
# PRINCE WILLIAM SOUND RCAC
# LONG TERM ENVIRONMENTAL MONITORING PROGRAM REVIEW

## OCTOBER 2007

**Submitted to:**

**Joe Banta**
**Project Manager**
**Prince William Sound Regional Citizens' Advisory Council**
**3709 Spenard Road**
**Anchorage, Alaska 99503**

**Submitted by:**

**BGES, INC.**
**750 West 2nd Ave., Suite 104**
**Anchorage, Alaska 99501**
**Ph: (907) 644-2900**
**Fax: (907) 644-2901**

**Eagle River Office**
**(907) 696-BGES (2437)**

# PRINCE WILLIAM SOUND RCAC
# LONG TERM ENVIRONMENTAL MONITORING PROGRAM
# PRELIMINARY REVIEW

PWSRCAC CONTRACT NUMBER 960.07.2

**Joel H. Reynolds, Ph.D.**
Solutions Statistical Consulting
6601 Chevigny Street
Anchorage, Alaska, 99502
(907) 248-4776
solutionsconsulting@ak.net

**Nick R. Braman**
BGES, Inc.
750 West 2nd Avenue, Suite 104
Anchorage, Alaska, 99501
(907)-644-2900
nick@bgesinc.com

# TABLE OF CONTENTS



## LIST OF TABLES



## LIST OF APPENDICES

## LIST OF ACRONYMS

**Stations:**

| | |
|---|---|
| AMT | Alyeska Marine Terminal, Port Valdez |
| AIB | Aialik Bay, west of Seward |
| DII | Disk Island, Knight Island Group, western PWS |
| GOC | Gold Creek, Port Valdez |
| KNH | Knowles Head, Eastern PWS |
| SHB | Sheep Bay, Eastern PWS |
| SHH | Shuyak Harbor, Kodiak |
| SLB | Sleepy Bay, LaTouche Island, Western PWS |
| WIB | Windy Bay, Outer Kenai Peninsula |
| ZAB | Zaikof Bay, Montague Island, Central PWS |
| ABL | NOAA/NMFS Auke Bay Laboratory, Juneau AK |

**Other:**

| | |
|---|---|
| AHC | aliphatic hydrocarbons |
| ANS | Alaskan North Slope |
| BCF | biological concentration factor |
| BWTF | Alyeska Terminal's Ballast Water Treatment Facility |
| CFR | Code of Federal Regulations |
| DSI | Dissolved Signal Index |
| EVOS | *Exxon Valdez* oil spill |
| EVTHD | *Exxon Valdez* Trustees Hydrocarbon Database |
| EMAP | US EPA Environmental Monitoring and Assessment Program |
| EMP | Environmental Monitoring Program |
| FFPI | fossil fuel pollution index |
| GC/FID | gas chromatography/flameionization detector |
| GC/MS | gas chromatography/mass spectroscopy |
| GERG | Geochemical and Environmental Research Group, Texas A&M |
| GOA | Gulf of Alaska |
| $H_0 1$ | Hypothesis 1 |
| KLI | Kinnetic Laboratories, Inc., Anchorage AK |
| LTEMP | Long-Term Environmental Monitoring Program |
| MDL | analytic method detection limit |
| MQL | Method quantitation limit |
| NIST | National Institute of Standards and Technology |
| NMFS | National Marine Fisheries Service |
| NOAA | National Oceanic and Atmospheric Administration |
| NS&T | National Status and Trends |
| PAH | polycyclic (or polynuclear) aromatic hydrocarbons |
| PECI | Payne Environmental Consultants, Inc., Encinitas, CA |
| PGS | particle grain size |
| PSI | Particulate Signal Index |
| PWS | Prince William Sound |
| QA/QC | quality assurance/quality control |
| RCAC | Regional Citizens' Advisory Council |
| SHC | saturated hydrocarbons (same as AHC: n-alkanes + pristane and phytane) |
| SIM | selected ion monitoring |
| SPMD | semi-permeable membrane device |

| | |
|---|---|
| SRM | NIST standard reference material |
| TAHC | total AHC |
| TALK | total n-alkanes |
| TIC | total inorganic carbon |
| TOC | total organic carbon |
| TPAH | total PAH |
| TSHC | total saturated hydrocarbons (same as TALK) |
| T/V | Tank Vessel |
| UCM | unresolved complex mixture |

## 1.0 INTRODUCTION

The Prince William Sound Regional Citizen's Advisory Council (RCAC) established the Long Term Environmental Monitoring Program (LTEMP) in 1993 with the overarching goal of

> "...identify[ing] present and potential future adverse impacts on the ecosystems of Prince William Sound (PWS) and the Gulf of Alaska (GOA) ...as a consequence of oil transportation," (KLI 1993).

The LTEMP provides the monitoring component for a larger adaptive management program for oil transportation system in PWS (KLI 1993). Periodic peer review is required to maintain an effective long-term monitoring program (MacDonald and Smart 1993, Lovett et al. 2007) as aspects of the program naturally tend to change through time, both in terms of explicit evolution driven by past findings, a sign of program success, and implicit changes in objectives, strategies, and protocol components, such as analytical procedures. This appears to be the first explicit review of LTEMP as an effective long-term monitoring program.

## 2.0 REVIEW FRAMEWORK

This review addresses how well LTEMP is (i) fulfilling its role as an integral part of an adaptive management program (Nichols and Williams 2006, Williams et al. 2007) and (ii) meeting the requirements of an effective monitoring program (MacDonald and Smart 1993, Noon 2003, Lovett et al. 2007). Necessary components of an effective monitoring program are presented in Table 1. Specific attention is given to identifying missing or weak components of the monitoring program and to performing a technical review of protocol components including sampling design, choice of indicator measures and summary statistics, data analysis and visualization, and data management and communication of results. Recommendations are provided in each section and summarized at the end of the report.

**Table 1**. Monitoring program components and effective characteristics, synthesized from Noon 2003, Lovett et al. 2007, Williams et al. 2007.

| Monitoring Component | Characteristics of an Effective Program |
|---|---|
| Goals | Clearly defined, based on compelling scientific or management questions; monitoring included as part of a scientific research or adaptive management program |
| Objectives | Specific, measurable, achievable, results oriented (Williams et al. 2007) |
| Protocols | Each component detailed in writing, modifications archived including their justifications |
| Sampling Design | Appropriate for desired inference scale, efficient |
| Revisitation Design | Appropriate frequency for process of interest and management response |
| Laboratory Analyses | Maintain quality and consistency of the data |
| Indicator Measures / Summary Statistics | Careful, logical selection based on theory or empirical research with consideration given to future information needs; linkage to cause-and-effect interpretation of signals; definable critical threshold values to trigger management response (if appropriate) |
| Data Management & Archiving | Provides for long-term data accessibility and sample archiving |
| Data Analysis & Visualization | Continually examine, interpret, and disseminate the monitoring data |
| Reporting & Communication | Include review, feedback, and adaptation in the monitoring program |

The goals of the monitoring component in an adaptive management framework should include insight into system status, feedback on effects of management actions, and learning about the process being monitored in order to improve both management and the monitoring program itself (Williams et al. 2007, Noon 2003). LTEMP's goals, as stated in the original study plan (KLI 1993) and periodic reviews (Payne et al. 1998 a,b, 2006), were compiled and assessed with regard to these three general goal classes.

The goals motivate the specific monitoring objectives, which should exhibit a number of properties summarized in Table 1 (Williams et al. 2007). LTEMP's objectives were compiled from the original study plan and periodic reviews, explicitly listed, and each was assessed with regard to these properties. If the suite of objectives did not adequately address the goals, missing objectives were identified and recommended for future explicit inclusion.

The monitoring program objectives dictate the survey protocols, which, in turn, specify the details of data collection, laboratory analyses, summary statistics and indicators, data analysis, data management, and reporting (Table 1). The majority of this review focused on the key aspects of the protocols, especially the choice of indicators/summary statistics, data analysis, treatment of observations below the

detection limit, and data visualization.

**LTEMP Goals**

The original LTEMP study plan explicitly identified three goals for the LTEMP monitoring program (KLI 1993):

> Goal 1: Identify present adverse impacts from oil transportation activities on the ecosystems of Prince William Sound (PWS) and the Gulf of Alaska (GOA),
> Goal 2: Identify future adverse impacts from oil transportation activities on the ecosystems of PWS and the GOA
> Goal 3: Provide an information basis for recommending future mitigation measures.

In terms of the broad goal classes for effective monitoring, Goal 1 focuses on system status while Goal 2 and Goal 3 focus on feedback to the larger management process. Missing is an explicit goal focused on improving the monitoring process itself, in terms of efficiency and effectiveness, by increasing insight into the transport and uptake processes by which hydrocarbon releases impact the ecosystems. For example:

Goal 4: Improve LTEMP sampling, indicators/summary measures, and analysis methods to increase efficiency and effectiveness at detecting release events and their sources.

Though not explicitly stated, this goal has been implicitly recognized given RCAC's support of efforts to use past data or supplemental studies to provide insight into the pathways of hydrocarbon impacts and thus assess sampling designs, indicators/summary measures, and analytical methods for detecting release events and their sources (e.g., Salazar et al. 2001, Payne et al. 2001).

In accordance with the three explicitly stated goals, LTEMP was designed to statistically test four null hypotheses (KLI 1993). These null hypotheses ($H_0$) are stated explicitly in the original project plan:

> $H_0$1: There are no changes in biological, chemical, or physical variables with time at various monitoring sites.
>
> $H_0$2: Observed changes in biological, chemical, or physical variables at various monitoring sites are not correlated with oil transport activities.
>
> $H_0$3: There are no differences in biological, chemical, or physical variables with time between various monitoring sites.
>
> $H_0$4 - Observed differences between monitoring sites are not correlated with oil transportation activities.

Null hypotheses 1 and 3 ($H_0$1 and $H_0$3) require relatively accurate quantification of analytes as well as a relatively low degree of variability between replicate measurements for hypothesis testing. Null Hypotheses 2 and 4 ($H_0$2 and $H_0$4) require adequate analysis of hydrocarbon fingerprints and key marker analytes to differentiate between oil transport activities and other sources of hydrocarbons. Typically this has been approached in the LTEMP project by attempting to differentiate biogenic, petrogenic, and pyrogenic hydrocarbon signatures. In the case of pyrogenic hydrocarbon sources a variety of sources, such as unreported diesel spills and natural hydrocarbon seeps present in PWS, can be further differentiated.

Contrary to the description often employed in the LTEMP reports, 'measuring hydrocarbon background levels in the Exxon Valdez Oil Spill (EVOS) region as long as oil flows through the pipeline' is neither a goal nor an objective and thus should not be given as the underlying motivation for the program.

**Recommendations:**

- The Regional Citizens' Advisory Council (RCAC) Science Advisory Council (SAC) should explicitly identify the goals of the LTEMP program and clarify that 'taking measurements' is neither a goal nor an objective, but rather is a means to achieving various objectives.
- The SAC should consider explicitly identifying a fourth LTEMP goal focusing on improving program efficiency and effectiveness by using past data and possible supplemental studies to improve and analysis methods for detecting release events and their sources. For example: Goal 4: Improve LTEMP sampling, indicators/summary measures, and analysis methods to increase efficiency and effectiveness at detecting release events and their sources

**LTEMP Objectives**

The original LTEMP study plan explicitly identified eight objectives derived from Goals G1 - G3 (KLI 1993):

- G1: Identify present adverse impacts
  - O1.1: Monitor recovery of sites impacted by the T/V *Exxon Valdez* in 1989 (EVOS), via comparison to appropriately selected paired control sites
  - O1.2: Estimate recovery rates of EVOS impacted sites
- G2: identify future adverse impacts
  - O2.1: Develop a baseline of background hydrocarbon levels (and sources) at study sites
  - O2.2: Develop a baseline of annual variation in background levels (and sources) at study sites
  - O2.3: Develop baseline of spatial variation in background levels (and sources) among sites
  - O2.4: Detect release events
  - O2.5: Identify sources of release events
  - O2.6: Monitor long-term inputs from the Alyeska Pipeline Service Company's Ballast Water Treatment Facility (BWTF)

- G3: provide an information basis for recommending future mitigation measures
  - O1.2, 0.2.1-3

Objectives O1.1-2 and O2.1-5 were each defined with regard to hydrocarbon levels in two specific substrates at each study site: the tissues of intertidal mussels and near-shore, sub-tidal, sediments. Mussel tissue sampling has been maintained throughout the duration of the program, while sediment sampling now only occurs at the Alyeska Pipeline Service Company's BWTF.

The objectives ostensibly require three distinct site selections: (i) O1.1 requires selection and monitoring of a representative set of impacted sites and a set of control sites (un-impacted), ideally in a paired set up for efficiency; (ii) O2.1 requires selection and monitoring of control (un-affected) sites representative of the range of conditions in the area (PWS, GOA); (iii) O2.6 requires selection and monitoring of representative sites at the BWTF.

The other objectives identify analyses to be conducted on the measurements resulting from whatever strategies and protocols are chosen for achieving objectives O1.1 and 02.1.

**Recommendations:**

The SAC should explicitly identify one or more objectives associated with Goal 4: Improve LTEMP sampling, indicators/summary measures, and analysis methods to increase efficiency and effectiveness at detecting release events and their sources, perhaps based on analyses of historic data such as provided by Payne et al. (2006).

## 3.0 LTEMP Survey Protocols

While numerous objectives have been defined, only objectives O1.1 (monitoring of recovery of EVOS-impacted sites), O2.1 (developing a baseline of background hydrocarbon levels), and O2.6 (monitor BWTF) establish distinct site selection requirements. O1.1 and O2.1 overlap to the degree that the control sites chosen for assessing recovery of EVOS-impacted sites were chosen to represent background conditions in a particular region and as such are de facto representative control sites for some portion of PWS/GOA. All other objectives define analysis goals to be based on the observations from these sites. Attention is thus primarily focused on LTEMP's sampling and monitoring design, field and lab protocols, and then, secondarily, on indicator or summary statistics, data analysis, and data management. Given the limited sediment sampling associated with the LTEMP project, attention is focused on the mussel sampling.

### 3.1 Sampling Design

A sampling design consists of an explicitly defined target frame, and an explicitly defined sample frame, a sample selection procedure, sampling protocols and effort, and for long-term monitoring, a re-visitation schedule (Cochran 1977, McDonald 2003) .

#### Target Frames

The target frame is the collection of entities you want to make inferences about.

- 1.1: All EVOS-impaired sites (and 'matched' controls of unimpaired sites).
- 2.1: Coastlines of PWS and GOA (within an implicit but unspecified range).

#### Sample Frames

The sample frame is the collection of entities you can actually sample from.

- 1.1: Safely accessible known EVOS-impaired sites with adequate mussel beds to support regular destructive sampling[1] (and similarly 'matched' assumed control sites).
- 2.1: Safely accessible coastlines of PWS and GOA (within an implicit, but unspecified range) with adequate mussel beds to support regular destructive sampling.

It should be noted that the requirement for mussel beds that will support regular destructive sampling could potentially limit the selection of impaired sites; any sites with formerly abundant mussels that had all died as a result of EVOS would no longer be eligible for selection as monitoring sites. The degree to which this bias is plausible should be explicitly discussed and recorded.

The use of destructive sampling in any long-term monitoring program raises the issue of whether or not the sample sites will continue to meet the standards of adequacy through time. The most recent report (Payne 2006) notes issues of missing, declining, or totally absent mussel beds for five of the ten sites (Table 2); four of them are 'EVOS-heavily impaired' sites. At three of the sites, the presence of predators was noted, suggesting that the declining beds are not necessarily directly attributable to the long-term impacts of EVOS.

#### Recommendations:

RCAC should:

- Identify viable mussel beds near the currently declining sites that meet the original selection requirements and that can be considered candidates for future sampling;
- For each declining site, consider sampling both the declining site and at least one of the nearby viable sites so as to assess the comparability of their hydrocarbon signals and build up a calibration period before the declining site disappears;

[1] Originally, sites had to meet the additional condition of appropriate nearby inter-tidal sediments to support sediment sampling requirements.

- Explore the ramifications of adopting some form of cross-sectional, rotating panel, or split-panel monitoring design to lessen the response burden on the existing mussel beds and potentially extend their duration as viable monitoring sites (see below).

**Site Selection Process**

Possible site selection processes range from subjective 'expert judgment', where sites for sampling are selected without any assurance against unrecognized biases or confounding factors, to probabilistic selection methods ('statistical sampling') that are specifically developed to control against unsuspected bias and confounding.

The National Oceanic and Atmospheric Administration (NOAA) National Status and Trends (NS&T) Mussel Watch Program lists several specific considerations in regard to the selection of sites from which to collect mussel tissue samples (NOAA 1993):

- Indigenous populations of mollusks must exist at a potential sampling site since the monitoring effort and NOAA methodologies do not use caged mussels.
- The site should contain indigenous bivalves of a suitable size (5 - 8 cm for mussels, 7 - 10 cm for oysters) available for collection.
- The NS&T Program is not intended to quantify contaminants in hot spots. Rather, mollusk collection sites are selected to be representative of the body of water sampled. Therefore, Mussel Watch sites should not be knowingly located near waste discharge points or in poorly flushed industrial waterways.
- Sample substrates should be limited to natural substrates or structures made of natural materials such as rock (including rip-rap and jetties), sand, or mud. Collection of samples on buoys and preserved wooden structures can yield artificially high results for some contaminants being quantified by the NS&T Program.
- The site should be suitable for follow-up sampling (i.e., it is not anticipated that the site will be physically disrupted by development or that the mollusk population will be depleted by sampling).

In the context of O1.1 (impaired sites), the goal is actually a comparison of two scenarios, EVOS impaired and not, making it an experimental design situation. The analogue of randomized assignment of treatments in the case of observational studies such as the LTEMP program is random selection of impaired sites and random selection of control sites, after defining a suitably matched subset of available control sites.

The LTEMP sites appear to have been selected more for logistical convenience than with regard to any probabilistic scheme to avoid bias. It should be noted that the LTEMP program diverges from the mussel watch program especially with regard to monitoring of known hot spots and discharge points. This does not impair the ability to draw valid *site-specific* inferences, but any extrapolation to the larger

PWS or GOA must be based on logic, i.e., what is known about the underlying processes and the patterns of spatial/temporal variability of these processes in PWS/GOA, and rhetoric, not statistical sampling ('design-based inference').

For O1.1, the pairing of sites ameliorates, to some extent, confounding in the assessment of site-specific recovery, but does not provide any safeguards for wider regional inferences. Further, the lack of complete pairing in Western PWS or Western GOA (Kodiak) (Table 2) eliminates any *direct* assessment of recovery for impaired sites in these regions via comparison to a nearby, relatively matched, control site. It should be noted that the relatively tight tracking of control and impaired sites within each region, and especially their fairly simultaneous declines in total polynuclear aromatic hydrocarbons (TPAH) magnitudes since 1998 (Figure 19, Payne et al. 2006), suggests either (i) that the impaired sites have relatively equilibrated with the control sites and hence reached some definition of 'background', or (ii) the control sites were more impaired than originally expected and both sites are 'recovering' at similar rates within each region.

**Recommendations:**

RCAC should consider

- Explicitly defining the region of the GOA that they are interested in making inferences regarding (the regional target frame);
- Review the findings of Mudge (2002), the recent LTEMP analyses (Payne et al. 2006) and any other work on regional background signals to try and refine the regional partitioning in terms of common background influences. This may provide justification for simplifying the current partition into Port of Valdez, East PWS, West PWS, and the GOA, lending support to logical application of the limited control sites to other nearby regions. Alternatively, these results may caution against such comparisons and reinforce the need to expand the set of control sites to regions such as West PWS that currently don't have any control sites.
- Obtaining the PWS/GOA coastline segment classifications, assuming NOAA/National Marine Fisheries Service (NMFS) or some other party has developed them (likely as a result of EVOS), and generate the summary distribution of coastline segments by type. This should then be compared to the coastline segment types currently being sampled (Table 2). This information could be used to guide any future selection of new sample sites (expand representation by coastline types), though this factor is likely secondary to the regional influence discussed above.

**Sample Size**

In terms of mussel tissue hydrocarbon burdens, LTEMP gathers information on two spatial scales: the site is the sample unit, the replicate mussel sample (25-30 mussels) from along the transect is the observational unit, and the 10 gm aliquot tissue sample is the measurement unit. There are three replicate observations per sample unit (site). Variation among the replicate observations tells one something about within-the-mussel-bed variation in hydrocarbon signal, variation among sites tells one

about spatial variation among mussel beds, though this information is of limited value since in the absence of probabilistic site selection it is impossible to properly estimate site-to-site variation (Thompson 1992).

As with most long-term monitoring programs, the number of sample sites is relatively small, especially for the 'control' sites (Table 2) – of which there are only three. However, as sites converge to regional 'background' levels for the summary statistics LTEMP has chosen to focus on, e.g. TPAH, the distinction between impaired and control sites dissipates with respect to the selected hydrocarbons and substrates (mussel tissues). This convergence appears to be occurring, at least in terms of TPAH and within the three broad regions of Port of Valdez, PWS, and GOA (Figure 19, Payne et al. 2006), though this convergence is difficult to assess given the limited control information.

**Table 2**: Current study sites selected for each objective. Mussels are sampled at each site; sediments are only sampled at Alyeska Marine Terminal and Gold Creek.

| **Study Area Region** | **Impacted (O1.1, O2.6)** | **Control (O1.1, O2.1)** |
|---|---|---|
| Port of Valdez | Alyeska Marine Terminal near Ballast Water Treatment Facility (BWTF) within mixing zone | Gold Creek – judged suitable far from outfall and terminal berths to be reference for this part of study area |
| Eastern PWS | Knowles Head – tanker anchorage | Sheep Bay (E. PWS); considered Olsen Bay as replacement, stuck w/ Sheep Bay (sediment composition) |
| Undefined | Zaikof Bay (Hinchinbrook Entrance) (July 1999) | None |
| Western PWS | Northwest Bay – unsuitable mussel bed; Disk Island (heavily impacted by EVOS) | None |
| Undefined | Sleepy Bay (N tip Latouche Island), heavily oiled by EVOS | None |
| Gulf of Alaska / Kenai Peninsula | Windy Bay – heavily oiled by EVOS | Harris Bay – unsuitable mussel bed: Aialik Bay |
| Gulf of Alaska / Kodiak | Pervalnie Passage – Shuyak Island EVOS oiled site; replaced by Shuyak Harbor (better substrate, safer sampling) | None |

**Revisitation Design**

Every monitoring program has a revisitation design (McDonald 2003). The most common revisitation designs are (i) a *panel* design where sites are selected once and then the same sites are revisited every time (also called *repeated measures* or *longitudinal data*), and (ii) *cross-sectional* designs where a new selection of sites is chosen for sampling every time (Table 3). There are hybrid designs such as *split-*

*panel* designs where some sites are visited every time and some are selected anew each time or *rotating panel* designs where a subset of sites are visited, say, every odd year and another subset visited every even year, etc.

LTEMP uses a panel design. This provides the most precise, hence efficient, estimate of site-specific trends but at the cost of (i) a high response burden on each site due to the destructive sampling and (ii) limited insight into what is going on at any other sites in the target frame. Estimating regional trends is somewhat more involved as it requires longitudinal data analysis methods to account for the within-site dependence in observations.

**Table** 3. Revisit designs: (a) classic panel or 'paired' design, (b) cross-sectional design (new selection of sites each event), (c) rotating panel design, (d) split-panel design.

(a)

| Sampling Event | 1 | 2 | 3 | 4 | … |
|---|---|---|---|---|---|
| Panel A | X | X | X | X | X |

(b)

| Sampling Event | 1 | 2 | 3 | 4 | … |
|---|---|---|---|---|---|
| Panel A | X | | | | |
| Panel B | | X | | | |
| Panel C | | | X | | |
| Panel D | | | | X | |
| … | | | | | … |

(c)

| Sampling Event | 1 | 2 | 3 | 4 | … |
|---|---|---|---|---|---|
| Panel A | X | X | X | | |
| Panel B | | X | X | X | |
| Panel C | | | X | X | X |
| Panel D | | | | X | X |
| … | | | | | … |

(d)

| Sampling Event | 1 | 2 | 3 | 4 | … |
|---|---|---|---|---|---|
| Panel A | X | X | X | X | X |
| Panel B | X | X | | | |
| Panel C | | X | X | | |
| Panel D | | | X | X | |
| … | | | | | … |

**Recommendations:**

Given the apparent convergence among sites to 'background' levels, RCAC may want to consider shifting from a strict panel design to

- A split-panel design (Table 3 d), supplementing the long-term sites with randomly chosen sites each year, as suggested in Payne et al. (1998, pg 88). This would provide greater power to detect events since there would be broader regional coverage each year. It would also allow LTEMP to slowly develop a body of (temporally limited) baseline data throughout the PWS/GOA region, providing insight into patterns and boundaries within these large regions. However, this monitoring design would not support site-specific estimation of trends at the sites selected anew each year, though one could still estimate site-specific trends from the panel(s) that were continually revisited. Efficient analysis also becomes much more complex.
- Alternatively, RCAC could consider shifting to a rotating panel design (Table 3 c) where the current panel of sites is split into two panels (subsets) and each panel is sampled on a rotating basis, e.g. odd years/even years. The liberated funds could be used to supplement the panel sites with a one-time random sample of sites, somewhat increasing the size of each panel but but not exceeding the current number of sites annually visited. This would lessen the response burden and overall cost, but may not expand coverage as much as a split panel design. Proper analysis

also becomes much more complex if there is interest in estimating any regional trends.

Either approach would, through time, maintain the consistency of the long-term sites *and* provide greater insight into regional patterns. The cost would be a much more complicated analysis for regional trends that would require a statistical consultant. However, given that LTEMP has operated for 14+ years without apparently directly estimating any regional trends, and there appears to be within-region variation that obviates the need for any regional trend estimates, this may not be such a price to pay.

### 3.2 Response Indicator/Measurement Selection

The adequacy of the implicit assumption that measurement of hydrocarbon levels in mussel tissue and sub-tidal sediment are sufficient representations of the hydrocarbon levels in the larger ecosystem should be assessed in light of the findings and discussion of Payne et al. (2001). While variations in the concentration of PAHs in mussel tissues may provide an important, and statistically significant, indication of a change in the hydrocarbon load in a specific location, LTEMP implicitly assumes that the concentration of PAHs in mussel tissue yield information about the health of the ecosystem as a whole. It is thus important to correlate the concentration of analytes in mussel tissues with the concentration of analytes found in the water column to demonstrate how PAH concentration in mussel tissue responds to different hydrocarbon releases and sources, as well as the relative degree of persistency of these constituents in mussel tissue versus that in the general environment, and the relative toxicity of the environment. Indeed, conclusions about the health of PWS ecosystems at large are drawn from the concentration of various or total PAHs found in mussel tissues, even though the relationship between the concentrations of PAH observed in mussel tissues and the concentration of PAH present in the water column in which those mussels are living is poorly established (Payne 1998, Axleman et al. 1999). Ideally, concentrations of PAHs observed in mussel tissue could be correlated to concentrations of PAHs in the water column.

Unfortunately not only are biological concentration factors (BCF's) poorly established from mussels, they are known to be highly variable even amongst populations of mussels exposed to the same environment, and are known to vary significantly depending on the concentration and type of PAH present in the water column, and the local environmental characteristics. To complicate things further, even if the BCF remains constant, uptake and depuration rates can change throughout the season, or depending on mussel health, growth, or reproductive condition. In addition, uptake and depuration rates are influenced by the partitioning coefficient of a particular hydrocarbon source, and whether or not it is in the dissolved, particulate, or colloidal phase, and whether or not this phase has reached equilibrium

with the water column[2] (Axelman et al. 1999).

This inability to accurately assess the concentration of analytes in the environment based on organism response confounds some of the LTEMP objectives, and raises concern about others. Objectives O1.2, O2.2, O2.4, and O2.6 can be evaluated via the mussel's observed site-specific responses, though it should be explicitly stated that there is an underlying assumption that the BCF, uptake, and depuration rates are relatively consistent between field events. It should be noted that uptake and depuration rates become particularly important after release events and near point sources; situations in which equilibrium partitioning is an issue. Objectives O2.1 and O2.3 are confounded by some of the fundamental limitations of mussels as an indicator species. Objective 2.1 seeks to develop a baseline of background hydrocarbon levels. To accomplish this objective would likely require an intense investigation into the variables at play at a given site. BCF would need to be calculated and correlated with concentrations of PAH in the water column.

In reality, LTEMP is measuring the response of a population of mussels to a particular, relatively unknown, concentration of PAH present in multiple phases. While it is relatively easy to describe a typical hydrocarbon load in mussel tissue that represents a *response* to background concentrations, it is significantly more difficult to relate this response to the actual concentration of PAH present in the water column, and this response will likely vary from site to site.

This site-specific variation in response confounds O2.3, which seeks to develop a model of spatial variation in background levels. O2.5 is partially confounded by these concerns as well. A release could be detectable based on the mussel response to hydrocarbons present in the water column above the background levels. Source identification should be undertaken with the awareness that uptake and depuration rates become particularly important after release events and near point sources; situations in which equilibrium partitioning is an issue because PAH is likely to be present in multiple phases. Once again, mussel response in terms of the accumulation of certain hydrocarbons will likely vary depending on the phases in which hydrocarbons are present, and the local environment.

Measurements are currently taken of two classes of hydrocarbons in mussel tissues and sediments:

[2] While this is not a concern with regard to the mussels, which are present in the environment for a sufficient period of time that all phases are likely to reach equilibrium with the water column, it does affect the ability to calibrate mussel BCF to water column concentrations, because water column sampling techniques can eliminate hydrocarbons and phases that are slow to reach equilibrium, especially near point sources.

saturated hydrocarbons, hydrocarbon chains from 9 to 40 carbons in length, including pristine and phytane; and a suite of 43 PAHs (Payne 2006). While these measurements provide a basis for fingerprinting and tracking hydrocarbon sources, other analytes could be targeted to provide even greater specificity. An example would be biomarkers, such as steranes and terpanes in sediment samples. Various sterane and terpane biomarkers have a long history of established use as indicators of particular hydrocarbon sources (Wang 2006), and are used as part of Alyeska Environmental Monitoring Program (EMP) for fingerprinting Alaskan North Slope (ANS) crude (Payne 2006), though only in the immediate vicinity of Port Valdez.

The usefulness of analyzing mussel tissue samples for aliphatic hydrocarbons (AHC) has been called to question in past reports (Payne 1998). The current diagnostic indices make little use of AHC, with the exception of the Marine Biogenic Index, which is a partial measure of dietary uptake of hydrocarbons (Payne 2004). None of the other indices that apply to mussel tissue make use of aliphatic hydrocarbons. In contrast, many of the sediment indices use aliphatic hydrocarbons, but sediments are currently only sampled in the immediate vicinity of port Valdez.

**Recommendations:**

Given the difficulty of establishing the relationship between mussel tissue concentrations and water column concentrations of PAHs, and the difficulties of fingerprinting hydrocarbon sources when (i) multiple sources are present, including natural hydrocarbon sources such as biological sources, coal, and oil seeps, and (ii) hydrocarbons can be highly weathered by the time of monitoring, the RCAC may want to consider the following approaches to better achieve project goals:

- Evaluate the correlation between hydrocarbons observed in mussel tissues and hydrocarbons actually present in the water column and sediments on a site-specific basis. While there are several factors that complicate this process, the use of direct water sampling, sediment sampling, and semi-permeable membrane devices (SPMDs) could be used to estimate background hydrocarbon concentrations at a particular site, and generate data regarding relative mussel response to hydrocarbons that could be accumulated over time.
- Set explicit, site-specific (and perhaps season-specific) thresholds for release event detection; considerations could include known hydrocarbon sources in specific areas, water column PAH concentrations, release events, and known toxic effect levels.
- Refine source identification methods, and re-assess the suite of chemical analytes chosen for measurement. Specifically: occasional sediment sampling could be conducted to test for sterane and terpane biomarkers. This could be accomplished with a rotating panel, or other type of sampling design, or could be conducted in conjunction with site-specific thresholds relating to release events.
- Review the usefulness of analyzing tissues samples for aliphatic hydrocarbon constituents. Large variations have been observed in mussel lipid content and biological aliphatics are

difficult to separate from petrogenic aliphatic, especially at background concentrations. Aliphatic analysis would likely remain useful in sediment sampling.

### 3.3 Field Collections

Field collection was carried out in general accordance with established mussel watch methods put forth in NOAA Technical Memorandum 71 (NOAA 1993) and as explicitly stated in LTEMP reports. Current sample collection methods appear to be adequate to achieve project goals. Field methods are not described in the 1993-1996 summary report, though several suggestions for improvement of field methods were given (Payne 1998). The 1997-1998 LTEMP report simply says that field sampling procedures were carried out in accordance with methods stated in earlier reports, but these reports were not available for review; however it is assumed that these studies followed the procedures outlined in the original study plan (KLI 1993). The original study plan included the collection of gonadal tissue to evaluate the reproductive condition of sampled mussels. This can be an important consideration when collecting mussel samples because certain contaminants can accumulate in gonadal tissue, making spawning an important depuration route (NOAA 1993, 1998). The collection of gonadal tissue was discussed in the 1999-2000 report, but is not discussed at all in subsequent reports. It appears that sampling of gonadal tissue is no longer carried out as part of LTEMP.

Studies carried out with caged mussels (Payne 1997) indicate that the use of caged mussels to monitor a particular region or area of concern may be acceptable. This might be of particular use in regions where declining mussel populations limit the ability of field personnel to collect mussel samples from the population of native mussels, but the RCAC wishes to preserve the sampling location, or in locations were monitoring is desired, but no native mussel populations exist.

**Recommendations**

Field sampling procedures associated with the LTEMP project appear to be in general accordance with the standards set forth in the NOAA Mussel Watch program. The RCAC may want to consider

- Explicitly stating the reasons for dropping the collection of gonadal tissue from the LTEMP program, as this is a divergence from the NOAA protocol;
- Utilizing caged mussels, for observation of particular areas of concern that do not have native mussel populations, to provide a test group for comparisons of control sites to impacted sites, or to supplement declining mussel populations in current site locations. In this event, a blank(s) mussel sample (prior to deployment) should be analyzed as a quality control procedure to evaluate potential prior exposure to hydrocarbons; and
- Use SPMD's to supplement mussel collection as an additional source of information about hydrocarbon concentrations, or to monitor sites with declining mussel populations.

### 3.4 Laboratory Practices

In general Laboratory methods seem to agree with those set forth in the newest (NOAA 2007) guidelines for laboratory analysis standards with regard to the mussel watch program. Methods explicitly stated in the LTEMP annual reports don't include a detailed description of calibrations curves, continuing calibration, or quality controls samples, but it is assumed that these internal quality controls have been instituted in accordance with good laboratory practice and the most recent technical guidance.

It should be noted that current guidelines indicate that detections of analytes in method blanks are considered to be acceptable as long as two or more analytes are not detected at a concentration greater than three times the method detection limit (MDL) in the method blanks. Many of the samples in the LTEMP project contain concentrations of analytes that are also less than 3 times the MDL, and so the presence of low concentrations of analytes in method blanks is of particular concern, and special attention should be applied to method blanks if very small concentrations of analytes are to be reported. Indeed, the 1993-1996 summary report (Payne 1998b) identifies problems with Geochemical and Environmental Research Group (GERG) method blanks; blanks were not considered to be a problem unless they were greater than 3 times the MDL, but data were reported at or below this level. Contamination was often present in blanks at levels that approach concentrations reported in mussel tissues and sediments in early samples. Fortunately, concentrations of analytes at or below 3 times the MDL do not pose a major problem when present in samples that contain relatively high PAH concentrations, as the relative error diminishes as the absolute concentration of the analyte increases.

Methods could potentially be developed that would further reduce the MDL. One possible means would be by using methods of calculating the MDL other than those set forth in 40 Code of Federal Regulations (CFR) 136 (Geiss 2001), although this would differ from the method explicitly stated in NOAA Technical Memorandum 30. Other techniques might include increasing the quantity of tissue extracted, increasing the degree to which the sample is concentrated prior to injection, and increasing the injection volume; these techniques are specific to the extraction methods used, the laboratory equipment available, and the instruments the samples are analyzed on and will have to be evaluated by the laboratory team. In addition any change in methodology will require an MDL study.

### 3.5 Treatment of observations below detection limits

As hydrocarbon levels in the sampled substrates decline, more and more analyte concentrations are at or below detection limits, thus greater caution should be accorded to LTEMP's handling of these

observations.

The MDL is the minimum concentration of a substance that can be measured and reported with 99 percent confidence that the actual analyte concentration is greater than zero and the most standard method is set forth in 40 CFR 136, though there is flexibility even within this method. This procedure relies on the analysis of replicate samples using standard deviations and t-values to generate a 99 percent probability that the signal observed is distinguishable from background noise; the concentration of analytes that yields this probability is reported as the MDL. This means many detectable signals lie below the MDL value, but they are (statistically) difficult to separate from background noise. The procedure is meant to minimize type I error, that is to say, it is meant to minimize the possibility of a false positive to less than 1 percent. It is experimentally established for each analyte and should be re-established every time there is a significant change in standard operating procedures or a change in analytical instruments (Jones and Clark 2005). The MDL is not established using a field matrix, and thus does not guarantee exactly the same detection ability for the field samples.

The method quantitation limit (MQL) is a set multiple of the MDL, at least three, but usually five to ten times, the MDL. This is the limit above which the laboratory can reliably measure targets depending on the prescribed performance criteria (EPA 2004, Jones and Clark 2005). The MQL is the limit above which measured analytical concentrations can be reported without any qualification as to their accuracy (EPA 2004, Jones and Clark 2005). In the course of the LTEMP project data have been reported and analyzed not only at concentrations below the MQL, but also at concentrations below the MDL. While there may be adequate reasons to quantify and analyze data above the MDL, it is unreasonable to quantify data below the MDL. The possibility is still open that some of these data may still be analyzed using methods that do not require quantification. While it is possible to qualitatively identify analytes below the MDL based on ratios of ions and the relationship of these ions to the baseline noise, the MDL for a procedure provides a measure of the amount of variation that occurs when low concentrations of analytes are introduced onto the instrument. While it is possible to demonstrate that an analyte is qualitatively present, there is an unacceptable amount of variation below the MDL (and arguably below the MQL) to accurately or precisely quantify the analyte. This variation has been determined experimentally by the analyst using the method and instrument upon which the analysis is carried out.

While a variety of statistical methods exist for estimating an analyte's mean concentration when limited portions of the observations are below detection levels (e.g., EPA 2006), we've found no currently accepted methods for accommodating censored observations (= below detection limits) in estimating

relative ratios of concentrations for different analytes, the summary statistics commonly used in hydrocarbon fingerprinting and weathering estimation methods (Wang et al. 1999, Alimi et al. 2003, Wang et al. 2006). Analyzing data reported below the MQL, for instance summing individual PAHs into a TPAH value, is questionable because these values have a relatively large variation associated with them, and trying to trace source signals by comparing ratios is likewise questionable, because concentrations can randomly vary enough to severely distort these ratios (though the probability goes up in a complicated way if the signal has been observed before, involves multiple analytes, involves multiple ions in specific ratios, conforms to a known source, etc.).

While it is not necessarily appropriate to simply write-off concentrations of analytes measured below the MDL as "non-detectable", these data certainly cannot be dealt with in a quantitative way. While they may yield information about hydrocarbon sources, data present below the MDL are simply too variable to be useful for the purposes of resolving the four null hypotheses presented in the original LTEMP Study Plan when the sites that are being compared have concentrations of analytes below the MDL. For this reason it is important to clearly define threshold levels of analytes that would be considered deleterious or indicative of a release event, and natural variation that may occur between sites that are exhibiting background responses.

## 4.0 Summary Statistics, Data Analysis, Visualization

The two analysis objectives of *detecting* hydrocarbon releases and identifying their sources (*fingerprinting*) generally use different summary statistics formed from the reported analyte concentrations. Detection summaries generally focus on the hydrocarbon burden, often summarized by summing concentrations across groups of analytes as an estimate of total burden (Wang et al. 1999). Source identification or apportionment focuses on the distribution of relative concentration across analytes (e.g. Wang et al. 1999, 2003, 2006; Driskell et al. 2005, Payne et al. 2006). Relative proportion vectors, where the component values have been normalized by dividing each by the total sum of the original concentrations are a form of compositional data (Aitchison 2003) and generally require special treatment, though this is not specifically addressed further below.

### 4.1 Event Detection

The main summary statistics used to detect release events in LTEMP analyses (e.g., KLI 2003, Payne et al. 1998B, Payne et al. 2006) have been *total polycyclic aromatic hydrocarbons* and *total saturated (aliphatic) hydrocarbons,* or TPAH and TSHC (TAHC), respectively. These are naturally interpretable summary statistics that clearly express the total burden, hence they are widely used in the literature (e.g.,

Short et al. 1999, Wang et al. 1999).

The analyses for LTEMP (e.g., KLI 2003, Payne et al. 2006) have regularly included estimated concentrations below the MDL in these summaries without any special treatment such as replacing the censored values with zero or half of the MDL (EPA 2006 pg 130). For logical consistency in the statistical analysis, only analytes whose concentrations exceed detection levels should be included in these calculations. However, their inclusion or not is more a matter of consistency than pragmatic impact – if the evidence for concluding that a release has occurred is so equivocal as to be influenced by the decision to include concentrations below the detection limits, then the evidence is of the same magnitude as the background laboratory errors – very weak indeed. If there is interest to pursue such a situation, one should employ a more refined statistical analysis to estimate the probability of exceeding the threshold background level of TPAH or TSHC (e.g., Barnett and O'Hagan 1997), which in turn requires defining the threshold level a priori.

The *unresolved complex mixture* (UCM) has also been used as an indicator of spill events as the underlying compounds creating the UCM are generally not found in organisms (see Payne et al. 2006 for definition). However, it is most useful in detecting heavily-weathered oils. Similarly, certain petroleums have distinct biomarker compounds that can act as 'silver bullets' for their detection (Wang et al. 2006).

An alternative approach for detecting spill events is to assess changes in the source phase composition of the PAH profile (Driskell et al. 2005, Payne et al. 2006): the percent of TPAH attributable to *dissolved-phase* sources vs. the percent attributable to *particulate-phase* sources. The chemistry and rates of the transfer process from particulate-phase to dissolved-phase varies in *broadly* predictable ways both across PAH families, with the lighter naphthalenes dissolving more easily than the other families, and across components within each PAH family, as the parent component and the C1- and C2-alkyl homologues are more water soluble than the other components (Driskell et al. 2005, Payne et al. 2006).

The broadly distinct profile patterns of dissolved-phase vs. particulate-phase signals have the potential to provide a much more powerful summary statistic for detecting spill events via a shift in source-phase composition to a large(r) particulate-phase portion, relatively independently of the TPAH concentration. Thus, it may allow detection of much smaller releases, given sufficient data for a reliable estimate of

'background' source-phase composition.

The current method of estimating the source proportions uses a predefined set of classification tree models (Breiman et al. 1984), a form of pattern recognition or machine learning, to assign whole components to either a dissolved or particulate sources (Payne et al. 2006). This could be made more objective by using the dissolved-phase and particulate-phase profiles available from the BWTF study (Payne et al. 2005a, b), and their overall non-partitioned PAH profile, to directly fit a supervised partitioning model using whichever of the recent machine learning methods seem most appropriate (Hastie et al. 2001). This algorithmic approach would eliminate the subjectivity of the predefined rules and potentially reveal a simpler set of rules. While acknowledging the heuristic goal of the predefined rules to 'capture the experience and skill of a trained chemistry expert', independently using an objective algorithmic approach would provide a useful assessment, potential insight, and possible cautions. Recommending a specific approach would require further research that was tangential to the overall goal of this review. The key is that the BWTF study provides known dissolved-phase and particulate-phase profiles. If other such known-phase profiles were available from other sites, the chosen modeling approach should be repeated on each available data set to (i) check for common model structures or rules, and (ii) assess variation in the rules across sites, weathering, etc. If the rules were found to vary widely from site to site, or could not be broadly generalized for use at new sites, then the approach's assumptions would need to be revisited.

While a machine learning approach would be more objective than the current method, and be based on the rather deeply developed computer science literature, it would not provide statistical estimates of uncertainty, though error rates could be estimated via cross-validation methods. Uncertainty estimates would only be provided by developing a probabilistic mixture estimation method, which would overcome the other limitation of the current estimation of source phase composition: that it completely allocates each component, or in some cases a whole family of components, to one or the other source when they actually result from a mixture of concentrations from both sources. The current estimates are thus a 'first order' approximation to the composition. The true profile is a mixture of contributing concentrations from both the dissolved and the particulate phases. Estimating the mixture proportions directly requires solving a multinomial mixture problem where there are two (broad) potential sources (dissolved, particulate) and the multinomial distribution of each are broadly defined in terms of constraints between PAH families and constraints among components within PAH families (these constraints underlay the logic 'rules' of the current estimation method). On the face of it, this is likely a

solvable statistical mixture estimation problem using Bayesian multinomial mixture analysis (e.g. Gilks et al. 1996, Pella and Masuda 2001) and appropriately constrained hierarchical priors on the parameters of each source phase's prior distribution. If this proves solvable, it would eliminate the need for the pattern recognition allocation process, be a more efficient use of information in the PAH profile, and provide estimates of uncertainty for the composition proportions, enabling better assessment of change detection.

Both the machine learning and mixture estimation approaches should treat those components with observed values less than the MDL with care. If the signal source is known (either in terms of phase or actual source materials), there is clear information on the expected concentration of a component in the concentrations of the other components – the component concentrations are not independent (unrelated) since a given source has an expected profile or relative distribution of concentration across all the components. Thus knowing this relative distribution (knowing something about the source), only having knowledge of the concentrations of some components may be sufficient to infer the concentrations of the other components. The ability to make such inference will depend, loosely phrased, on the consistency of the relative source profile, more specifically in the multiple correlations among the components to be inferred and those acting as the information basis. The fact that the profiles tend to be fairly consistent is what allows one to do fingerprinting (Wang et al. 1999, 2003, 2006).

At low concentrations, this means that knowing something about the concentrations of some components, and their likely source, may allow one to predict whether another component is expected to occur at a concentration above or below its MDL. In this sense, observing a concentration pattern among a group of components, even with some of the components below the MDL, still provides for valid inference regarding the source. More specifically, it may provide information for *eliminating* possible sources. For example, if (i) all of the components observed above the MDL exhibit the relative distribution associated with source Z, and (ii) if, when the source truly is Z and the release magnitude matches this overall concentration level (TPAH), component X is *also* expected to be well above the MDL, but (iii) X was observed below the MDL, then (iv) this would suggest Z was not the actual source. Thus the information being used is whether or not X was above/below the MDL, not the actual estimated concentration of X. The caution is that while this level of information (above/below the MDL) is reliable given quality assurance/quality control (QA/QC) standards, the actual estimated concentration of X when it is below the MDL is **not** reliable and thus should not be used (except to

denote 'below the MDL'). Specifically, it should not be used in any quantitative calculation involving ratios of concentrations among two components, and especially not between two components, both of which are below the MDL.

The underlying statistical points are (i) component concentrations do not vary independently of one another given some knowledge of the source and weathering, etc., (ii) thus, there is information in the *marginal* distribution of the concentrations of the quantified analytes (greater than the MDL) regarding the expected *marginal* distribution of the concentrations of another subset of analytes (which may or may not be less than the MDL). Such information may be sufficient for the question at hand, assuming that the purported source identification is not contradicted by the qualitative (above or below the MDL) information in the second subset of analytes.

The machine learning approach employed by Payne et al. (2006) follows these guidelines to the extent that their allocation rests strictly on comparisons among analytes with concentrations greater than the MDL; when the allocation rests on decisions of patterns among analytes all of which are below the MDL, this guideline is abrogated. This could be assessed by rewriting the allocation rules to ignore any values less than the MDL, rerunning the analyses and seeing if an approximately similar composition is returned. In effect, this is redefining the rules to be more flexible and to do the best they can with only the analytes currently above MDLs.

Thus RCAC appears to have three choices regarding MDLs and source-phase composition: (i) take whatever steps are available to lower the MDLs, (ii) only pursue source-phase composition when TPAH is sufficient to warrant it (i.e., define a TPAH threshold below which the more intensive analyses will not be conducted), (iii) develop an objective algorithm or mixture estimation method for source-phase composition that is flexible enough to operate on only those subsets of analytes that are above MDLs, which may work for only particular subsets of analytes and be impossible in other situations. While a strong argument can be made for the validity of source-phase allocation patterns involving analytes below MDLs, coming to rely on such summaries could produce legal problems if the results lead to policy actions whose foundation would be difficult to justify in court.

### 4.2 Detecting Change

When historical measurements for any of these summary statistics are available, event detection can be based on either the exceedance of an established threshold, preferably using a statistical test to control for error rates (Barnett and O'Hagan 1997), or by testing for significant change from recent levels using,

for example, control chart methods from industrial process control (Montgomery 2000) or more refined statistical change detection methods.

LTEMP is in a unique position, given its historic data set, to estimate site-specific thresholds of background hydrocarbon burdens. However, doing so requires a priori deciding 'how low is small enough to be considered background'? One approach to this would be to define, a priori, acceptable error rates in estimating sites-specific trends, then when robust trend estimation methods have revealed a sufficient recent period of no significantly detectable trend, assert that that period represents background conditions and use it to estimate expected background levels, as well as temporal variation and sampling variation, for the summary statistic of choice. From this, one could develop an exceedance standard and statistical assessment (Barnett and O'Hagan 1997). An alternative method is to define that background has been reached when there is convergence among a known EVOS-impaired site and its associated control site, if such a control site has been defined and monitored. Doing so assumes that the control site is, actually, a good control and has not been impaired by the activities LTEMP is focusing on detecting.

Both of these approaches define background as the absence of a detectable difference (in time or between impaired/not impaired sites); there do not appear to be any other means of defining 'background'. Thus the declaration of 'background' level will vary with the detection summary statistic. For example, the concentrations of TPAH suggest background may have been reached for a number of sites while the current source-phase composition may still be detecting subtle changes. The choice of summary statistic should be driven by RCAC's overall goals for LTEMP; specifically, there needs to be a clearly defined lower threshold on the type, or at least size, of release events LTEMP should be detecting. The event detection summary statistics discussed above generally provide little or no information on potential sources.

**Table 4**. Summary statistics for detecting release events (after Table 3 in Payne et al. 2006).

| | |
|---|---|
| TPAH (mussel tissue and sediments) | Total PAH as determined by high resolution gas chromatography/mass spectroscopy (GC/MS) with quantification by selected ion monitoring; defined as the sum of 2- to 5-ring polycyclic aromatic hydrocarbons:<br>Naphthalene + fluorene + dibenzothiophene + phenanthrene/anthracene + chrysene + their alkyl homologues +other PAHs (excluding perylene). |
| TSHC (sediments) | Total Saturated Hydrocarbons as determined by high resolution gas chromatography with flameionization detection (GC/FID):<br>total n-alkanes (n-C10 to n-C34) + pristane + phytane |
| DSI (mussel tissue and sediments) | Dissolved Signal Index sums the soluble PAH fractions of the PAH profile; defined as the sum of dissolved(naphthalenes+fluorene+C1-fluorene)+dissolved(phenanthrenes) + dissolved(dibenzothiophenes)<br>where 'dissolved' is determined by characteristic ordering patterns among the relative concentrations of particular families of components and among parent and alkyl-homologue components with each family (see Payne et al. 2006). |
| PSI (mussel tissue and sediments) | Particulate Signal Index sums the less soluble PAH components and any water-washed groups; defined as the sum of (C2- + C3-fluorene) + particulate (anthracenes & phenanthrenes) + particulate (dibenzothiophenes) + particulate (fluoranthene/pyrenes)+particulate (chrysenes)<br>where 'particulate' is determined by characteristic ordering patterns among the relative concentrations of particular families of components and among parent and alkyl-homologue components with each family (see Payne et al. 2006). |

### 4.3 Visualization

The observations, as well as the source-phase composition results (Payne et al. 2006), suggest systematic seasonal differences in the analyte signals. Visualization of patterns and trends would be aided by displaying the winter sample results separately from the summer sample results, removing this source of systematic variation and allowing better focus on sampling and background variation. If source-phase composition focuses strictly on dissolved versus particulate contributions, displaying a time series of just one reveals all the information (rather than needing to plot both on the same graph).

If the triple composition (DSI percent, PSI percent, pyrogenic index percent) is the focus, a more effective graphing method for detecting change is to use ternary plots (Aitchison 2003). For example, Figure 20 of Payne et al. (2006) is likely more informative as a ternary plot. These could also be used to good avail to display the relative composition of the TPAH families, though it would require combining families to a collection of three or four groups. The latter could be plotted using tetrahedron plots (for example, see Reynolds and Templin 2004). Other minor adjustments to the current data displays that would improve their effectiveness include (Cleveland 1993, 1994):

- Time series plots using one type of symbol for the 'control' sites and another for the 'impaired' sites, such as open symbols for control and filled symbols for impaired;
- Consistency in these symbols types across figures;
- Adding error bars to the summary statistic values (from calculating each summary statistic on each laboratory replicate);
- Profile plots adding summary plots, such as tetrahedron plots or stacked bar plots or parallel plots of the relative concentration distribution across PAH families;
- Plotting the relative concentration distribution across PAH families (total naphthalene, etc.) then separately plot the parent and alkyl-homologues relative distribution for each family, as appropriate, to aid detection of subtle differences in time or space. In effect, partition the current 40-bin histograms into a cut across-family summary and then separate (i.e. visually digestible) family-specific plots; and
- If particular cross-component comparisons are of regular interest, create plots specifically focused on those comparisons.

### 4.4 Fingerprinting/Source Identification

There are two types of oil spill fingerprinting problems: the assignment problem, where the goal is to identify the (most likely) source of a spill from a set of potential sources (e.g. Short and Heintz 1997, Daling and Faksness 2002), and the mixture estimation problem, where the goal is to estimate the contribution proportions from a predefined set of potential sources (e.g. Page et al. 2002). These two classes of problems occur quite frequently in other situations, such as commercial fisheries management where there is a long history of methods development (Pella and Milner 1987). Assignment methods are sometimes used as an approximation to the generally more difficult mixture estimation problem (e.g. Mudge 2002), though doing so can introduce bias and is not as informationally efficient as mixture estimation.

Within each problem class there are an ever-expanding range of methods (Wang et al. 1999, 2003, 2006), from non-statistical expert 'pattern recognition' (Daling and Faksness 2002) using simple indices or the relative ratios of particularly informative analytes, termed *diagnostic ratios* and *double ratios* (Wang et al. 1999, 2003, 2006), to optimization-based methods built around minimizing a *distance* between the observed PAH relative concentration profile and those of the potential sources (e.g., Short

2002, Page et al. 2002, Mudge 2002). The majority of methods in the literature use very little of the information in the PAH profile, focusing on small sets of specific analytes (Wang et al. 1999, 2003, 2006) and generally ignoring the rest of the profile. While this makes for simpler plots and, in many cases, is sufficient to achieve the goals desired, more effective methods would utilize the full profile. This appears to be slowly being realized in the literature, e.g., the multiple criteria methods (Wang et al. 2002, Page et al. 2002). There is a large body of statistical methods for both assignment and mixture estimation that utilize the full profile (e.g., Gilks et al. 1996, McLachlin and Peel 2000, Pella and Masuda 2001), but these have not yet been applied in this context. This may in part be due to the added feature of variability in source profiles due to weathering, biogenic uptake, etc.

So far LTEMP has tended to rely on the expert pattern recognition methods based on indices calculated from various weighted sums of concentrations among different analyte groups, such as the Fossil Fuel Pollution Index or the CRUDE index (Payne et al. 1998b). This appears to be adequate for the program's needs in determining recent events. Pursuing more refined fingerprinting methods may be of limited value to LTEMP given the possibly large spatial variability in background mixtures, the moderately large number of potential sources influencing PWS and GOA (coal, seep oil, eroding shales, rivers, EVOS), and the apparent similarity of some of these source's fingerprints (Mudge 2002).

The development of the source-phase partitioning (Driskell et al. 2005, Payne et al. 2006) generated some cautions that the RCAC should consider in clarifying the fingerprinting goals of LTEMP. More statistically advanced source fingerprinting methods using mussel tissues should probably be restricted to using PAH profiles from the particulate-phase as the transfer into dissolved-phase can produce relatively similar dissolved-phase PAH profiles from very different particulate-phase profiles (Driskell et al. 2005, Payne et al. 2006). This places a premium on adequate estimation of the source-phase profiles if these more advanced methods are to be pursued from this source. It also may imply that more refined source fingerprinting may be most effective if focused on sediments, possibly utilizing specific biomarkers; sediments are no longer actively sampled across the region. These issues should be discussed further with subject field experts (i) to clarify this interpretation and (ii) to discern likely countervailing issues.

The recommendations and comments raised above regarding the current machine learning approach to estimating source-phase composition equally apply to the broader task of estimating dissolved, particulate, and pyrogenic components of the source signal (Payne et al. 2006).

**Recommendations:**

- RCAC should clarify LTEMP's event detection goals; specifically, there needs to be a clearly defined lower threshold on the type, or at least magnitude, of release events LTEMP wants to detect.
- This will allow RCAC to then determine the most appropriate summary statistics for event detection – the less powerful but simpler TPAH/TSHC or the more refined source-phase composition. If the latter is chosen, then effort should be expended on improving the estimation method, ideally by developing a mixture estimation approach, at a minimum by refining the current tree model by letting standard model fitting algorithms reformulate, and hence assess, the classification model.
- RCAC should require analysis contractors to display the winter observations separately from the summer observations as a simple step to improve visual recognition of trends in the summary statistics.
- RCAC should clarify their definition of *background* levels. If the resulting definition implies that some study sites have reached that level, effort should be given to developing more formal detection methods using the baseline data (e.g., summary statistic-specific thresholds). Similarly, if any impacted sites are thus deemed to have returned to background, their recovery rates should be estimated and published in the literature.
- RCAC should explicitly adopt a tiered detection approach, where fingerprinting is only undertaken when an event of sufficient magnitude, as determined by the clarified detection goals, is detected.
- RCAC should clarify LTEMP's source fingerprinting goals – recent events or remnant signals of past events (which introduces the added difficulty of weathering)? Then RCAC should decide whether there is a need to pursue development of more statistically refined fingerprinting methods. At the moment, the current rough indices that appear adequate for the implicit goals may not be suitable if the explicitly decided goals differ.

## 5.0 Data Management and Dissemination

A key feature of an effective long-term monitoring program is a well maintained, well documented data archiving system that is readily accessible to potential users. Such a system encourages broad use of the monitoring results, raises program visibility and generates broad support for program continuance.

While we have not specifically assessed the LTEMP database, the very fact that it exists is a positive sign. The RCAC may want to discuss making this available over the web using a web-interface to the database that would allow the public to download data sets. If this is a legally viable option it could greatly broaden the pool of users and supporters.

**Feedback into the larger Adaptive Management Framework**

If LTEMP *is* the monitoring component of a larger management framework, its role should be clarified

and the pathways for information flow distinguished. Where do the LTEMP-generated reports go? Who reads them? What influence do they have? Is there anything else LTEMP could be doing in terms of disseminating information that would increase its influence in the PWS/ Port of Valdez management community? We didn't address these questions, but raise them as an important set of points to be considered in making LTEMP an effective long-term management program.

# References


Alimi, H., T. Ertel, and B. Schug. (2003). Fingerprinting of hydrocarbon fuel contaminants: literature review. *Environmental Forensics* **4**: 25-38.

Aitchison, J. (2003). *The Statistical analysis of Compositional Data*. The Blackburn Press, Caldwell, New Jersey.

Axleman, J., K. Naess, C. Naf, D. Broman. (1999). Accumulation of polycyclic aromatic hydrocarbons in semi-permeable membrane devices and caged mussels in relation to water column phase distribution. *Environmental Technonology* **18**, 2454-2461.

Barnett, V., and A. O'Hagan. (1997). *Setting environmental standards: the statistical approach to handling uncertainty and variation.* Chapman & Hall, London.

Breiman, L., J. H. Freidman, R. A. Olshen, and C. J. Stone. (1984). *Classification and Regression Trees*. Wadsworth.

Carls, M. G. (2006). Nonparametric identification of petrogenic and pyrogenic hydrocarbons in aquatic ecosystems. *Environmental Science and Technology* **40**, 4233-4239.

Cleveland, W. S. (1994). *The elements of graphing data.* Chapman & Hall, NY.

Cleveland, W. S. (1993). *Visualizing Data.* Hobart Press, Summit, New Jersey.

Cochran, W. G. (1977). *Sampling Techniques (2nd ed.).* Wiley & Sons, NY.

Daling, P. S. and L.-G. Faksness. (2002). Improved and standardized methodology for oil spill fingerprinting. *Environmental Forensics* **3**, 263-278.

Driskell, W. B., J. R. Payne, and G. Shigenaka. (2005). Revisiting source identification weathering models, and phase discrimination for *Exxon Valdez* oil. Proceedings of Arctic Marine Oil Spill Conference 2005, Calgary, Alberta, Canada. Pp 33-58.

EPA. (2006). Data Quality Assessment: Statistical methods for Practitioners. U.S. EPA. EPA QA/G-9S, EPA/240/B-06/003, February 2006. http://ww.epa.gov//quality1/qs-docs/g9s-final.pdf (June 2007).

EPA. (2004). Revised Assessment of Detection and Quantitation Approaches. Engineering and analysis division, Office of Science and Technology, Office of Water (4303T), U.S. EPA. EPA-821-B-04-005. http://www.epa.gov/waterscience/methods/det/red.pdf (July 2007).

Geiss, E. (2001). Comparison of detection limits in environmental analysis. *Fresenius' Journal of Analytical Chemistry* **370**, 673-678.

Gilks, W. R., S. Richardson, D. Spiegelhalter. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, NY.

Hastie, T., R. Tibshirani, and J. Friedman. (2001). *Elements of Statistical Learning: data mining, inference, and prediction.* Springer, NY, NY.

Jones, R. P. and J. U. Clarke. (2005). Analytical chemistry detection limits and the evaluation of dredged sediment. ERDC/TN EEOP-04-36, U. S. Army Engineer Research and Development Center, Vicksburg, MS. http://el.erdc.usace.army.mil/dots/pdfs/eedp04-36.pdf (July 2007).

Kinetic Laboratories, Inc. (KLI). (1993). Final project plan Prince William Sound Regional Citizens' Advisory Council Environmental Monitoring Program. Prince William Sound Regional Citizens' Advisory Council Scientific Advisory Committee. Anchorage, AK. 38 pp.

Kinetic Laboratories, Inc. (KLI). (2003). 2000-2002 LTEMP monitoring report. Prince William Sound Regional Citizens' Advisory Council Long-Term Environmental Monitoring program. 94 pp. Plus.

Lovett, G. M., D. A. Burns, C. T. Dirscoll, J. C. Jenkins, M. J. Mitchell, L. Rustad, J. B. Shanley, G. E. Likens, and R. Haeuber. (2007). Who needs environmental monitoring? *Frontiers in Ecology and the Environment* **5** (5): 253-260.

MacDonald, L. H., and A. Smart. (1992). Beyond the *Guidelines*: practical lessons for monitoring. *Environmental Monitoring and Assessment* **26**: 203-218.

McDonald, T. L. (2003). Review of environmental monitoring methods: survey designs. *Environmental Monitoring and Assessment,* **3**: 277 - 292.

McLachlan, G. J., and D. Peel. 2000. Finite Mixture Models. Wiley & Sons, NY, NY.

Montgomery, D. C. (2000). *Introduction to Statistical Quality Control (4th ed).* Wiley & Sons, NY, NY.

Mudge, S. M. (2002). Reassessment of the hydrocarbons in Prince William Sound and the Gulf of Alaska: identifying the source using partial least squares. *Environmental Science and Technology* **36**, 2354-2360.

Nichols, J. D., and B. K. Williams. 2006. Monitoring for conservation. *Trends in Ecology & Evolution* **21**(12): 688-674.

NOAA (1993) Technical Memorandum 71: Sampling and analytical methods of the national status and trends program benthic surveillance and mussel watch projects 1984-1992, NOAA NOS ORCA 71, (July 1993)

NOAA (1998) Technical Memorandum 130: Sampling and analytical methods of the national status and trends program mussel watch projects 1993-1996 update, NOAA NOS ORCA 130, (March 1998)

NOAA (2007) Technical Memorandum 30: Organic contaminant analytical methods of the national status and trends program 2000-2006, NOAA NOS NCCOS 30, (June 2007)

Noon, B. R. (2003). Conceptual issues in monitoring ecological resources. In *Monitoring Ecosystems: interdisciplinary approaches for evaluating ecoregional initiatives*, D. E. Bush and J. C. Trexler, eds. Island Press, Washington D. C. pp.27-72.

Page, D. S., A. E. Bence, W. A. Burns, P. D. Boehm, J. S. Brown, and G. S. Douglas. (2002). A holistic approach to hydrocarbon source allocation in the subtidal sediments of Prince William Sound, Alaska, embayments. *Environmental Forensics* **3**, 331-340.

Payne, J. R., W. B. Driskell, and D. C. Lees. (1998a). Long term environmental monitoring program data analysis of hydrocarbons in intertidal mussels and marine sediments, 1993-1996. Non-technical summary report for the Prince William Sound Regional Citizen's Advisory Council, Anchorage, Alaska 99501 (PWS RCAC Contract No. 611.98.1).

Payne, J. R., W. B. Driskell, and D. C. Lees. (1998b). Long term environmental monitoring program data analysis of hydrocarbons in intertidal mussels and marine sediments, 1993-1996. Final report for the Prince William Sound Regional Citizens' Advisory Council, Anchorage, Alaska 99501 (PWS RCAC Contract No. 611.98.1). 97 pp plus appendices.

Payne, J. R., W. B. Driskell, M. G. Barron, and D. C. Lees. (2001). Assessing transport and exposure pathways and potential petroleum toxicity to marine resources in Port Valdez, Alaska. Prince William Sound RCAC (http://www.pwsrcac.org/projects/EnvMonitor/ltemp.html).

Payne, J. R., W. B. Driskell, J. W. Short, and M. L. Larsen. (2006). 2004-2005 LTEMP oil monitoring report, Exxon Valdez oil spill restoration project final report (Restoration Project 040724). Prince William Sound Regional Citizens' Advisory Council, Anchorage, AK. 149 pp.

Pella, J. J., and M. Masuda. (2001). Bayesian methods for analysis of stock mixtures using genetic characters. *Fishery Bulletin* **99** (1): 151-167.

Pella, J. J., and G. B. Milner. (1987). Use of genetic markers in stock composition analysis. In *Population Genetics in Fisheries Management*, N. Ryman and F. W. Utter, eds. University of Washington Press, Seattle, Washington. Pp

247-276.

Reynolds, J. H. and W. D. Templin. (2004). Comparing Mixture estimates by parametric bootstrapping likelihood ratios. *Journal of Agricultural, Biological, and Environmental Statistics*, **9** (1): 57-74.

Salazar, M., J. R. Payne, and J. W. Short. (2001). Draft 2001 Port Valdez integrated monitoring report. Submitted to John S. Devens, Prince William Sound RCAC, Valdez, Alaska. 129 pp plus appendices.

Short, J. W. and R. A. Heintz. (1997). Identification of *Exxon Valdez* oil in sediments and tissues from Prince William Sound and the Northwestern Gulf of Alaska based on a PAH weathering model. *Environmental Science and Technology* **31**: 2375-2384.

Short, J. W., K. A. Kvenvolden, P. R. Carlson, F. D. Hostettler, R. J. Rosenbauer, and B. A. Wright. 1999. Natural hydrocarbon background in benthic sediments of Prince William Sound, Alaska: Oil vs Coal. *Environmental Science and Technology* **33**: 34-42.

Short, J. W. (2002). Oil identification based on a goodness-of-fit metric applied to hydrocarbon analysis results. *Environmental Forensics* **3**, 349-355.

Thompson, S. K. (1992). *Sampling.* Wiley & Sons, NY.

Wang, Z. D., M. Fingas, and D. S. Page. (1999). Oil spill identification. *Journal of Chromatography A,* **843**: 369 – 411.

Wang, Z. D. and M. Fingas. (2003). Development of oil hydrocarbon fingerprinting and identification techniques. *Marine Pollution Bulletin,* **47**: 423 – 452.

Wang, Z. D., S. A. Stout, and M. Fingas. (2006). Forensic fingerprinting of biomarkers for oil spill characterization and source identification. *Environmental forensics*. **7**: 105-146.

Williams, B. K., R. C. Szaro, and C. D. Shapiro. (2007). Adaptive Management: The U. S. Department of the Interior Technical Guide. Adaptive Management Working Group, U. S. Department of the Interior, Washington, DC. ISBN 1-411-31760-2. http://www.doi.gov/

# APPENDIX A

# STRUCTURED SUMMARY OF MAJOR RECOMMENDATIONS

## STRUCTURED SUMMARY OF MAJOR RECOMMENDATIONS

Clarifying Goals & Objectives (SAC Tasks presumably; some requiring technical input)

1. Explicitly identify the goals of the LTEMP program and clarify that 'taking measurements' is neither a goal nor an objective, but rather is a means to achieving various objectives.
2. Explicitly identify a fourth LTEMP goal focusing on improving program efficiency and effectiveness by using past data and possible supplemental studies to improve sampling designs, indicators/summary measures, and analysis methods for detecting release events and their sources. For example:
   G4: to learn about the fate and transport of hydrocarbons released into PWS waters and how best to identify their source signatures in mussel tissue and sediments.
3. Explicitly define the region of the GOA that PWSRCAC is interested in making inferences regarding (the regional target frame);
4. Clarify LTEMP's event detection goals; specifically, there needs to be a clearly defined lower threshold on the type, or at least magnitude, of release events LTEMP wants to detect.
5. Explicitly adopt a tiered detection approach, where fingerprinting is only undertaken when an event of sufficient magnitude, as determined by the clarified detection goals (below), is detected.
6. Clarify the definition of *background* levels. If the resulting definition implies that some study sites have reached that level, effort should be given to developing more formal detection methods using the baseline data (e.g., summary statistic-specific thresholds). Similarly, if any impacted sites are thus deemed to have returned to background conditions, their recovery rates should be estimated and published in the literature.
7. Clarify LTEMP's source fingerprinting goals – recent events or remnant signals of past events (which introduces the added difficulty of weathering)? Then RCAC should decide whether there is a need to pursue development of more statistically refined fingerprinting methods. At the moment the current rough indices that appear adequate for the implicit goals may not be sufficient if the explicitly decided goals differ.
8. Explicitly state and record the reasons for dropping the collection of gonadal tissue from the LTEMP program.
9. [Policy + Data based] Set explicit, site-specific (and perhaps season-specific) thresholds for release event detection; considerations could include hydrocarbon sources in specific areas, water column PAH concentration, release events, and known toxic effect levels.

Method Refinement: Sampling

*Refining Sampling Strata*

10. Review the findings of Mudge (2002), the recent LTEMP analyses (Payne et al. 2006) and any other work on regional background signals to try and refine the regional partitioning in terms of common background influences. This may provide justification for simplifying the current partition into Port of Valdez, East PWS, West PWS, and GOA, lending support to logical application of the limited control sites to other nearby regions. Alternatively, these results may caution against such comparisons and reinforce the need to expand the set of control sites to regions such as West PWS that currently do not have any control sites.

11. Obtain the PWS/GOA coastline segment classifications, assuming NOAA/NMFS or some

other party has developed them (likely as a result of EVOS), and generate the summary distribution of coastline segments by type. This should then be compared to the coastline segment types currently being sampled (Table 2). This information could be used to guide any future selection of new sample sites (expand representation by coastline types), though this factor is likely secondary to the regional influence discussed above. Identify viable mussel beds near the currently declining sites that meet the original selection requirements and that can be considered candidates for future sampling.

*Modify Monitoring Design*

12. Explore the ramifications of adopting some form of cross-sectional, rotating panel, or split-panel monitoring design to lessen the response burden on the existing mussel beds and potentially extend their duration as viable monitoring sites (see below).

    1. Consider shifting to a rotating panel design (Table 3 c) where the current panel of sites is split into two panels (subsets) and each panel is sampled on a rotating basis, e.g. odd years/even years. The liberated funds could be used to supplement the panel sites with a random sample of sites. This lessens the response burden and overall cost, but does not provide any expanded coverage of the study area. Proper analysis also becomes much more complex if there is interest in estimating any regional trends.

    2. Alternatively, evaluate using split-panel design (Table 3 d), supplementing the long-term sites with randomly chosen sites each year, as suggested in Payne et al. (2006). This would provide greater power to detect events since there would be broader regional coverage each year. It would also allow LTEMP to slowly develop a body of (temporally limited) baseline data throughout the PWS/GOA, providing insight into patterns and boundaries within these large regions. However, this monitoring design would not support site-specific estimation of trends at the sites selected anew each year (though one could still estimate site-specific trends from the panel(s) that were continually revisited) efficient analysis also becomes much more complex.

    3. Note that under Subpoint 2 immediately above, the long-term sites could be revisited following a rotating panel design, allowing adoption of both subpoints but at a clear cost of increased management complexity (organizational ovesight) and analysis complexity.

*Response Burden/Declining mussel beds (may depend on decisions regarding Monitoring Design)*

13. Identify viable mussel beds near the currently declining sites that meet the original selection requirements and that can be considered candidates for future sampling.

14. For each declining site, consider sampling both the declining site and at least one of the nearby viable sites so as to assess the comparability of their hydrocarbon signals and build up a calibration period before the declining site disappears.

15. Evaluate utilizing caged mussels, for observation of particular areas of concern that do not have native mussel populations, to provide a test group for comparisons of control sites to impacted sites, or to supplement declining mussel populations in current site locations.

16. Evaluate utilizing SPMD's to supplement mussel collection as an additional source of information about hydrocarbon concentrations, or to monitor sites with declining mussel

populations.

Method Refinement: Sampling

*Substrate sources, analytes of interest, summary indices*

17. Evaluate the correlation between hydrocarbons observed in mussel tissues and hydrocarbons actually present in the water column and sediments on a site-specific basis. While there are several factors that complicate this process, direct water sampling, sediment sampling, and SPMDs could be used to estimate background hydrocarbon concentrations at a particular site, and data regarding mussel response could be accumulated over time.

18. Refine source identification methods, and conduct a re-assessment of the suite of chemical analytes chosen for measurement. Specifically, occasional sediment sampling to test for sterane and terpane biomarkers could be employed. This could be accomplished with a rotating panel, or other type of sampling design, or in conjunction with site-specific thresholds relating to release events.

19. Determine the most appropriate summary statistical method for event detection – the less powerful but simpler TPAH/TSHC or the more refined source-phase composition. If the latter is chosen, then effort should be expended on improving the estimation method, ideally by developing a mixture estimation approach, at a minimum by refining the current tree model by letting standard model-fitting algorithms reformulate, and hence assess, the classification model.

Other

20. Require analytical contractors to display the winter observations separately from the summer observations as a simple step to improve visual recognition of trends in the summary statistics.

**Author's Comments regarding Recommended Tasks:**

Tasks 1-7, 9, 17, and 19 are essential to the long-term effectiveness and success of LTEMP and should be considered high priority. It should be noted that tasks 1-7 predominantly require action on behalf of PWSRCAC while Tasks 17 and 19 will also involve technical reviews, field studies, or other activities presumably conducted by contractors. The decisions resulting from these tasks will likely result in identification of a few short-term technical tasks (reviews, methods development, etc.) requiring technical assistance outside of PWSRCAC.

# APPENDIX B

# POTENTIAL PEER-REVIEWED PUBLICATION TOPICS

Based on the bulleted list in the scope of services, and in light of the current review, the following appear to be the most immediate topics for possible peer-reviewed publication:

*Analysis and Reporting of LTEMP data/methods refinement*

1. Background, temporal and geographical trends
2. Hydrocarbon source signatures – data visualization
3. Hydrocarbon source signature – logical algorithm

*Methodology*

4. Mussel and sediment monitoring methodology
5. Environmental variables affecting hydrocarbon concentration and signatures

Topic 1 is of primary importance as it provides the context for all of the other topics and will enlighten future discussions of methodologies and environmental variables in Prince William Sound. It requires that PWSRCAC implement/resolve a number of the initial Goals/Objectives recommendations given in Appendix A, i.e. 5, 6, and 19. If there was interest, and sufficient resources, this could include an exploration into appropriate process-control methods for statistically detecting change events. However, that is perhaps left as a possible separate topic as it could become quite involved. It should be noted that this requires resolution of how PWSRCAC wants to handle the problematic pre-1997 data. Topic 2 could be brought into Topic 1 in passing.

After PWSRCAC has read this Review, we would like to meet and discuss the tasks and goals in the scope of services in light of our recommendations. It would likely be easiest to complete and publish a paper based on data analysis rather than methodology review.

Please note that:

(i) Reviewing and addressing the issues raised regarding the relationship between hydrocarbon concentrations in mussel tissues and in the water column, and

(ii) Developing either a 'supervised learning' machine algorithm or statistical mixture analysis of the laboratory results to partition hydrocarbon signatures into phase sources contributions, are both beyond the scope of the original request for proposal.