Effectiveness of RIVPACS Predictive Models to Evaluate Diatom Response to Nutrient Stress in Coastal California Streams

Charles Ritz
California State University, Monterey Bay

Follow this and additional works at: https://digitalcommons.csumb.edu/sns_theses

Recommended Citation
https://digitalcommons.csumb.edu/sns_theses/19

This Master's Thesis (Open Access) is brought to you for free and open access by the School of Natural Sciences at Digital Commons @ CSUMB. It has been accepted for inclusion in SNS Master's Theses by an authorized administrator of Digital Commons @ CSUMB. For more information, please contact digitalcommons@csumb.edu.
EFFECTIVENESS OF RIVPACS PREDICTIVE MODELS TO EVALUATE DIATOM RESPONSE TO NUTRIENT STRESS IN COASTAL CALIFORNIA STREAMS

A Thesis
Presented to the
Faculty of the
Division of Science and Environmental Policy
California State University Monterey Bay

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
in
Coastal and Watershed Science and Policy

by
Charles Ritz
Spring 2010
CALIFORNIA STATE UNIVERSITY MONTEREY BAY

The Undersigned Faculty Committee Approves the

Thesis of Charles Ritz:

EFFECTIVENESS OF RIVPACS PREDICTIVE MODELS TO EVALUATE DIATOM
RESPONSE TO NUTRIENT STRESS IN COASTAL CALIFORNIA STREAMS

Marc Los Huertos, Committee Chair
Division of Science and Environmental Policy, CSUMB

Fred Watson,
Division of Science and Environmental Policy, CSUMB

Peter Ode,
California Department of Fish Game Staff Environmental Scientist

Marsha Moroh, Dean
College of Science, Media Arts and Technology, CSUMB

May 2010
Copyright © 2010

by

Charles Ritz

All Rights Reserved
You cannot step twice into the same stream. For as you are stepping in, other waters are ever flowing on to you.

Heraclitus of Ephesus (c.535 - 475 BC)
ABSTRACT

Effectiveness of RIVPACS Predictive Models to Evaluate Diatom Response to Nutrient Stress in Coastal California Streams

by

Charles Ritz

Master of Science in Coastal and Watershed Science and Policy
California State University Monterey Bay, 2010

The goal of this project was to determine if predictive models of diatom assemblages would provide an effective method to report on biological degradation in streams along the Central Coast of California. This project focused on nutrient stress to evaluate stream water-quality degradation. I employed the River Invertebrate Prediction and Classification System (RIVPACS) model with diatom assemblages. Diatoms were an accessible indicator of nutrient stress occurring in abundance on Central Coast streams. Diatom samples from 190 stream sites were used to construct and test the RIVPACS model. The RIVPACS methodology used a reference condition approach to compare assemblages at reference sites to observed assemblages at degraded test sites. Reference sites were used to train the predictive model and develop an expected taxa count. A ratio of observed taxa to expected taxa (OE) was the concluding measure of biological integrity at each site. I used the OE scores to test the postulate that degraded sites had diatom assemblages dissimilar from the reference site diatom assemblages. The RIVPACS model did not perform well. The model suffered from low precision of reference site OE scores (mean SD = 0.22) and lack of accuracy to consistently predict low OE scores at known degraded sites. However, the model was able to identify likely trends. For example, agricultural land use sites trended toward lower OE scores indicating possible biological degradation. The uncertainty in the RIVPACS model did not provide a definitive measure of model effectiveness. I concluded the assessment model was limited by the quality of reference streams and the temporal variability and spatial patchiness of diatom assemblages. I recommended further evaluations the explore the application of diatom assemblages to assess streams on the Central Coast.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT .......................................................................................................................... V</td>
</tr>
<tr>
<td>LIST OF TABLES .................................................................................................................... VIII</td>
</tr>
<tr>
<td>LIST OF FIGURES ................................................................................................................. IX</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS ............................................................................................................ X</td>
</tr>
<tr>
<td>CHAPTER 1 ............................................................................................................................ 12</td>
</tr>
<tr>
<td>WHY USE BIOASSESSMENTS TO TEST WATER QUALITY? ......................................................... 12</td>
</tr>
<tr>
<td>STREAM BIOASSESSMENTS ..................................................................................................... 12</td>
</tr>
<tr>
<td>Appropriate biological indicators: .................................................................................... 13</td>
</tr>
<tr>
<td>BIOASSESSMENT RATIONALE ............................................................................................... 13</td>
</tr>
<tr>
<td>BIOASSESSMENT APPLICATION ............................................................................................ 14</td>
</tr>
<tr>
<td>BIOASSESSMENTS AND NUTRIENT ENRICHMENT: A POLITICAL PERSPECTIVE ......................... 15</td>
</tr>
<tr>
<td>Legal Origins for Clean Water ............................................................................................ 15</td>
</tr>
<tr>
<td>Nutrients Enrichment in California .................................................................................... 17</td>
</tr>
<tr>
<td>BIOASSESSMENT AND NUTRIENT ENRICHMENT: A SCIENTIFIC PERSPECTIVE ...................... 18</td>
</tr>
<tr>
<td>Algal Characteristics ........................................................................................................... 18</td>
</tr>
<tr>
<td>Ecosystem Processes ......................................................................................................... 19</td>
</tr>
<tr>
<td>Diatoms and Nutrients ....................................................................................................... 19</td>
</tr>
<tr>
<td>RIVPCAS MODEL .................................................................................................................. 20</td>
</tr>
<tr>
<td>IMPLICATION OF A RIVPACS APPLICATION IN COASTAL CALIFORNIA ............................... 22</td>
</tr>
<tr>
<td>CHAPTER 2 ............................................................................................................................ 23</td>
</tr>
<tr>
<td>EFFECTIVENESS OF RIVPACS TO EVALUATE DIATOM RESPONSE TO NUTRIENT STRESS IN COASTAL CALIFORNIA STREAMS .................................................................................. 23</td>
</tr>
<tr>
<td>INTRODUCTION .................................................................................................................... 23</td>
</tr>
<tr>
<td>METHODS ............................................................................................................................. 23</td>
</tr>
<tr>
<td>Study Area ......................................................................................................................... 27</td>
</tr>
<tr>
<td>Sampling Design and Sample Collection .......................................................................... 29</td>
</tr>
<tr>
<td>Predictor Variables ............................................................................................................ 30</td>
</tr>
<tr>
<td>Reference Site Selection .................................................................................................... 31</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1: Variables to consider when identifying effective biological indicators ....................13
Table 2: Glossary of terms for reference streams and expected biological conditions ............15
Table 3: Environmental factors that affect diatom growth ....................................................18
Table 4: Predictor variables employed for associating reference-site biological groups and site environmental characteristics to predict expected taxa at test sites. ............32
Table 5: Commonly occurring species at reference sites....................................................40
Table 6: Occurrence of predictor variable in the best performing models. ............................41
Table 7: RIVPACS model performance from three scenarios for probability of capture (Pc) threshold as measured by OE values. ..................................................42
LIST OF FIGURES

Figure 1: Example of diatoms from California Central Coast .................................................. 18
Figure 2: Process overview of RIVPACS method ................................................................ 21
Figure 3 Process overview of RIVPACS method .................................................................. 26
Figure 4: Central Coast region and diatom sample-site locations ........................................ 28
Figure 5: Dendrogram of calibration sites ............................................................................. 39
Figure 6: Boxplot showing OE score comparison of validation, calibration and test sites .......... 43
Figure 7: Evaluation of reference site performance measured with O and E scores ............... 44
Figure 8: Nutrient comparison to OE scores, total phosphorus (TP) total nitrogen (TN) compared to test sites OE scores ................................................................. 46
Figure 9: Results of land use comparison to OE scores ....................................................... 47
Figure 10: Root mean squared errors (RMSE) from sixty one (61) top performing model OEs .................................................................................................................. 61
Figure 11: Percentage of sites correctly associated with the taxa dendogram groups ............ 61
Figure 12: $P_C=0.5$ boxplot ............................................................................................... 62
Figure 13: $P_C=0$ boxplot ................................................................................................. 62
ACKNOWLEDGEMENTS

This project is the result of the dedication and hard work of many people. I would like to thank Dr. Marc Los Huertos for his mentorship, inclusion in this project and his willingness to share knowledge. I would like to thank Dr. Peter Ode for his incredible insight into RIVPACS modeling. I would like to thank Dr. Fred Watson for sharing so much of his modeling and statistics experience and Dr. Scott Rollins for helping to conceive this project. My thanks to Dr. Kristy Morris and Dr. Jan Stevenson for their lab support and Karen Worcester at the State Water Resources Control Board for supporting and help funding this project. I would like to convey a heartfelt thank you to everyone on the periphyton team, a terrific staff to work with. To all my colleagues and peers especially, William Perry, Lisa Dillon, Shana Carmichael, Kelleen Harris, thank you for your help on this project and making it fun to go to work. Also, thank you to CSUMB faculty and staff, there are too many to list but you have all helped me at every step. Lastly, to my wife, Gretchen, I would not be here without you. Your support, love and your question, “forget the mortgage, forget the job, forget everything else – what would you love to be doing?” helped me achieve so much, thank you.
STRUCTURE OF THIS DOCUMENT

I divided this study into three main sections. Chapter 1 is a discussion of biological assessments, policy background and background on RIVPACS assessments. Chapter 2 is the main thesis project intended as a stand-alone manuscript. The appendix contains additional material included to compliment chapter two.
CHAPTER 1

WHY USE BIOASSESSMENTS TO TEST WATER QUALITY?

Stream Bioassessments

Aquatic bioassessments interpret the ecological condition of a waterbody by directly measuring the resident, surface-water biota (USEPA 1996). Bioassessments often utilize communities of organisms to communicate broad meaning beyond the measurement of a single organism (Karr 1981; Norris and Hawkins 2000). The inferences of indicator species can aid scientific knowledge, policy and management decisions and communicate the condition of a waterbody to a larger audience (Norris and Hawkins 2000). Biocriteria can provide the narrative guidelines or the numeric targets used to evaluate the biological integrity of a waterbody (USEPA 2000). States commonly designate the beneficial uses for a waterbody, such as important fisheries or critical habitats for species of concern. Biocriteria help evaluate and protect these aquatic life uses (USEPA 1999, 2000).

Researchers have made considerable progress to develop sophisticated techniques identifying the chemical constituents of water quality and potential sources of pollution (Cude 2001); however, traditional monitoring of chemical water quality and toxicological data can underestimate biological degradation by failing to assess the extent of ecological damage in streams (USEPA 1996; Yagow et al. 2006). Compounding the challenge to define ‘clean’ water is the complex and dynamic nature of lotic systems and the range of characteristics such as biological, physical, and chemical attributes of stream environments (Vannote et al. 1980; Resh et al. 1988; Dodds et al. 1998; Allan and Castillo 2007). Sole reliance on stream chemistry monitoring may be an incomplete indication of stream health whereas, biological indicators provide a more effective tool to monitor the ecological response to chemical stressors in the environment (Barbour et al. 1999; Karr 1999; Karr and Chu 2000; Yagow et al. 2006).
Due to varying political perspectives, fiscal challenges and the dynamic nature of streams, effective assessments must incorporate multiple factors relevant to policy regulations, management activity, economics impact and sound science (Noss 1990; Norris and Hawkins 2000; Spellerburg 2005). Norris and Hawkins (2000) outline six variables to consider when identifying appropriate biological indicators for stream integrity (Table 1). By evaluating the suitability of particular taxa for stream biological assessments can aid project design and situational application.

Table 1: Variables to consider when identifying effective biological indicators (adapted from Norris and Hawkins 2000).

<table>
<thead>
<tr>
<th>APPROPRIATE BIOLOGICAL INDICATORS:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characterize and measure simple and complex ecological systems</td>
</tr>
<tr>
<td>Offer straightforward interpretable results</td>
</tr>
<tr>
<td>Respond in a predictable manner to anthropogenic changes</td>
</tr>
<tr>
<td>Geographically relevant for region(s) being assessed</td>
</tr>
<tr>
<td>Consistent with resource management objectives</td>
</tr>
<tr>
<td>Offer valid and defensible scientific meaning</td>
</tr>
</tbody>
</table>

Bioassessment Rationale

Biological assessments and the associated biocriteria evaluate the integrity of freshwater streams. Stream taxa, such as fish, invertebrates or diatoms, have the potential to assimilate the effects from anthropogenic changes into their population structure (Karr 1981; Wright et al. 1984; Barbour et al. 1999; Stevenson and Pan 1999). Changes in assemblage composition thus effectively measures the biological integrity of streams, including changes in stream chemistry and changes unrelated to stream chemistry such as physical modifications (Davis and Simon 1995; Barbour et al. 1999; Bailey et al. 2004; Magurran 2004). Biological integrity, in this instance, refers to the unimpaired condition and the ability of aquatic taxa, communities and guilds to respond and recover from natural fluctuations (Angermeier and Karr 1994; Karr 1999). As part of the long-term national goals for clean water, the United States Congress incorporated a concept of biological integrity into United States water quality policy. The Federal Water Pollution Control Act Amendments of 1972 and 1987, referred to as the Clean Water Act (CWA)
requires federal and state governments to restore and maintain the "biological integrity of
the Nation's waters" (USEPA 2002). The CWA established the need to preserve and
protect the biological integrity of aquatic resources and institute the appropriate
biocriteria to assess water quality.

**Bioassessment Application**

To interpret the relative scores of bioassessments, researchers often compare
sampled sites against an expected or reference condition (Stoddard et al. 2006). The
reference condition approach (RCA) used during many biological assessments can
quantify the biological integrity of aquatic resources (Hughes et al. 1986; Moss et al.
1987; Reynolds et al. 1997; Stoddard et al. 2006). The RCA evaluates indicator
organisms at reference sites and compares the reference sites to test sites. Bioassessments
using a RCA can measure the deleterious effects anthropogenic stressors have on
indicator organisms by first measuring stream integrity at sites unaffected by human
influence. Early development of reference condition applications had varying definitions
of the reference condition (Hughes 1995), and definitive classifications, reference or non-
reference. However, in application, a gradient of reference conditions exist and range
from high-integrity undisturbed sites to lower integrity disturbed sites (Stoddard et al.
2006). Stoddard et al. (2006) outlined a lexicon of terms to define this gradient of terms
and the expected biological conditions for reference sites (Table 2). Several studies in
California have successfully used a RCA approach to bioassess changes in invertebrate
assemblages (Hawkins et al. 2000; Ode et al. 2005; Herbst and Silldorff 2006; SWRCB
2006).

Resource managers will frequently use a suite of biological indicators and
multiple stream chemistry measurements to assess biological integrity rather than sole
reliance on one measurement or indicator (Karr 1999; Karr and Chu 2000; Bain et al.
2000; Norris and Hawkins 2000; Yagow et al. 2006). Several applied bioassessment
methods use the RCA approach. For example, Multi-Metric Indexes (MMIs) assign
values (metrics) to multiple biological attributes and compare results of reference streams
to test streams.
Table 2: Glossary of terms for reference streams and expected biological conditions (adapted from Stoddard et al. 2006)

<table>
<thead>
<tr>
<th>Glossary of Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Reference Condition (RC(BI)) – reserved to exemplify true “naturalness” and meant to preserve goals and objectives outline in the Clean Water Act.</td>
</tr>
<tr>
<td>• Historical Conditions (HC) – describes conditions at some point in history, e.g. pre-intensive agriculture or pre-settlement, may represent RC(BI).</td>
</tr>
<tr>
<td>• Minimally Disturbed Condition (MDC) – represents best approximation of biological integrity. Recognizes no stream or river is completely free of human disturbance, such as from atmospheric deposition. Nonetheless, accounts for natural variability in the absence of significant anthropogenic disturbance.</td>
</tr>
<tr>
<td>• Least Disturbed Condition (LDC) – denotes best available conditions for a particular region. Recognizes some level of predetermined disturbance but considered the “best of” for an area.</td>
</tr>
<tr>
<td>• Best Attainable Condition (BAC) – symbolizes potential for biological conditions to recover if best management practices were implemented.</td>
</tr>
</tbody>
</table>

In contrast, the RIVPACS (River Invertebrate Prediction and Classification System) and AusRivAs (Australian River Assessment Scheme) use multivariate models to predict how much a test site would support the biota as compared to reference sites (Reynoldson et al. 1997; Karr and Chu 1999; Hawkins et al. 2000). These approaches assist researchers and resource managers to clarify and understand the definition of a “clean” stream and quantify the influence stressors have on the biological integrity.

Bioassessments And Nutrient Enrichment: A Political Perspective

Legal Origins For Clean Water

Policy makers in California have enacted water-pollution control legislation as required from federal mandates and from state sponsored initiatives. From the federal side, the CWA is the foundation for regulating the release of pollutants into waters of the United States (USEPA 2002). Section 303(d) requires states to identify beneficial use for streams and determine water quality standards needed to meet those demands. The standards, defined as the Total Maximum Daily Loads (TMDLs), require states to set up programs to monitor and assess streams and rivers. Section 319, Nonpoint Source Management Program, requires states to assess and establish programs to address
problems associated with nonpoint source pollution. Additional federal environmental laws also influence the management of stream systems. The Coastal Zone Management Act of 1972, as amended in 1987, addresses issues in a marine context but includes sections on runoff and nonpoint source pollution. The Endangered Species Act protects endangered plants and species including their habitat. Lastly, the Safe Drinking Water Act (SDWA) of 1974 calls for states to assess drinking water resources, such as rivers and lakes, ensuring acceptable and establishes safe water quality for consumption.

The California Water Code is the body of legislative policy that regulates all water related activities in the state. Within the code, the California Porter-Cologne Act of 1969 established California's early response to environmental concerns about water protection and preserving beneficial uses with streams and rivers. The act established responsibilities for the State Water Resources Control Board (SWRCB) and for semi-autonomous Central Coast Regional Water Quality Control Boards (CCRWQCB) to assess and implement quality improvement strategies. The state board is responsible for managing statewide issues, whereas the regional boards have responsibilities for creating regional plans or basin plans. The California Coastal Act of 1976 authorized the California Coastal Commission to assess coastal and marine ecosystem health. Included are provisions to work in conjunction with the SWRCB and the regional boards to implement nonpoint source programs. California Environmental Quality Act (CEQA) 1970 requires the state to take actions to help preserve and mitigate plans and projects that may have an environmental impact. In addition, the Public Trust Doctrine of 1928 ensures state stewardship to preserve the environmental resources of sovereign lands for present and future generations.

The SWRCB controls, protects, and manages the beneficial uses of streams and rivers. In addition to monitoring urban influences and the health status of the streams resources, the board manages the discharge of agricultural wastes such as fertilizers, pesticides and sediment. Historically, along the Central Coast the CCRWQCB granted waivers to agricultural operations for waste discharge into surface waters (CCRWQCB 2006a); however the CCRWQCB adopted a conditional agricultural waiver program in 2004 requiring education, monitoring and adoption of best practices for irrigated farming operations in the central coast region (CCRWQCB 2006b).
NUTRIENTS ENRICHMENT IN CALIFORNIA

California receives considerable benefit from streams and rivers. Streams are vital economic and natural resources. Often characterized as “renewable-but-limited resources” (Tietenberg and Lewis 2009), streams help sustain human populations and provide habitat for aquatic and riparian species. Broad resource examples include water consumption, habitat for fish, water for crops, hydroelectric power, and waste removal. However, as previously discussed, the biological integrity of aquatic life uses in California may be in question. Examples of pathways for pollutants entering the waterway include point source and nonpoint source. These two types of sources can degrade surface waters and ultimately lower the biological integrity, thus the benefits derived from streams (Carpenter et al. 1998; Dodds et al. 1998).

Non-point source surface water impairment commonly occurs in California due to excessive inputs of nutrients, such as phosphorous (P) and nitrogen (N). Sources of P and N include natural occurrences, runoff from agriculture and other urban activities (USEPA 2002, 2005, 2006). Cultural eutrophication can lead to excessive algal growth and a reduction in dissolved oxygen resulting in a negative influence to biological assemblages (USDA 1999; Dodds et al. 2002). Current policy for the levels of nutrients allowed in Californian Central Coast streams may not adequately protect aquatic life uses as required by the California Porter-Cologne Water Quality Control Act and the CWA. The numeric objectives for inorganic nitrates in many coastal California streams are set to drinking water standards of 10 milligrams per liter (nitrate-N = 10mg/L) and there are no numeric standards for inorganic phosphorous (ortho-P). N and P are limiting nutrients in aquatic ecosystems (Tilman et al. 1982; Carpenter et al. 1998); thus, relatively small increases in nutrient loads can have significant affects on aquatic ecosystems (Dodds et al. 2002). Implementation of TMDLs for nutrients on California Central Coast streams (e.g., Chorro Creek, Los Osos Creek and Pajaro River) illustrates a problem with nutrient over-enrichment in the region. However, few assessments exist in California and specifically on the California Central Coast to evaluate the effects of nutrient stress on aquatic habitat.
Bioassessment and Nutrient Enrichment: A Scientific Perspective

Table 3: Environmental factors that affect diatom growth (Weitzel 1979)

<table>
<thead>
<tr>
<th>Availability of light</th>
<th>Solar incidence</th>
<th>Turbidity</th>
<th>Substrate type</th>
<th>Depth</th>
<th>Currents</th>
<th>Water Velocity</th>
<th>pH</th>
<th>Alkalinity</th>
<th>Nutrients</th>
<th>Dissolved metals</th>
</tr>
</thead>
</table>

Algal Characteristics

Generally, algal assemblages grow in a variety of streams from mountainous, low-order streams to relatively flat, high-order rivers. Algal assemblages contain a diverse collection of plant-like organisms constituting the basis of stream food webs and are important elements in the stream ecosystems (Cushing and Allan 2001). Diatoms (Bacillariophyceae) make up part of the micro-flora of submerged, benthic organisms, commonly referred to as periphyton (Weitzel 1979). Though microscopic, periphyton can be “seen” and felt as the greenish or brownish slippery substance covering substrate material in many streams. The unicellular eukaryotic diatoms contain photosynthetic pigmentation and silica infused cell walls (Figure 1). Multiple environmental factors affect diatom growth (Table 3). Light and nutrients have been identified as the main factors regulating primary productivity (Weitzel 1979).

Figure 1: Example of diatoms from California Central Coast (Dillon 2008, printed with permission)
ECOSYSTEM PROCESSES

Several ecosystem processes influence plant growth and water quality. The hydrologic cycle, describes the interaction of climatic features, such as precipitation and evaporation, with biological variables and the flow of surface and ground water. Stream work in this cycle occurs as a function of slope, elevation and the ability of streams to transport runoff and sediment (Leopold et al. 1964). In addition to the hydrologic cycle, two chemical processes essential to plant growth include the nitrogen cycle and the phosphorous cycle (Allan et al. 2007). The nitrogen cycle occurs when nitrogen rich compounds in soils, such as areas of agricultural cultivation, decompose and oxidize leaching plant accessible nutrient, such as ammonium ($\text{NH}_4^+$), through surface runoff. Similarly, the phosphorous cycle includes the release of organic phosphorous from multiple sources, such as fertilizers, manures and industrial wastes, into stream systems from surface runoff and erosion.

DIATOMS AND NUTRIENTS

Many investigators have documented the use of algal assemblages, specifically diatoms, to characterize the effects from anthropogenic changes (Patrick 1968; Hansmann and Phinney 1973; Pan et al. 1996; McCormick and Stevenson 1998; Chessman et al. 1999; Carpenter and Wait 2000; Fore and Grafe 2002; Passy and Bode 2004; Cao et al. 2007). Furthermore, multiple researchers have established relationships between diatom assemblages and levels of nitrogen and phosphorous (Pan et al. 1996; McCormick and Stevenson 1998; Leland et al. 2001; Munn et al., 2002; Weihoefer and Pan 2006; Ponader et al. 2007; Lavoie et al. 2008). As indicator taxa, diatoms have multiple benefits because diatoms are short-lived organisms; diatoms rapidly assimilate stream nutrients, a relatively abundant and important component in the food web (McCormick and Stevenson 1998).

Availability of nitrogen and phosphorous limit diatom biomass and growth (Smith et al. 1999; Dodds et al. 2002). The availability of these inputs and other environmental conditions influence the abundance and composition of diatom assemblages (Sigee 2005). McCormick and Stevenson (1998) argued diatom abundance, rapid growth and early senescence allowed assemblages to quickly integrate environmental changes into their community structure.
RIVPCAS Model

The predictive-type model, RIVPACS, interprets the biological integrity of stream sites using invertebrate assemblages. Stream researchers first developed the RIVPACS method in Great Britain to establish the baseline health of streams and rivers (Wright et al. 1984; Moss et al. 1987). Researchers evaluated the process in the United States and a similar process in Australia (Norris 1996; Hawkins et al. 2000). RIVPACS compares the expected occurrence of macroinvertebrate species at reference sites with observed occurrence at test sites (Hawkins et al. 2000). The strength of the predictive models relies partly on how effectively the reference sites represent the gradient of conditions found at the test sites (Norris and Hawkins 2000). Model construction first clusters reference sites biologically, grouping like sites according to the occurrence of assemblages. Discriminant analysis attempts to associate the biological groupings with major environmental attributes of the reference sites. In an effort to isolate potential stressors, discriminant modeling only utilizes non-anthropogenic environmental attributes, for example latitude, elevation and precipitation. Lastly, an appraisal of test sites assigns each test site a probability of membership in each of the environmentally grouped reference clusters (Moss et al. 1987; Hawkins et al. 2000).

The endpoint indices consist of observed to expected ratios (OE) for stream test sites. Impairment is a measurement of how far the assemblages of a test site deviate from the assemblages of a reference site. For example, an OE value significantly less than one (OE << 1) would indicate the absence of assemblages at the test site, thus a degraded site. A non-impaired score of an OE equal or close to one (OE ≈ 1) indicates the observed occurrence of assemblages at a test site is approximately equal to the expected occurrence at reference sites. Model construction commonly excludes the occurrence of assemblages at the 95% level and 5% level (Hawkins et al. 2000). This exclusion increases the sensitivity of the models by removing taxa occurring at nearly all the reference sites, and decreases exaggerated exclusivity by eliminating rare occurrences. Thus, the OE metric can represent a precise measurement of biological integrity. Post OE processing, a comparison of chemical levels, such as nitrogen and phosphorous, present at the test sites and the OE index can relate the effect changes in stream chemistry have on the resident
biota. Figure 2 shows an overview of their entire RIVPACS process from reference site selection to OE index endpoints.

Instead of invertebrates, several researchers have employed benthic diatoms (Bacillariophyceae) to assess streams using RIPACS-type predictive models (Chessman et al. 1999; Mazor et al. 2006; Cao et al. 2007). Their results have been somewhat mixed. Environmental conditions on the California Central Coast and diatom life history attributes may lend themselves to a RIVPACS diatom evaluation on the Central Coast. Conditions such as the Mediterranean climate can account for multiple annual growth cycles, and the ephemeral status of some streams can support quick growth populations and potential for stream flashiness, allowing diatoms to incorporate chemical fluctuations into their assemblage structure. However, multiple and variable growth cycles may serve to confound sampling data when comparing assemblages at various levels of growth.

---

**Figure 2: Process overview of RIVPACS method.**
Implication of a RIVPACS application in Coastal California

Stream health on the California Central Coast affects many individuals including farmers, residents and outdoor enthusiasts. Streams in this region provide a mix of beneficial uses such as replenishment groundwater recharge, drainage, endangered species habitat (e.g. Steelhead, *Oncorhynchus mykiss*) and scenic destinations. Detection of human caused degradation, in this region, can be difficult to detect against a background of normal chemical and biological variations and the pervasive and historic anthropogenic influences.

A diatom RIVPACS investigation adds a line of evidence available for interpreting the biological integrity and impact on aquatic life uses. A suite of evaluation techniques, such as indicator assessments and water quality monitoring can help discern the overall health and status of Central Coast streams. A diatom assessment can inform resource managers on the potential effects from biological stressors due to nutrient over-enrichment. The results of this project may have a significant bearing on the agricultural community and other land-use stakeholders. A review of numeric nutrient objectives and OE scores could have policy and economic ramifications, such as assessing CWA compliance, prioritizing monitoring and remediation efforts or measuring management effectiveness.
CHAPTER 2

EFFECTIVENESS OF RIVPACS TO EVALUATE DIATOM RESPONSE TO NUTRIENT STRESS IN COASTAL CALIFORNIA STREAMS

Introduction

The goal of this research was to evaluate the effectiveness of diatom-based, predictive models to assess streams with degraded water quality on the California Central Coast. In this thesis, I described the development and performance of diatom predictive models to generate an OE metric (ratio of observed diatom taxa to expected diatom taxa) to measure biological integrity. The OE metric provided the concluding measure of biological integrity at stream sites. For these models, an OE score of one represented high biological integrity, whereas and OE score considerably different than one represented degraded biological integrity. I organized the predictive model construction into four major components. 1) Sampling all stream sites and identifying degraded sites and reference sites on the Central Coast in order to describe difference of reference quality assemblages as compared to the degraded sites. 2) Construction of predictive models from reference sites to identify environmental variables used to predict the expected taxa at impaired sites. 3) Utilization of the predictive models to generate an OE score for all sampled sites, degraded and reference; and 4) analysis of precision and accuracy of OE scores to successfully identify degraded and reference sites.

Impaired water quality can have numerous effects such as reduced biological diversity, habitat destruction, economic losses, legal implications and other social and biological impacts (Karr and Chu 2000; Poff et al. 2003; Baron and Poff 2004). In addition, excessive levels of nutrients associated with poor water quality in stream ecosystems are well established as significant ecological stressors in the Western United States (USEPA 2005, 2006). In the California Central Coast region, non-point source pollution from urban and agricultural areas is present in multiple streams (Los Huertos et al. 2001; Anderson et al. 2003; Dowd et al. 2008). Problematic areas for contaminated runoff often include the lower portions of river valleys, such as the Salinas, Pajaro and
the Santa Maria Rivers (Los Huertos et al. 2001; Anderson et al. 2003; SWRCB 2006; Dowd et al. 2008). Ultimately, this project assessed these problematic regions and evaluated the effectiveness of an assessment technique using diatoms as indicators of nutrient stress.

Several environmental assessment techniques exist to assess the degree of stream impairment. Examples of assessments include sampling for chemical and toxicity levels, paleoecologic studies, landscape and stream-form analysis, and biological assessments (Dodds et al. 1998; Bain et al. 2000; Yagow et al. 2006). However, all of these assessments have potential drawbacks. Strict reliance on nutrient chemistry and toxicity monitoring in stream ecosystems can be problematic and fail to detect the effects of pollutants on biological systems (Karr and Yoder 2004; Yagow et al. 2006). One-time water quality sampling and pollutant concentrations may not detect stressor signals with temporal variability such as water pulses, which inappropriately characterize conditions from fluctuating chemical concentrations. Water quality samples may also fail to determine whether pollutant levels are harmful to resident biota. In contrast, where assessment tools are available, scoring ecological conditions or analyzing assemblages of aquatic organisms, such as diatoms, invertebrates or fish, can provide sensitive methods for evaluating biological integrity (Karr 1981; Wright et al. 1984; McCormick and Stevenson 1998; Barbour et al. 1999; Stevenson et al. 2008). A significant relationship between algal assemblages and water quality is well-established (Kolkwitz and Marsson 1908; Patrick et al. 1968; Tilman et al. 1982; Stevenson et al. 2006). For stream bioassessments, diatoms are effective ecological indicators due to their variability, wide distribution, relative abundance and ability to integrate changes in water quality rapidly (Pan et al. 1996; McCormick and Stevenson 1998; Sabater and Admiraal 2005; Cao et al. 2007; Stevenson et al. 2008).

A multivariate, predictive modeling approach developed in Britain, known as the River InVertebrate Prediction And Classification System (RIVPACS), measures biological integrity by quantifying the taxonomic completeness of biological assemblages at stream sites (Wright et al. 1984; Moss et al. 1987; Wright 1995; Marchant et al. 1997; Clark et al. 2003; Bailey et al. 2004). Taxonomic completeness measures the observed set of organisms relative to that expected to occur in the natural state relatively free of the
stressor of concern. The predicted assemblages are determined statistically using a set of control sites referred to as reference sites to generate weighted averages of taxa lists. Conceptually, the predicted inventories are created by adding the weighted frequencies of species occurring at reference sites. The weighting is determined by the probability of a test site belonging to a group of reference sites (Wright et al. 1984). In essence, each site receives a site-specific expected species list based on 1) the potential membership to reference groups and 2) the proportion of site species occurring at the reference groups. Reference sites are representative of regional stream sites determined to have high biological integrity. Reference sites also are descriptive of the range of conditions similar to the known degraded sites, referred to as test sites. I applied a reference condition approach to identify the ‘least disturbed’ or ‘best available’ streams (Hughes et al. 1986; Stoddard et al. 2006).

Ultimately, the model process uses the OE score to measure degradation. The OE measurement compares the observed assemblages of diatoms at test sites as compared to the assemblages expected in the absence of anthropogenic disturbance. A site would be considered non-degraded if it did not depart significantly from one (OE ≈ 1). A score of one would indicate observed assemblage composition equals reference assemblage composition. The model creates expected taxa assemblages for each site (sites sampled in this study include reference and test sites). Figure 3 shows the general process I used to develop the RIVPACS model. The OE score was based on an exact match of the statistically generated species and counts in the expected value when compared to the actual observed species counts from the test sites. Biological integrity represents the proportion of expected taxa present in a test-site stream sample (Hawkins 2009). The OE ratio, in theory, ranges from zero to one and greater than one. An OE value considerably less than one or substantially greater than one (OE << 1 or OE >> 1) would indicate a possible degraded site or low biological integrity, whereas a score close to one (OE ≈ 1) was inferred as reference-state, high biological integrity. Sites with high OE scores (OE >> 1) indicate more species were counted at the test sites than were expected. This may indicate greater biological diversity at the test site or possible enrichment causing an assemblage shift (Bailey et al. 2004).
I investigated the assumption underlying the RIVPACS approach that differences in observed versus expected taxa are related to harmful environmental conditions not associated with natural variations. I postulated the RIVPACS-type predictive models were suitable within the Central Coast region using diatom assemblages for assessment of stream integrity. The goal of this research was to evaluate the effectiveness of diatom based RIVPACS-type predictive models to show biological degradation at impaired water quality sites. This study focused on excessive nutrient stress because the biological effects from eutrophication on California Central Coast streams were not well documented. In other regions, researchers have utilized diatom based RIVPACS-type predictive models to measure biological integrity on stream ecosystems (Chessman et al. 1999; Mazor et al. 2006; Cao et al. 2007; Carlisle et al. 2008). Ultimately, this study continues research on the development of diatom based RIVPACS models and will aid resource managers in establishing biological assessment tools on the Central Coast.
Methods

STUDY AREA

Individual diatom samples (n=190) were collected from wadeable streams along the California Central Coast region during the 2007 and 2008 summer and fall sampling seasons, with the exception of a small number of samples collected in March 2008 from intermittent-type streams. The majority of sample sites were located in a State Water Resources Control Board Region 3, which is the region overseen by the Central Coast Regional Water Quality Control Board (Figure 4). This region covers 29,200 square kilometers, includes approximately 3,798 kilometers of perennial and annual streams and 378 miles of coastline (SWCRB 2002). The area encompasses portions of Santa Cruz County on the coast, inland to the counties of Santa Clara, Monterey, San Benito, San Luis Obispo and south to parts of Santa Barbara County and Ventura County. Multiple north-south trending mountain ranges populate the region, such as the Santa Cruz Mountains, Diablo Range and Santa Lucia Range. The mountains are steep but relatively low in elevation with the highest peaks less than 1800 m. Runoff events from the watersheds typically have short lag times after rainfall events and high peaks due to the relative size and steepness of the surrounding mountains (Mount 1995). Unstable rock and soil types, such as alluvium and sandstone, large rates of uplift owlands separate the mountains such as the Salinas and Santa Maria river valleys. Characterized by a Mediterranean climate, the Central Coast contains several ecological regions. Ecoregions include Coast Range, California oak woodland and California chaparral (Omernik 1987). Climatic attributes for the region include mild wet winters, dry hot summers and mild coastal temperatures (Sugihara et al. 2006). Precipitation patterns vary greatly from 1700 mm mean annual precipitation in the Santa Cruz Mountains to 250 mm mean annual precipitation the dryer interior Salinas River valley (PRISM 2004).
Figure 4: Central Coast region as defined by Central Coast Regional Water Quality Control Board (California Interagency Watershed Map of 1999); diatom sample-site locations including reference and degraded sites and National Land Cover Dataset (2001); shaded relief derived from USGS National Elevation Dataset
SAMPLING DESIGN AND SAMPLE COLLECTION

In conjunction with California State University Monterey Bay and a state-funded project studying periphyton-based bioassessments, a team of researchers performed fieldwork and sample collection. We developed our sampling plan with two main objectives. 1) We sampled streams with known impairment in order to test the capacity of the predictive models to detect departure from reference conditions, and 2) we located and sampled reference streams to model the expected taxa. Staff used landscape analysis with geographic information systems (GIS) to generate a random set of possible sample locations throughout the region. Sites were originally identified in part by calculating accessibility (proximity to public roads) and stream order. However, field teams were unable to utilize some of the randomized sites. Limited accessibility, logistical considerations and a multi-year drought constrained the ability of teams to sample from pre-identified locations. Field crew leaders used best professional judgment and consultation with area experts to identify the majority of sample locations. We sampled wadeable streams with varying morphological features and a range of ecological characteristics. This included headwater streams, mid-valley streams, and low-valley streams with diverse land uses in the surrounding watershed. Land uses examples such as urban areas, forests, recreation and agricultural settings were sampled. In addition to sampling impaired test sites, we sampled sites with minimal disturbance in the watershed such as state parks, reserves and undeveloped regions of the Central Coast.

Field personnel used rapid assessment techniques consistent with methods described in Ode (2007) and a modified algae collection method from Barbour et al. (1999) and Peck et al. (2006) to record and collect samples. Sampling consisted of 150m reaches for streams less than 10m wide and 220m for streams greater than 10m wide. Each reach was subdivided into 11 transects of 10m or 20m respectively. Crews collected benthic diatom samples, physical measurements and stream habitat observations at each transect (e.g. depth, substrate type, velocity, riparian cover, etc.). Field notes for geomorphic and riparian features included sediment deposition, stream incision, herbivory, water clarity, channel slope (%) and evidence of fire. We collected water samples prior to diatom collection, placed the samples on ice, and processed for nutrient content at California State University Monterey Bay and University of California Santa
Cruz water quality laboratories. Laboratory samples were colorimetrically analyzed with a Lachat QuickChem 8000 series analysis system (Hach Company, Loveland, Colorado) for nutrient levels including dissolved and total phosphates and nitrates.

Diatom sampling consisted of gathering the benthic substrate at each transect location. Field crews systematically collected substrate material from the left, middle or right of the stream channel. The collection technique included sampling rocks or loose substrate material at each subsection. Personnel processed diatom collection by using a circular template (12cm²) to scrape rocks with a plastic spatula and toothbrush. Crews collected fines, sand and gravel type substrates with a similarly sized circular cup (12cm²) and spatula. In rare cases, bedrock and large boulder sampling for diatoms was not performed. If needed, substrata in close proximity to these substrate types were used as a proxy. Field crews rinsed the template region or the collected loose material into a container bucket. The total liquid volume was measured (ml), transferred into a 45ml aliquot sample bottles and placed on ice. Field personnel added a solution of glutaraldehyde within a 12-hour holding time to preserve samples. Diatom samples were refrigerated and sent to Center for Water Sciences at Michigan State University for identification to lowest possible taxonomic level, usually genus or species, hereafter referred to as operational taxonomic units (OTU). Relative abundances for OTUs were established the Center for Water Sciences from a count of 600 individuals.

**Predictor Variables**

RIVPACS-type predictive models utilize environmental variables to characterize reference sites. Discriminant function analysis was used to associate environmental variables with reference-site biologic groups (Wright et al. 1984). This association of environmental variables with reference-assemblages allows the model to make future predictions for expected taxa. The predictor variables were used to develop the OE metric by establishing a strong association to biological groups at reference sites and comparing those environmental characteristics at test sites to make expected taxa predictions. I chose 13 environmental predictor variables from the reference sites with a focus on variables expected to influence diatom assemblages. To avoid problems with circularity, I chose variables least related to nutrient stress (Reynoldson et al. 1997; Bailey et al. 2004). Ambient stream conditions, other water chemistry variables and various physical
attributes, such as canopy cover, were not used to avoid calculating reference-condition predictions based on human influenced predictors (Reynoldson et al. 1997). To ensure the discriminant analysis met assumptions for normality, I tested all the variables for normal distribution using graphical quantile plots, transformed as needed to ensure normality. I applied a correlation criterion (R<0.9), to exclude correlated variables. The list included climate, geomorphology and stream measurements at site locations (Table 4). I chose a limited number of variables based on recommendations for RIVPCS model development (Van Sickle et al. 2006).

**REFERENCE SITE SELECTION**

An initial step in the development of the RIVPACS-type predictive model was determining which streams from the entire pool of sampled sites represented the reference state. Reference selection was made after diatom sampling by evaluating landscape attributes with geographic information systems (GIS) and analyzing the field data. I used this modeling to determine a relative range of least degraded conditions. The relative range provided a practical method to identify reference streams in a landscape with known human development (Hughes et al. 1986; Bailey et al. 2004, Stoddard et al. 2006). On the Central Coast, farming, grazing, urban development, hydro-modification and oil production have had significant roles in the development of the region (Newman and Watson 2003). Moreover, changes in climatic patterns and atmospheric deposition potentially eliminated absolute or pristine-like stream conditions. As defined by Stoddard et al. (2006), the reference selection included a mix of minimally disturbed condition (MDC) sites, which represented sites in a near “natural” state, and least disturbed condition (LDC) sites, which represented reference conditions relative to the region. The MDC sites characterized the archetypal ‘healthy’ streams or streams with high biological integrity, whereas LDC sites denoted healthy streams only relative to the region’s land-use history (Stoddard et al. 2006).
Table 4 Predictor variables employed for associating reference-site biological groups and site environmental characteristics to predict expected taxa at test sites. Variables were chosen based on various criteria including potential influence on diatom life cycle and independence from human influence.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Transformation</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Seasonal / climatic attribute</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day of year (day of sampling 1-365)</td>
<td>number</td>
<td>raw</td>
<td>Calendar</td>
</tr>
<tr>
<td>Mean annual precipitation (sample point)(^1)</td>
<td>cm(^2)</td>
<td>raw</td>
<td>Map</td>
</tr>
<tr>
<td><strong>Basin geography and geomorphology</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latitude</td>
<td>decimal degrees</td>
<td>raw</td>
<td>Field and map</td>
</tr>
<tr>
<td>Longitude</td>
<td>decimal degrees</td>
<td>raw</td>
<td>Field and map</td>
</tr>
<tr>
<td>Site elevation(^2)</td>
<td>m</td>
<td>square root</td>
<td>Field and map</td>
</tr>
<tr>
<td>Catchment area (above sample point)</td>
<td>m(^2)</td>
<td>log (_{10})</td>
<td>Map</td>
</tr>
<tr>
<td>Sedimentary sandstone rocks(^3)</td>
<td>% area</td>
<td>arcsine square root</td>
<td>Map</td>
</tr>
<tr>
<td><strong>In stream reach attributes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reach gradient</td>
<td>% slope</td>
<td>square root</td>
<td>Field</td>
</tr>
<tr>
<td>Wetted width(^4)</td>
<td>m</td>
<td>square root</td>
<td>Field</td>
</tr>
<tr>
<td>Minimum depth(^5)</td>
<td>cm</td>
<td>log (_{10})</td>
<td>Field</td>
</tr>
<tr>
<td>Maximum depth(^6)</td>
<td>cm</td>
<td>log (_{10})</td>
<td>Field</td>
</tr>
<tr>
<td>Minimum velocity(^5)</td>
<td>m(^3)/sec</td>
<td>log (_{10})</td>
<td>Field</td>
</tr>
<tr>
<td>Maximum velocity(^6)</td>
<td>m(^3)/sec</td>
<td>log (_{10})</td>
<td>Field</td>
</tr>
</tbody>
</table>

\(^1\) PRISM Precipitation Maps (2004)
\(^2\) USGS national elevation model
\(^3\) Derived from, Division of Mines and Geology, CD-ROM 2000-007 (2000), GIS Data for the Geologic Map of California
\(^4\) Mean wetted width per transect
\(^5\) Minimum depth and velocity mean value of five lowest values from transect subsections
\(^6\) Maximum depth and velocity mean value of five highest values from transect subsections

I used multiple factors to define reference conditions (Appendix A). Land use and land cover data were provided from National Land Cover Data (NLCD 2001). I reclassified land-use categories using ArcGIS 9.2 (Environmental Systems Research Institute, Redlands, California) to broaden categories and facilitate interpretation (Appendix A). I evaluated the watershed above the sample locations using my modified land-use categories at varying scales (1k, 5k, and entire watershed). I combined several aspects of human activity (or lack of) to determine stream status. Variables such as the...
land-use categories (urban, light urban and agricultural densities), site conditions (physical and water quality characteristics) and best professional judgment were employed to evaluate reference status. The goal for reference selection was a balance between MDC, high integrity sites, versus more numerous LDC sites of lower integrity. The challenge to achieve this goal was balance of lower quantity MDC sites versus lower quality LDC to effectively represent the Central Coast region. Sites eliminated during the reference selection made up the pool of non-reference, potential degraded test sites, used to evaluate the OE metric of biological integrity. The pool of 190 total sample sites was reduced to 115 test sites (n=115) and 75 reference sites (n=75).

**Predictive Model Construction**

I provided descriptions of model construction below. The majority of these steps were developed by Van Sickle et al. (2006) using R (R Development Core Team 2009) and other RIVPACS procedures. More in-depth discussions of the statistical steps for RIVPACS model construction are described elsewhere (Wright et al. 1984; Moss et al. 1987; Kaufman and Rousseeuw 1990; Wright 1995; Marchant et al. 1997; Hawkins et al. 2000; McCune and Grace 2002).

**Step One: Organize reference sites.** From the pool of 75 reference sites, I separated out a small number of reference sites (n=23) to validate model performance and assess the accuracy of the predictive model. These reference-validation sites were not used to build the model. Instead, I used them post model construction to validate the accuracy of the predictive model. I separated out another set of reference sites (n=52), for calibration and construction of the predictive model. I evaluated model performance by generating an OE score for the calibration sites, and for the validation sites and reviewed how close to one, or high biological integrity, they scored (Hawkins et al. 2000; Van Sickle et al. 2006). Optimal model performance would be indicated by obtaining OE scores at validation sites of one or very close to one.

**Step Two: Biological clustering.** This step grouped reference sites together into like-assemblage clusters. In later steps, these reference clusters provided the basis for associating environmental variables to biological groups in order to create a predictive model. The clusters were employed to develop predictive models by clustering reference sites into taxonomically self-similar assemblages and to determine environmental
predictor variables to relate with the self-similar groups. Use of the RIVPACS method assumes that species composition and abundance within assemblages varies and conforms along changing environmental gradients and settings (McCune and Grace 2002). I started by removing rare species (those occurring at fewer than 5% of the reference sites) prior to the biological clustering (Hawkins et al. 2000). Rare taxa removal had two purposes. Removal decreased the “noise” from rarely occurring species (McCune et al. 2000), and reduced the need to transform the species abundance data (Michie 1982). After clustering, I added the previously removed taxa back into the data used for final OE predictions. These clusters of self-similar assemblages were used to find predictor variables strongly associated with the cluster groups in order to predict assemblages at test sites. I accomplished this by using discriminant analysis. These strongly associated predictor variables would be used to predict expected taxa at degraded sites.

To achieve the clustering of sites into groups based on their taxonomic composition, I created a hierarchical dendrogram using an agglomerative nesting technique (AGNES). The agglomerative nesting constructed a tree-like dendrogram by resolving individual sites at one end and one cluster containing all sites at the other end (Kaufman and Rousseeuw 1990; McCune and Grace 2002). A flexible, unweighted, pair-group average method (UPGMA) used untransformed relative abundance data in conjunction with a Bray-Curtis dissimilarity coefficient to determine ordination distances (McCune and Grace 2002; Van Sickle et al. 2006). Calibration sites were linked with a flexible-β method, where $\beta = 1 - 2\alpha$ (Hawkins et al. 2000; McCune and Grace 2002; Van Sickle et al. 2006). To reflect an ordination strategy similar to Ward’s linkage method (Ward 1963), which minimized sum of square errors derived from Euclidean distances, I followed McCune and Grace (2002) recommendations by setting $\beta = -0.25$. Once the dendrogram was created, I “pruned” the tree to establish cluster groups. Cluster groups were formed by creating a cut-off point on the dendrogram to maximize the formation of taxonomically self-similar groups with at least five reference sites per cluster (Hawkins et al. 2000).

**Step Three: Predictive modeling with environmental variables.** This portion of model construction associated environmental characteristics with the previously
established biological clusters. After model construction, this step enabled the model to predict references assemblages any site based on the similar environmental characteristics. I used linear discriminant analysis (DA) to perform the procedure. Linear DA analogous to multiple regression analysis, employs predictor variables to determine the best fitting classification of a sample set to a group (Williams 1983). The R program used DA to identify predictors with the strongest association to the biological clusters to classify and group the calibration sites to match the dendrogram of biological clusters (Wright et al. 1984; Marchant et al. 1997; Hawkins et al. 2000; Van Sickle et al. 2006). I executed a best-subset algorithm to analyze every possible combination of predictor variables. The discriminant algorithm executed every possible linear function by evaluating the suitability of each model from the set of all the \((2^p-1)\) models combined from a set of \(p\) predictor variables (Van Sickle et al. 2006). By analyzing all possible permutations, the prediction model could identify areas of over-fitting, erroneous significance and potentially avoid step-wise biases (Van Sickle et al. 2006; Poquet et al. 2009).

The best-subset R routine utilized Wilks' \(\lambda\) to calculate the strength of group separation. Wilks' \(\lambda\) described the variances for objects not explained by the discriminant functions (McCune and Grace 2002); thus, a small value, close to zero, indicated greater group separation, whereas a value close to one indicated no separation (McCune and Grace 2002). Van Sickle et al. (2006) opted for Wilks' \(\lambda\) because it was a popular test for significance regularly used in multivariate discriminant analysis (Tatsuoka and Tiedeman 1954; McCune and Grace 2002; Van Sickle et al. 2006); however, Wilks' \(\lambda\) was not used as a concluding measure of statistical significance rather a determinant for model prediction.

The best-subset routine ranked the top performing models (linear equations) using bins based on the number variables in the linear equation. For example, order one models included one predictor variable, order two models contained two predictor variables in every possible combination, order three contained three predictor variables in every possible combination, etc… until the 13\(^{th}\) order, which only contained one model with every predictor variable. The program calculated the orders separately ranking the models
in each order with Wilks’ λ. For choosing the strongest predictor models, I retained the top five performing models from each order (Van Sickle et al. 2006).

To increase the sensitivity of the model for predictor variable selection, I programmed the discriminant procedure with a probability threshold ($P_T$) of ≥0.25 (Hawkins et al. 2000; Van Sickle et al. 2006; Van Sickle et al. 2007). $P_T$ represents a modeling threshold to exclude rare and uncommon taxa from the predictive calculations. Discussion among modelers, as to the most effective level to remove rare species (0% to 70%) for best model performance, remains unresolved (Van Sickle et al. 2007). Accordingly, I performed a sensitivity analysis by adjusting the $P_T$ value (not shown here) to reduce error, increase accuracy and make final $P_T$ selection; however, overall model performance appeared somewhat insensitive to $P_T$ adjustment.

**Step Four: Repeat discriminant analysis for membership probability and determine taxon frequency.** This step determined the probability of any site belonging to a reference group. I used these probabilities during modeling to help generate the expected taxa lists. DA had a dual purpose for model development by first grouping the reference site data (step 3 above), and second by assigning the probability of any site (test or reference) being a member of any one of the classified reference groups ($P_j$). DA was used to accomplish this by maximizing the separation between a fixed number of groups (previously discerned from biological clusters) along an orthogonal scale in ordination space and calculated the probabilities of each site belonging to each group (Mahalanobis distance in multidimensional space between each site and the centroid of cluster groups) (McCune and Grace 2002; Poquet et al. 2009). A frequency of occurrence for each taxon ($k$) was established within each cluster group ($g$). The average proportion of each taxon within the member-established reference cluster groups ($g_{j,k}$) was calculated (Marchant et al. 1997).

**Step Five: Probability of capturing observed taxa at reference sites.** Final taxa counts were established using statistical and mathematical operations to generate the expected diatom assemblages. This step enabled me to exclude rare species, as needed, to improve model performance. To facilitate prediction of taxa at each site, the program summed the product of $g_{j,k}$ and $P_j$ to determine the ‘probability of capture’ ($P_C$) for each taxon. $P_C$ represents a similar probability to the probability threshold ($P_T$) used for
predictor selection. In this case, $P_C$ uses the final set of selected predictor variables and predicts the expected taxa for all sites. For this investigation, I compared a $P_C$ of 0.5, 0.25 and 0. The comparison of various $P_C$ values helped determine the effectiveness of removing rare species for modeling. Previous studies have shown rare species exclusion had the potential to improve model performance (Van Sickle et al. 2007).

**Step Six: Expected prediction and OE calculation.** These final model construction steps calculated total expected taxa for a site and produced the observed to expected ratio (OE) metric. The OE score was then used as a measure of biological integrity at stream sites. I executed the program for $P_C$ prediction. The program calculated the expected taxa (E) by summing the $P_C$ across all taxa at each site. Observed taxa from the test sites were counted only if the species were identified at reference sites. Species observed but not part of the expected lists were not incorporated into the OE metric. The procedure calculated observed taxa (O) at all the sites by summing the total of each expected taxon (derived from either $P_C$ 0, 0.25 or 0.5) observed in the actual sample data. The program also reported sites which had predictor variables attributes determined to be outside the statistical population of the reference site predictor variable attributes. Referred to as a chi-squared ($\chi^2$) test, this was a measure in multivariate space of the Mahalonabis distance between a test site and the classification groups (Hawkings et al. 2000). The distance was a measure of how similar a test site was to a cluster of reference sites. If the value exceeded an outlier test, $\chi^2 = 0.01$, the site was removed from OE consideration. If failed, the test site was too dissimilar and not appropriate for prediction models. The program reported a matrix of the inverse of the pooled covariance for each predictor variable. The pooled covariance was an indicator of how much the predictor values were correlated. In addition, a parametric two sample t-test was used to distinguish differences between the OE scores for calibration-validation dataset and the OE scores for the calibration-test dataset. The outcome of the predictive models was an OE score. The OE score was used to determine degradation by establishing and upper baseline score and lower baseline score for OE values. OE scores near one were identified as non-degraded. OE scores outside the upper and lower bands were identified as degraded. I deemed a site degraded based on the 0.10 and 0.90 percentiles of calibration OE results (Van Sickle et al. 2005).
SELECTION AND ASSESSMENT OF PREDICTIVE MODEL

Performance measures were reported for the top models during model construction to in order to evaluate and select one predictor model. The best subset routine calculated the OEs for the calibration sites and the validation sites separately. The top models were plotted where each top performing model was a point for either the calibration or validation data. I used root mean squared error (RMSE) to assess magnitude of prediction errors. RMSE was selected to account for any bias in the validation sites (Van Sickle et al. 2006). Resubstitution and leave-one-out cross-validation were reported for both sets of data to assess the percent of models correctly placed within the cluster groups (g).

The performance evaluation included an unbiased null-model test. Null-model tests essentially excluded the experience of the predictive model to evaluate predictor performance and establish a baseline of precision (Van Sickle et al. 2005). Null model OE predictions showed the expected number of species (E) as fixed by summing the frequency of taxa across reference sites without cluster groups (Van Sickle et al. 2005). The null test ensured top model selection was not biased from an overall set of poor performing models. The upper baseline of precision was evaluated by assessing the standard deviation of the null model calibration-reference sites (Van Sickle et al. 2005).

I selected a final model based on multiple assessments of the calibration and validation sites and simultaneous sensitivity analyses. These assessments included: 1) percentage of correctly classified sites occurring without overfitting (Van Sickle et al. 2006), 2) low standard deviation and Wilk’s λ values relative to other models, 3) relative value of F-statistic, 4) low RMSE and below null model baseline, 5) low inverse of pooled covariance figures, 6) number of sites retained within the experience of the model (χ^2), 7) significant separation of reference site means and test site means, and 8) ease of determining model predictors for new test sites.
Results

In order to identify self-like taxonomic clusters, I pruned the dendrogram tree to four (4) classification groups (Figure 5). I ensured each group contained greater than five reference sites (group1= 11 sites, group2= 21 sites, group3= 18 sites, group4= 5 sites) (Hawkins et al. 2000). I reviewed multiple pruning routines and attempted to stratify the biological groups to a greater extent to improve overall model performance (not shown here); however, other attempts yielded poor biological groupings for modeling. Two hundred and fifty nine (259) OTUs were identified from the reference dataset. Thirteen (13) of those OTUs occurred at more than 50% of reference sites (Table 5).

The model binning system reported sixty one (61) best performing models (Appendix B) from a total of 8191 models. Model performance for the entire group of best performing models was computed (calibration sites mean=1.0, sd=0.235, validation sites mean=1.062, standard deviation=0.172).
Table 5: Commonly occurring species at reference sites, greater than fifty percent (>50%) occurrence at reference sites.

<table>
<thead>
<tr>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
<th>Species Nawqa2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pennales</td>
<td>Achnanthaceae</td>
<td>Planothidium</td>
<td>Planothidium frequentissimum (Lange-Bertalot)</td>
</tr>
<tr>
<td>Pennales</td>
<td>Naviculaceae</td>
<td>Amphora</td>
<td>Amphora pediculus (Kützing) Grunow</td>
</tr>
<tr>
<td>Pennales</td>
<td>Achnanthaceae</td>
<td>Cocconeis</td>
<td>Cocconeis placentula var. lineata (Ehrenberg) Van</td>
</tr>
<tr>
<td>Pennales</td>
<td>Achnanthaceae</td>
<td>Planothidium</td>
<td>Planothidium lanceolatum (Brébisson) Lange-Bertalot</td>
</tr>
<tr>
<td>Pennales</td>
<td>Naviculaceae</td>
<td>Rhoicosphenia</td>
<td>Rhoicosphenia abbreviata (Agardh) Lange-Bertalot</td>
</tr>
<tr>
<td>Pennales</td>
<td>Naviculaceae</td>
<td>Navicula</td>
<td>Navicula gregaria Donkin</td>
</tr>
<tr>
<td>Pennales</td>
<td>Nitzschiae</td>
<td>Nitzschia</td>
<td>Nitzschia inconspicua Grunow</td>
</tr>
<tr>
<td>Pennales</td>
<td>Achnanthaceae</td>
<td>Achnanthidium</td>
<td>Achnanthidium minutissimum (Kützing) Czarnecki</td>
</tr>
<tr>
<td>Pennales</td>
<td>Nitzschiae</td>
<td>Nitzschia</td>
<td>Nitzschia dissipata (Kützing) Grunow</td>
</tr>
<tr>
<td>Pennales</td>
<td>Naviculaceae</td>
<td>Navicula</td>
<td>Navicula cryptotenella Lange-Bertalot ex Krammer</td>
</tr>
<tr>
<td>Pennales</td>
<td>Naviculaceae</td>
<td>Reimeria</td>
<td>Reimeria uniseriata Sala Guerrero et Ferrario</td>
</tr>
<tr>
<td>Pennales</td>
<td>Diatomaceae</td>
<td>Synedra</td>
<td>Synedra ulna (Nitzsch) Ehrenberg</td>
</tr>
</tbody>
</table>

I reviewed the best performing models using statistical measures recommended by Van Sickle et al. (2006) to assess the precision and accuracy of OE predictions. An unweighted proportion of variables in the best models was reported (Table 6).

I selected the final predictor model: area, latitude and precipitation. This model had a Wilk’s $\lambda = 0.357$, F-statistic = 6.999, low inverse of pooled covariance figures, and identified 7 sites for being outside the experience of the model (based on $\chi^2 = 0.01$).

Output of the R program included an evaluation RMSE of the calibration sites (0.22) and the null model (0.23) and the percentage of sites correctly classified with resubstitution (64%) and leave-one-out cross validation (64%) (Appendix C).
Table 6: Occurrence of predictor variable in the best performing models, proportion is not weighted by model quality.

<table>
<thead>
<tr>
<th>Count</th>
<th>Variable</th>
<th>Proportion of Best Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Maximum depth</td>
<td>86.9</td>
</tr>
<tr>
<td>2</td>
<td>Latitude</td>
<td>82.0</td>
</tr>
<tr>
<td>3</td>
<td>% Sandstone</td>
<td>77.0</td>
</tr>
<tr>
<td>4</td>
<td>Day of year</td>
<td>65.6</td>
</tr>
<tr>
<td>5</td>
<td>Area of watershed</td>
<td>62.3</td>
</tr>
<tr>
<td>6</td>
<td>Slope of reach</td>
<td>62.3</td>
</tr>
<tr>
<td>7</td>
<td>Longitude</td>
<td>54.1</td>
</tr>
<tr>
<td>8</td>
<td>Maximum velocity</td>
<td>54.1</td>
</tr>
<tr>
<td>9</td>
<td>Wetted width</td>
<td>42.6</td>
</tr>
<tr>
<td>10</td>
<td>Precipitation</td>
<td>29.5</td>
</tr>
<tr>
<td>11</td>
<td>Elevation</td>
<td>14.8</td>
</tr>
<tr>
<td>12</td>
<td>Minimum depth</td>
<td>14.8</td>
</tr>
<tr>
<td>13</td>
<td>Minimum velocity</td>
<td>14.8</td>
</tr>
</tbody>
</table>

MODEL PREDICTION

Exclusion of rare taxa had been shown to improve predictive model performance (Van Sickle et al. 2007). The modeling procedures included an adjustment to vary thresholds of rare species segregation. By adjusting these thresholds, I could identify the best performing exclusion level. I ran three probability of capture ($P_c$) scenarios (0, 0.25 and 0.5) to determine an effective threshold for modeling OE (Table 7). Predictions with a $P_c$=0 yielded an OE with a wider range between the mean of the calibrations sites (0.97) and the mean of the validation sites (1.08). OE predictions made with $P_c$=0.25 showed a slightly closer to one value for the mean of the calibration sites OE (1.01) than the $P_c$=0.5 scenario mean of calibration sites OE (1.02). Both, $P_c$=0.25 and $P_c$=0.5, had the same standard deviation of the null model calibration sites (0.24). The $P_c$=0.25 scenario had a lower test site mean (0.87) than the $P_c$=0.5 scenario test site mean (0.91) and $P_c$=0 scenario test site mean (0.99). I made expected taxa prediction with predictor variables (area, latitude and precipitation) and $P_c = 0.25$. 
Table 7: RIVPACS model performance from three scenarios for probability of capture (Pc) threshold as measured by OE values. In an optimal model, OE scores should be near 1 for calibration and validation sites, in addition, the standard deviation scores for calibration sites should be noticeably below those of the null model. The table shows weak performance due to the scores considerably above 1 for validation sites and for near standard deviation scores of calibration models and null models.

<table>
<thead>
<tr>
<th>Pc</th>
<th>OE Mean</th>
<th>OE SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration Sites</td>
<td>0.97</td>
<td>0.24</td>
</tr>
<tr>
<td>Validation Sites</td>
<td>1.08</td>
<td>0.30</td>
</tr>
<tr>
<td>Null (Calibration)</td>
<td>1.0</td>
<td>0.27</td>
</tr>
<tr>
<td>Test Sites</td>
<td>0.99</td>
<td>0.26</td>
</tr>
<tr>
<td>Pc = 0.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration Sites</td>
<td>1.01</td>
<td>0.22</td>
</tr>
<tr>
<td>Validation Sites</td>
<td>1.08</td>
<td>0.16</td>
</tr>
<tr>
<td>Null (Calibration)</td>
<td>1.0</td>
<td>0.24</td>
</tr>
<tr>
<td>Test Sites</td>
<td>0.87</td>
<td>0.23</td>
</tr>
<tr>
<td>Pc = 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration Sites</td>
<td>1.02</td>
<td>0.22</td>
</tr>
<tr>
<td>Validation Sites</td>
<td>1.09</td>
<td>0.11</td>
</tr>
<tr>
<td>Null (Calibration)</td>
<td>1.0</td>
<td>0.24</td>
</tr>
<tr>
<td>Test Sites</td>
<td>0.91</td>
<td>0.25</td>
</tr>
</tbody>
</table>

The mean OE score for the Pc = 0.25 scenario between calibration and validation sites did not significantly differ (p = 0.1374). This is one indicator of good model performance because the reference sites excluded from the model build scored similar OE values as the reference sites used to train the model. A parametric two sample t-test from models with Pc = 0.25 was used to measure significance between the calibration and validation sites at the 95% confidence interval. The same test between the calibration sites and test sites showed a significant difference (p = 0.001406). This indicated favorable performance because the degraded sites were included in the test population; thus, a significant difference for OE scores from reference-type sites and test-type sites showed the model was able to distinguish between the two types of sites.

Examination of the OE scores on the boxplot graph indicated several areas of concern for model performance (Figure 6). The plot showed overlapping OE scores
between calibration and test sites. This indicated subpar performance and lack of precision from the model. The large range of reference OE scores shows some ambiguity when the model attempts to identify reference quality sites. The range of OE values for the calibrations sites was 0.35 to 1.46. Optimal models would score calibration sites near one; these scores are considerably different from one. The same was true for the validation sites. The OE range was 0.74 to 1.37. The test sites OE range appeared to score appropriately given the degraded and non-reference status of these sites. The scores ranged was 0.22 to 1.35. Reviewed separately, the calibration sites showed a lack of precision. Comparison at calibration sites between observed taxa counts and expected taxa counts showed considerable scatter (Figure 7). Less than optimal performance of calibration sites was observed because the model sometimes grossly over-predicted or under-predicted taxa counts. This lack of predictive capabilities directly affected overall
model performance because expected taxa counts for all sites were derived from the calibration sites.

Common practices for RIVPACS assessment showed the creation of boundaries to identify upper and lower OE values to evaluate degradation (Bailey et al. 2004). These boundary levels would identify site status as degraded or non-degraded. For optimal model performance, these modeled test sites should mirror the known degraded sites. The

![Figure 7: Evaluation of reference site performance measured with O and E scores. Optimal performance should demonstrate plots close to a 1:1 ratio, where a well-trained model could predict observed taxa at reference sites. This scatter plot showed points distributed away from the 1:1 line (blue dashed line).](image)

OE model identified degraded sites falling outside the boundaries whereby all other test sites (within the boundaries) should be considered non-degraded. I established levels to use as indicators of degraded sites by taking the 10% and 90% percentiles of the calibration sites (Van Sickle et al. 2005). The below-one impairment OE score was 0.77 and the above-one impairment OE score was 1.21. OE scores falling between 0.77 and 1.21 were considered non-graded. All OE scores were reported for all sites (Appendix E). Sixty six (66) sites were rated as possibly degraded (44 sites below 0.77 OE threshold and 21 sites above the 1.21 OE threshold). Seventeen (17) of these sites were considered
reference sites (6 sites below the 0.77 OE threshold and 11 sites above the 0.77 OE threshold). The large number of reference sites identified as possibly degraded indicated a lack of model accuracy. The remaining 49 test sites were classified as possibly degraded (39 sites below the 0.77 OE threshold and 10 from the above 1.21 OE threshold).

Sites scoring above the 1.21 threshold indicated test sites contained more reference quality taxa than the reference sites. An OE score greater than the upper threshold band (OE >1) indicated more taxa species were found than expected. These higher scores were anticipated to occur because an OE score equal to one (OE =1) represented the center of the reference distribution. However, high OE scores were not a definitive measure of biological degradation. The OE >1 scores indicated the test sites were more biologically diverse than the reference sites. High score OE sites should not automatically be classified as biologically degraded but should be identified nonetheless. These scores represent sites with possible high biological diversity, moderate organic enrichment or on-going irrigation discharge into an intermittent-type stream (Bailey et al. 2004). Furthermore, considerable issues were raised with modeling effectiveness because of numerous high scoring OE sites (n=21). The above 1.21 OE score may indicate an issue of over-fitting the expected model.

This RIVPACS study utilized diatoms because of the strong connection of diatoms to nutrient influence (Stevenson et al. 2008). In theory, optimal model performance would yield a strong relationship between OE scores and nutrient concentrations of water samples. Upon review, the model demonstrated a lack of correlation between OE scores and nutrient levels. A scatter plot evaluation of OE scores and amounts total phosphorus (TP) and total nitrogen (TN) revealed no discernable pattern (Figure 8). High nutrient values were somewhat obscured and inconsistent for OE scores based on concentration levels of total phosphorus (TP) or total nitrogen (TN). This lack of diatom and nutrient relationship indicated the diatom RIVPACS-type predicted model performed less than expected.
Figure 8: Nutrient comparison to OE scores, total phosphorus (TP) total nitrogen (TN) compared to test sites OE scores. Dashed lines represent 10% (lower limit) and 90% (upper limit) percentiles of the reference data. Charts include all reference and test sites. Sites located between the lower and upper OE band were identified as unimpaired. Sites above or below the OE bands were classified as degraded.

I compared OE scores of the reference-calibration sites to the land-use categories used during reference selection process (Figure 9). Reference selection criteria excluded sites from the reference pool based on land use and land cover categories. I classified sites as agriculture (n=33) or urban (n=39) if more than 5% of the watershed area above the sample locations at various distances from the sample site (1k, 5k or whole watershed) were deemed either agriculture or urban respectively. I classified light urban (n=44) and an unnatural index (n=68) if more than 15% of the watershed above the sample locations at various distances from the sample site (1k, 5k or whole watershed) were determined light urban or a combination of any other category (agriculture, urban and light urban). Based on the bounds determined by the 10% and 90% quartiles (boundary OE range 0.77 to 1.21), sites in the agriculture category generally fell into the degraded range (agriculture OE mean= 0.73). This indicates that diatom community composition was somewhat sensitive to agricultural influence. However, the data were not conclusive for the agricultural sites. This indicates the model was somewhat sensitive to agricultural influence. However, the data were not conclusive for the agriculture sites. Examination of the agriculture dataset revealed overlapping values with the reference site data in the upper ranges, differences in median value and less overlap in the lower values. Urban, light urban and the index categories generally scored within the bounds of
Figure 9: Results of land use comparison to OE scores indicated a slight trend for lower OE scores at agriculture sites and no trend at urban, light urban and index sites. Dashed lines represent OE boundaries for identifying degradation. Sites with an OE score between the lower limit (blue line) and the upper limit (red line) were rated as non-degraded. Boxplot shows 25th and 75th quartile, median, Tukey whiskers 1.5 interquartile range [IQR], and outliers points for OE values.

unimpaired sites (OE means= 0.88, 0.95 and 0.87 respectively). Thus, either impacted landscapes (urban, agricultural, etc.) did not differ in diatom community composition from reference sites, or the model generally failed to detect these differences (with the partial exception of agricultural impacts).

**Discussion**

The goal of this project was to evaluate the effectiveness of a diatom RIVPACS-type predictive model by demonstrating biological degradation at impaired water quality sites. Several studies have shown varying success employing a diatom RIVPACS model (Chessman et al. 1999; Mazor et al. 2006; Cao et al. 2007). This study attempted to advance this research of diatom-based assessments and aid in decision making for California Central Coast resource managers. I utilized the RIVPACS methodology by identifying reference sites, deriving expected taxa counts and comparing observed taxa counts to reference-site derived expected taxa counts to determine biological integrity
I used the OE metric to score sites and to evaluate the accuracy and precision of the predictive models. The OE scores for sites with known nutrient stress or other water quality impairment were reviewed. The results of the RIVPACS model and OE scores on Central Coast streams showed limited success. My research illustrated the challenges of employing a diatom RIVPACS model and performing a reference condition approach for stream assessments in a region with known urban and agricultural development.

Life history and biological characteristic of diatom assemblages, such as rapid growth, make them ideal indicator taxa (Pan et al. 1996; Stevenson and Pan 1999). However, researchers have demonstrated the need to account for diatom assemblage variability (Stevenson et al. 2008). The results from this study showed the diatom-based models suffered because the same ideal indicator characteristics of diatom assemblages hindered the effectiveness for predictive modeling. For example, rapid growth of diatom assemblages affected the spatial and temporal variability, which lowered the precision and reduced the accuracy of the model. My results showed the model had less taxa composition overlap at the reference sites. This indicated assemblages in reference clusters were too widely distributed which lowered model effectiveness. Consequently, there was limited success distinguishing differences between reference sites and degraded sites. Other RIVPACS studies utilizing diatoms noted the similar variability issues with diatom assemblages (Chessman et al. 1999; Cao et al. 2007).

The results highlighted the need for resource managers on the Central Coast to identify high quality reference sites for use in studies employing a reference condition approach. High quality reference sites provide useful information for biological assessments (Hughes 1995; Stoddard et al. 2006). My findings showed the reference sites on the Central Coast were of lower quality. This diminished the effectiveness of the RIVPACS models by confounding the results of the OE metric. For example, the OE scores for many reference sites were indistinguishable from impaired sites. This issue was explained by the wide distribution of OE scores for reference sites and indicated a potential problem utilizing semi-degraded reference sites. Recommendations to use the relative scaled LDC sites for predictive model development was appropriate in this region of historic agricultural practices and development (Bailey et al. 2004; Stoddard et al.
The sampling design strategy attempted to identify and sample as many reference sites as possible; however, large quantities of high quality reference sites posed a difficult challenge to locate. I selectively expanded the criteria I used to identify reference sites in order to increase the total number of sites; however, by increasing the number of lower quality reference sites I potentially reduced the overall ability of the model to assess degradation.

Despite the weaknesses, the predictive models achieved intermittent success in characterizing degraded biological conditions at known low water-quality sites. I identified a possible trend with agricultural land use. Previous studies have identified agricultural land use in the Central Coast region has the potential to influence stream water-quality (Los Huertos et al. 2001; Anderson et al. 2003; Dowd et al. 2008). I utilized the OE scores of sites identified during the model build as non-reference due to anthropogenic influence. I analyzed the OE scores for sites containing agricultural land used and compared these sites to the reference site OE scores (Figure 9). From the land-use comparison, the OE data showed a trend towards a nutrient non-point source signal (Figure 9). The agriculture land-use OE scores indicated possible degradation but was not conclusive. The agriculture OE median scored below the lower degradation band line and the majority of OE scores were below the reference site OE scores. The land-use classification had limited capabilities due to unrefined agricultural categories (Appendix A); however, this trend mirrored known algal response to nutrient input (McCormick and Stevenson 1998; Pan et al. 2006). These observations may represent a link on the Central Coast between diatom indicator taxa and agriculture practices. Future analyses with greater precision of diatom-based models or diatom indices may validate this trend.

In order to test the model against known causes of degradation, I compared OE scores of the test sites to nutrient levels, TN and TP (Figure 8), and test sites OE scores identified in land-uses classification to reference sites (Figure 9). Previous studies have identified a relationship to nutrient levels and diatoms (Pan et al. 1996; McCormick and Stevenson 1998; Leland et al. 2001; Munn et al. 2002; Ponader et al. 2007; Lavoie et al. 2008); however, no discernable pattern was readily apparent with OE scores and nutrient values. This may have indicated the diatom RIVPACS model was unable to identify a nutrient stress signal. The weakness in these results may oblige resource managers to
utilize other assessment tools for nutrient stress evaluation. The results of TN and TP comparison to OE were consistent with the assumption one-time water samples were poor characterization stream status; however, given the ambiguity of these RIVPACS results this will require future testing.

Generally, the predictive portion of the diatom model performed better as the number of species employed for modeling decreased. Although the effectiveness of excluding species is still debated by researchers, my results were consistent with other RIVPACS studies (Van Sickle 2007). I observed $P_c$ thresholds greater than zero improved model accuracy, precision and ability to identify degraded sites (Table 5). Removal of rare species corresponded to the species lists (OTUs) generated from the reference sites, which indicated 259 OTUs were identified yet only 13 OTUs occurred at more than 50% of the reference sites. The lack of more OTU overlap within reference sites may indicate a problem for diatom RIVPACS-type predictive models on the Central Coast. Diatom distribution and life-history characteristics include rapid seasonal growth cycles and spatially dependent succession and replacement (Pan et al. 1996; Leland et al. 2006; Stevenson 2008). Spatial patchiness associated with this temporal growth and replacement scattered the species lists and counts of diatom assemblages at reference sites. The RIVPACS model relied on distinctive diatom assemblages at reference sites in order to measure a compositional change at stressed sites. Paradoxically, this distinction reduced the success of the predictive models because there was considerable assemblage disparity among expected taxa. Additionally, diatoms have shown strong relationships with multiple environmental conditions (Stevenson et al. 2008); in this study, excessive nutrients may not be distinguishable from other conditions affecting diatom assemblage growth. I recommend an increase in reference sites and repeated sampling of the same sites to reduce the random occurrence of rare species and account for the temporal and spatial variability of diatom assemblages.

The diatom prediction models were based on data from a range of wadeable-stream habitat. The rapid collection technique employed by the field teams sampled diatoms from multiple stream types and habitat features such as riffles and pools. It is unknown if this type of combined habitat sampling may have reduced the ability of the predictive models to detect impairment. Parsons and Norris (1996) demonstrated
isolating habitat to a single habitat type for macroinvertebrate RIVPACS-type predictions improved the detection of impairment. They concluded the inclusion of more than one habitat type may confound rather than help with the assessment of biological impairment.

The confounding effects of temporal variability and patchiness associated with reference assemblages, lower quality reference sites and multiple habitat sampling may help explain the problems with training diatom predictive models on the Central Coast. The problems identified during model construction included: a low number of biological groups (n=4) established during the dendrogram-build phase (Figure 4), the wide range of reference OE values 0.35 to 1.46, and the SD value of the OE reference sites (SD = 0.22). In this study, the reduced number of biological classification groups, quality of sites and the number of sites may have limited the prediction success of the model. Future modeling efforts may attempt to stratify the biological groups to greater degree possibly ignoring the five reference sites per biological group rule. This may allow the predictive models to account for differences in assemblages not observed in this study.

Further challenges of this model, included the lack of a lower baseline for precision and the process for final model selection. The lower baseline for precision could be established by reviewing the error from replicate samples of individual sites (Van Sickle et al. 2005). No replicate samples were available for evaluation at the time of modeling; however future models derived from these data may be updated to include replicate samples. The process for final model selection appeared somewhat arbitrary. Van Sickle et al. (2006) provided many statistical tools for predictor model evaluation in addition to those reviewed in this methods section and reported in the results section above. However, they recognized selecting one models from the set of all the \(2^p-1\) models (in this case, 8191 models) though practical may be somewhat ad hoc. They did not recommend any one particular tool as a definitive process for model selection; however, they suggested developing comparable weighting tools such as Akaike information criterion (AIC) for future model selection.
Conclusion

I concluded the RIVPACS-type predictive model did not perform well and degraded site identification was not consistent. However, the uncertainty presented in this research was not wholly conclusive. I observed several trends including degraded OE scores for sites with agricultural land use. I recommend further examination of diatom assessments to determine the effectiveness of diatom models. Overall, the evaluation of the predictive models indicated the following:

- The advantages of diatoms as an indicator species that respond rapidly to changing conditions also proved to be paradoxically a detriment to model performance. High rates of succession and replacement, as well as, spatial patchiness of assemblages generally reduced model effectiveness to predict expected taxa.

- Identification of reference sites on the Central Coast was problematic. The reference sites were critical during the modeling process to establish the expected taxa. A lack of high quality reference sites led to poor model precision thus, OE scores did not clearly identify degraded sites.

- A low number of biological clusters were identified from the dendrogram of reference site assemblages. This affected the ability of the discriminant analysis to strongly associate predictor variables without over-fitting the model. As a result, model precision and accuracy suffered. This negatively influenced the strength of OE scores to represent biological integrity.
REFERENCES


R Development Core Team. 2009. R: A language and environment for


Wright JF, Moss D, Armitage PD, Furse MT. 1984. A preliminary classification of running water sites in Great Britain based on macro-invertebrate species and the prediction of community type using environmental data. Freshwater Biology, 14:221-256.

APPENDIX A

REFERENCE SELECTION PROCESS

Reference selection process, removed non-reference sites based on field notes and GIS analyses. 75 reference sites were identified from a pool of 190 total sites.

<table>
<thead>
<tr>
<th>Stresor/Confounding Effect</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual/ Seasonal Independence</td>
<td>One reference site per multiple site visits</td>
<td>One sample per site</td>
</tr>
<tr>
<td>Land Use (GIS Analysis)</td>
<td>Unnatural Index (LU + Urb + Ag)</td>
<td>&lt;15%</td>
</tr>
<tr>
<td>Evaluate each class</td>
<td>Light Urban (LU)</td>
<td>&lt;15%</td>
</tr>
<tr>
<td>at 1km, 5km and watershed</td>
<td>Urban (Urb)</td>
<td>&lt;10%</td>
</tr>
<tr>
<td></td>
<td>Agriculture (Ag)</td>
<td>&lt;10%</td>
</tr>
<tr>
<td>Physical Habitat / Management Activities</td>
<td>Erosional Deposition</td>
<td>Mass wasting or mass wastings</td>
</tr>
<tr>
<td></td>
<td>Stream Incisement</td>
<td>Active downcutting, new floodplain</td>
</tr>
<tr>
<td></td>
<td>Observed Livestock Herbivory</td>
<td>&lt;25%</td>
</tr>
<tr>
<td></td>
<td>Water Clarity</td>
<td>Very turbid</td>
</tr>
<tr>
<td></td>
<td>Evidence of fire within past 5yrs</td>
<td>Yes/no</td>
</tr>
<tr>
<td></td>
<td>Primary Landuse - Crops</td>
<td>Immediate area</td>
</tr>
<tr>
<td></td>
<td>Primary Landuse - Herbivory</td>
<td>Immediate area</td>
</tr>
<tr>
<td></td>
<td>Primary Landuse - Stream Diversions</td>
<td>Immediate area</td>
</tr>
<tr>
<td></td>
<td>Primary Landuse - Mining</td>
<td>Immediate area</td>
</tr>
<tr>
<td></td>
<td>Primary Landuse - Logging</td>
<td>Immediate area</td>
</tr>
<tr>
<td>Spatial Independence</td>
<td>Remove sites within same catchment</td>
<td>&lt; 3km</td>
</tr>
<tr>
<td></td>
<td>Remove sites within same catchment</td>
<td>&lt; 3 tributaries</td>
</tr>
<tr>
<td>Best Professional Judgment</td>
<td>Remove sites with known stressors</td>
<td>BPJ</td>
</tr>
</tbody>
</table>

Reclassification categories derived from National Land Cover Dataset (NLCD, 2001).

<table>
<thead>
<tr>
<th>NLCD(2001)</th>
<th>Category</th>
<th>Reclassify - Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Open Water</td>
<td>Water</td>
</tr>
<tr>
<td>21</td>
<td>Developed Open Space</td>
<td>LightUrb</td>
</tr>
<tr>
<td>22</td>
<td>Developed Low Intensity</td>
<td>Urb</td>
</tr>
<tr>
<td>23</td>
<td>Developed Med Intensity</td>
<td>Urb</td>
</tr>
<tr>
<td>24</td>
<td>Developed High Intensity</td>
<td>Urb</td>
</tr>
<tr>
<td>31</td>
<td>Barren Land</td>
<td>Barren</td>
</tr>
<tr>
<td>41</td>
<td>Deciduous Forest</td>
<td>ForShrub</td>
</tr>
<tr>
<td>42</td>
<td>Evergreen Forest</td>
<td>ForShrub</td>
</tr>
<tr>
<td>43</td>
<td>Mixed Forest</td>
<td>ForShrub</td>
</tr>
<tr>
<td>52</td>
<td>Shrub/Scrub</td>
<td>ForShrub</td>
</tr>
<tr>
<td>71</td>
<td>Grassland</td>
<td>Grass</td>
</tr>
<tr>
<td>81</td>
<td>Pasture / Hay</td>
<td>Ag</td>
</tr>
<tr>
<td>82</td>
<td>Cultivated Crops</td>
<td>Ag</td>
</tr>
<tr>
<td>90</td>
<td>Woody Wetlands</td>
<td>Wetland</td>
</tr>
<tr>
<td>95</td>
<td>Emergent Wetlands</td>
<td>Wetland</td>
</tr>
</tbody>
</table>
### APPENDIX B

#### BEST MODELS

<table>
<thead>
<tr>
<th>Model Order</th>
<th>E-Statistic</th>
<th>Wilks Lambda</th>
<th>Resubstitution Learn-on-test classification</th>
<th>Max OE Calibration</th>
<th>SOE Calibration</th>
<th>RMLE OE Calibration</th>
<th>Model Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>0.98</td>
<td>0.25</td>
<td>0.27</td>
<td>0.27</td>
<td>longitude, max depth, max velocity, sedimentary sandstone, precipitation</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.86</td>
<td>0.97</td>
<td>0.21</td>
<td>0.26</td>
<td>0.26</td>
<td>latitude, precipitation</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.05</td>
<td>0.95</td>
<td>0.19</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, max depth, sedimentary sandstone, slope</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.95</td>
<td>0.96</td>
<td>0.19</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, max depth, max velocity, sedimentary sandstone</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.06</td>
<td>0.95</td>
<td>0.19</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, max depth, max velocity, sedimentary sandstone, slope</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.86</td>
<td>0.97</td>
<td>0.21</td>
<td>0.26</td>
<td>0.26</td>
<td>latitude, precipitation</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2.05</td>
<td>0.95</td>
<td>0.19</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, max depth, sedimentary sandstone, slope</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.95</td>
<td>0.96</td>
<td>0.19</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, max depth, max velocity, sedimentary sandstone</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2.09</td>
<td>0.98</td>
<td>0.16</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, precipitation</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.96</td>
<td>0.97</td>
<td>0.19</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, max depth, sedimentary sandstone, slope</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2.08</td>
<td>0.97</td>
<td>0.16</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, precipitation</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1.98</td>
<td>0.97</td>
<td>0.19</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, max depth, sedimentary sandstone, slope</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2.04</td>
<td>0.98</td>
<td>0.16</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, precipitation</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1.96</td>
<td>0.97</td>
<td>0.19</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, max depth, sedimentary sandstone, slope</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2.09</td>
<td>0.98</td>
<td>0.16</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, precipitation</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1.97</td>
<td>0.97</td>
<td>0.19</td>
<td>0.21</td>
<td>0.21</td>
<td>latitude, max depth, sedimentary sandstone, slope</td>
<td></td>
</tr>
</tbody>
</table>

Note: The table continues with additional models that are not listed here for brevity.
APPENDIX C

BEST MODEL EVALUATION

Figure 10: Root mean squared errors (RMSE) from sixty one (61) top performing model OEs. Symbol 'C' denotes calibration site, 'V' denotes validation site. Model order indicates the number of variables per model. Solid and dashed lines represent null models.

Figure 11: Percentage of sites correctly associated with the taxa dendogram groups (Figure 4) using resubstitution ‘R’ and leave-one-out cross validation ‘C’. Dashed and solid lines connect mean for each model order.
Figure 12: \( P_c = 0.5 \) boxplot shows 25th and 75th quartile, median, Tukey whiskers 1.5 interquartile range [IQR], and outliers points for OE values from the validation, calibration and test sites.

Figure 13: \( P_c = 0 \) boxplot shows 25th and 75th quartile, median, Tukey whiskers 1.5 interquartile range [IQR], and outliers points for OE values from the validation, calibration and test sites.
## APPENDIX E

### OE PREDICTIONS

<table>
<thead>
<tr>
<th>Date</th>
<th>Sitecode</th>
<th>OE</th>
<th>Site Name</th>
<th>Calibration 1</th>
<th>Validation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/3/2007</td>
<td>305MUILR</td>
<td>0.223</td>
<td>Millers Canal at Frazier Lake Road</td>
<td>0.922</td>
<td>0.0564</td>
</tr>
<tr>
<td>8/16/2007</td>
<td>304BUCFR</td>
<td>0.352</td>
<td>Little Butano Creek at Olmo Fire Road pullout</td>
<td>0.3175</td>
<td>0.102404</td>
</tr>
<tr>
<td>9/27/2007</td>
<td>309SALDAV</td>
<td>0.361</td>
<td>Salinas River at Davis Road</td>
<td>30.15</td>
<td>0.0354</td>
</tr>
<tr>
<td>10/2/2007</td>
<td>305AGDTRA</td>
<td>0.423</td>
<td>Trafton Road Ag ditch on Trafton Road</td>
<td>32.2</td>
<td>0.0765</td>
</tr>
<tr>
<td>7/23/2008</td>
<td>309SANDOR</td>
<td>0.439</td>
<td>San Antonio River @ Fort Hunter Liggett</td>
<td>0.02933</td>
<td>0.0383</td>
</tr>
<tr>
<td>5/20/2008</td>
<td>312CUV166</td>
<td>0.469</td>
<td>Cuyamn River @ Hwy 166</td>
<td>0.2355</td>
<td>0.070477</td>
</tr>
<tr>
<td>7/14/2007</td>
<td>309SALDAV</td>
<td>0.469</td>
<td>Salinas River at Davis Road</td>
<td>47.5</td>
<td>0.08285</td>
</tr>
<tr>
<td>7/31/2008</td>
<td>305CLECCA</td>
<td>0.506</td>
<td>Clear Creek in BLM Clear Creek Mngmnt Area</td>
<td>1</td>
<td>0.175</td>
</tr>
<tr>
<td>7/2/2007</td>
<td>309SALH58</td>
<td>0.507</td>
<td>Salinas River at Hwy 58</td>
<td>0.656</td>
<td>0.03785</td>
</tr>
<tr>
<td>7/16/2008</td>
<td>305AGDTRA</td>
<td>0.518</td>
<td>Trafton Road Ag ditch on Trafton Road</td>
<td>50.54875</td>
<td>0.01685</td>
</tr>
<tr>
<td>9/24/2007</td>
<td>309SANBWR</td>
<td>0.564</td>
<td>San Lorenzo Creek @ Bitter Water Road</td>
<td>0.978</td>
<td>0.0598</td>
</tr>
<tr>
<td>5/7/2008</td>
<td>315HONMIG</td>
<td>0.566</td>
<td>Honda Creek @ San Miguelito Road</td>
<td>0.825</td>
<td>0.0562</td>
</tr>
<tr>
<td>7/7/2008</td>
<td>304COYOC</td>
<td>0.573</td>
<td>Coyote Creek @ Cochrane Rd.</td>
<td>1.17755</td>
<td>0.00526</td>
</tr>
<tr>
<td>3/19/2008</td>
<td>309CHLCWF</td>
<td>0.576</td>
<td>Chalone West Fork</td>
<td>0</td>
<td>-88</td>
</tr>
<tr>
<td>4/22/2008</td>
<td>309SAVR</td>
<td>0.577</td>
<td>Salinas River @ Indian Valley</td>
<td>1</td>
<td>0.175</td>
</tr>
<tr>
<td>7/3/2007</td>
<td>309SARSAN</td>
<td>0.582</td>
<td>San Antonio River at San Antonio Lake Road</td>
<td>0.7385</td>
<td>0.05</td>
</tr>
<tr>
<td>9/18/2007</td>
<td>305CORSCU</td>
<td>0.589</td>
<td>Corralitos Creek at Scrunch Lane</td>
<td>5.95</td>
<td>0.0825</td>
</tr>
<tr>
<td>5/21/2008</td>
<td>314SANGRE</td>
<td>0.613</td>
<td>Santa Ynez River @ Greco Crossing</td>
<td>0.6915</td>
<td>0.0873</td>
</tr>
<tr>
<td>7/14/2007</td>
<td>304LLOMLM</td>
<td>0.616</td>
<td>Lompico Creek at Lompico Creek Road</td>
<td>0.80224</td>
<td>0.129</td>
</tr>
<tr>
<td>8/18/2008</td>
<td>305SALPIL</td>
<td>0.617</td>
<td>Salinas River at Pillitas Road</td>
<td>0</td>
<td>0.09675</td>
</tr>
<tr>
<td>6/3/2008</td>
<td>309SALAV</td>
<td>0.617</td>
<td>Salinas River @ Indian Valley</td>
<td>3.1</td>
<td>0.61192</td>
</tr>
<tr>
<td>9/25/2007</td>
<td>305SICANZ</td>
<td>0.641</td>
<td>San Juan Creek at Anzar Road</td>
<td>39.6</td>
<td>0.736</td>
</tr>
<tr>
<td>5/22/2008</td>
<td>312SANSUE</td>
<td>0.649</td>
<td>Santa Maria River @ Suey Crossing</td>
<td>1.82</td>
<td>0.421558</td>
</tr>
<tr>
<td>7/1/2008</td>
<td>309SALDAV</td>
<td>0.650</td>
<td>Salinas River at Davis Road</td>
<td>1</td>
<td>-88</td>
</tr>
<tr>
<td>5/8/2008</td>
<td>314SALIAL</td>
<td>0.651</td>
<td>Salsipueules Creek @ Jalama Road</td>
<td>0.279</td>
<td>0.155</td>
</tr>
<tr>
<td>6/19/2008</td>
<td>305PAJROG</td>
<td>0.659</td>
<td>Pajaro River at Rogge Lane</td>
<td>17.65</td>
<td>0.1975</td>
</tr>
<tr>
<td>9/4/2008</td>
<td>305SBRPR2</td>
<td>0.663</td>
<td>San Benito @ Cienaga Road</td>
<td>0</td>
<td>0.294375</td>
</tr>
<tr>
<td>9/15/2008</td>
<td>309SALPIL</td>
<td>0.689</td>
<td>Salinas River at Pillitas Road</td>
<td>0</td>
<td>0.4635</td>
</tr>
<tr>
<td>7/5/2007</td>
<td>305LLAGLE</td>
<td>0.692</td>
<td>Llagas Creek at pullout on Oak Glen Ave</td>
<td>0</td>
<td>0.8305</td>
</tr>
<tr>
<td>9/15/2007</td>
<td>309SALPOZ</td>
<td>0.701</td>
<td>Salinas River at High Mtn Rd</td>
<td>0</td>
<td>0.1595</td>
</tr>
<tr>
<td>4/16/2008</td>
<td>312BREFR</td>
<td>0.722</td>
<td>La Brea Creek at Rancho Sisquoc</td>
<td>0</td>
<td>0.295</td>
</tr>
<tr>
<td>10/4/2007</td>
<td>309NATCAS</td>
<td>0.733</td>
<td>Natividad Creek at Las Casitas Road</td>
<td>2</td>
<td>0.863</td>
</tr>
<tr>
<td>7/31/2008</td>
<td>304BRAOCE</td>
<td>0.736</td>
<td>Branciforte Creek @ Ocean Street</td>
<td>0</td>
<td>0.0926</td>
</tr>
<tr>
<td>8/9/2007</td>
<td>305LGMKAI</td>
<td>0.739</td>
<td>Los Gatos Creek at Main Street</td>
<td>0</td>
<td>0.3375</td>
</tr>
<tr>
<td>9/12/2007</td>
<td>305LURHIG</td>
<td>0.741</td>
<td>San Lorenzo River at Highland Park</td>
<td>1</td>
<td>1.155</td>
</tr>
<tr>
<td>8/21/2007</td>
<td>305SBBH1W</td>
<td>0.742</td>
<td>Soberanes Creek at Hwy 1</td>
<td>2</td>
<td>0.2945</td>
</tr>
<tr>
<td>9/25/2007</td>
<td>305SALL29</td>
<td>0.745</td>
<td>Salsipueules Creek at Hwy 129</td>
<td>3.83</td>
<td>0.241</td>
</tr>
<tr>
<td>7/17/2008</td>
<td>304MILSWA</td>
<td>0.753</td>
<td>Mill Creek @ Swanton Rd</td>
<td>1</td>
<td>0.69433</td>
</tr>
<tr>
<td>6/9/2008</td>
<td>309GABOLD</td>
<td>0.755</td>
<td>Gabial Creek at Old Stage Road</td>
<td>0</td>
<td>0.464</td>
</tr>
<tr>
<td>4/16/2008</td>
<td>312SISRNS</td>
<td>0.758</td>
<td>Sisquoc River @ Rancho Sisquoc</td>
<td>0</td>
<td>3.51</td>
</tr>
<tr>
<td>7/7/2008</td>
<td>309SALPIL</td>
<td>0.762</td>
<td>Salinas River at Pillitas Road</td>
<td>0</td>
<td>0.6335</td>
</tr>
<tr>
<td>5/19/2008</td>
<td>305WILWIL</td>
<td>0.767</td>
<td>Williams Canyon Creek @ Williams Canyon Trail</td>
<td>1</td>
<td>9.98</td>
</tr>
<tr>
<td>8/2/2007</td>
<td>304APTMAR</td>
<td>0.775</td>
<td>Aptos Creek at Margaret Bridge</td>
<td>1</td>
<td>0.2135</td>
</tr>
<tr>
<td>6/26/2007</td>
<td>309SALPOZ</td>
<td>0.775</td>
<td>Salinas River at High Mtn Rd</td>
<td>0</td>
<td>0.257</td>
</tr>
<tr>
<td>6/19/2008</td>
<td>304KINROC</td>
<td>0.783</td>
<td>Kings Creek @ Castle Rock Falls Trail</td>
<td>1</td>
<td>0.2885</td>
</tr>
</tbody>
</table>
## APPENDIX E (CONT.)

### OE PREDICTIONS

<table>
<thead>
<tr>
<th>sample date</th>
<th>sitecode</th>
<th>OE</th>
<th>sitename</th>
<th>calibration=1</th>
<th>validation=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/13/2008</td>
<td>305ALACP</td>
<td>0.784</td>
<td>Paloma Creek @ Paloma Creek Park (City Park)</td>
<td>0.801</td>
<td>0.839</td>
</tr>
<tr>
<td>7/30/2007</td>
<td>200SEA8EA</td>
<td>0.798</td>
<td>Santa Ynez River @ Refugio Road</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>6/11/2008</td>
<td>307ROBCYN</td>
<td>0.793</td>
<td>Bear Gulch Near Visitor Center</td>
<td>0.798</td>
<td>0.805</td>
</tr>
<tr>
<td>7/24/2008</td>
<td>308RARH2A</td>
<td>0.800</td>
<td>Salinas River at Bradley</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>6/26/2008</td>
<td>304PETPR5</td>
<td>0.800</td>
<td>Peters Creek @ Portola State Park</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>8/13/2008</td>
<td>309BCGVN5</td>
<td>0.800</td>
<td>Bear Gulch Creek at Pinnacles established MC1</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>6/11/2008</td>
<td>307ROBCYN</td>
<td>0.800</td>
<td>Bear Gulch Creek @ Bear Gulch Rd</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>6/25/2008</td>
<td>304ALPAP</td>
<td>0.800</td>
<td>Alpine Creek @ Alpine Road</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>9/14/2007</td>
<td>3095ALSR6</td>
<td>0.800</td>
<td>Salinas River at Hwy 198</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>9/18/2007</td>
<td>305Duanam</td>
<td>0.800</td>
<td>Ramsey Creek at Ramsey Road</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>7/11/2007</td>
<td>305BROWT5</td>
<td>0.800</td>
<td>Browns Creek at Watsonville uptake</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>7/24/2008</td>
<td>308BR5H1W</td>
<td>0.800</td>
<td>Lower Arroyo Seco</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>6/11/2008</td>
<td>3085AENY1</td>
<td>0.800</td>
<td>San Carporororo Creek @ HWY 1</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>5/28/2008</td>
<td>304RABMR</td>
<td>0.800</td>
<td>Branchline Creek @ Market St.</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>8/14/2008</td>
<td>308PESWRA</td>
<td>0.800</td>
<td>Pescadero Creek at Stage Road</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>9/12/2006</td>
<td>309GABACSC</td>
<td>0.800</td>
<td>Gabilan Creek at Constitution Soccer Complex</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>3/18/2008</td>
<td>309CHAPVF</td>
<td>0.800</td>
<td>Boulder Creek at Rte. 9</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>6/12/2008</td>
<td>307POTCH3</td>
<td>0.800</td>
<td>Potrero Creek @ Chamisal Rd</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>5/7/2008</td>
<td>307JLACAM</td>
<td>0.800</td>
<td>Jalama Creek @ Jalama Beach Campground</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>7/16/2007</td>
<td>304VARLAL</td>
<td>0.800</td>
<td>Valencia Creek @ Valencia Road</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>7/5/2008</td>
<td>200LGCMN</td>
<td>0.800</td>
<td>Los Lagos Creek at Main Street</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>8/14/2008</td>
<td>304PESESSA</td>
<td>0.800</td>
<td>Pescadero Creek at Stage Road</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>7/26/2008</td>
<td>309GBAVPC</td>
<td>0.800</td>
<td>Bear Gulch Creek next to Pinnacles visitor center</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>9/26/2006</td>
<td>309GABACSC</td>
<td>0.800</td>
<td>Gabilin Creek at Constitution Soccer Complex</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>3/18/2008</td>
<td>309CHAPVF</td>
<td>0.800</td>
<td>Chaline near Peaks View parking</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>6/11/2008</td>
<td>307LASSLP</td>
<td>0.800</td>
<td>Las Garzas Crk below Santa Lucia Lake Preserve</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>9/4/2007</td>
<td>305SBPR1</td>
<td>0.800</td>
<td>San Berito at Paicines Ranch</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>9/6/2007</td>
<td>308MILHW1</td>
<td>0.800</td>
<td>Mill Creek at Hwy 1</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>8/20/2008</td>
<td>301LSCUIE</td>
<td>0.800</td>
<td>San Luis Obispo Creek at Cuesta Park</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>6/16/2008</td>
<td>200MIDHEN</td>
<td>0.800</td>
<td>Middle Fork Coyote Creek</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>6/4/2008</td>
<td>301TORCTY</td>
<td>0.800</td>
<td>Toro Creek off Toro Creek Road</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>4/21/2008</td>
<td>305BNRCCS</td>
<td>0.800</td>
<td>Baldy Ryan Creek @ Casa Loma Rd</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>4/17/2008</td>
<td>3095ALATS</td>
<td>0.800</td>
<td>Salinas River @ Atascadero</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>9/19/2007</td>
<td>309NSHANS</td>
<td>0.800</td>
<td>Upper Arroyo Seco River</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>9/11/2007</td>
<td>305CRAWAT</td>
<td>0.800</td>
<td>Corralitos Creek, WB, Watsonville City locked site</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>7/10/2007</td>
<td>308IOSCRR</td>
<td>0.800</td>
<td>Joshua Creek at Ken Eukland's house</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>8/6/2008</td>
<td>308BR5H1W</td>
<td>0.800</td>
<td>Big Sur River @ Hwy1</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>7/10/2008</td>
<td>312DAVARV</td>
<td>0.800</td>
<td>Davy Brown Creek @ Davy Brown Camp</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>7/8/2008</td>
<td>305UVASWA</td>
<td>0.800</td>
<td>Uvas Creek at Uvas Canyon County Park</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>4/9/2008</td>
<td>309ARSG16</td>
<td>0.800</td>
<td>Arroyo Seco @ G16 Bridge</td>
<td>0.800</td>
<td>0.815</td>
</tr>
</tbody>
</table>
## APPENDIX E (CONT.)

### OE PREDICTIONS

<table>
<thead>
<tr>
<th>Sample Date</th>
<th>Site Code</th>
<th>OE</th>
<th>Site Name</th>
<th>Test 0</th>
<th>ch(X.01)</th>
<th>TN</th>
<th>TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/11/2007</td>
<td>304ZAYGRA</td>
<td>0.957</td>
<td>Zayante Creek at Graham Hill Road</td>
<td>0</td>
<td>1.045</td>
<td>0.3822</td>
<td></td>
</tr>
<tr>
<td>5/28/2008</td>
<td>305BATMAI</td>
<td>0.959</td>
<td>Bates Creek @ N. Main St.</td>
<td>0</td>
<td>0.605</td>
<td>0.35255</td>
<td></td>
</tr>
<tr>
<td>7/2/2008</td>
<td>305PAH2S</td>
<td>0.964</td>
<td>Pajaro River at Hwy 25</td>
<td>0</td>
<td>13.95</td>
<td>0.1535</td>
<td></td>
</tr>
<tr>
<td>5/21/2008</td>
<td>314NOINOJ</td>
<td>0.972</td>
<td>Nojol Creek @ County Park</td>
<td>1</td>
<td>0.0892</td>
<td>0.069276</td>
<td></td>
</tr>
<tr>
<td>7/11/2007</td>
<td>305GREHAZ</td>
<td>0.975</td>
<td>Green Valley Creek at Hazel Dell Road</td>
<td>2</td>
<td>1.145</td>
<td>0.391</td>
<td></td>
</tr>
<tr>
<td>5/19/2008</td>
<td>314SALGAR</td>
<td>0.981</td>
<td>Salsiquedues Creek @ Gazas Trail</td>
<td>0</td>
<td>0.0968</td>
<td>0.080231</td>
<td></td>
</tr>
<tr>
<td>7/9/2008</td>
<td>314OSOUPO</td>
<td>0.981</td>
<td>Los Osos Creek @ Upper Oso</td>
<td>1</td>
<td>0.753735</td>
<td>0.00542</td>
<td></td>
</tr>
<tr>
<td>7/9/2008</td>
<td>304SQBRI</td>
<td>0.991</td>
<td>Soquel Creek @ Bridge Rd</td>
<td>0</td>
<td>0.74447</td>
<td>0.1205</td>
<td></td>
</tr>
<tr>
<td>7/16/2008</td>
<td>304UTSWA</td>
<td>0.993</td>
<td>Little Creek at Swanton Rd</td>
<td>2</td>
<td>0.55263</td>
<td>0.0946</td>
<td></td>
</tr>
<tr>
<td>9/13/2007</td>
<td>304UTSWA</td>
<td>0.993</td>
<td>Little Creek at Swanton Rd</td>
<td>2</td>
<td>0.289</td>
<td>0.141</td>
<td></td>
</tr>
<tr>
<td>7/19/2007</td>
<td>304CARCAR</td>
<td>0.993</td>
<td>Carbonera Creek at Carbonera Road</td>
<td>0</td>
<td>1.72</td>
<td>0.1325</td>
<td></td>
</tr>
<tr>
<td>7/9/2008</td>
<td>304CARCAR</td>
<td>0.993</td>
<td>Carbonera Creek at Carbonera Road</td>
<td>0</td>
<td>1.153356</td>
<td>0.0964</td>
<td></td>
</tr>
<tr>
<td>7/31/2007</td>
<td>305BRCOA</td>
<td>0.993</td>
<td>San Benito, below Hernandez Res, Coalinga Rd pull-off</td>
<td>0</td>
<td>0.3695</td>
<td>0.065</td>
<td></td>
</tr>
<tr>
<td>6/16/2007</td>
<td>304MORSSI</td>
<td>0.994</td>
<td>Moore's Gulch Creek off Old Soquel San Jose Rd</td>
<td>0</td>
<td>0.3955</td>
<td>0.394275</td>
<td></td>
</tr>
<tr>
<td>8/20/2007</td>
<td>308MCWP</td>
<td>0.994</td>
<td>McWay Creek, Julia Pfeiffer SP Canyon Trail bridge</td>
<td>1</td>
<td>0.1735</td>
<td>0.0234</td>
<td></td>
</tr>
<tr>
<td>7/7/2008</td>
<td>305LADAK</td>
<td>0.995</td>
<td>Lasgas Creek @ Oak Glen Ave</td>
<td>0</td>
<td>1.090255</td>
<td>0.0481</td>
<td></td>
</tr>
<tr>
<td>6/4/2008</td>
<td>310OLDOLD</td>
<td>0.997</td>
<td>Old Creek @ Old Creek Rd</td>
<td>1</td>
<td>0.2575</td>
<td>0.0694</td>
<td></td>
</tr>
<tr>
<td>7/17/2008</td>
<td>304COSWA2</td>
<td>1.002</td>
<td>Scott Creek @ Upper Swanton Rd</td>
<td>0</td>
<td>0.61258</td>
<td>0.0387</td>
<td></td>
</tr>
<tr>
<td>8/21/2007</td>
<td>308CARCK</td>
<td>1.003</td>
<td>Garrapata Creek above Joshua Creek Confluence</td>
<td>2</td>
<td>0.3345</td>
<td>0.0456</td>
<td></td>
</tr>
<tr>
<td>7/10/2007</td>
<td>308CARCK</td>
<td>1.003</td>
<td>Garrapata Creek above Joshua Creek Confluence</td>
<td>2</td>
<td>0.29045</td>
<td>0.04155</td>
<td></td>
</tr>
<tr>
<td>8/10/2007</td>
<td>304PAGAZ</td>
<td>1.004</td>
<td>Opa Creek at Gazos Creek Road</td>
<td>1</td>
<td>0.101</td>
<td>0.425</td>
<td></td>
</tr>
<tr>
<td>7/30/2008</td>
<td>304BEADLD</td>
<td>1.007</td>
<td>Bear Creek at Old Bear Road</td>
<td>0</td>
<td>0.0278</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>7/22/2008</td>
<td>304SABRA</td>
<td>1.010</td>
<td>Salinas River at Bradley</td>
<td>0</td>
<td>1.8639</td>
<td>0.006345</td>
<td></td>
</tr>
<tr>
<td>8/9/2007</td>
<td>200LGALD</td>
<td>1.011</td>
<td>Los Gatos Creek at Aldercraft Heights Road</td>
<td>0</td>
<td>0.154</td>
<td>0.09205</td>
<td></td>
</tr>
<tr>
<td>7/8/2008</td>
<td>305PAH2BET</td>
<td>1.011</td>
<td>Pajaro River @ Betsel Road</td>
<td>0</td>
<td>23.838</td>
<td>0.209</td>
<td></td>
</tr>
<tr>
<td>6/4/2008</td>
<td>307CARLOS</td>
<td>1.018</td>
<td>Carmel River @ above Los Padres Dam</td>
<td>0</td>
<td>0.016155</td>
<td>0.0356</td>
<td></td>
</tr>
<tr>
<td>3/18/2008</td>
<td>309SAPIC</td>
<td>1.021</td>
<td>Sandy Creek @ Pinnacles Campground</td>
<td>2</td>
<td>-88</td>
<td>-88</td>
<td></td>
</tr>
<tr>
<td>7/23/2008</td>
<td>305GREHAZ</td>
<td>1.027</td>
<td>Green Valley Creek at Hazel Dell Road</td>
<td>1</td>
<td>1.147517</td>
<td>0.0361</td>
<td></td>
</tr>
<tr>
<td>7/23/2008</td>
<td>305RAMRAM</td>
<td>1.029</td>
<td>Ramsey Creek at Ramsey Road</td>
<td>0</td>
<td>-88</td>
<td>-88</td>
<td></td>
</tr>
<tr>
<td>7/2/2008</td>
<td>200STECO</td>
<td>1.033</td>
<td>Stevens Creek @ Cooley Picnic Area</td>
<td>1</td>
<td>0.2585</td>
<td>0.0808</td>
<td></td>
</tr>
<tr>
<td>7/24/2008</td>
<td>304WAPRED</td>
<td>1.044</td>
<td>Waddell Creek at Redwood Camp in Big Basin</td>
<td>2</td>
<td>0.0438</td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td>5/1/2008</td>
<td>307CARRSC</td>
<td>1.045</td>
<td>Carmel River @ Ranch San Carlos Rd</td>
<td>0</td>
<td>1.44</td>
<td>0.044</td>
<td></td>
</tr>
<tr>
<td>6/24/2008</td>
<td>304BURSHU</td>
<td>1.048</td>
<td>Burns Creek @ Schultes Road</td>
<td>0</td>
<td>0.373</td>
<td>0.213682</td>
<td></td>
</tr>
<tr>
<td>7/9/2008</td>
<td>314ACHAP</td>
<td>1.048</td>
<td>Cachuma Creek @ Happy Cyn Rd.</td>
<td>1</td>
<td>0.95266</td>
<td>0.03685</td>
<td></td>
</tr>
<tr>
<td>9/13/2007</td>
<td>304COSWA</td>
<td>1.051</td>
<td>Scott Creek at fish gate off Swanton Rd</td>
<td>0</td>
<td>0.351875</td>
<td>0.08765</td>
<td></td>
</tr>
<tr>
<td>6/3/2008</td>
<td>307ULCRA</td>
<td>1.054</td>
<td>Tularcitos Creek @ Carmel Valley Rd</td>
<td>1</td>
<td>0.113</td>
<td>0.1215</td>
<td></td>
</tr>
<tr>
<td>7/23/2008</td>
<td>309NACOF</td>
<td>1.056</td>
<td>Nacimiento River @ Fort Hunter Uggert</td>
<td>1</td>
<td>0.1606</td>
<td>0.0322</td>
<td></td>
</tr>
<tr>
<td>8/3/2008</td>
<td>308CARCK</td>
<td>1.058</td>
<td>Garrapata Creek above Joshua Creek Confluence</td>
<td>2</td>
<td>0.173</td>
<td>0.2285</td>
<td></td>
</tr>
<tr>
<td>7/21/2008</td>
<td>304ZM23B</td>
<td>1.060</td>
<td>Semprevivens Creek at Hwy 236</td>
<td>1</td>
<td>0.01835</td>
<td>0.0281</td>
<td></td>
</tr>
<tr>
<td>7/20/2007</td>
<td>305UYVSAW</td>
<td>1.063</td>
<td>Uvas Creek at Uvas Canyon County Park</td>
<td>1</td>
<td>0.1475</td>
<td>0.0222</td>
<td></td>
</tr>
<tr>
<td>6/12/2008</td>
<td>309PASBET</td>
<td>1.063</td>
<td>Paso Robles Creek @ Bethal Rd</td>
<td>0</td>
<td>0.3035</td>
<td>0.02785</td>
<td></td>
</tr>
<tr>
<td>5/13/2008</td>
<td>304SANSTA</td>
<td>1.070</td>
<td>San Gregorio Creek @ Stage Rd</td>
<td>0</td>
<td>0.624</td>
<td>0.163</td>
<td></td>
</tr>
<tr>
<td>5/7/2008</td>
<td>314SANMIG</td>
<td>1.072</td>
<td>San Miguelito Creek @ San Miguelito Road</td>
<td>0</td>
<td>-88</td>
<td>-88</td>
<td></td>
</tr>
<tr>
<td>7/22/2008</td>
<td>309SALANX</td>
<td>1.083</td>
<td>Salinas River @ King City</td>
<td>0</td>
<td>1.0028429</td>
<td>0.0536</td>
<td></td>
</tr>
<tr>
<td>3/19/2008</td>
<td>309BCGNVS</td>
<td>1.087</td>
<td>Bear Gulch Near Visitor Center</td>
<td>1</td>
<td>-88</td>
<td>-88</td>
<td></td>
</tr>
<tr>
<td>7/22/2008</td>
<td>305SBRH2S</td>
<td>1.090</td>
<td>San Benito at Hwy 25</td>
<td>0</td>
<td>-88</td>
<td>-88</td>
<td></td>
</tr>
<tr>
<td>9/6/2007</td>
<td>308WILH1</td>
<td>1.094</td>
<td>Willow Creek @ Hwy 1</td>
<td>1</td>
<td>0.0795</td>
<td>0.0226</td>
<td></td>
</tr>
</tbody>
</table>
## APPENDIX E (CONT.)

### OE PREDICTIONS

<table>
<thead>
<tr>
<th>Sample Date</th>
<th>Site Code</th>
<th>Site Name</th>
<th>Sample Date</th>
<th>Site Code</th>
<th>Site Name</th>
<th>Sample Date</th>
<th>Site Code</th>
<th>Site Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/18/2007</td>
<td>305GMABD</td>
<td>Gamecock Creek off Browns Valley Rd</td>
<td>0/0</td>
<td>0.3995</td>
<td>0.127</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/7/2007</td>
<td>304WSTSOQ</td>
<td>Soquel Creek, WestB, unnamed st before Olson Rd</td>
<td>0/0</td>
<td>0.215</td>
<td>0.2065</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/2/2007</td>
<td>305WATLEE</td>
<td>Watsonville Slough at Lee Road</td>
<td>0/0</td>
<td>10.8</td>
<td>0.227</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/16/2008</td>
<td>304BIGBG</td>
<td>Big Creek @ Big Creek Rd</td>
<td>1/0</td>
<td>0.518295</td>
<td>0.0144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/29/2008</td>
<td>304APTMAR</td>
<td>Aptos Creek at Margaret Bridge</td>
<td>0/0</td>
<td>0.073406</td>
<td>0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/8/2008</td>
<td>304SEM234</td>
<td>Semprevires Creek at Hwy 234</td>
<td>2/0</td>
<td>0.2015</td>
<td>0.0784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/12/2007</td>
<td>304AYGR</td>
<td>Zayante Creek at Graham Hill Road</td>
<td>0/0</td>
<td>0.8155</td>
<td>0.3855</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/10/2007</td>
<td>304BIGB</td>
<td>Big Creek @ Big Creek Rd</td>
<td>0/0</td>
<td>0.518295</td>
<td>0.0144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/29/2008</td>
<td>304APTMAR</td>
<td>Aptos Creek at Margaret Bridge</td>
<td>0/0</td>
<td>0.073406</td>
<td>0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/8/2007</td>
<td>304SEM234</td>
<td>Semprevires Creek at Hwy 234</td>
<td>2/0</td>
<td>0.2015</td>
<td>0.0784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/12/2007</td>
<td>304AYGR</td>
<td>Zayante Creek at Graham Hill Road</td>
<td>0/0</td>
<td>0.8155</td>
<td>0.3855</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/10/2007</td>
<td>304BIGB</td>
<td>Big Creek @ Big Creek Rd</td>
<td>0/0</td>
<td>0.518295</td>
<td>0.0144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/29/2008</td>
<td>304APTMAR</td>
<td>Aptos Creek at Margaret Bridge</td>
<td>0/0</td>
<td>0.073406</td>
<td>0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/8/2007</td>
<td>304SEM234</td>
<td>Semprevires Creek at Hwy 234</td>
<td>2/0</td>
<td>0.2015</td>
<td>0.0784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/12/2007</td>
<td>304AYGR</td>
<td>Zayante Creek at Graham Hill Road</td>
<td>0/0</td>
<td>0.8155</td>
<td>0.3855</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/10/2007</td>
<td>304BIGB</td>
<td>Big Creek @ Big Creek Rd</td>
<td>0/0</td>
<td>0.518295</td>
<td>0.0144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/29/2008</td>
<td>304APTMAR</td>
<td>Aptos Creek at Margaret Bridge</td>
<td>0/0</td>
<td>0.073406</td>
<td>0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/8/2007</td>
<td>304SEM234</td>
<td>Semprevires Creek at Hwy 234</td>
<td>2/0</td>
<td>0.2015</td>
<td>0.0784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/12/2007</td>
<td>304AYGR</td>
<td>Zayante Creek at Graham Hill Road</td>
<td>0/0</td>
<td>0.8155</td>
<td>0.3855</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/10/2007</td>
<td>304BIGB</td>
<td>Big Creek @ Big Creek Rd</td>
<td>0/0</td>
<td>0.518295</td>
<td>0.0144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/29/2008</td>
<td>304APTMAR</td>
<td>Aptos Creek at Margaret Bridge</td>
<td>0/0</td>
<td>0.073406</td>
<td>0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/8/2007</td>
<td>304SEM234</td>
<td>Semprevires Creek at Hwy 234</td>
<td>2/0</td>
<td>0.2015</td>
<td>0.0784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/12/2007</td>
<td>304AYGR</td>
<td>Zayante Creek at Graham Hill Road</td>
<td>0/0</td>
<td>0.8155</td>
<td>0.3855</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/10/2007</td>
<td>304BIGB</td>
<td>Big Creek @ Big Creek Rd</td>
<td>0/0</td>
<td>0.518295</td>
<td>0.0144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/29/2008</td>
<td>304APTMAR</td>
<td>Aptos Creek at Margaret Bridge</td>
<td>0/0</td>
<td>0.073406</td>
<td>0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/8/2007</td>
<td>304SEM234</td>
<td>Semprevires Creek at Hwy 234</td>
<td>2/0</td>
<td>0.2015</td>
<td>0.0784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/12/2007</td>
<td>304AYGR</td>
<td>Zayante Creek at Graham Hill Road</td>
<td>0/0</td>
<td>0.8155</td>
<td>0.3855</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/10/2007</td>
<td>304BIGB</td>
<td>Big Creek @ Big Creek Rd</td>
<td>0/0</td>
<td>0.518295</td>
<td>0.0144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/29/2008</td>
<td>304APTMAR</td>
<td>Aptos Creek at Margaret Bridge</td>
<td>0/0</td>
<td>0.073406</td>
<td>0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/8/2007</td>
<td>304SEM234</td>
<td>Semprevires Creek at Hwy 234</td>
<td>2/0</td>
<td>0.2015</td>
<td>0.0784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/12/2007</td>
<td>304AYGR</td>
<td>Zayante Creek at Graham Hill Road</td>
<td>0/0</td>
<td>0.8155</td>
<td>0.3855</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/10/2007</td>
<td>304BIGB</td>
<td>Big Creek @ Big Creek Rd</td>
<td>0/0</td>
<td>0.518295</td>
<td>0.0144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/29/2008</td>
<td>304APTMAR</td>
<td>Aptos Creek at Margaret Bridge</td>
<td>0/0</td>
<td>0.073406</td>
<td>0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/8/2007</td>
<td>304SEM234</td>
<td>Semprevires Creek at Hwy 234</td>
<td>2/0</td>
<td>0.2015</td>
<td>0.0784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/12/2007</td>
<td>304AYGR</td>
<td>Zayante Creek at Graham Hill Road</td>
<td>0/0</td>
<td>0.8155</td>
<td>0.3855</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/10/2007</td>
<td>304BIGB</td>
<td>Big Creek @ Big Creek Rd</td>
<td>0/0</td>
<td>0.518295</td>
<td>0.0144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/29/2008</td>
<td>304APTMAR</td>
<td>Aptos Creek at Margaret Bridge</td>
<td>0/0</td>
<td>0.073406</td>
<td>0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/8/2007</td>
<td>304SEM234</td>
<td>Semprevires Creek at Hwy 234</td>
<td>2/0</td>
<td>0.2015</td>
<td>0.0784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/12/2007</td>
<td>304AYGR</td>
<td>Zayante Creek at Graham Hill Road</td>
<td>0/0</td>
<td>0.8155</td>
<td>0.3855</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX F

R Code

### Cory Ritz, California State University Monterey Bay, April 25 2009
# RIVPACS model build and all-subset sample routine
# Adopted from John VanSickle et al. 2006, code available from:
# USEPA Western Ecology Division [internet]

# Includes computing dissimilarity matrix, clustering, cluster pruning, and
# discriminant function analysis (DFA)
# Version 1

STEP 1 -- SETUP -- Import and sort diatom and environmental predictor data;
# Below used for building California Central Coast CSUMB diatom predictive models.

"Input data are predictor data (all sites) and a (site x spp) matrix for all bugs at
all sites.
# The bug matrix is the output of subsampling and matrify programs; see 'matrify'
rcode
# Assume that predictor data file includes a column to ID the calibration, validation
and test sites";

# "Input the predictor file, tab delimited;
#assume predictors have already been appropriately transformed";

predall <- read.csv(file.choose(),header=T)
#attach(predall)

## Input the site by taxa matrix of bugs (diatoms) for all samples(sites);
# ** For site grouping analysis input file should only have species occurring
# ** 95% > x > 5%. Remove diatom species occur too frequently or too rarely
# ** species will need to be added in later for final OE analysis
# ref column indicates 0=test sites, 1=reference sites, 2=validations sites
bugall<- read.csv(file.choose(),header=T) #use precluster.csv (removed <5%bugs)
#ensure sample(row) alignment of diatom and predictor data; all values = true
row.names(bugall)==row.names(predall);
# see Van Sickle for code to correct alignment
#Begin model build;

#Presence/Absence (1/0) site by species matrix for the diatoms;
bugall.pa<-bugall;
bugall.pa[bugall.pa>O]<-1;

#Extract subsets of diatom and predictor data for the reference ("1") sites;
#note: C=reference sites, 0=test sites, V=validations sites;
predcal<-predall[predall[,"ref"]=='C',];
bugcal<-bugall[bugall[,"ref"]=='C',];  #Abundance matrix;
bugcal.pa<-bugcal.pa[predall[,"ref"]=='C',];  #P/A matrix;

#Continue processing: data sets created and aligned;

#STEP 2 -- DISSIMILARITIES AND CLUSTER ANALYSIS;

# Compute dissimilarity matrix for reference site diatoms;
# The code below calculates Sorensen dissimilarities;
# Sorensen was chosen over Bray-Curtis for CSUMB diatom data
# VanSickle used "the generalized outer product function dapply();
# and choose the desired dissimilarity measure as a called function;
# dapply() output is an (n(n-1)/2) length vector storing;
# the lower triangle of the site dissimilarity matrix in column major order;
# can be input directly to R clustering functions;"
#source dapply, Van Sickle et al. (2006); source("C:/mypath/dapply.r");

# Option 2 (from Van Sickle) -- "Sorensen dissimilarity for P/A data;
# function computes Sorensen P/A dissimilarity between one site pair, siti and sitj;
#input can be P/A data or abundance data;"

sornfun<-function(siti,sitj) {
    shared<-sum((siti>0)&(sitj>0));
    uniquei<-sum((siti>0)&(sitj==0));
    uniquej<-sum((siti==0)&(sitj>0));
    l=(2*shared/(2*shared+uniquei+uniquej));   #return Sorensen dissimilarity;
} #end of function;

#Sorensen dissimilarities are calculated from the reference sites
dissim<-dapply(bugcal,1,bugcal,1,sornfun);

#END OF MODEL BUILD;
# Option 3 -- Bray-Curtis (Sorenson) dissimilarity for abundance data;
# in this example, use untransformed relative abundance;

# first compute site by spp matrix of relative abundance;
totabun<-apply(bugcal,1,sum); # vector of total abundance, each site;
rel.abun<-sweep(bugcal,1,totabun,FUN="/"); # relative abundance matrix;

# function below computes BC dissim within dapply();
# Instead, could use gdist() in mpart package, to do Bray-Curtis;
# siti, sitj are vectors of abundances for 2 sites;
# if zero abundance at both sites, then dissimilarity = 0;

bcfun<-function(siti, sitj) {
  bcnum<-sum(abs(siti-sitj));
  bcdenom<-sum(siti+sitj);
  ifelse(bcdenom>0, (bcnum/bcdenom), 0); # return BC dissimilarity;
} # end of function;

# compute Bray-Curtis dissimilarity;
dissim<-dapply(rel.abun,l,rel.abun,l,bcfun);
# Proceed to clustering;

# Clustering of calibration sites;
# Use flexible-Beta method, with Beta=0.6; Note: Beta=0.6 appears unique for
# CSUMB data (other authors use a negative Beta, however this failed to yield
# usable clusters
# Method is an option that is available in agnes() function of "cluster" package,
# but only in R version 2.0.1 and later;

# load "cluster" package. See R documentation on agnes();
# in using agnes(), note that:
# For Flexible Beta strategy, Beta=(1-2*Alpha) in Lance-Williams formula;
# A single value for par.method value specifies alpha, so alpha=0.8 gives Beta=-0.6;

clus1<-agnes(x=dissim, diss=T, method="flexible", par.method=0.625, keep.diss=F, keep.data=F);

## Various plots of cluster outcome. Leaf labels are row numbers of dissim matrix;
# that is, the order of sites in the calibration data set;
plotree(clus1); # Or else can just plot the dendrogram;
plotree(clus1, main = paste(labels = NULL, xlab = "Reference Sites"))
#plot(clusl); #lst plot is banner. 2nd is the dendrogram;
write.csv(clusl, file = "c:\CurrentWork\l00329Results_cluster_ref_sites_BC.csv")

#Pruning the dendrogram to create a small number of groups;
# level pruning can be done by specifying the number of groups (k parameter);
#Also can prune at a specified height. See cutree help;
#result is a vector of site group assignments;
#can repeat this process to generate several candidate groupings from a single
dendrogram;
grps<-cutree(clusl,k=4); #vector of group assignments is in the order of sites in the
clustered data;
table(grps); #count number of sites in each group;
cbind(row.names(predcal),grps); #list calibration sites and their group assignments;
#candidate site groups complete;
#alternative is non-level pruning. Use the following to interactively;
#pick out desired clusters and store their observation numbers;
#experimental, not fully developed;
#ccc<-identify(as.hclust(clusl)); #interactive ID of clusters on dendrogram;

# Post CLUSTERING need to add original bugall data back in
# initially bugall was cropped of spp occurring less than 5% of streams
#bugall<- read.csv(file.choose(),header=T) #use bray_curtis2.csv
#row.names(bugall)==row.names(predall);
#bugall.pa<-bugall;
#bugall.pa[bugall.pa>0]<-1;
#bugcal<-bugall[predall[, 'ref']=='l',]; #Abundance matrix;
#bugcal.pa<-bugall.pa[predall[, 'ref']]=='l',]; #P/A matrix;
#dissim<-dapply(bugcal,1,bugcal,1,sornfun);
#dissim<-dapply(rel.abun,1,rel.abun,1,bcfun);

#STEP 3 -- DISCRIMINANT FUNCTION ANALYSIS (DFA);
# Instead of DFA, consider using classification tree model (R packages "tree" or
"rpart");
# or a random forest model (R package "randomForest");
#Below, I have options for stepwise DFA and also for all-subsets DFA;

#First, put the names of candidate predictors in a vector;
candvar <-c("wet", "slopeT", "elevT", "doy",
"rain", "lat", "long", "areaT", "sed_sandT", "mindepthTln", "maxdepthTln",
"maxvelTlog","minvelTlog") #cory

#Option 2 -- All subsets DFA:
# Feasible for up to about 15 candidate predictors;  
# User specifies a small number of best models for selected model orders;  
# Wilks lambda, classification accuracy, and statistics of O/E are reported for each best model;  
# If user supplies an independent set of validation data (bug data and predictor data), then;  
# O/E statistics also computed for validation set;  
# set up data. Calibration data already set up;  
# Need to specify the validation data;  
# pred.vld<-predall[substr(as.character(predall[, 'ref']),1,1)=='2',];  
# bug.vld.pa<-bugall.pa[substr(as.character(predall[, 'ref']),1,1)=='2',];

pred.vld<-predall[substr(as.character(predall[, 'ref']),1,1)=='V',];  
bug.vld.pa<-bugall.pa[substr(as.character(predall[, 'ref']),1,1)=='V',];

# Optional -- visually check distributions of validation and calibration predictors;  
cand.cont<-c("wet", "slopeT", "elevT", "doy", "rain", "lat", "long", "areaT", "sed_sandT", "mindepthTln", "maxdepthTln", "maxvelTlog", "minvelTlog");
par(mfrow=c(4,4));  
lapply(cand.cont, function(x) boxplot(list(clb=predcal[,x], vld=pred.vld[,x]), ylab=x));

# Specify a vector describing how many models of each order to keep;  
# The following example specifies keeping 5 models each for;  
# orders 1, 2, ..., 13 and the single (saturated) model of order 14;  
nkeep<-c(rep(5,12),1);  
# Load the all subsets DFA function;
source("c:/mypath/dfa.allsub.v3.r");  
source(file.choose()); # look for dfa.allsub.v3.r in Van Sickle  
# LOAD "MASS" and "GTOOLS" packages;  
# execute the following block of code. dfa.allsub.v3() is surrounded;  
# by code that records and prints the execution time;  
# Execution may take several minutes;  
# In example below, Pc is set to a very small value, to retain all taxa in O/E and BC;  
# Another alternative is Pc=0.5;  
start.time=proc.time();  
dfm.best<-
  dfa.allsub.v3(bug.cal=bugcal.pa, bug.vld=bug.vld.pa, pred.cal=predcal, pred.vld=pred.vld,  
  grps=grps, candvar=candvar, numkeep=nkeep, Pc=0.25);

elaps<-proc.time()-start.time;  
print(c("elapsed time = ", elaps));
dfm.best.5grp.25<-dfm.best; #store result under a new name, indicating the Pc value used;

dfm.best.5grp.5<-dfm.best; # Store result of a second run, which had Pc=0.5;

# Various ideas for exploring the set of best DFA models;

# A) - Results list contains the set of best models for a single;
# candidate site group assignment. Rename this data frame for future analysis;

#rename results list for analysis, and extract best-model data frame;

dfm.best<-dfm.best.5grp.25;
bestmods<-dfm.best$subset.stats;

# B) - look at all the models, sorted by a chosen criterion;
#for example, sort the best models by SD(O/E) at calibration sites;
format(bestmods$order(bestmods$RMSE.cal),,digits=3);
write.csv (bestmods, file = "C:\CurrentWork\bestmods_100329.csv")

# C) plot a measure of model performance against model size (ie, model order);
#For example, plot RMSE(O/E) against model order separately for calibration and
validation sites;

plot (bestmods$order,bestmods$RMSE.cal,ylim=c(0.10,0.26 ),type='p',pch='C',
  cex=.7,xlab='Model order',ylab='RMSE(O/E) ');
points (bestmods$order,bestmods$RMSE.vld,pch='V',cex=.7) ;

#put null model RMSE as a baseline, separate for Calibration and validation sites.;
abline(bestmods$null.stats$"RMSE.cal"),0,1ty=1);

abline (dfm.best$null.stats$"RMSE.vld"),0,1ty=2);
# identify the C and V points for one model on the plot;
# the Cal point is marked with a solid box, Vld with a solid triangle, and the model
is printed;
cc<-identify (bestmods$order,bestmods$RMSE.vld,n=l,plot=F);
points (bestmods$order[cc],bestmods$RMSE.vld[cc],pch=17,cex=1);
#cc<-identify (bestmods$order,bestmods$RMSE.cald,n=l,plot=F);
points (bestmods$order[cc],bestmods$RMSE.cal[cc],pch=15,cex=1);
print (bestmods$model[cc]);

#following lines put a title and legend on the plot;
legend(locator(l),legend=c('Calibration sites','Validation sites'),pch=c('C','V'));
title(main=list('ORDEQ models: RMSE(O/E) from 5 best models of each model order',cex=.9));

# Can also experiment with similar plots for BC statistics. "Better" models will have;
# smaller BC90;
# D) Plot the two classification accuracy measures against model order;
# DFM overfitting starts occurring where the CV accuracy flattens out;
plot(bestmods$order,bestmods$cls.crct.resub,ylim=c(20,80),type='p',pch='R',
     cex=.8,xlab='Model order',ylab='Percent correct');
points(bestmods$order,bestmods$cls.crct.cv,pch='C',cex=1.0, col='blue');
lines(predict(loess(bestmods$cls.crct.cv-bestmods$order))-bestmods$order,lty=3,
     col='blue')
lines(predict(loess(bestmods$cls.crct.resub-bestmods$order))-bestmods$order,lty=2)
legend(locator(1),legend=c('Resubstitution', 'Crossvalidation'),pch=c('R', 'C'));
title(main=list('Classification accuracy',cex=.9));

#E) PREDICTOR IMPORTANCE. Calculate the proportion of best models that include;
# each of the predictors. Proportion is not weighted by model quality;
round((100*table(unlist(strsplit(bestmods$model," ") )
       )/dim(bestmods)[[1]],1); predperc <- round((100*table(unlist(strsplit(bestmods$model," ") )
       )/dim(bestmods)[[1]],1);
plot(predperc)

#F) plot the geographic locations of the site clusters;
plot(predcal$long[grps==1], predcal$lat[grps==1],col='black',
     type='p',xlim=c(-124,-118.5),ylim=c(34,37.6));
points(predcal$long[grps==2], predcal$lat[grps==2],col='red')
points(predcal$long[grps==3], predcal$lat[grps==3],col='green')
points(predcal$long[grps==4], predcal$lat[grps==4],col='blue')
points(predcal$X_coord[grps==5], predcal$Y_coord[grps==5],col='blue')

#G) scatterplot matrix of model size and performance on validation and calibration sites;
pairs(as.matrix(bestmods[,c('order','RMSE.cal','RMSE.vld')]));
scatterplot3d(x=bestmods$RMSE.cal,y=bestmods$RMSE.vld,z=bestmods$order)

# End of model development code. By iterating the above pieces, you;
# can choose the "final" model(s), which consist of a desired classification
# for reference sites, and one or more "best" DFA models for predicting class membership;
# Once these have been decided, go to model.predict.r, for code that makes predictions;
# at new sites;
# R code to make predictions of O/E for a new set of sites,
# based on a single 'final' predictive model;
# Program assumes that you have run model.build.r, so that its data sets are
# available in the R workspace;
#Version 3, June 25, 2007 -- Includes BC index;

bugall<- read.csv(file.choose(),header=T) #use postcluster.csv (added back <5%bugs)
bugall.pa<-bugall;
bugall.pa[bugall.pa>0]<-1;
bugcal<-bugall[predall[, 'ref']=='C',]; #Abundance matrix;
bugcal.pa<-bugall.pa[predall[, 'ref']=='C',]; #P/A matrix;

#STEP 1 -- # Set up the needed data objects;
# Use code like that in model.build.r to set up these objects;
# In particular, Step 1 of model.build.r shows how the rows and columns of;
# the data frames must be aligned;
# The needed objects are:
"predall" = data frame containing predictor variables (columns) at all sites #(rows);
# at which predictions are desired (e.g., reference plus 'test' sites);
"bugall" = corresponding data frame of sites (rows) by species (columns) of observed
# presence/absence (coded 1/0) for all sites;
#This program (model.predict.r) rebuilds the chosen predictive model from calibration
#data ;based on the 'final' site groups and chosen predictor variables;
#To do this, the following objects are needed, which are available following runs of
#model.build.r.

"predcal" = Predictor variables(columns) for calibration sites (rows). Usually a
#subset of predall;
# "bugcal.pa"= Corresponding data frame of observed presence/absence (1/0) at
calibration sites. Usually a subset of bugall;
"grps.final" = Corresponding vector identifying the cluster membership of
calibration sites;
# For example, grps.final<-grps.5;grps.final<-grps
"preds.final" = Vector with names of the chosen predictor variables, all of which
# must be available;
# in predall and predcal. The prediction code assumes that there
# are no missing
# values in predall or predcal, for any of these variables;
# Here are 3 options for specifying preds.final. ;

# OPTION A -- Choose a single DF model from the subset of best ;
# models that were identified by the dfa.allsub.v3 function (all subsets DFA);
# These best models are stored in the "bestmods" data frame;
# The following example employs the model in row 27 of the "bestmods" data frame;
preds.final<-unlist(strsplit(bestmods[14,'model']," "));
#OPTION B -- Use the final DF model selected by the stepwise regression function, dfa.step;
#In model.build.r, the output of dfa.step was called "step.res";
# preds.final<-attr(terms(step.res),"term.labels");
# OPTION C -- Directly list the names of chosen predictors;
# preds.final<-c("DAYNUM","X_coord","Bsnrgeco","Lithol", "El_m_sqg");
# preds.final<-c ("lat", "maxdepthTln", "sed_sandT", "doy")
# preds.final<-c("doy", "lat", "rain", "areaT")
# preds.final<-c ( "sed_sandT", "lat", "maxdepthTln", "areaT" )
# "lat", "minvelTlog", "wet")
#"lat", "long", "maxvelTlog", "sed_sandT", "slopeT") # cory
# preds.final<-c("doy","slopeT")

#STEP 3 -- MAKE THE PREDICTIONS;
# First, follow instructions below under "COMPILATION OF MAIN CODE";
# This only needs doing once in an R session;

#Next, run the prediction function that you have just compiled, as shown in the next statement;
# To include all reference taxa having nonzero occurrence probs,;
# set Pc equal to a very small positive number, such as .000001;
#Load the prediction compilation; cory
#source("c:/mypath/predictionCompile.r");
#O.E.final.pr<-predict.OE.v3(grps=grps.final,predvars=preds.final,Pc=1.E-14,
#bugall <- bugall[-1] #deletes column 1 'ref'
#bugall.pa <- bugall.pa[-1]
#bugcal.pa <- bugcal.pa[-1]
#O.E.final.pr<-predict.OE.v3(grps=grps.final,predvars=preds.final,Pc=0.25,
#predcal=predcal,bugcal=bugcal.pa, predall=predall,bugall=bugall.pa);
#O.E.id <- cbind(predall$id, O.E.final.pr)
#O.E.all <- cbind (predall$ref, O.E.id)
##O.E.all <-O.E.final.pr[substr(as.character(O.E.final.pr[,"outlier.01"]),1,190)=='0',]
##O.E.all
#O.E.all.ref<-O.E.all[substr(as.character(predall[,'ref'])),1,190]=='C',]
#mean (O.E.all.ref$OoverE)
#range ( O.E.all.ref$OoverE)
#sd (O.E.all.ref$OoverE)
#hist(O.E.all.ref$OoverE, breaks=10, xlim=c(0.1,1.6))
#O.E.all.test<-O.E.all[substr(as.character(predall[,'ref'])),1,190]=='T',]
#mean (O.E.all.test$OoverE)
#range ( O.E.all.test$OoverE)
#sd (O.E.all.test$OoverE)
hist(OE.all.test$OoverE, breaks=10, xlim=c(0.1,1.6))
OE.all.vld<-OE.all[substr(as.character(predall[, 'ref']),1,190)=='V',]
mean (OE.all.vld$OoverE)
range ( OE.all.vld$OoverE)
sd (OE.all.vld$OoverE)
hist(OE.all.vld$OoverE, breaks=10, xlim=c(0.1,1.6))
sd (OE.all.ref$OoverE.null)
rangle (OE.all.ref$OoverE.null)
mean (OE.all.ref$OoverE.null)
mean (OE.all.test$OoverE.null)
mean (OE.all.vld$OoverE.null)
x=c("Validation", "Calibration", "Test")
boxplot(OE.all.vld$OoverE, OE.all.ref$OoverE, OE.all.test$OoverE,
names=(x), ylab="Observed to Expected Ratio OE", cex.lab=1.25 )
#boxplot(OE.all.vld$OoverE, OE.all.ref$OoverE, OE.all.test$OoverE )
title("Pc = .25")
#OE.all
# Quantiles #
quantile(OE.all.vld$OoverE,probs=0.1)
quantile(OE.all.ref$OoverE,probs=0.9)
# plot the land use stuff
OE.lu<- read.csv(file.choose(),header=T) #test histograms of prediction of landuse
versus ref sites
x=c("Reference", "Agriculture", "Urban", "Light Urban", "Index")
boxplot(OE.lu$OE_Ref, OE.lu$OE_Ag, OE.lu$OE_Urb, OE.lu$OE_LtUrb, OE.lu$OE_Index,
names=(x),
ylab="Observed to Expected Ratio OE", cex.lab=1.25)
abline (quantile(OE.all.ref$OoverE,probs=0.1),0, lty=5, lwd=2, col="blue")
abline (quantile(OE.all.ref$OoverE,probs=0.9),0, lty=5, lwd=2, col="red")
hist( list (OE.all.vld$OoverE, OE.all.ref$OoverE, OE.all.test$OoverE) )
write.csv (OE.all, file = "c:\CurrentWork\100802OE_25percent.OE.all_model_15.csv")
write.csv(OE.all.ref$E, file = "c:\CurrentWork\100409OE_refsites_model_14_pc25_E.csv")

# plot nutrient stuff
OE.nut<- read.csv(file.choose(),header=T) # under C/CurrentWork
#na.omit(OE.nut)
par(mfrow=c(2,2))
plot(log(OE.nut$tp),OE.nut$oe, xlim=c(-10,5) ,cex=1.2,
  ylab="Test Sites OE", xlab="Total Phosphorus TP",
  font.lab=1, cex.lab=1.6)
abline ( quantile(OE.all.ref$OoverE,probs=0.1) ,0, lty=5, lwd=2, col="blue")
abline ( quantile(OE.all.ref$OoverE,probs=0.9) ,0, lty=5, lwd=2, col="red")
plot(log(OE.nut$tn),OE.nut$oe, xlim=c(-5,5) ,cex=1.2,
  ylab="Test Sites OE", xlab="Total Nitrogen TN",
  font.lab=1, cex.lab=1.6)
abline ( quantile(OE.all.ref$OoverE,probs=0.1) ,0, lty=5, lwd=2, col="blue")
abline ( quantile(OE.all.ref$OoverE,probs=0.9) ,0, lty=5, lwd=2, col="red")
quinnorm (OE.nut$tp)

#plot(lm(OE.all.ref$O ~ OE.all.ref$E))
summary (lm(OE.nut$tp ~ OE.nut$oe))
t.test(OE.all.test$OoverE ~ OE.all.test$OoverE, paired=False, var.equal=False)
t.test(OE.all.ref$OoverE)
t.test(oe.cal, oe.test)
t.test(oe.val, oe.cal)
#lapply(cand.cont, function(x)boxplot(list(clb=predcal[,x],vld=pred.vld[,x]),ylab=x));
merge(anova.lm(OE.all.test), anova.lm(OE.all.ref), by=0, all=T)
fit <- lm(OE.all.test$OoverE ~ OE.all.ref$OoverE, data = OE.all.ref$OoverE)
> fit <- lm(OoverE ~ ., data = OE.all.ref)
fit <- lm(OoverE ~ OoverE.null, data = OE.all.test)
anova(fit)
plot(OE.all.test$OoverE,OE.all.test$OoverE.null,ylim=c(0,2),type='p',pch='R',
  cex=.8,xlab='something',ylab='OE');
points(OE.all.test$OoverE,bestmods$cls.crct.cv,pch='C' ,cex=1.0, col='blue');
lines(predict(loess(bestmods$cls.crct.cv-bestmods$order))~bestmods$order,lty=3,
  col='blue')
lines(predict(loess(bestmods$cls.crct.resub-bestmods$order))~bestmods$order,lty=2)

### WRITE FILE TO CSV FOR FURTHER ANALYSIS ###
write.csv(OE.all.test, file = "c:\CurrentWork\100329Results_testsites_PC_0.csv")

### plotting the OE Null ###
plot(OE.all.ref$OoverE,OE.all.ref$predall$id,ylim=c(0.5,1.6),type='p',pch='C',
cex=.7,xlab='Reference Sites',ylab='OE');
points(bestmods$order,bestmods$RMSE.vld,pch='V',cex=.7);  
hist(OE.all.ref$OoverE)
#put null model RMSE as a baseline, separate for Calibration and validation sites.;
abline(OE.all.ref$OoverE.null,0,lty=1);
abline(dfm.best$null.stats["RMSE.vld",0,lty=2);  
##  
OE.final.pr
hist(OE.final.pr$OoverE)
write.csv (OE.all, file = "C:\CurrentWork\OE_090504.1900.OE.all.csv")  #cory

plot(OE.final.pr$O, OE.final.pr$E)
plot(OE.final.pr$E,pnorm(OE.final.pr$E, mean=mean(OE.final.pr$E),sd=sd(OE.final.pr$E)))
barplot(OE.final.pr$OoverE)
barplot(OE.all.test$OoverE)
summary (OE.final.pr)
OE.refsites<-grps
plot(OE.final.pr$OoverE,OE.refsites)

#oe.refsites
#prediction complete. ;
# OE.final.pr is the output data frame contains O and E, and the O/E and BC indices,
# for the selected best model and the null model, for all sites;
# The columns named "outlier.xx" contain 1, if the site is an outlier ;
# at the chi-squared probability level of xx;


predict.OE.v3<-function(grps,predvars,Pc,predcal,bugcal, predall,bugall) {;

    names(grps)<-row.names(predcal);
    flush.console();
    print("Number of calibration samples in each group",quote=F);
    print(table(grps));

    #STEP 1 -- construct linear DFA predictor, for known model, using calibration data;
    # Assumes MVN (multi variate normal ditribution), equal covariance.
    # See Johnson & Wichern, pp 505 ff;
    # Also assumes equal priors for group membership;
    nsite.cal<-length(grps); #number of calibration sites;

# extract desired predictor variables for calibration sites and configure as a matrix;
datmat<-as.matrix(predcal[,predvars]);
npreds<-dim(datmat)[[2]]; # number of predictor variables;
# check site alignment of group ID vector and the predictor data;
row.names(datmat) == names(grps);
# calculate matrix of group means for all predictors;
grpmns<-apply(datmat,2,function(x)tapply(x,grps,mean));
print('Table of group means',quote=F);
print(grpmns);
# Next block calculates and displays the inverse of the pooled covariance matrix;
# ?? within group variance
# first is a list of covariance matrices for each group;
covlist<-lapply(split.data.frame(datmat,grps),cov);
# pooled cov matrix is weighted average of group matrices, weighted by group size.
Johnson & Wichern, 11-64;
grpsiz<-table(grps);
ngrps<-length(grpsiz);
# zero out an initial matrix for pooled covariance;
covpool<-matrix(rep(0,npreds*npreds),nrow=npreds,dimnames=dimnames(covlist[[1]]));
# weighted sum of covariance matrices;
for(i in 1:ngrps){covpool<-covpool+(grpsiz[i]-1)*covlist[[i]]};
covpool<-covpool/(sum(grpsiz)-negrps); # renormalize;
covpinv<-solve(covpool); # inverse of pooled cov matrix;
print('Inverse of pooled covariance matrix',quote=F);
print(covpinv);
print ('DFM prediction model components are complete',quote=F);

# STEP 2 -- predict the group (cluster) membership for ALL sites. ;
# In this step, follow RIVPACS assumption of weighting;
# the membership probabilities by Calibration group size, as a prior;
# Also, flag any outlier sites, using chi-squared statistic;
dmat<-as.matrix(predall[,predvars]); # matrix of predictor data for ALL sites and/or samples;
flush.console();
print ('Number of sites with complete predictor data',quote=F);
print(sum(complete.cases(dmat))); # count samples that have nonmissing data;
npreds<-dim(dmat)[[2]]; # number of predictors;
# predict group membership probs for every site, based on preds.final variables;
# group size is used as a prior;
# store probs in matrix, sites are rows, columns are groups;
# use mahalanobis function, where new vector is taken as the 'center', mu,
and matrix of means is taken as the 'data matrix', x;

Prelim step A -- compute the critical chi-squared values for flagging outlier sites;
# uses the MINIMUM of (a)(number of groups-1), and (b) number of predictor variables;
# will flag each site at P-value = .05 and also P-value = .01 level;
dff<-(min(c(npreds,(ngrps-1))));
crit.05<-qchisq(0.95,df=dff);
crit.01<-qchisq(0.99,df=dff);

#construct empty matrix for predicted membership probabilities;
nsit.all<-dim(dmat)[[1]]; #number of ALL sites;
grpprobs<-matrix(rep(0,nsit.all*ngrps),nrow=nsit.all,
dimnames=list(dimnames(dmat)[[1]],dimnames(grpmns)[[1]]));

Also construct data.frame for outlier flag;
# include site type vector and minimum (squared)distance;
# Each site is either a PASS (denote by 0) or FAIL (denote by 1) for the outlier test;
outlier.flag<-
data.frame(outlier.05=rep(0,nsit.all),outlier.01=rep(0,nsit.all),dismin=rep(0,nsit.all ),
row.names=dimnames(dmat)[[1]]);
#ready to compute group membership probs;
#loop over ALL sites, compute vector of group membership probs and flag outliers;
#execute the following code piece as a single block;
##;
for(i in 1:nsit.all){

vector of squared Mahal. dist from current site to each group mean;
dist<-mahalanobis(grpmns,dat[i,],covpinv,inverted=T); #vector of distances;
grpprobs[i,]<-grpsiz*exp(-0.5*dist); # see Clarke et al. (2000);
grpprobs[i,]<-grpprobs[i,]/sum(grpprobs[i,]);

#check for outlier;
outlier.flag$dismin[i]<-min(dist); #save minimum distance;
if(outlier.flag$dismin[i]>crit.05)outlier.flag[i,'outlier.05']<-1;
if(outlier.flag$dismin[i]>crit.01)outlier.flag[i,'outlier.01']<-1;
}; #finish site loop;

#print outlier count;
print('Group membership probabilities complete',quote=F)
print('Count of OK (=0) and outlier (=1) sites, assessed at P=0.01 level of chi-square',quote=F);
print(table(outlier.flag[, 'outlier.01']));
print('',quote=F);
print('Please wait ... ',quote=F);
flush.console();
### site membership probabilities complete;

---------------------;

**STEP 3** -- Compute predicted occurrence probabilities for each taxon at each site;

To do this, need occurrence freqs of all calibration-site taxa in the Calibration site groups;

matrix of relative occurrences of each spp at sites in each group of reference sites;

grocc<-apply(bugcal,2,function(x) tapply(x,grps,function(y){sum(y)/length(y)}));

finally, compute the matrix of predicted occurrence probabilities, for all sites and all spp;

site.prd.dfa<-grpprobs*grocc;

MODEL PREDICTIONS ARE COMPLETE;

--------------------;

**STEP 4.** Compute O, E, O/E and BC for all sites. ;

Also compute O/E and BC for the null model;

temporary data frame to holdnonnull results for all sites. ;

OE.stats<-data.frame(OBS=rep(NA,nsit.all),
E.prd=rep(NA,nsit.all),BC.prd=rep(NA,nsit.all),row.names=row.names(bugall));

loop over all sites. Compute O, predicted E, predicted BC for each site. ;

for(i in 1:nsit.all) {

  i<-1;
  cur.prd<-site.prd.dfa[i,]; #vector of taxon probs for current site;
  spdyn<-names(cur.prd)[cur.prd>=Pc]; #subset of taxa with Pi>=Pcutoff for current site;
  cur.prd<-cur.prd[spdyn]; #vector of p for species subset, current site;
  cur.obs<-bugall[i,spdyn]; #vector of OBS for those species;
  OE.stats$OBS[i]<-sum(cur.obs); #observed richness (O);
  OE.stats$E.prd[i]<-sum(cur.prd); #Expected richness (E);
  OE.stats$BC.prd[i]<-sum(abs(cur.obs-cur.prd))/(OE.stats$OBS[i]+OE.stats$E.prd[i]);
  }

compute Null model expected richness and null model O/E and BC;

first, compute vector of null-model occurrence probabilities;

pnull<-apply(bugcal,2,sum)/dim(bugcal)[[1]];

Compute Expected richness (E) and BC for null model using taxa >= Pc.

Note that the set of taxa included in the null model is fixed for all sites;

nulltax<-names(pnull[pnull>=Pc]); #subset of taxa with Pnull >= Pc;

Enull<-sum(pnull[nulltax]);

print(c('Null model expected richness = ',Enull),quote=F);

null model taxa';

print(nulltax);
Obsnull<-apply(bugall[,nulltax],1,sum); #vector of Observed richness, all sites, under null model; 
BC.null<-apply(bugall[,nulltax],1,function(x)sum(abs(x-pnull[nulltax])/)(Obsnull+Enull)); #vector of null-model BC;
#Final data frame contains values of O, E, O/E, Onull, Enull, Onull/Enull, BC.prd and BC.null for all sites;
#Also includes outlier flags;
OE.final<-data.frame(O=OE.stats$OBS,E=OE.stats$E.prd,
                      OoverE=OE.stats$OBS/OE.stats$E.prd,
                      Onull=Obsnull,Enull=rep(Enull,length(Obsnull)),OoverE.null=Obsnull/Enull,
                      BC=OE.stats$BC.prd,BC.null=BC.null,
                      outlier.05=outlier.flag$outlier.05,outlier.01=outlier.flag$outlier.01,
                      row.names=row.names(bugall));
print('','quote=F);
print(' All predictions are finished',quote=F);
OE.final; #return data frame as final object;
}; #end of function;