monthly sampling will be compared against? What will the Giardia sample results be compared against?

IV. Other Comments

Earlier programs
In what specific ways were former/current monitoring programs "not adequate"? (QAPP, p. 12). Was there a report that evaluated the programs and identified specific deficiencies and made recommendations for improvement? If so, it would be helpful to address how this plan makes up for prior monitoring program deficiencies.

Water Quality Objectives.
What are the time frame definition for "acute" and "chronic" in the WQO or WQC (QAPP, p. 17)? Many of the samples in the Specific Monitoring Designs are monthly grab samples, so it is not clear that the sampling timeframes are consistent with the evaluation criteria. If they are not, then how is Delta RMP to be used for its primary objective, to assess whether Beneficial Uses are being impaired?

Maps and tables.
Sampling location numbers should be given in all maps and tables. Much time can be wasted trying to link them.

Lab measurements (QAPP p. 48.)
Is the plan to compare concentrations in water to water quality objectives/criteria or other benchmarks? Are these reporting limits and method detection limits sufficiently below the benchmarks that there is confidence in the quantification of the concentration?

What are the detection limits/limits of quantification for the analyses (QAPP p. 93)? These limits can be lab specific. It is not clear from the information provided in QAPP, whether the stated analytical methods are able to accurately detect concentrations at or near the WQO or WQC.

Adaptive design.
QAPP (p. 78) says "Collected data are used to evaluate future data needs and adjust the sampling and analysis plan as needed to optimize data collection in an adaptive manner. The program will be continually adjusted to optimize data collection." There seems to be nothing on how this is to be done.

Graphs.
Pie charts should not be used: a table or bar graph is always better. Fake dimensions should not be used. The main value of plots is to convey much information clearly and succinctly, but thought and explanatory text are often needed; **MDS**, p. 28, contains much information but is uninterpretable (other than high scores for Diuron). Plots on p. 52 are better, but still need summarization of both the messages and their reliability.

List of Acronyms

ASC: Aquatic Science Center (aka SFEI-ASC)
AHPL: Aquatic Health Program Aquatic Toxicology Lab
CCWD: Contra Costa Water District
CEDEN: California Environmental Data Exchange Network
CEQA: California Environmental Quality Act
CMP: Coordinated Monitoring Program
CRMP Certified Reference Materials
CVDWPWG: Central Valley Drinking Water Policy Workgroup
CVWQCB: Central Valley Water Quality Control Board
Delta RMP/DRMP: Delta Regional Monitoring Program
DPR: Department of Pesticide Regulation
DTMC: Delta Tributary Mercury Council
DWR: (California) Department of Water Resources
DO: Dissolved oxygen
DOC: Dissolved organic carbon
DON: Dissolved organic nitrogen
DSM2: Delta Simulation Model II
DSP: Delta Science Program
EC: Electrical conductivity
ELAP: Environmental Laboratory Accreditation Program
EMP: Environmental Monitoring Program
ESWTR: Enhanced Surface Water Treatment Rule
FWS: US Fish and Wildlife Service
FY: Fiscal Year
Hg: Mercury
IEP: Interagency Ecological Program
IEP-EMP: Interagency Ecological Program Environmental Monitoring Program
ILRP: Irrigated Lands Regulatory Program (part of Central Valley Regional Water Quality Control Board)
LT2: EPA's Long Term Enhanced Surface Water Treatment Rule outlining monitoring requirements for Cryptosporidium.
LRM: Lab Reference Material
MeHg: Methylmercury
MDL: Method Detection Limits
MLML: Moss Landing Marine Laboratory
MPSL: Marine Pollution Studies Lab at Moss Landing Marine Lab
MQO: Measurement Quality Objectives
MS: Matrix Spikes
MSD: Matrix Spike Duplicate
MST: Microbial source tracking
MWQI: Municipal Water Quality Investigations (a Department of Water Resources program)
NELAP: National Environmental Laboratory Accreditation Program
NPDES: National Pollutant Discharge Elimination Systems (US EPA permit program)
NWQL: National Water Quality Laboratory
OCRL: Organic Carbon Research Laboratory
PCR: Polymerase chain reaction
POC: Particulate organic carbon
POD: Pelagic Organism Decline
POTW: Publically Owned Treatment Works (sewage treatment facilities)
QA: Quality Assurance

QAO: Quality Assurance Officer
QAPP: Quality Assurance Program Plan
QC: Quality Control
RL: Reporting Limit
RMP: Regional Monitoring Program
RPD: Relative Percent Difference
RWQCB: Regional Water Quality Control Board
SC: Steering Committee (of the Delta RMP)
SDWA: South Delta Water Agency
SFEI: San Francisco Estuary Institute (now SFEI-ASC)
SFRWQCB: San Francisco Regional Water Quality Control Board
SPoT: Stream Pollution Trends monitoring (a program of the California Department of Water Resource's Surface Water Ambient Monitoring Program)
SRWTP: South Delta Water Agency
SWAMP: Surface Water Ambient Monitoring Program (a California Department of Water Resource's division)
TAC: Technical Advisory Committee (of the Delta RMP)
TDN: Total Dissolved Nitrogen
TIE: Toxicity Identification Evaluation
TL3: Trophic Level 3
TMDL: Total Maximum Daily Load
TN: Total Nitrogen
TSS: Total Suspended Solids
TST: Test of Significant Toxicity
UCD: University of California, Davis
USEPA: US Environmental Protection Agency
USGS: US Geological Survey
UVA: Ultra-violet absorbance
WDR: Waste Discharge Requirements
WPCL: Water Pollution Control Lab
WTP: Waste Treatment Plant

References

Gitzen, R.A., J.J. Millspaugh, A.B. Cooper, and D.S. Licht (eds). 2012. "Design and Analysis of Long-term Ecological Monitoring Studies". Cambridge University Press, 2012.

Nichols, J.D., and B. K. Williams. 2006. Monitoring for conservation. TRENDS in Ecology and Evolution 21:668-673.

Pollock, K.H., J.D. Nichols, and T.R. Simmons. 2002. Large scale wildlife monitoring studies: statistical methods for design and analysis. Environmetrics 13:105-119.

Stevens, D.L., Jr. and A. R. Olsen. 2004. Spatially balanced sampling of natural resources. J. American Statistical Association 99:262-278

Urquhart, N.S., and T. M. Kincaid. 1999. Designs for detecting trend from repeated surveys of ecological resources. J. Agricultural, Biological and Environmental Statistics 4:404-414.

APPENDIX 1

This section discusses sampling plans intended to describe an area over time. While some sites might be chosen for their importance (e.g., drinking water intakes), their information may apply only to small neighborhoods. The plans we discuss here aim to provide reliable estimates for the entire area, at single points in time and over periods.

Overview

In the following we briefly outline some basic principles of successful and defensible environmental monitoring programs. To our knowledge, the single best source that discusses these principles in detail is a 2012 book entitled: "Design and Analysis of Long-term Ecological Monitoring Studies", edited by R.A. Gitzen, J.J. Millspaugh, A.B. Cooper, and D.S. Licht. In chapter one of that book, the editors state that "… inadequate attention to qualitative and quantitative design issues has been reported to be a common problem in environmental monitoring programs…" We agree with this statement and believe it characterizes several key weaknesses in the Delta Regional Monitoring Program, particularly the lack of quantitative design and analysis details in the RMP monitoring design document. The reality is that development of effective sampling design and analytical methods for monitoring programs involves complex quantitative issues that require extensive engagement by an environmental statistician.

General Principles

Monitoring programs must be efficiently administered, adequately funded, supported by the clients of the monitoring program, have effective data management procedures and regular reporting schedules. However, our focus here is on the essential analytical components for environmental monitoring. Fortunately, there is a strong consensus in the scientific literature on the essential components of monitoring programs designed to assess status and trend. The key requirements are to:

1) Specify objectives in terms of measurable attributes
2) Identify the monitoring state variables (e.g., indicators) and why they were selected
3) State the spatial and temporal domain of the population of interest (i.e., the sample frame)
4) State the type of change to detect
5) Specify the magnitude of change to detect (effect size; essential for sample design decisions)
6) Following (5), specify desired precision for the trend estimate (requires pilot data and a components of variance analysis)
7) Generate estimates of uncertainty
8) Specify 'trigger point' (thresholds) that will lead to a management response
9) Specify the management action that will occur
10) Determine (monitor) the effects of the management actions
11) Update design as needed (adaptive monitoring)

All of the above steps are important but program components cannot compensate for inadequate attention to design and analytical issues. Specifically, we believe that the statistical model(s) to be used for analysis must be decided upon early in the process. Given specific monitoring state variables (indicators), sampling objectives such as desired statistical power, effect sizes, and statistical precision require a priori identification of specific statistical methods. Failure to do this makes it impossible to perform basic sample size calculations and to optimally allocate sampling effort across time and space. This also ensures that limited project funding is used in the most efficient way and is not wasted. Decisions on sample designs, methods of analysis, and variance components analysis go hand-in-hand and should occur before major data collection begins.

To clarify the components of variance concept, we assume a design in which each site is visited in each of a set of years. Given this assumption, the key components of variation are (see expanded discussion by Scott Urquhart in chapter 7 in Gitzen et al. 2012):

1) Spatial: variation among sample units (sites); treated as a random effect in an ANOVA model
2) Temporal: how much the state variable varies from year-to-year across all sample units; treated as a random effect
3) Space by time interaction: how much the state variable changes across time within a sample unit independent of changes in other sample units
4) Error variance

Partitioning the total variance is expressed as: $\sigma^2_{Total} = \sigma^2_{site} + \sigma^2_{time} + \sigma^2_{site\ x\ time} + \sigma^2_{error}$

To estimate trend, we must first assume a model for how the response variable (e.g., indicator value at sample unit i) changes over time. For example, if we assume a simple linear time-trend model for the indicator, y, our model is:

$$y_{ij} = \mu + S_i + T_j + \varepsilon_{ij}$$

where,

$y_{ij} =$ the value of the state variable at site i in year j

S_i = effect of site i

T_j = effect of year j; {j = 1, 2, … , t}

$\varepsilon_{ij} =$ error term

Then our estimation model for a linear trend, assuming a common trend across sample sites, is:

$$\hat{y}_{ij} = \beta_0 + \beta_{1j} + \varepsilon_{ij}$$

where,

β_1 estimates trend

$\beta_0 + \beta_1(t+1)/2$ estimates 'status'

The null and alternative hypotheses of interest are, respectively: H_0: $E[\beta_1] = 0$; H_a: $E[\beta_1] \neq 0$. That is, to detect trend we test the null hypothesis that no trend is present in the indicator against the alternative hypothesis that a trend is present. The ability of a monitoring program to detect trend when it is truly present is referred to as its statistical power.

The best source of information for a component of variance analysis is from preliminary survey data. These preliminary data also provide information essential for sample-size calculations and determination of an optimal sampling design.

Design-based or Model-based

There are two broad categories of environmental monitoring programs—design-based and model-based. Both require that the target population and the sample frame be clearly defined in order to avoid the potential for confounding arising from changing frame errors. Those programs that use design-based inference use the selection probabilities of the sample units to calculate an estimate for the statistical population and provide estimates of uncertainty. In contrast, programs that use model-based inference assume an a priori statistical model for the distribution of indicator values and do not require a probability based sample design. The following discussion develops this distinction further.

At each sample site i there is an observable value Z_i for the indicator attribute. In a designed-based view, Z_i is a fixed quantity. Any probabilistic process that may have produced Z_i is unknown and irrelevant. The probabilistic component of the data arises from the sample design itself (i.e., a simple random sample with equal probability of inclusion for each sample unit).

In contrast, in a model-based view, Z_i is a random variable—a random realization from a statistical model, such as a normal, with mean μ and variance σ^2. The values $Z_1, Z_2, \ldots, Z_N$ at any time t are just one outcome of many possible outcomes under the statistical model. Under this model, the sample design that provides the data is irrelevant.

In the design-based view, if the goal is to estimate the population mean, then we simply compute:

$$\hat{\bar{Z}} = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

Even if the entire population, N, had been sampled and the mean was based on a census, the estimate provides no insights to μ since we have observed only one realization from the statistical distribution. Generally, n << N, and there is uncertainty about <u>both</u> the realized mean (due to sampling variance) and the parameters of the statistical model that generated the Z's.

In contrast to a design-based approach, if we use a model-based approach and re-compute the mean, as above, from the sample of size n (where the sample design is irrelevant) then the expected value of the sample mean is:

$$E\left[\hat{\bar{Z}}\right] = \mu$$

Designed-based inference makes three assumptions: 1) the values, Z_i, that are measured at each sample unit are fixed quantities; 2) the only source of error in the population estimate is due to sampling variation—that is, no distributional assumptions are made about the data; and 3) all values are measured perfectly.

In contrast to designed-based, model-based inference assumes: 1) there is some statistical process that generated the observed data—the super-population model; 2) we have an approximating model—that is, an a priori hypothesis that we can translate into a well-defined model; 3) our approximating models lies close to truth. In general, analyses for model-based programs are considerably more complex than for design-based programs.

Many environmental attributes, including indicator values in the RMP, are likely generated by dynamic processes. Because of their inherent dynamics, measured indicator values have two sources of uncertainty—uncertainty arising from the sampling process and uncertainty about the underlying statistical processes that generate the observed values. Thus, model-based designs may seem most appropriate because they better characterize the generating process for the indicator values. However, based on our knowledge of environmental monitoring programs, design-based approaches are most common. The primary reason is that there is seldom sufficient knowledge of the system to develop a strong a priori hypothesis about the statistical generating model for the data. The generating process is likely to be extremely complex due to the complexity of natural systems, particularly those disturbed by human drivers. It is usually very difficult to identify all of the un-modeled (and unknown) environmental factors that affect the assumed statistical model for the data.

In practice, many environmental monitoring programs are a hybrid of design-based and model-based components. For example, in wildlife and fishery studies, estimating the abundance, and temporal trend

in abundance, of a harvested species is a common objective (Pollock et al. 2002). In this case, abundance in sample unit *i* is most often assumed to be fixed during the survey period (designed-based) but it is recognized that abundance is estimated with error. As a result, an observation model is adopted to model uncertainty in the measurement process. This model estimates the probability of detection, *p*, conditioned on the animal's presence in the sample unit. Based on the number of animals counted in a sample unit (C_i), the adjusted estimate of abundance is then given by:

$$\hat{N}_i = \frac{C_i}{\hat{p}}$$

Inference to the Target Population

The goal of environmental monitoring programs is to make inference to the status and trend of the entire target population based on a sample of that population. Making inference to indicators values at un-sampled locations is inherently a model-based task. If the program for indicator estimation is model-based to begin with, then extrapolation from the sample data to un-sampled locations is more direct than for designed-based programs.

Because design-based monitoring is grounded in a random sample design where all potential sampling units have a non-zero inclusion probability, inferences can be made to the entire sample frame. However, this extrapolation is not spatially explicit—that is, it does not allow prediction at the scale of un-surveyed sample units. Extrapolation to this scale can be accomplished by measuring one or more covariates at the sample locations. This is followed by estimating a statistical model that relates spatial variation in the indicator values—for example, by means of multiple regression—to the covariates. Prior knowledge, or measurement, of the covariate values at the un-sampled locations allows one to predict (with uncertainty) indicators values throughout the study area.

Extended Discussion

This section is a short account of model-based and design-based sampling plans. It also uses a simple design-based plan to show how the intended data analysis can help guide choices of sites and times.

To outline the problem, suppose we are to monitor a variable, Z, the level of a contaminant, over a region, R, for a time interval, T. At each site s = (x, y) = (Latitude, Longitude) in R, and each time t in T, there will be a value of Z, say Z(s, t). Our goal is to estimate some summary of these values, like the average over both space and time, say $\bar{Z}(\blacksquare, \blacksquare)$. (The "■" indicates we have taken the mean over all values of the missing variable.) We cannot observe Z for all sites and times. We need a set of (s, t) choices, say (s_1, t_1), (s_2, t_2), …, (s_n, t_n), so we can get a good estimate of $\bar{Z}(\blacksquare, \blacksquare)$, say $\hat{Z}$, by applying a formula (which we must devise) to the values $Z(s_1, t_1)$, …, $Z(s_n, t_n)$.

The analysis (inference) step is to measure the reliability of $\hat{Z}$. The most common measure for unbiased estimates is the standard error (SE). This is a hypothetical value obtained by imagining the entire process (the area obtains its Z values, we choose sites and times, get the Z values and apply the formula) being repeated over and over. If sites, times and Z values were the same for each repetition, then all $\hat{Z}$ values would be the same and the measure would be useless. Thus chance must enter into the (s, t) choices or the Z values, or both. (This account assumes that Z(s, t) is observed exactly; otherwise there is observation error which can be estimated from individual samples and is often smaller.)

In a model-based approach, the value of Z at a given site s and time t is treated as a random variable, resulting from natural processes occurring over space and time. These are described by a model, called a superpopulation model, as if the full collection of Z values is randomly chosen from a set of possible

collections. Even for fixed (s, t) choices, imaginary repetitions of the sampling process will give different Z(s, t) values. Each Z(s, t) has a variance and each pair, Z(s, t) and Z(u, w), has a covariance. If these are known, the SE of $\hat{Z}$ can be calculated for any set of (s, t) choices.

However, the conditions where this approach is effective, and the questions it answers best, are different from those of the Delta Regional Monitoring Program. It focuses on modeling the processes represented by the data and predicting what they will do in future. It estimates summaries of the actual Z(s, t) values, such as $\overline{Z}(\blacksquare, \blacksquare)$, only in passing: its real targets are the parameters of the underlying process. For this, it must have a suitable model which is detailed enough to use values at one set of times and sites to help predict values at another set. Usually this requires models of specific physical, chemical or biological processes which operate at many scales but combine to have effects at the larger scale. It also requires large amounts of data to help distinguish between competing models. These questions and conditions apply to the study of climate, but not to the Delta program. Here the questions concern current (actual) status and trends - which are descriptions of (actual) data over time. Concerns about the future are not based on specific causative models but on the belief that the trends are the result of continuing human activity. There are no detailed models for the Delta-wide processes and not enough detailed past data to generate and assess them. We therefore do not discuss model-based designs further.

A "pure" <u>design-based</u> approach is to choose (s, t) pairs randomly, using computer-generated random numbers. The full set of Z(s, t) values is assumed fixed though unknown. When the process is repeated (in imagination) to get the SE, the values of Z are observed at a different set of (s, t) choices. Thus $\hat{Z}$ will vary between repetitions. Its SE depends on the variation of the full set of Z(s, t) values, and the chance comes entirely from the random (s, t) choices.

Usually the selection is more structured. We separately choose random sites and a set of times sufficiently spaced so Z values at different times can be assumed to be independent. If the sites are s(1), s(2), …, s(m) and the times are t(1), t(2), …, t(n), then our observations are the values Z(s(i), t(j)) for each of the mn combinations of an i and a j.

A natural estimate of the average of <u>all</u> Z(s, t) values (giving all sample sites and times equal weight) is the average of <u>observed</u> Z(s(i), t(j)) values:

$$\text{Average of sample values} = \Sigma\Sigma Z(s(i), t(j))/mn.$$

The variance of this average is (after some algebra):

$$\text{Var}\{\Sigma\Sigma Z(s(i), t(j))/mn\} = \sigma_S^2/m + \sigma_T^2/n + \sigma_{Int}^2/mn \qquad ***$$

where,

σ_S^2 = Variance over all sites, s, of $\overline{Z}(s, \blacksquare)$ which is the mean over all times of Z(s, t).

In other words, get the mean over time of each site; then get the variance of these means.

σ_T^2 = Variance over all times, t, of $\overline{Z}(\blacksquare, t)$ which is the mean over all sites of Z(s, t).

σ_{Int}^2 = Variance due to interaction.

One way (of many) to describe σ_{Int}^2 is as "variance due to non-additivity". If the values of Z(s, t) were additive over sites and times, then all sites would change over time in unison. If one site went up by 5 from year 1 to year 2, then they all would. If that were the case, then

$$Z(s, t) = \overline{Z}(\blacksquare, \blacksquare) + [\overline{Z}(s, \blacksquare) - \overline{Z}(\blacksquare, \blacksquare)] + [\overline{Z}(\blacksquare, t) - \overline{Z}(\blacksquare, \blacksquare)]$$

= overall mean + site effect + time effect.

Variance due to interaction is the mean squared difference between the left and right sides = the mean square of the error you would make if you assumed Z(s, t) was additive.

The message of the starred variance formula is:

If sites vary more than times ($\sigma_S^2 > \sigma_T^2$), choose more sites (large m);
If times vary more than sites ($\sigma_T^2 > \sigma_S^2$), choose more times (large n).

This message is oversimplified, but is still a useful guide. It ignores the interaction term, but this is reduced by increasing either m or n, and usually plays a smaller role: σ_{Int}^2 is unlikely to be larger than both σ_S^2 and σ_T^2 (since sites are likely to go up or down similarly over time, though not exactly) and its divisor (mn) is larger.

The model is also oversimplified. In practice, the random selection of sites would be "spatially balanced" so that sites will not be chosen too close together. However, this and the even spread of times, are responses to variation. They separate the range of sites, R, or the range of times, T, into strata that are more homogeneous than R or T as a whole. The number of strata needed for a factor (sites or times) will tend to be higher when the factor is more variable.

We don't know σ_S^2 or σ_T^2 (or σ_{Int}^2). However, we often have some idea at the outset as to whether Z varies more over space or over time, especially if there are preliminary data. If the over simple analysis is biased, it may be in favor of adding times. If multiple observations are taken each year, it might be possible to reduce the variance over time by including a small number of parameters to describe seasonal effects. If so, fewer times would be needed.

In realistic situations there are additional problems. There is usually more than one "Z" (contaminant) and more than one goal (e.g., averages, trends or spikes overall or in subregions). The area of interest can be irregular and poorly defined: for example, a map of "perennial streams (may include) many ephemeral or intermittent streams, or long-dry channels". Sites are not usually equally important or equally accessible. (Strictly, design-based inferences cannot apply to sites that could not have been chosen.)

The design-based approach can be modified to deal with such problems, usually with the aid of an informal model. On average, Z values will differ less between sites that are closer together, so the region can be divided into strata and an appropriate number of sites randomly selected from each stratum. The strata could be defined by other characteristics too. The probability of selection can vary among sites if "there are … scientific, economic, or political reasons for sampling some portions of a resource more intensively than others". The design might allow for updating when the data or other new information cause "the 'important' subpopulations (to) change, necessitating a corresponding change in sampling intensity" or we find that some planned sites are unusable. These quotes and a design with these three features (generalized random-tessellation stratified: GRTS) are given by Stevens and Olsen (2004). Standard methods allow estimates of means and other simple (linear) summaries; estimates of SE are harder. They illustrate the design in four surveys of Indiana river systems. See also Chapter 6 in Gitzen et al. (2012).

The GRTS design is over space. When sampling is over time as well, new decisions are needed. In practice, as above, times may be equally spaced (perhaps within a season). Observations at different times are assumed to be independent, but small time gaps may cause dependence which needs to be modeled. "Panel plans" have different visiting schedules for different groups of sites (panels), so more sites are covered but less often. For example, two panels may be visited in years 1, 3, 5, … and 2, 4, … respectively. There are many such plans: e.g., see Urquhart and Kincaid (1999).

The aim of this discussion is not to urge adoption of some design off the shelf. It is to make two points. The first is that all useful monitoring plans

(a) have goals that require linking observations taken at different times or sites into estimates of summaries, like means or trends, which can help determine management actions, and
(b) give the reliability of these estimates a major role in the selection of sampling times and sites.

The second is that methods for achieving these goals have been studied for several decades by many able people. None have developed designs specifically for the Delta RMP, but even the simple models can provide guidance (as above), and it is likely that some of their more detailed work can be used. The references below may help, especially the book by Gitzen et al (2012). However, this is a very short list. More important is a team member who can use these and other references to work with the rest of the team to develop a monitoring plan that attends to items (a) and (b) above and clearly addresses the management and assessment questions.