

**CAPP 09-913-50**  
**Alternative Process for Developing Tier 2**  
**SSROs**

FINAL REPORT



Prepared for:  
**Petroleum Technology Alliance Canada**  
Suite 400, Chevron Plaza  
500-5<sup>th</sup> Avenue SW  
Calgary, Alberta T2P 3L5 Canada

Prepared by:  
**Stantec Consulting Ltd.**  
Suite 1 – 70 Southgate Drive  
Guelph, Ontario N1G 4P5 Canada

In association with:  
**Dr. Eric Lamb, University of Saskatchewan**  
51 Campus Dr.  
Saskatoon, Saskatchewan S7N 5A8 Canada

**Barry Zajdlik, Zajdlik & Associates**  
RR1, 5675 5<sup>th</sup> Line  
Rockwood, Ontario N0B 2K0 Canada

**Dr. Melissa Whitfield-Aslund**  
6605 Hurontario Street, Suite 500  
Mississauga, Ontario L5T 0A3 Canada

File 122160069  
September 9, 2013

**DISCLAIMER:** PTAC does not warrant or make any representations or claims as to the validity, accuracy, currency, timeliness, completeness or otherwise of the information contained in this report , nor shall it be liable or responsible for any claim or damage, direct, indirect, special, consequential or otherwise arising out of the interpretation, use or reliance upon, authorized or unauthorized, of such information.

The material and information in this report are being made available only under the conditions set out herein. PTAC reserves rights to the intellectual property presented in this report, which includes, but is not limited to, our copyrights, trademarks and corporate logos. No material from this report may be copied, reproduced, republished, uploaded, posted, transmitted or distributed in any way, unless otherwise indicated on this report, except for your own personal or internal company use.

## Executive Summary

---

The current Tier 2 pass/fail process described for *in situ* contamination in the Alberta remediation guidelines involves a post-remediation eco-toxicological assessment to demonstrate minimal risk to ecological receptors exposed via the soil direct contact pathway. Tier 2 site-specific remedial objectives (SSROs) may be established to assist in site management, but are not accepted for regulatory closure unless supported by additional lines of evidence. The objective of this project was to investigate the feasibility of four statistical approaches that might form the basis of an SSRO derivation procedure acceptable for regulatory closure. The impetus for the project came from conducting a carefully designed and robust ecotoxicological assessment of a site where some soil samples failed to satisfy the Tier 2 pass/fail criteria and some passed. The assessment generated many toxicological data (36 biological endpoints) that could be used to derive an SSRO. The initial challenge was to develop an acceptable approach for the SSRO derivation. A subsequent challenge was to determine if models describing the relationships among pedological characteristics, contaminant concentrations, and biological responses could be used successfully to predict “effects” in soil at other sites when the contaminant concentrations and pedological characteristics were known.

Four approaches were investigated. The defined GMR approach entailed calculating the geometric mean of the no-observable-adverse-effects concentration (NOAEC) and the lowest-observable-adverse-effects concentration (LOAEC) for each biological response variable (i.e., measurement endpoint). The distribution of the geometric means was used to determine threshold effect concentrations. The 25th percentile of the distribution of the ranked geometric means would provide fraction-specific remedial objectives for agricultural/residential land uses for these soils that would protect 75 % of the “species”, while the 50th percentile would provide the remedial objective for commercial/industrial land uses that would protect 50% of the “species”. Although there are different ways to generate the distributional data, one was selected in consideration of reliability, repeatability, uncertainty, and the degree of conservatism.

The remedial objectives derived using a distribution of the bounded geometric means for soils contaminated with residual PHCs (F3) were lower than current Tier 1 standards for F3 in soil, despite several of these soils passing a Tier 2 Pass/Fail Assessment. Thus, this approach for developing Tier 2 SSROs was not pursued further because the degree of conservatism in the approach remained high.

Some of the challenges to constructing an alternative Tier 2 process are due to the interactions between the physical and chemical characteristics of the site soils and the biological responses as well as among the site physical and chemical characteristics themselves. Therefore, a second approach (e.g., data reduction and model averaging or DRAMA approach) involved the exploration and critical evaluation of the data using a series of established statistical procedures to assess the relative importance of potential explanatory variables as well as the interaction and potential redundancy between and among site physical and chemical characteristics. The

latter was addressed through the creation of synthetic variables using ordination. Correlations between site physical/chemical characteristics, synthetic variables, contaminants and toxicity tests responses were explored using a suite of ecotoxicologically plausible model structures. Due to the nature of the exposure data, mixed effects models were used to account for sub-sample variation and the non-Gaussian distribution of many of the responses was addressed using generalized linear models. Rather than select a single “best” model, contributions from individual models were “averaged” to create a single model using model averaging. The advantages of this approach were that undue reliance was not placed on a single model and that the true model uncertainty was acknowledged rather than ignored. For the DRAMA approach, the dataset was explored and redundancy as measured by covariation among the non-PHC variables was examined using principal components analysis of the correlation matrix. Some variables were excluded and some retained. The matrix of retained explanatory variables was multiplied by the ordination eigenvectors to create site-specific scores for each principal component. The scores were considered for use as synthetic explanatory variables in lieu of some directly measured explanatory variables. The first, second and third principal components of the non-PHC variables measured consistently across three different studies comprising the preliminary dataset represented 85% of the total variability in the dataset. The ordination suggests using only PC1 scores as a heuristic for soil particle size and concomitants such as nutrients with the additional individual variables pH, soil moisture and clay as potential explanatory variables in subsequent modeling. Models were constructed for 17 toxicological endpoints. Three sets of additive linear models allowed for site effects, site and variable interactions and soil physico-chemical variables only. The model structures were driven by the following questions:

1. What is the relative importance of contaminant and non-contaminant heuristics as descriptors of toxicity, i.e., the biological responses?
2. In addition to the soil texture heuristic (PC1), clay, pH and moisture were flagged as major sources of variability in the studies examined. Are these three variables important descriptors of toxicity?
3. Does the relative importance of PHCs, clay and non-contaminant heuristics vary by study?

The relative importance of variables was assessed using model averaged coefficients and p-values for Wald tests of significance for each parameter within a model. The relative importance of variables was also assessed by summing the AICc (second order Aikake’s Information Criterion for small sample sizes) weights for a model in which a parameter occurs.

The parameters of the models were averaged by weighting the parameters of multiple models fit to the same data using the model-specific AICc rather than shrinkage estimators which include zeros for parameters in models where the parameter does not appear. This latter procedure was avoided because the degree of shrinkage is a function of the model structures considered for model averaging. Unconditional variance estimates were used in Wald tests ( $H_0$ : parameter = 0) of model averaged parameter estimates. A measure of relative variable importance was measured as the sum of AICc weights over all models including the explanatory variable.

Because the studies used for this feasibility assessment were conducted for other purposes, test species changed from study to study and the pedological variables that were measured were inconsistent among studies. This reduced the number of pedological variables and test species that could be evaluated. Generally, it was concluded that the dataset would benefit from improvement to the study and experimental design.

A third approach, the Partial Least Squares (PLS) Regression approach, was investigated because it had been used previously with success to investigate large “noisy” datasets in order to identify and assess the existing signals. Models predicting various ecotoxicological endpoints resulting from exposure of ecological receptors to contaminated soil were developed, by combining multiple predictors in the same model through the application of multivariate statistical methods. Since multivariate statistics consider many variables simultaneously, they can detect meaningful trends that might not be identified by traditional univariate analyses. PLS regression is a multivariate statistical method that was used to model the relationship between multivariate predictor matrix (X) and a response matrix (Y), which could include either single or multiple responses. Analogous to simple linear regression models, PLS provides an assessment of the strength of the relationship between X and Y (i.e., the percent of variation in Y that can be explained in terms of the variation of X), and can also be used as a foundation for predicting the “Y values” of future unknown observations based on their known X data (which can be measured). By using soil physico-chemical properties, non-exhaustive chemical extraction results, and measured bioaccumulation values in the X matrix and either individual toxicity endpoints or a matrix of multiple toxicological endpoints as the Y matrix, a multivariate model was constructed that is capable of predicting the relative toxicity of various soils to key ecological receptors based on purely physico-chemical measurements. The SSROs for a site could be derived then based on the distribution of the predicted relative toxicities.

The matrix analyses comprising data from only one site with PHC- and metals-contamination used pedological characteristics as the ‘X’ matrix of multiple predictors and each of the biological responses (endpoints) as a separate ‘Y’ matrix. PLS models were cross validated using leave-one-class-out cross validation (LOCOCV) and the number of components that maximized the internally cross-validated  $R^2Y$  value (reported as  $Q^2Y$ ) was selected as the number of components for each final PLS model. For each PLS model, the explained variation of X and Y ( $R^2X$  and  $R^2Y$ ) were reported to indicate how well the model fit the training data and  $Q^2Y$  was reported as a preliminary measure of the predictive ability of the model. In addition, the significance of each PLS model was estimated through response permutation testing. Statistical significance was assessed at  $\alpha \leq 0.05$ .

Significant models were created for several endpoints; however, for a number of models the significance of the model was generally fairly low. That said, it was clear that the non-contaminant variables were either more, or as, important as explanatory variables as the contaminant variables and that PHCs were important explanatory variables in all of these models which corroborated the results of the DRAMA and SEM approach.

The results of the PLS approach demonstrated that it was possible to link multivariate soil properties to certain ecotoxicity endpoints. However, the analyses also highlighted that the predictive power of these models is likely to be inadequate for soils with soil properties that vary substantially from the soils used to build the initial model. The small sample size might be responsible. Predictive power of the models might be improved by increasing the number of site soils in the model-building exercise and model averaging might strengthen the cross-site predictive applicabilities.

Structural Equation Modeling (SEM) was the fourth approach investigated as a potential alternative Tier 2 process. Data for multiple species and endpoints from toxicity tests were incorporated into a single analysis through use of a latent variable and subsequent EC/IC25 and EC/IC50 values were estimated; models were developed for predicting toxicological responses that incorporated both contaminant levels and environmental covariates. The focus of this investigation was the relationships among endpoint responses, as represented by an “aggregate response” latent variable; and, constructing structural models to describe the causal relationships among the non-contaminant and contaminant variables in the model. Upon construction of the models, based on these relationships, covariate models were derived in order to predict “effects” or “impacts” to ecological receptors for sites for which toxicity data were either minimal or lacking. Cross-site models were investigated with the intent to implement them as predictive tools.

SEM has two components, the measurement model and the structural model. The structural model consists of the paths between variables, while the measurement model consists of a latent variable and its associated observed indicator variable(s). Latent variable modeling has two major advantages: 1) it can be used to estimate the general species response across a range of toxicant concentrations; and 2) it allows estimates of measurement error to be incorporated into the model and measurement error is implicitly included as imperfect correlations among the indicator variables. Measurement error is rarely explicitly considered, yet is nearly always present to some extent in data. In the model fitting process unacknowledged measurement error can cause problems in the estimation of path coefficients. For example, if measurement error is present in an explanatory variable, the residual error variance will contain both prediction error and measurement error, and as a result the true strength of the relationship between the response and explanatory variables will be underestimated. This underestimation of the true strength of the relationship can cause a downward bias in both the unstandardized and standardized estimates of path coefficients in the structural model. The structural (path) model describes the causal relationships among the variables in a model. The structural model consists of either the paths between latent variables or, in an observed variable model, direct relationships among observed variables.

The utility of the SEM approach showed promise; most of the challenges encountered were related to the size of the preliminary dataset. This was not surprising given the “data hungry” nature of the approach. The major outcome of this investigation was the prospective use of confirmatory factor analysis to aggregate multiple endpoints into a single latent variable that could then be incorporated into standard non-linear modeling procedures to estimate IC25

values. This provides a direct solution to the problem of reconciling divergent ICp estimates from individual endpoints. In particular, the confirmatory factor analysis was uniquely able to identify endpoints that may not be responding in the same manner as the majority (variables with weak and/or non-significant loadings on the latent variable). With this approach the toxicologist can determine with confidence whether all of the endpoints are providing equivalent information and, if so, develop a single IC25 estimate from the latent variable using standard procedures. The overall goal of this project was to develop analytical methods that could incorporate environmental covariates into analyses of toxicological responses and to develop cross-site predictive models that could be used to estimate provisional remediation targets based on readily measured environmental variables. Models were constructed that linked an aggregate species response variable based on two earthworm endpoints, two collembolan endpoints, and four northern wheatgrass endpoints to toxicant concentrations and measures of soil quality. These cross-site models are promising, but not ready for implementation in a predictive mode. The models successfully explained the aggregate species responses ( $R^2 > 0.7$ ), but failed many tests of model adequacy (significant  $\chi^2$ , low CFI etc.). A small sample size relative to the complexity of the models was a major impediment to the implementation of these models.

Recommendations to improve the validity of three of the Tier 2 alternative approaches were made. Recommendations for standardization of the choice of toxicity endpoints and environmental co-variables were made that would improve the utility of these data for these types of assessments. Modifications were recommended to optimize sampling designs to improve the utility of the predictive models.

## **ACKNOWLEDGEMENTS**

This research was supported by an NSERC Engage grant to Dr. Eric Lamb with support from the R&D Fund from Stantec Consulting Ltd. The research was part of a collaborative initiative (Stantec Consulting Ltd., Guelph, ON; Zajdlik & Associates, Rockwood, ON, Intrinsic Environmental Sciences Inc., Mississauga, ON; and, University of Saskatchewan, Saskatoon, SK) facilitated by the Petroleum Technology Alliance Canada (PTAC) through their Alberta Upstream Petroleum Research Fund (AUPRF). The project was funded by the Canadian Association of Petroleum Producers (CAPP), the Program for Energy Research and Development (PERD) of Environment Canada, the National Science and Engineering Research Council, and Stantec Consulting Ltd. (R&D Fund). We thank Gordon Dinwoodie for providing comments on the initial draft report and Kathryn Bessie for technical input.



---

## Table of Contents

EXECUTIVE SUMMARY .....	E.1
<hr/>	
<b>1.0 INTRODUCTION .....</b>	<b>1.1</b>
1.1 BACKGROUND .....	1.1
1.2 OBJECTIVE(S) .....	1.2
1.3 SCOPE OF REPORT .....	1.2
<hr/>	
<b>2.0 OVERVIEW OF APPROACHES INVESTIGATED FOR THE DERIVATION OF TIER 2 SSROS .....</b>	<b>2.1</b>
2.1 GMR APPROACH: DISTRIBUTION OF THE GEOMETRIC MEANS OF THE NOAECS <sup>1</sup> AND LOAECS <sup>2</sup> .....	2.1
2.2 DRAMA APPROACH: DATA REDUCTION AND MODEL AVERAGING .....	2.1
2.3 PLS APPROACH: PARTIAL LEAST SQUARE REGRESSION .....	2.3
2.4 SEM APPROACH: STRUCTURAL EQUATION MODELING .....	2.3
<hr/>	
<b>3.0 GMR APPROACH: DISTRIBUTION OF THE GEOMETRIC MEANS OF THE NOAECS AND LOAECS .....</b>	<b>3.1</b>
3.1 RATIONALE .....	3.1
3.2 MATERIALS AND METHODS .....	3.2
3.3 RESULTS AND DISCUSSION .....	3.2
3.4 CONCLUSIONS .....	3.3
<hr/>	
<b>4.0 DRAMA APPROACH: DATA REDUCTION AND MODELING AVERAGING .....</b>	<b>4.1</b>
4.1 RATIONALE .....	4.1
4.2 MATERIALS AND METHODS .....	4.1
4.2.1 Available Data and Data Manipulation .....	4.1
4.2.2 Data Analyses .....	4.2
4.3 RESULTS AND DISCUSSION .....	4.5
4.3.1 Data Exploration and Reduction .....	4.5
4.3.2 Model Averaging .....	4.9
<hr/>	
<b>5.0 PLS APPROACH: PARTIAL LEAST SQUARES REGRESSION .....</b>	<b>5.1</b>
5.1 RATIONALE .....	5.1
5.2 MATERIALS AND METHODS .....	5.1
5.3 RESULTS AND DISCUSSION .....	5.2
5.3.1 PLS Predictions of the Number of Earthworm Progeny Produced .....	5.2
5.3.2 PLS Predictions of the Dry Mass of Earthworm Progeny .....	5.5
5.3.3 PLS Prediction of Organism Responses .....	5.6
5.4 CONCLUSIONS .....	5.18
<hr/>	
<b>6.0 SEM APPROACH: STRUCTURAL EQUATION MODELING .....</b>	<b>6.1</b>
6.1 INTRODUCTION AND RATIONALE .....	6.1



---

6.2	MATERIALS AND METHODS .....	6.7
6.2.1	Confirmatory Factor Analysis and Aggregation of Multiple Endpoints .....	6.7
6.2.2	Estimation of IC25 and IC50 Values from a Latent Variable .....	6.12
6.2.3	Structural Equation Modeling.....	6.15
6.2.4	Modeling of Toxicological Data – Cross Site Field Data.....	6.16
6.2.5	Models Incorporating Contaminant Levels and Multiple Environmental Predictors .....	6.18
6.3	CONCLUSIONS.....	6.20
6.3.1	Utility of SEM for Toxicological Data.....	6.20
<hr/>		
7.0	RECOMMENDATIONS.....	7.1
7.1	GMR APPROACH.....	7.1
7.2	DRAMA APPROACH .....	7.1
7.3	PLS APPROACH .....	7.1
7.4	SEM APPROACH .....	7.1
7.5	RECOMMENDED MODIFICATIONS FOR DATA COLLECTION.....	7.2
7.5.1	Choice of Endpoints and Environmental Covariates .....	7.2
7.5.2	Sampling Design .....	7.2
7.6	RECOMMENDATIONS FOR FUTURE WORK .....	7.3
<hr/>		
8.0	SIGN-OFF.....	8.1
<hr/>		
9.0	CITED REFERENCES.....	9.1

---

## LIST OF FIGURES

---

Figure 3.1:	F3 Geomean of the NOEC/LOECs for all site soils and endpoints relative to the Control.....	3.3
Figure 4.1:	Loadings Plot for First Principal Component using Consistently Measured Non-PHC Variables. ....	4.6
Figure 4.2:	Loadings Plot for Second Principal Component, using Consistently Measured Non-PHC Variables.....	4.7
Figure 4.3:	Loadings Plot for Third Principal Component using Consistently Measured Non-PHC Variables .....	4.8
Figure 5.1:	Actual number of progeny produced plotted against predictions for each class based on ‘leave one class out’ cross validation. ....	5.3
Figure 5.2:	Actual number of progeny produced plotted against predictions for each class based on ‘leave one class out’ cross validation. ....	5.4
Figure 5.3:	Actual progeny dry mass produced plotted against predictions for each class based on ‘leave one class out’ cross validation. ....	5.5
Figure 6.1:	An example of a latent variable with multiple indicators. ....	6.2
Figure 6.2:	Initial measurement models for Study 4. ....	6.9

---

Figure 6.3:	Logistic (panels a and b) and exponential (panels c and d) models showing the relationship between aggregate species responses and spiked (a and b) and observed (c and d) contaminant levels. ....	6.13
Figure 6.4:	Cumulative distribution of IC25 values against contaminant concentration. ....	6.15
Figure 6.5:	Initial structural equation models with hydrocarbon contamination as predictors of the aggregate species response. ....	6.16
Figure 6.6:	Confirmatory factor analysis Model J for the cross-site data. ....	6.17
Figure 6.7:	Model L - a structural equation model relating species responses to environmental covariates. ....	6.18

---

## LIST OF TABLES

---

Table 3.1:	CCME reference method Hexane:Acetone extracted PHC concentrations at test setup reported as mean mg/kg dry weight in soil $\pm$ one standard deviation of the mean (n=3) and Tier 2 Pass/Fail Assessment results. ....	3.1
Table 4.1:	Suite of Linear Models Considered for Model Averaging.....	4.4
Table 4.2:	Details of Measurement Response - Specific Models Used .....	4.9
Table 4.3:	Summary of the Model Averaging Results .....	4.9
Table 4.4:	Wald Tests of Significance for Model Averaged Parameters.....	4.12
Table 4.5:	Relative Importance of Explanatory Variables.....	4.14
Table 5.1:	VIP values for PLS model predicting earthworm number of progeny using 'Study 3' results with soil 4-3 excluded.....	5.4
Table 5.2:	VIP values for PLS model predicting earthworm progeny dry mass using 'Study 3' results with soil 4-3 excluded. ....	5.6
Table 5.3:	Parameters of fitted PLS models using multivariate soil property information* ...	5.7
Table 5.4:	Average predictions of y-values ( $y_i$ ) for earthworm endpoints .....	5.8
Table 5.5:	Average predictions of y-values ( $y_i$ ) for springtail endpoints*.....	5.10
Table 5.6:	Average predictions of y-values ( $y_i$ ) for plant endpoints* .....	5.11
Table 5.7:	Top 20 VIP values for PLS model predicting RC root dry mass using 'Study 3' results (full dataset).....	5.16
Table 5.8:	Top 20 VIP values for PLS model predicting RC shoot dry mass using 'Study 3' results (full dataset).....	5.17
Table 5.9:	Top 20 VIP values for PLS model predicting RC shoot length using 'Study 3' results (full dataset).....	5.17
Table 6.1:	Example Variance – Covariance Matrix .....	6.5
Table 6.2:	Commonly Used Structural Equation Model Fit Indices <sup>1</sup> .....	6.6
Table 6.3:	Measurement Model Results for Study 4 .....	6.10
Table 6.4:	Full results for Model C including unstandardized path coefficients (col. 2), standard error of the unstandardized coefficients (col. 3), ratio of the unstandardized estimate and standard error (col. 4), test of path significance (col. 5), and standardized path coefficient estimates (col. 6) for Model C .....	6.10
Table 6.5:	Estimated IC25 and IC50 values and 95% confidence intervals (CI) for the four nonlinear models fit.....	6.13
Table 6.6:	Measurement model results for the combined site data .....	6.17
Table 6.7:	Results of a cross site models incorporating environmental covariates with (Model M) and without (Model N) three multivariate outlying data points. ....	6.19

Table 6.8:	Measurement model paths (paths between latent and composite variables and their indicators) for Model N .....	6.19
Table 6.9:	Structural model paths in Model N .....	6.20

---

## LIST OF APPENDICES

---

Appendix A:	Final DRAMA Approach Report (B. Zajdlik, 2013)
Appendix B:	Final PLS Approach Report (Dr. M. Whitfield-Aslund, 2012)
Appendix C:	SEM Approach Report (Dr. E. Lamb <i>et al.</i> , 2012)

## 1.0 Introduction

---

### 1.1 BACKGROUND

In Canada, soils contaminated with petroleum hydrocarbons are managed on the basis of four hydrocarbon fractions (Canadian Council of Ministers of the Environment [CCME], 2006). Tier 1 Canada-wide soil standards for each fraction protect ecological receptors exposed via the direct contact exposure pathway. If these standards are exceeded by fraction-specific concentrations in soil, then the proponent has the option to conduct a Tier 2 assessment to demonstrate that 1) the exposure pathway can be excluded; 2) PHC residuals are stable and represent minimal risk to soil organisms; or 3) data generated can be used to derive site-specific remedial objectives (SSRO) (Alberta Environment, 2010). In practice and in Alberta, Tier 2 ecotoxicological assessments are conducted primarily to demonstrate that PHC residuals are stable and represent minimal risk to soil organisms. The data from the toxicity assessment must satisfy criteria established for different land-use classes. This is called the Tier 2 Pass/Fail approach (Alberta Environment, 2007). If a site passes, then no further remediation or action is required.

However, if the site soils fail to satisfy the criteria, the proponent must select management alternatives to mitigate risk or conduct an risk assessment using weight of evidence. Alternatively, the data generated from the Tier 2 ecotoxicity assessment can be used to derive site-specific remedial objectives (SSROs) to guide future remediation. The current challenge facing regulators, assessors, and managers alike is the lack of a framework or process for the derivation of these Tier 2 SSROs.

Currently, there is a provision in the draft Alberta Tier 2 guidance document for development of a Tier 2 site-specific remedial objective(s) (SSROs) for PHCs in soil to guide remediation; however, to our knowledge and in practice, only the pass/fail approach applied to remediated PHC-contaminated soils has been used to date. The current pass/fail process involves a post-remediation eco-toxicity assessment to demonstrate minimal risk to ecological receptors via the soil contact exposure pathway for *in situ* contamination. Although there is a provision in the guidelines for deriving SSROs at Tier 2, there is no process for deriving these clean-up values.

Depending on the size and complexity of a site, members of the oil and gas industry can pay up to \$160K for an eco-toxicity assessment of a remediated site with PHC contamination. Should the eco-toxicity assessment indicate that the soils on the site do not satisfy the "pass/fail" criteria (Alberta Environment, 2007) further remedial activities are required for those soils with the hope that they will then pass a second eco-toxicity assessment. Alternatively, a proponent might elect to conduct a risk assessment and assess the relative risk of exposure to PHCs in soil for ecological receptors and develop risk-based remedial objectives (Tier 3).

A typical Tier 2 eco-toxicity assessment of remediated PHC-contaminated site soils generates between 11 (minimum requirement) and 36 endpoints, depending primarily on the number of species and site soil samples tested, from which an assessment can be made. Should the results of the assessment fail to satisfy the Tier 2 pass criteria; there remains in hand a sufficient amount of information and data from which we believe a remedial objective can be derived.

The aim of this project was to investigate different potential methods for developing Tier 2 site-specific remedial objectives. These Tier 2 SSROs would serve as clean-up values that provide the same level of protection as the Tier 1 standards, but they would be derived on a site-specific basis. Site-specific conditions and weathering and aging processes influence the bioavailability of the PHCs in soil to ecological receptors. Conceivably, the site-specific remedial objectives at Tier 2, where there is historical contamination of site soils, would be less stringent, providing relief from forced clean up to the Tier 1 standards. It is important to note that Tier 1 standards initially were never intended to serve as clean-up standards. The Tier 2 SSRO derivation process under development should overcome some of the current limitations inherent in the current Tier 2 pass/fail approach and reduce remedial costs. If accepted and effective, the process could impact the current regulatory remediation guidelines.

## 1.2 OBJECTIVE(S)

The specific objectives of this research project are to: 1) investigate three or four different approaches that could be used to develop a site-specific remedial objective for lands contaminated with petroleum hydrocarbons; and 2) develop the framework for using data generated from a Tier 2 ecotoxicological assessment to derive the SSRO(s).

The project will address, in part, policy data gaps associated with lack of closure mechanism for risk-managed sites. In addition, it will address risk assessment cost reduction and mitigate remediation costs. More specifically, the result should be the development of a practical and pragmatic approach to using site-specific data for deriving Tier 2 SSROs that are demonstrably as protective of ecological receptors exposed to PHCs in soil via the direct contact exposure pathway as the Tier 1 standards.

## 1.3 SCOPE OF REPORT

**Section 1.0** of this report provides the background information and rationale for the project as well as the aims and objectives of the project. Four potential approaches were investigated for the derivation of Tier 2 SSROs; a general description of each of these approaches is provided in **Section 2.0**. **Section 3.0** provides the methodology and detailed description of the GeoMean Response (GMR) approach with a case study provided as an example. The subsequent chapters (**Sections 4.0, 5.0 and 6.0**) provide detailed descriptions of each of the other three approaches (DRAMA – Data Reduction and Model Averaging; PLS – Partial Least Squares Regression; and SEM – Structural Equation Modeling, respectively) including the methodology that was used to develop each approach, investigations into the utility of the method or process,

conclusions and recommendations. **Section 7.0** summarizes the findings to date and the recommendations for future investigations in year 2 of this project. The references (**Section 8.0**) are followed by appendices. The appendices comprise the detailed reports for three of the statistical approaches that were investigated for developing Tier 2 SSROs.

## 2.0 Overview of Approaches Investigated for the Derivation of Tier 2 SSROs

---

### 2.1 GMR APPROACH: DISTRIBUTION OF THE GEOMETRIC MEANS OF THE NOAECs<sup>1</sup> AND LOAECs<sup>2</sup>

An alternative process for deriving SSROs is necessary in order to accurately assess sites that fail to satisfy the criteria for Alberta Environment's Tier 2 pass/fail approach. One of the possible alternative approaches for deriving soil clean-up criteria is the use of species sensitivity distributions (SSDs) generated with data derived from the "failed" Tier 2 pass/fail ecotoxicity assessment. Because site soil contamination levels and toxicity test results are often not amenable to regression analyses, we determined the endpoint-specific and species-specific NOAECs<sup>\*</sup> and LOAECs<sup>†</sup> for the site soils using the toxicity data generated from the Tier 2 ecotoxicity assessment. The geometric mean of these responses (GMR) was then calculated for each bounded NOAEC-LOAEC combination for each endpoint and each species. The geometric means were ranked and the distribution of the ranks plotted to generate a sensitivity distribution. The NOAEC/LOAECs are derived from statistical comparison to a reference control site soil. The 25th percentile of the distribution of the ranked geometric means would provide fraction-specific remedial objectives for agricultural/residential land uses for these soils, while the 50th percentile would provide the remedial objective for commercial/industrial land uses. Although there are different ways to generate the distributional data, one was selected in consideration of reliability, repeatability, uncertainty, and the degree of conservatism. The methods and procedures for implementing this approach are outlined in detail in **Section 3.0**; a summary with recommendations was included.

### 2.2 DRAMA APPROACH: DATA REDUCTION AND MODEL AVERAGING

A second alternative process is also proposed for soils where no clear monotonic exposure concentration-response relationship is observed for the biological responses to PHCs in soil. A lack of monotonic response to the contaminant(s) of concern (COCs) is often due to the influence of confounding factors such as soil texture, organic carbon, organic matter content, cation exchange capacity, etc. As a suite, these non-contaminant variables generally co-occur across sample locations at a given site, and again, as a suite will induce "noise" or variability in the "signal" (biological responses). The influence of this non-contaminant signal can in many cases be filtered using a combination of multivariate data exploration, reduction and regression methodologies, thus improving the clarity of the biological response signal to the point where an exposure concentration-response relationship is observed (Renoux *et al.*, 2012).

The statistical nature of the soil toxicity test response variables almost always precludes the defensible use of typical linear regression methodologies under the assumptions of normality

---

<sup>\*</sup> No observable adverse effect concentration(s)

<sup>†</sup> Lowest observable adverse effect concentration(s)



and homogeneity of variance. If challenged, these methodologies and any conclusions based thereon will fail. Thus, alternative regression models (generalized linear models) that defensibly address these issues were used to model the relationships between putative COCs and the biological responses after suitable filtering (i.e., data reduction).

Ordination analysis was used to explore potential explanatory variables for each of the biological responses (e.g. each measurement endpoint). Correlations within the different classes of analytes (e.g. chemical and non-contaminant and physical pedological variables, PAHs, PHCs) were examined. Rank-based Spearman correlations were used to obviate the assumption of bivariate normality implicit to the Pearson product-moment correlation. Finally, a modeling approach described below was used to distinguish between the strong non-contaminant structure in the data set and an exposure-concentration response with contaminants. The methods and procedures employed sought a balance between the purpose of the data modeling (i.e., to test whether the toxicological responses were associated with petroleum hydrocarbons) and challenges inherent to the available data. These included the ratio of putative explanatory variables relative to the number of experimental units; the co-linearity of the soil quality parameters, the distribution of the biological response variation within and among soils, and the presence of non-contaminant structure that clearly identifies the differences in soil samples on the basis of physico-chemical characteristics. Reduction of data dimensionality was undertaken through the creation of heuristic synthetic variables using ordination procedures. Further reduction was achieved by examining rank correlation between members of a class of analytes (e.g. physico-chemical non-contaminant and pedological variables). The models for hypothesis testing of the reduced data set were mixed-effects models for those responses approximately normally or transformably normally distributed and generalized linear models for non-normal responses. Because the purpose of the models was to test whether the toxicological responses were associated with petroleum hydrocarbons, models were built according to the following principles:

- Marginally useful predictors might be retained;
- Variables that exhibit undue influence might be preferentially selected for manual exclusion;
- Known toxicological modifying variables might be retained if only marginally significant; and
- Model distributional assumptions were emphasized.

Rather than identifying a single “best model”, model averaging procedures were applied in recognition of the contributions from individual models and to address model uncertainty. Given a pre-defined unacceptable biological effect level, fitted models that meet to-be-defined criteria for adequacy can be used to estimate SSROs at sites for which chemistry data but no biological data are available. These predictions may be used to delineate areas requiring remediation / cleanup. Statistical techniques could also be used to ensure a specified level of confidence that the excluded areas achieve the SSROs, but this is envisioned as beyond scope of this initiative. Should the results of this project require validation or verification, such an initiative would be

invaluable. The methods and procedures are summarized in **Section 4.0** and detailed in the report of **Appendix A**.

## 2.3 PLS APPROACH: PARTIAL LEAST SQUARE REGRESSION

Models predicting various ecotoxicological endpoints resulting from exposure of ecological receptors to contaminated soil were developed, by combining multiple predictors in the same model through the application of multivariate statistical methods. Since multivariate statistics consider many variables simultaneously, they can detect meaningful trends that may not be identified by traditional univariate analyses. The objective of this investigation was to develop such a model or models using partial least squares (PLS) regression procedures. PLS is a multivariate statistical method that can be used to model the relationship between multivariate predictor matrix (X) and a response matrix (Y), which could include either single or multiple responses. Analogous to simple linear regression models, PLS provides an assessment of the strength of the relationship between X and Y (i.e., the percent of variation in Y that can be explained in terms of the variation of X), and can also be used as a foundation for predicting the “Y values” of future unknown observations based on their known X data (which can be measured). By using soil physico-chemical properties, non-exhaustive chemical extraction results, and measured bioaccumulation values in the X matrix and either individual toxicity endpoints or a matrix of multiple toxicity results as the Y matrix, a multivariate model could be constructed that is capable of predicting the relative toxicity of various soils to key ecological receptors based on purely physico-chemical measurements. SSROs can be derived then based on the distribution of the predicted relative toxicities. The methods and procedures are summarized in **Section 5.0** and detailed in the report of **Appendix B**.

## 2.4 SEM APPROACH: STRUCTURAL EQUATION MODELING

Structural Equation Modeling (SEM) is a potential solution for many of the problems encountered in the analysis of site-specific toxicological data. The inter-correlated environmental variables that are problematic in the current methods used to develop SSROs are readily incorporated into a SEM framework (Grace, 2006; Kline, 2011; Lamb *et al.*, 2011). Further, SEM provides a natural way to incorporate data for multiple species and endpoints from toxicity tests into a single analysis through use of a latent variable (a general concept that is indirectly measured through observation of correlated variables). In toxicity testing, the multiple species and endpoints are effectively indirect measures of the formally unmeasured concept “toxicity” and hence ideal for analysis as a latent variable. Finally, SEM can incorporate measurement error (Grace, 2006; Kline, 2011; Lamb *et al.*, 2011) and thus account for variability in replicate toxicity tests with the same soils.

The application of SEM to a range of toxicological datasets was investigated to determine how SEM can be used to aggregate multiple species endpoints into a single synthetic variable that can then be used in standard nonlinear regression modeling to estimate IC25\* and IC50\* values;

---

\* Inhibitory concentration affecting emergence, growth, or reproduction by 25 percent

and, to develop models of toxicological responses that incorporate both contaminant levels and environmental covariates. This second approach is a critical step toward the development of effective cross-site predictive models of toxicological responses.

Structural equation models have two components, the measurement model and the structural model. The structural model consists of the paths between variables, while the measurement model consists of a latent variable and its associated observed indicator variables. A latent variable represents a concept or quantity that has not been measured directly, but is rather indicated indirectly through one or more observed variables presumed to be highly correlated with the latent variable. The focus of this investigation was the relationships among endpoint responses as represented by an “aggregate response” latent variable with its associated measurement error and the structural model, which describes the causal relationships among the variables in the model. Upon construction of the models, based on these relationships, covariate models were derived in order to predict “effects” or “impacts” to ecological receptors for sites for which toxicity data were either minimal or lacking. Cross-site models were investigated with the intent to implement them as predictive models. The methods and procedures are summarized in **Section 6.0** and detailed in the report of **Appendix C**.

---

\* Inhibitory concentration affecting emergence, growth, or reproduction by 50 percent

### 3.0 GMR Approach: Distribution of the Geometric Means of the NOAECs and LOAECs

#### 3.1 RATIONALE

Often for Tier 2 pass/fail assessments, the results of an ecotoxicity assessment do not monotonically correspond to the contaminant levels in the soils. In other words, the intensity of effects does not increase with increasing concentration. For one such site, a Tier 2 Pass/Fail assessment revealed high variability in the pass/fail results. Additionally, the soils that failed to satisfy the Tier 2 criteria included those with F3 concentrations below the Tier 1 soil standards for commercial/industrial land uses (**Table 3.1**). The site soils were surface soil samples collected as composite samples from the site and were considered to be representative of the potential hydrocarbon contamination across the site. An alternative approach using species sensitivity distributions of the NOAECs and LOAECs was used to derive soil remediation criteria for this site for F3, and the NOEC/LOECs were derived from statistical comparison with the Control results.

**Table 3.1: CCME reference method Hexane:Acetone extracted PHC concentrations at test setup reported as mean mg/kg dry weight in soil  $\pm$  one standard deviation of the mean (n=3) and Tier 2 Pass/Fail Assessment results.**

Guideline Values	F1 (mg/kg)	F2 (mg/kg)	F3 (mg/kg)	F4 (mg/kg)	--
Com/Ind Fine surface soil (AENV, 2010)	320	260	2500	6600	--
Res/Ag Fine surface soil (AENV, 2010)	210	150	1300	5600	--
Soil Type	--	--	--	--	Tier 2 Pass / Fail Result
Soil 1-1	14 $\pm$ 6	283 $\pm$ 23	1967 $\pm$ 306	1020 $\pm$ 164	Fail
Soil 1-2	<10	220 $\pm$ 44	1500 $\pm$ 436	780 $\pm$ 192	
Soil 1-3	245 $\pm$ 332	283 $\pm$ 90	1933 $\pm$ 551	893 $\pm$ 240	
Soil 2-1	317 $\pm$ 38	2233 $\pm$ 513	2233 $\pm$ 551	340 $\pm$ 123	Pass
Soil 2-2	237 $\pm$ 23	2067 $\pm$ 451	2100 $\pm$ 500	310 $\pm$ 130	
Soil 2-3	273 $\pm$ 42	2733 $\pm$ 153	2933 $\pm$ 115	567 $\pm$ 196	
Soil 3-1	<10	207 $\pm$ 12	917 $\pm$ 59	447 $\pm$ 67	Pass
Soil 3-2	<10	513 $\pm$ 51	1367 $\pm$ 115	527 $\pm$ 124	
Soil 3-3	<10	160 $\pm$ 46	717 $\pm$ 190	360 $\pm$ 92	

The derivation process followed the precedent set by the 2006 Canadian Council of Ministers of the Environment (CCME, 2006) protocol, which used rank species sensitivity analysis. The approach used here was a modified version of the CCME approach for deriving the Tier 1 standards for PHCs in surface soils. For the current CCME guidelines, EC/IC25s (effect or inhibitory concentration either affecting 25% of the test species or resulting in an inhibitory effect of 25% relative to the control) for the various species were used to generate species-sensitivity

**ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS**

GMR Approach: Distribution of the Geometric Means of the NOAECs and LOAECs

September 9, 2013

---

distributions (SSDs), from which the direct soil contact values for ecological receptors were derived for the different land-use classifications. For this analysis, the geometric mean was calculated and used to combine redundant endpoints (single endpoint wet and dry weights). Regression procedures were applied to the ranks, and the 25th percentile was used to derive soil contact values for agricultural/residential land-use areas; the 50th percentile was used for commercial/industrial land-use areas. In order to meet the Weight of Evidence method outlined by the CCME, a dataset is required to have  $\geq 10$  data points and  $\geq 2$  plant and 2 invertebrate taxa, and  $\geq 3$  studies.

The present approach modified the CCME precedent such that data from ecotoxicological assessments with site soils could be used to construct a SSD that would result in applicable remedial objectives for a site. Contaminated site soils are generally not amenable to the regression analysis required for E/IC25 calculation as tests with site soils frequently have PHC exposure concentrations in a narrow range, elicit a narrow range of toxicological responses (i.e., effects), and a wide range of physical-chemical characteristics. It was hypothesized that, by determining the NOAECs and LOAECs for contaminant(s) of concern in these soils (F3) and then calculating the geometric mean of the paired NOAECs and LOAECs, a SSRO could be derived from an SSD constructed from the distribution of the geometric mean values for each endpoint. The SSRO would retain a degree of conservatism yet reflect the reduced risk (represented by the site soils passing the Tier 2 Pass/Fail Assessment) of these residual PHCs in the site soils.

### **3.2 MATERIALS AND METHODS**

Data from a Tier 2 ecotoxicity assessment with 4 plant and 2 invertebrate species, and 8 site soil samples and 3 reference control soils failed to satisfy the Tier 2 pass/fail criteria. This data set was used to assess the efficacy of the GMR approach. NOAEC/LOAECs for each endpoint and F3 were determined for all site soils and endpoints relative to the control test results; the geometric mean of both the NOAECs and the LOAECs was determined for each toxicity test endpoint using a database comprised of the results for all tested site soils. Then, the geometric means for each geomean-NOAEC combination for all soils and geomean-LOAEC combination for all soils were determined and plotted by rank sensitivity. The 25th percentile (1153 mg F3/kg) and 50th percentile (1405 mg F3/kg) were determined; the 25th percentile provides the remedial objective for agricultural/residential land uses for F3 for these soils, while the 50th percentile provides the remedial objective for commercial/industrial land uses.

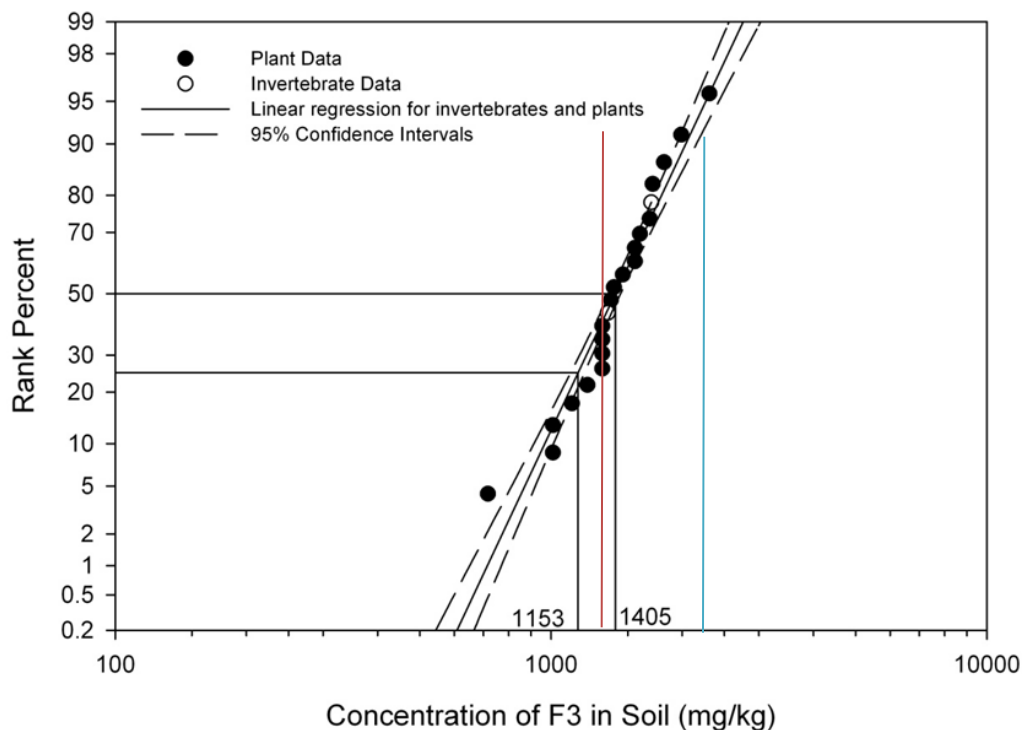
### **3.3 RESULTS AND DISCUSSION**

The species sensitivity distribution of the geometric means is presented in **Figure 3.1**.

**ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS**

GMR Approach: Distribution of the Geometric Means of the NOAECs and LOAECs

September 9, 2013



**Figure 3.1: F3 Geomean of the NOEC/LOECs for all site soils and endpoints relative to the Control.** The 25th percentile (1,153 mg F3/kg) and 50th percentile (1,405 mg F3/kg) were determined; the 25th percentile provides the remedial objective for agricultural/residential land uses for F3 for these soils, while the 50th percentile provides the remedial objective for commercial/industrial land uses. The Tier 1 standard for residential/agricultural lands is represented by the vertical red line and that for commercial/industrial is represented by the vertical blue line.

The potential Tier 2 SSROs for this site using this statistical approach would be 1,153 mg/kg for agricultural and residential areas and 1,405 mg/kg in soil for commercial and industrial areas. These soil standards are more restrictive than current Tier 1 CWS for PHC fractions in soil based on the ecotoxicological data for soil receptors exposed to F3 in soil (**Table 3.1**).

### 3.4 CONCLUSIONS

The remedial objectives derived using a species sensitivity distribution of toxicity test data for soils contaminated with residual PHCs (F3) were lower than current Tier 1 standards for F3 in soil, despite several of these soils passing a Tier 2 Pass/Fail Assessment. Thus, this approach for developing Tier 2 SSROs was not pursued further because the degree of conservatism in the approach remained high.

## **4.0 DRAMA Approach: Data Reduction and Modeling Averaging**

---

### **4.1 RATIONALE**

Ecotoxicological assessment of a site may be used to generate site-specific remedial objectives (SSROs). However the current challenge facing regulators, assessors, and managers alike is the lack of a framework or process for the derivation of these Tier II SSROs. Some of the challenges to constructing a process are due to the interactions between site physical and chemical characteristics and toxicity test responses as well as among the site physical and chemical characteristics themselves.

The DRAMA approach in this chapter explores and critically evaluates the data using established statistical procedures to assess the relative importance of potential explanatory variables as well as the interaction and potential redundancy between and among site physical and chemical characteristics. The latter is addressed through the creation of synthetic variables using ordination (Legendre and Legendre, 1998; King and Jackson, 1999). Correlations between site physical/chemical characteristics, synthetic variables, contaminants and toxicity tests responses were explored using a suite of ecotoxicologically plausible model structures. Due to the nature of the test exposures available, mixed effects models were used to account for sub-sample variation (Pinheiro and Bates, 2000) and the non-Gaussian distribution of many of the responses was addressed using generalized linear models (McCullagh and Nelder, 1989). Rather than select a single “best” model, contributions from individual models are “averaged” to create a single model using model averaging (Burnham and Anderson, 2002; Claeskens and Hjort, 2008). The advantages of doing so are that undue reliance is not placed on a single model and that the true model uncertainty is acknowledged rather than ignored. Chatfield (1995) comments on the failure of confidence intervals to achieve nominal coverage when standard errors are estimated conditionally upon a model that is assumed correct. Bailer *et al.* (2005) discuss the importance of acknowledging model uncertainty in the context of risk assessment.

### **4.2 MATERIALS AND METHODS**

#### **4.2.1 Available Data and Data Manipulation**

Ecotoxicity assessments were conducted at three sites contaminated with PHCs. The studies are described in detail in Appendix C. Soils in Study 3 were also contaminated with metals. Alberta Environment (2007) Tier 2 pass/fail criteria were used to focus remediation efforts (Studies 1 and 3) or to determine if environmental risks to ecological receptors were acceptable (Study 2). The suite of toxicity tests and chemical measurements were not consistent among studies. Thus, only the consistently measured toxicity test data were used herein. These include tests with a plant (*Elymus lanceolatus*), an earthworm (*Eisenia andrei*), and a springtail (*Folsomia candida*) species.



*E. lanceolatus* toxicity tests were conducted following Environment Canada methods (EC, 2005b). Twenty eight (28) distinct soil samples with five seeds per exposure container were used for a total of 140 emergence, shoot dry mass and length, root dry mass and length observations. *E. andrei* toxicity tests were conducted following Environment Canada methods (EC, 2004). Twenty six (26) distinct soil samples with either 10 or 5 organisms per exposure container were used for a total of 230 survival and progeny observations. As not all *E. andrei* survived, only 130 wet /dry mass observations were available. *F. candida* toxicity tests were conducted following Environment Canada methods (EC, 2007). Twenty eight (28) distinct soil samples with either 3 or 5 organisms per exposure container were used for a total of 126 survival and progeny observations. Each toxicity test was conducted on a subsample of soil collected from a location with a single vector of explanatory variable measurements. Each response within a location represents a subsample and not a replicate. The failure to acknowledge subsampling will result in statistical tests that are artificially more powerful than they should be. Thus, variation among responses within locations is modelled as a random variable.

The soil variables consistently measured across the three studies are soil moisture, pH, conductivity, water holding capacity, total N, total C, inorganic carbon, organic carbon, P, organic matter, gravel, sand, very fine sand, fine sand, medium sand, coarse sand, very coarse sand, silt, clay and texture. Each of the three studies is a distinct entity. Of interest is whether the toxicity of total PHCs can be explained across studies. Thus, “study” was included as a potential explanatory variable. Because there may be variations in how a particular independent variable is correlated with a response, interaction terms between “study” and selected variables were examined. Some explanatory variables were measured in soil subsamples but these were distributed inconsistently across studies. Because these subsamples were not paired with a specific toxicity test replicate, only a measure of central tendency is appropriate to represent these subsamples. Additionally some measurements reflect an analytical laboratory duplicate measurement that cannot be linked to one of the subsamples and thus cannot be used as a quality assurance measurement. These laboratory duplicate measurements were deleted from the dataset for future analyses.

The consistently measured contaminant variables were petroleum hydrocarbon fractions (PHC) (F1 <C10, F2:C10-C15, F3:C16-C33, F4:C34-C50). Alberta Environment (2007) states that PHC concentration must be expressed as either the single dominant fraction consistently detected above Tier 1 guidelines or the cumulative value of all individual fractions that are frequently detected above their respective Tier 1 values (e.g. F2+F3+F4). Because no single fraction was dominant among the three studies, the sum of F2 through F4 PHC fractions was used.

#### 4.2.2 Data Analyses

Redundancy as measured by covariation among the non-PHC variables consistently measured across the three studies listed above was examined using principal components analysis of the correlation matrix (Legendre and Legendre, 1998). The categorical variable “texture”, and

**ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS**

DRAMA Approach: Data Reduction and Modeling Averaging  
September 9, 2013

---

“organic carbon” and “organic matter content” were excluded, the latter two because the analytical methods used also include PHCs. The matrix of retained explanatory variables was multiplied by the ordination eigenvectors to create site-specific scores for each principal component. The scores were considered for use as synthetic explanatory variables in lieu of some directly measured explanatory variables. See Legendre and Legendre (1998) and King and Jackson (1999) for details.

Toxicity test responses were modelled using generalized linear mixed effects models as described in Pinheiro and Bates (2000). The distribution families used included Poisson, Gaussian and binomial. Rather than searching for a single “best fitting” model using one or more statistical criterion, an information theoretic approach, model averaging (Burnham and Anderson, 2002; Claeskens and Hjort, 2008) was used. Model averaging addresses the uncertainty implicit in model selection (Wang *et al.*, 2009).

Model averaging involved weighting the parameters of multiple models fit to the same data. Weights were chosen using an information theoretic approach such as the AICc (second order Akaike Information Criterion for small sample sizes) (Sugiura, 1978). Parameters were averaged across the suite of models in which the parameter appears using AICc weights rather than shrinkage estimators which include zeros for parameters in models where the parameter does not appear. This latter procedure was avoided because the degree of shrinkage is a function of the model structures considered for model averaging. Unconditional variance estimates following Buckland *et al.* (1997) and Burnham and Anderson (2002) were used in Wald tests ( $H_0$ : parameter = 0) of model averaged parameter estimates. A measure of relative variable importance was measured as the sum of AICc weights over all models including the explanatory variable following Burnham and Anderson (2002).

When model averaging is being considered, the suite of candidate models must be carefully selected (Burnham and Anderson, 2002). Guthery *et al.* (2005) are particularly vehement on this point. The suite of models considered herein included linear and additive models with a random effect to appropriately address the within-sample variation. Explanatory variables included pH, clay, soil moisture the first principal component following ordination of the non-PHC variables consistently measured over all three studies, the categorical variable “study” which indicates each of the studies, interactions between study and other explanatory variables. The model structures were driven by the following questions:

4. What is the relative importance of contaminant and non-contaminant heuristics as descriptors of toxicity i.e. the biological responses?
5. In addition to the soil texture heuristic (PC1), clay, pH and moisture were flagged as major sources of variability in the three studies examined. Are these three variables important descriptors of toxicity?
6. Does the relative importance of PHCs, clay and non-contaminant heuristics vary by study?

The model suite incorporating these questions is presented below in **Table 4.1**.

**ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS**

DRAMA Approach: Data Reduction and Modeling Averaging

September 9, 2013

**Table 4.1: Suite of Linear Models Considered for Model Averaging**

Model #	Model Form	Model Comment
Study Effects and Interactions with Studies		
1	$S_j + S_j^*(PHC + PC1 + pH + C2 + M3)4$	Global model includes study, interactions between study and all 5 putative explanatory variables.
2	$S_j + S_j^*(PC1 + pH + C + M)$	As above with subsets of 4 putative explanatory variables.
3	$S_j + S_j^*(PHC + pH + C + M)$	
4	$S_j + S_j^*(PHC + PC1 + C + M)$	
5	$S_j + S_j^*(PHC + PC1 + pH + M)$	
6	$S_j + S_j^*(PHC + PC1 + pH + C)$	
Study Effects Without Interactions		
7	$S_j + PHC + PC1 + pH + C + M$	Study and all 5 putative explanatory variables without any interactions.
8	$S_j + PC1 + pH + C + M$	As above with subsets of 4 putative explanatory variables.
9	$S_j + PHC + pH + C + M$	
10	$S_j + PHC + PC1 + C + M$	
11	$S_j + PHC + PC1 + pH + M$	
12	$S_j + PHC + PC1 + pH + C$	
No Study Effects		
13	$PC1 + pH + C + M$	No study effects and all combinations of 4 putative explanatory variables.
14	$PHC + pH + C + M$	
15	$PHC + PC1 + C + M$	
16	$PHC + PC1 + pH + M$	
17	$PHC + PC1 + pH + C$	

1 -  $S_j$  – study, categorical variable,  $j = 1 \dots 3$ 

2 - Clay

3 - moisture

4 - Main effects corresponding to interaction terms always included.

Finally it is critical to note that model averaging does not imply that the model-averaged estimates describe the data “well”; they are simply the estimates “best” supported by the data. The ability of the most heavily weighted models to describe the data was examined using traditional goodness of fit techniques that are not presented due to the very large number of models used ( $n = 132$ ). Pseudo coefficients of determination ( $R^2$  values) estimated following Magee (1990) are presented. Statistical analyses were conducted using R (R Development Core Team, 2012). Specific libraries containing specialized programming were used. These libraries were:

- MuMIn (Barton, 2012)
- AICcmodavg (Mazerolle, 2012)
- nlme (Pinheiro *et al.*, 2010)
- lme4 (Bates *et al.*, 2011)

Statistical assumptions for fitted models were assessed either formally or by visually examining diagnostic graphics. No egregious violations were noted.

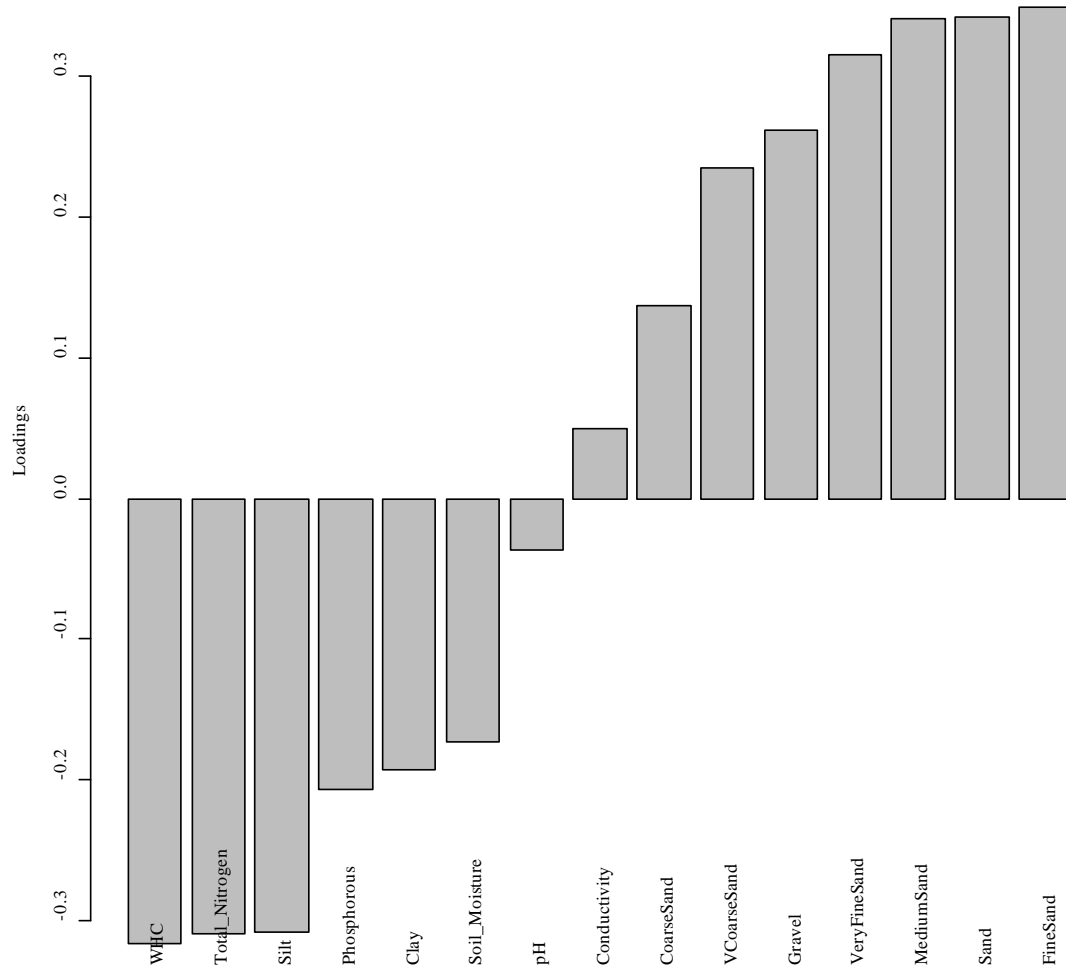
## 4.3 RESULTS AND DISCUSSION

### 4.3.1 Data Exploration and Reduction

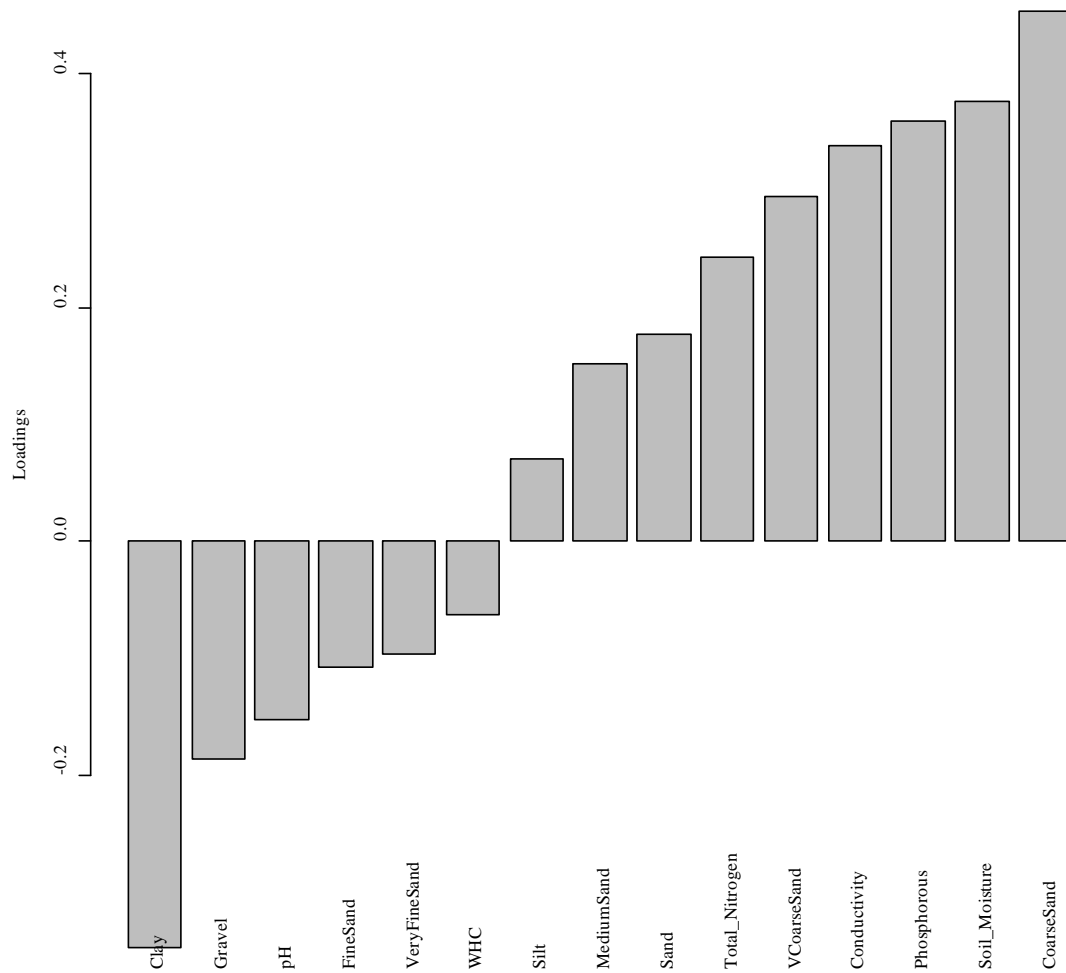
The following data reduction exercise used the non-PHC variables consistently measured across the three studies listed above, with the exclusion of the categorical variable texture and organic carbon and organic matter as the analytical methods used also include PHC concentrations in their measures. A principal components analysis of the correlation matrix is presented below in **Figure 4.1**.

The first principal component (PC1) describes 50% of total variability in the non-PHC variables in studies 1 through 3. **Figure 4.1** describes a contrast between coarse soils (very fine sand and coarser) and fine soils with higher nutrient concentrations. Examination of raw data in the context of soil sample scores shows that PC1 is a heuristic for soil particle size with scores increasing from finely textured nutrient rich soils to coarser nutrient poor soils.

The second principal component (PC2, **Figure 4.2**) describes an additional 25% of total variability in the non-PHC variables in studies 1 through 3. **Figure 4.2** describes a contrast between clay and coarser sands. Examination of raw data in the context of soil sample scores shows that PC2 separates stations with low clay content and relatively large percentages of coarse sand and paradoxically, relatively large concentrations of P and soil moisture, from other stations. It is not clear how this principal component should be interpreted.

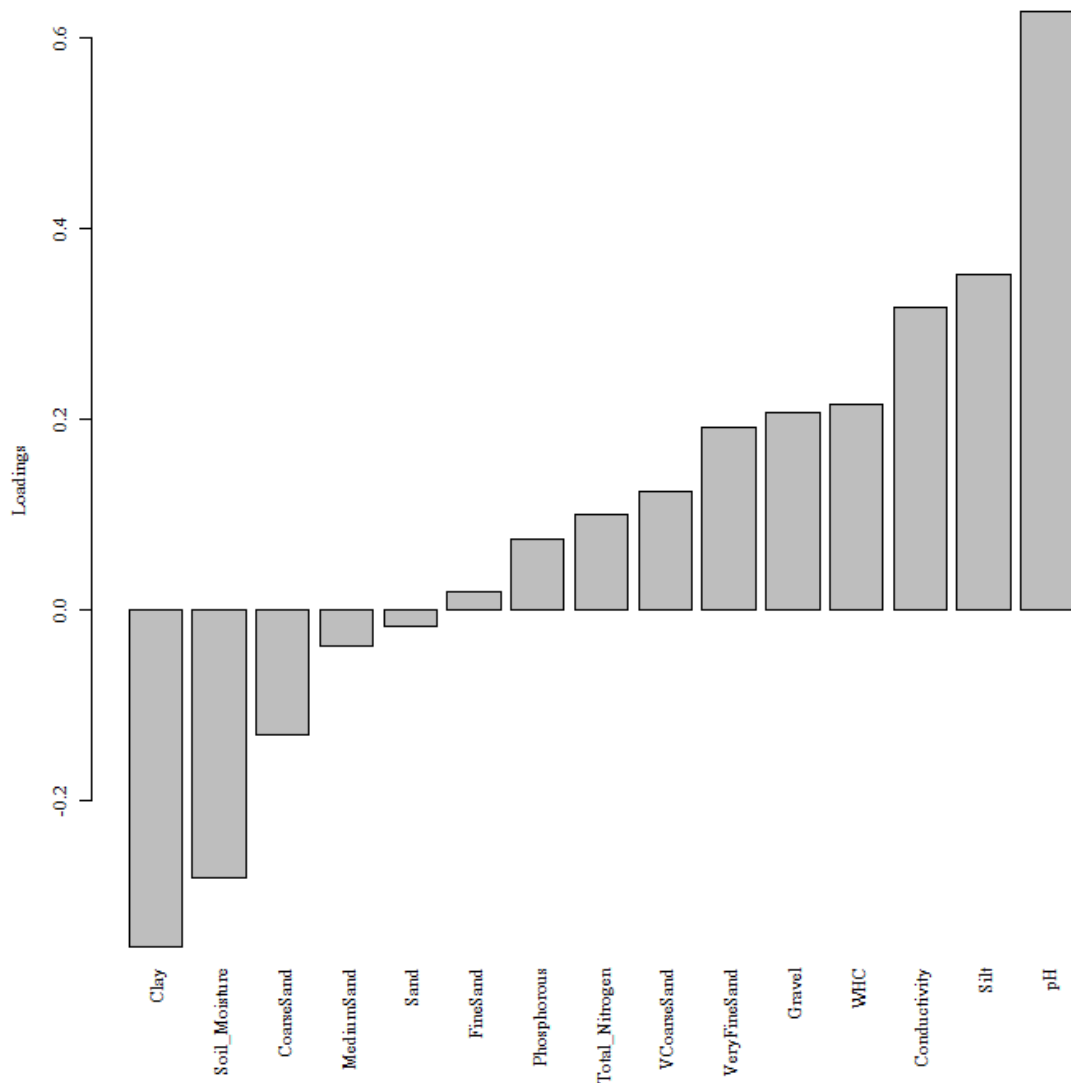


**Figure 4.1: Loadings Plot for First Principal Component using Consistently Measured Non-PHC Variables.**



**Figure 4.2: Loadings Plot for Second Principal Component, using Consistently Measured Non-PHC Variables**

The third principal component (**Figure 4.3**) describing an additional 11% of total variability separates stations with extreme values of high clay content and low pH (study 3, station C5 and to a lesser extent C1 from the same study).



**Figure 4.3: Loadings Plot for Third Principal Component using Consistently Measured Non-PHC Variables**

The first, second and third principal components of the non-PHC variables measured consistently across the three studies represent 85% of the total variability in the dataset. The ordination suggests using PC1 scores as a heuristic for soil particle size and concomitants such as nutrients with the additional individual variables pH, soil moisture and clay as potential explanatory variables in subsequent modelling.



#### 4.3.2 Model Averaging

Although general structural model forms are presented in **Table 4.1**, vagaries of the biological responses precluded modeling random effects in some instances. Also, the response distribution used necessarily varies across models. Details of the response-specific models used are presented in **Table 4.2**, below.

**Table 4.2: Details of Measurement Response - Specific Models Used**

Measurement Endpoint	Modeling Comments
<i>Elymus lanceolatus</i> emergence	Logistic models with random effect for sub-samples were used. The limited number of seeds (5) per exposure container limits the range to 20% increments.
<i>Elymus lanceolatus</i> root and shoot dry mass and length	Gaussian models with random effect for sub-samples were used.
<i>Eisenia andrei</i> survival	Only one study with less than 50% survival and 16/26 soils with complete survival. No modeling was conducted.
<i>Eisenia andrei</i> progeny	9 / 26 soil samples with no progeny results in no within-soil variability. Thus random effects within soils were not estimated. The within-soil counts were aggregated using medians and the subsequent data were modeled using Poisson regression.
<i>Eisenia andrei</i> progeny wet / dry mass	Due to mortality, the progeny wet and dry mass data are sparse. Only one observation was available for plot 4-2, Study 3 and was deleted. One negative wet mass was deleted. Due to this weak data structure some of the more complicated models presented in <b>Table 4.1</b> failed to converge. These models were omitted from the model averaging procedure. Gaussian models with random effect for sub-samples were used.
<i>Folsomia candida</i> survival	Logistic models with random effect for sub-samples were used.
<i>Folsomia candida</i> progeny	Large progeny numbers induced a Gaussian distribution. Gaussian models with random effect for sub-samples were used.

Summary criteria for response-specific models such that sum of AICc weights ( $w_i$ ) is  $> 0.95$  are presented **Table 4.3**.

**Table 4.3: Summary of the Model Averaging Results**

Response	Model	Degrees of Freedom	Log Likelihood	AICc	$w_i$	Pseudo $R^2$
<i>Elymus lanceolatus</i>						
Shoot Dry Mass	2	17	-204.752	448.520	0.947	0.840
	1	20	-203.613	454.285	0.053	0.843
Shoot Length	6	17	-553.625	1146.266	1.000	0.865
Root Dry Mass	1	20	-118.216	283.492	0.924	0.744
	4	17	-124.738	288.493	0.076	0.719
Root Length	4	17	-630.086	1299.189	0.261	0.702
	15	7	-642.218	1299.284	0.249	0.645
	16	7	-642.403	1299.654	0.207	0.644
	10	9	-640.982	1301.348	0.089	0.652
	13	7	-643.742	1302.333	0.054	0.638
	11	9	-641.494	1302.372	0.053	0.649
	7	10	-640.730	1303.165	0.036	0.653
	9	9	-641.982	1303.349	0.033	0.647
	1	20	-628.717	1304.492	0.018	0.708

**ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS**

DRAMA Approach: Data Reduction and Modeling Averaging

September 9, 2013

**Table 4.3: Summary of the Model Averaging Results**

Response	Model	Degrees of Freedom	Log Likelihood	AICc	wi	Pseudo R <sup>2</sup>
Emergence	13	6	-18.142	48.916	0.883	0.941
	8	8	-17.934	52.968	0.117	0.941
<b><i>Eisenia andrei</i></b>						
Progeny	16	5	-48.528	110.057	0.433	<0.010
	14	5	-48.714	110.428	0.359	<0.010
	11	7	-45.803	111.828	0.179	<0.010
	9	7	-47.614	115.450	0.029	<0.010
Wet Mass	13	7	-450.491	916.102	0.153	0.036
	15	7	-450.537	916.194	0.146	0.035
	17	7	-450.605	916.330	0.136	0.034
	16	7	-450.660	916.440	0.129	0.033
	8	9	-448.498	916.832	0.106	0.071
	10	9	-448.505	916.848	0.105	0.070
	12	9	-448.535	916.907	0.102	0.070
	14	7	-451.553	918.226	0.053	0.017
	11	9	-449.622	919.081	0.034	0.051
Dry Mass	7	10	-448.414	919.097	0.034	0.072
	10	9	-256.573	532.946	0.214	0.187
	12	9	-256.650	533.100	0.198	0.186
	8	9	-256.716	533.232	0.186	0.185
	15	7	-259.901	534.900	0.081	0.136
	16	7	-259.987	535.073	0.074	0.135
	13	7	-260.008	535.114	0.073	0.134
	7	10	-256.539	535.299	0.066	0.187
	11	9	-257.898	535.595	0.057	0.167
<b><i>Folsomia candida</i></b>	17	7	-260.356	535.811	0.051	0.129
	17	6	-161.480	335.667	0.408	0.964
	15	6	-161.597	335.901	0.363	0.964
	16	6	-162.938	338.581	0.095	0.964
	10	8	-161.348	339.926	0.048	0.965
	12	8	-161.563	340.357	0.039	0.964
	11	8	-162.052	341.335	0.024	0.964
	9	8	-162.117	341.465	0.022	0.964
	6	17	-881.163	1801.992	0.488	0.715
Progeny	3	17	-881.673	1803.013	0.293	0.713
	2	17	-882.428	1804.523	0.138	0.709
	5	17	-882.957	1805.581	0.081	0.707

The results for *E. andrei* (Table 4.3) illustrate the fallacy in relying on AICc weights that, by definition, range from 0 to 1 and correspond to models that are not useful descriptors of the data. The *E. andrei* responses were not predictable given the explanatory variables considered and are not further discussed.

The results in **Table 4.3** also show that, for some of the responses, the AICc criterion heavily weights one or two models. Burnham and Anderson (2002) suggest that, unless  $w_i \geq 0.9$ , alternative explanations offered by the data analyses should be considered. In the table above the only single model interpretation is for *Elymus lanceolatus* shoot length.

Models 2, 8 and 13 that do not include a PHC variable are heavily supported by *E. lanceolatus* shoot dry mass and emergence results. No obvious patterns in importance of study or interactions with study are apparent.

The relative importance of variables was assessed using model averaged coefficients and p-values for Wald tests of significance for each parameter within a model. P-values are colour coded according to generally accepted (but arbitrary) ranges. These are: dark green, p-value  $\leq 0.01$ ; light green,  $0.01 \leq \text{p-value} < 0.05$ ; and yellow, p-value  $\geq 0.05$  (**Table 4.4**).

**Table 4.4: Wald Tests of Significance for Model Averaged Parameters**

Variable	<i>E. lanceolatus</i>										<i>F. candida</i>			
	Shoot Dry Mass		Shoot Length		Root Dry Mass		Root Length		Emergence		Survival		Progeny	
	Estimate	Pr(> z )	Estimate	Pr(> z )	Estimate	Pr(> z )	Estimate	Pr(> z )	Estimate	Pr(> z )	Estimate	Pr(> z )	Estimate	Pr(> z )
Intercept	-4.024E+01	0.001	-5.208E+02	0.001	-3.213E+00	0.696	3.620E+02	0.202	1.223E+00	0.495	6.272E-01	0.897	-1.293E+04	<0.001
S2	5.108E+01	0.007	1.069E+03	<0.001	-6.381E+00	0.594	-3.457E+02	0.434	-2.162E-01	0.752	4.601E-02	0.992	2.084E+04	<0.001
S3	3.280E+01	0.018	6.371E+02	<0.001	8.264E+00	0.372	-2.314E+02	0.493	-6.628E-02	0.923	-7.779E-01	0.864	1.228E+04	0.002
PC1	1.922E+00	0.015	1.291E+01	0.125	-6.911E-02	0.878	-1.705E+01	0.388	-7.043E-02	0.096	-3.244E-01	0.020	-1.860E+02	0.299
PHC	-3.393E-05	0.865	-3.637E-03	0.044	-3.447E-04	0.002	-5.148E-03	0.458	-1.339E-05	0.479	-7.794E-05	0.008	-5.070E-02	0.233
pH	3.146E+00	0.123	7.214E+01	0.000	1.673E+00	0.199	-7.016E+00	0.803	-1.188E-01	0.569	4.126E-01	0.546	1.777E+03	0.000
C	5.183E-01	0.028	1.597E+00	0.589	-2.225E-01	0.124	-6.790E+00	0.470	-3.317E-03	0.707	-4.584E-02	0.150	2.547E+01	0.672
M	2.981E-01	0.155	9.759E-01	0.792	-6.172E-02	0.693	-3.911E+00	0.381	-1.974E-02	0.039	1.711E-02	0.696	-3.157E+00	0.962
PC1:S2	-2.749E+00	0.001	-3.018E+01	0.001	-1.944E-01	0.697	2.002E+01	0.448	-3.197E-01	0.472	3.558E-01	0.771	2.185E+02	0.276
PC1:S3	-2.761E+00	0.001	-1.789E+01	0.052	-5.664E-01	0.253	1.180E+01	0.651	-5.861E-01	0.171	5.561E-01	0.646	2.045E+02	0.317
pH:S2	-5.293E-01	0.855	-1.216E+02	<0.001	2.070E+00	0.277	8.625E+01	0.405	-8.224E-01	0.580	-4.854E+00	0.316	-2.596E+03	0.001
pH:S3	-8.288E-01	0.692	-6.533E+01	0.001	-1.465E+00	0.274	-2.503E+01	0.762	-8.957E-01	0.407	-5.633E+00	0.072	-1.634E+03	0.001
C:S2	-8.630E-01	<0.001	-3.331E+00	0.270	4.003E-02	0.791	1.541E+01	0.055	-5.844E-02	0.646	-6.064E-02	0.869	-1.184E+01	0.851
C:S3	-4.869E-01	0.041	-2.733E+00	0.359	2.147E-01	0.142	1.487E+01	0.057	-8.892E-02	0.470	-1.236E-01	0.727	-6.827E+00	0.912
M:S2	-1.009E+00	0.002	-2.606E+00	0.670	-3.640E-01	0.128	8.231E+00	0.404	-5.060E-02	0.754	-3.073E-01	0.592	-2.393E+01	0.830
M:S3	-4.152E-01	0.051	-7.537E-01	0.842	-6.452E-02	0.683	4.766E+00	0.453	-4.726E-02	0.662	-9.979E-02	0.798	-6.535E+00	0.923
PHC:S2	1.814E-04	0.535	-1.730E-03	0.499	2.516E-04	0.115	4.242E-03	0.573	8.057E-05	0.817	-2.058E-04	0.621	9.178E-02	0.123
PHC:S3	4.969E-06	0.980	2.097E-03	0.251	3.648E-04	0.001	1.518E-02	0.005	5.981E-05	0.811	-1.184E-04	0.656	4.645E-02	0.281

Dark green, p-value ≤ 0.01; light green, 0.01 ≤ p-value < 0.05; and yellow, p-value ≥ 0.05

The presence of significant interactions between study and at least one biological response for every explanatory variable illustrates that the effects of the variables examined differ among sites precluding simple summary statements. Total PHC concentrations were significantly correlated with at least one toxicity test metric for each species studied, but not all metrics. Of the four explanatory variables examined, PC1 and total PHCs were most frequently significantly correlated with at least one toxicity test metric. Examination of raw data in the context of soil sample scores showed that the large positive scores represent the coarse grained soils low in nutrient soils whereas the large negative first principal component scores represent the soils with fine-grain textures that are higher in nutrients. Thus, the two statistically significant negative interaction coefficients show that as soils become coarser and poorer in nutrients, *E. lanceolatus* shoot length and dry mass are adversely affected. As only the main effect of PC1 is statistically significant for *E. lanceolatus* emergence and *F. candida* survival, the negative main effects show that increasing coarseness and decreasing nutrients is negatively correlated with these two metrics. This observation may reflect either a soil texture/nutrient effect or increased bioavailability of hydrocarbons. The interpretation of total PHC coefficients is less consistent because *E. lanceolatus* shoot length and emergence and *F. candida* survival were adversely affected whereas *E. lanceolatus* root dry mass was positively correlated with total PHCs in at least Study 3. These model-averaged results using data from three studies show that the observed responses are partially or primarily driven by edaphic variables.

The relative importance of variables was also assessed by summing the AICc weights for a model in which a parameter occurs. Note that this approach is only valid if each variable appears the same number of times across the suite of models being compared. Examination of the main effects in **Table 4.1** shows that each parameter appears 14 times.

The results in **Table 4.5** show the relative importance of a variable as a correlate of a biological response over the suite of 17 models assessed. For example the importance of PC1 as a correlate of *E. lanceolatus* shoot dry mass was 1 whereas the relative importance of total PHCs was only 0.055. The relative importance may be averaged over the biological responses conducted over three sites to assess the relative importance of the explanatory variables on a broader geographic scale. In this ranking, the edaphic variables clay, pH, moisture, and the soil texture/nutrient heuristic PC1 are relatively more important than total PHCs in describing the biological responses.

PHC toxicity can be affected by bioavailability (Wong *et al.*, 1999; Semple *et al.*, 2003 and Paton *et al.*, 2005). The synthetic variable PC1, clay, pH and soil moisture were assessed as a correlates of cyclodextrin-extracted hydrocarbon fractions (2, 3 and 4, and their sum) that were measured in Study 3 using Spearman rank correlations. The null hypothesis tested was that the Spearman rank correlation is 0 for each of the comparisons against the alternative hypothesis that correlations were greater than 0. The only null hypothesis rejected was for the synthetic variable PC1 (Spearman rank correlation = 0.643; p-value < 0.0001) suggesting that PC1 may be a heuristic for bioavailability.

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

DRAMA Approach: Data Reduction and Modeling Averaging

September 9, 2013

Table 4.5: Relative Importance of Explanatory Variables

Variable	<i>E. lanceolatus</i>					<i>F. candida</i>		Mean
	Shoot Dry Mass	Shoot Length	Root Dry Mass	Root Length	Emergence	Survival	Progeny	
Study	1.000	1.000	1.000	0.499	0.117	0.147	1.000	0.680
Clay	1.000	0.999	1.000	0.750	1.000	0.884	0.920	0.936
pH	1.000	0.997	0.927	0.427	1.000	0.601	1.000	0.850
Moisture	1.000	0.027	1.000	0.982	1.000	0.566	0.517	0.727
PC1	0.999	1.000	0.978	0.964	1.000	0.965	0.710	0.945
PHC	0.055	1.000	0.987	0.929	0.000	0.998	0.864	0.690

Model averaging has been criticized as being a glorified sensitivity analysis or a data mining exercise rather than a thoughtful consideration of hypotheses (Guthery *et al.*, 2007). This criticism is not relevant herein as the model structures were constructed to assess plausible specific hypotheses regarding practical application of soil toxicity tests in ecotoxicity assessments. One advantage of model averaging is that the bias in standard errors conditional upon a model that is assumed correct leads to overstated achieved levels of significance in hypothesis testing is obviated (Chatfield, 1995). Bailer *et al.* (2005) emphasize the desirability of acknowledging model uncertainty in risk assessment. Model averaged estimates directly include this model uncertainty in the estimated parameter standard errors (Buckland *et al.*, 1997; Burnham and Anderson, 2002) that are used to test hypotheses to achieve the stated levels of significance. A less obvious advantage is that the uncertainty in model selection is acknowledged through the inclusion of model terms that may not appear in a single “best” fitting model which can have important practical consequences.

In a Tier II risk assessment or site remediation context, the implications between using a single “best” fitting model or an average of plausible models weighted by the support the data gives to each model can affect management decisions regarding, for example, which site rehabilitation option occurs, if any, or the applicability of estimated remediation objectives. Consider for example, the two models with the highest data support for northern wheatgrass emergence or root length. The two “best” fitting models for each measurement endpoint vary only with respect to inclusion of terms that allow the response to vary by study or not. The single “best” fitting root length model suggests that effects of root lengths vary by site whereas the single “best” fitting emergence model for the same species suggests that there is no effect of site on emergence. Under the assumption that each “best” fitting model represent the two models are mutually implausible. The typical data analyst uses only coefficients of determination to choose among models. The wheatgrass emergence coefficients of determination are identical to the third decimal and so it is not clear which single model should be chosen. A more sophisticated data analyst might invoke the principle of parsimony to choose the model with fewer degrees of freedom. Again for the northern wheatgrass emergence results, there is little difference in the degrees of freedom among the two most heavily weighted models which precludes an obvious “choice”. In this case there is a degree of subjectivity among model choices. The subjectivity may have important implications as it affects applicability of the model should remediation

targets based on wheatgrass emergence (after adjustment for the PC1 synthetic variable, clay and moisture) be the same for all three sites or should they vary.

Rather than making subjective or less than obvious decisions regarding model applicability and by extension remediation objectives, model averaging allows the data to “speak” by providing a measure of support for each ecotoxicologically plausible model. The degree of support for each model allows for an objective synthesis of models. In the case of the wheatgrass emergence and root length data examined, model averaging reconciles the contradicting interpretations following the choice of single best fitting models for each measurement endpoint.

The study also illustrates the importance of non-contaminant soil quality variables such as the PC1 heuristic for grain size and specific edaphic variables relative to PHCs with respect to explaining the biological test variability. None of data supported only PHCs as a correlate of biological test responses and northern wheatgrass emergence over three studies was best described by non-contaminant variables. Using model averaging and the suite of candidate models the average relative importance of the putative explanatory variables is considered. Although the average importance of PHC as an explanatory variable was higher than the effect of study, the most important variable overall was the PC1 grain size heuristic followed by clay. The “importance” of non-contaminant variables may be due to induced effects on bioavailability, if cyclodextrin-extracted fractions measure bioavailability.

The model averaging and individual model results suggest that non-contaminant variables be considered in the estimation of Tier II SSROs. The finding corroborates Efroymson *et al.* (2004) who state “Existing toxicity data are not sufficient to establish broadly applicable TPH ecotoxicity screening benchmarks with much confidence, even for specific mixtures”. The detailed report for the ERM approach by Zajdlik *et al.* 2013 is provided as a draft in **Appendix A**.



## 5.0 PLS Approach: Partial Least Squares Regression

---

### 5.1 RATIONALE

PLS is a multivariate statistical method that can be applied to model the relationship between a multivariate predictor matrix (X) and a response matrix (Y). Analogous to simple linear regression models, PLS provides an assessment of the strength of the relationship between X and Y (i.e., the percent of variation in Y that can be explained in terms of the variation of X), and can also be used as a foundation for predicting the 'Y-values' of future unknown observations based on their known X-data (which can be measured).

The purpose of this initiative was to determine whether PLS modeling could be applied to predict individual traditional toxicological endpoints from multivariate soil property information (including physical properties and contaminant concentrations). Therefore, a preliminary examination of a dataset (Study 3 – description is provided in Appendix C) was selected to provide an initial indication of the utility of this approach. Study 3 was selected because contaminant concentrations and organism responses were generally highest in this study which should provide more information for model development.

### 5.2 MATERIALS AND METHODS

One of the first steps in determining how this technique could be applied was to examine a dataset for which the parameter data were complete (Study 3). The earthworm (*Eisentia andrei*) endpoints were selected as a starting point because contaminant concentrations and organism responses were generally complete for this study. The results for this preliminary assessment were summarized in the interim draft report (November 11, 2012). The analyses of the results for the toxicity tests with red clover, barley, northern wheatgrass and the soil arthropod, *Folsomia candida*, have now been completed and summarized in the corresponding subsection 5.3 below.

Study 3 comprises data and information relating to an ecotoxicity assessment completed as part of a field study to support a Tier 2 assessment of soils from a site. The main objective was to determine if the site soil samples would satisfy the Tier 2 pass/fail criteria in order to focus remediation efforts on the areas of greatest concern and to exclude from further consideration areas with soils that were not hazardous. The site soils were contaminated with petroleum hydrocarbons, and metals, including boron, copper, lead, and zinc. After initial (Phase 1) testing, two issues critical to the ecotoxicity assessment were identified. Characterization of the site soils and the reference control soils suggested that the control, which was from a small agricultural hay field/ungrazed tame pasture mix, was not representative of the majority of the upland areas nor representative of previous, future or adjacent land uses (upland forest and/or tame pasture). Secondly, some of the petroleum hydrocarbon (PHC) contaminated site soils that did not satisfy the Tier 2 pass/fail criteria were co-contaminated with concentrations of

metals (lead, zinc, boron, and copper). Therefore, a Phase 2 toxicity assessment with the reference control soil from Phase 1 ecotoxicity testing and two new reference control soils, as well as a PHC and metals contaminated site soil from Phase 1 testing, and the same contaminated site soil mock-thermally desorbed of PHCs, was conducted with a reduced species battery comprised of one plant, one earthworm, and one soil arthropod species to assess the test organism performance in these soils. Seven soil samples with F3 contamination ranging between 3650 and 33900 mg kg<sup>-1</sup> and one control sample were collected with three to five replicates of each species endpoint per sample. Additional soil samples were collected from this site, but did not include species endpoint data suitable for inclusion in the cross-site analysis.

PLS regressions, using the NIPALS PLS algorithm (Vendeginste *et al.*, 1998), were constructed using a matrix of measured soil characteristics as the 'X' matrix of multiple predictors and each of the ecotoxicity test endpoints as a separate 'Y' (response) matrix.

PLS models were cross validated using leave-one-class-out cross validation (LOCOCV) and the number of components that maximized the internally cross-validated R<sup>2</sup>Y value (reported as Q<sup>2</sup>Y) was selected as the number of components for each final PLS model. For each PLS model, the explained variation of X and Y (R<sup>2</sup>X and R<sup>2</sup>Y) were reported to indicate how well the model fit the training data (Eriksson *et al.*, 2006) and Q<sup>2</sup>Y was reported as a preliminary measure of the predictive ability of the model (Varmuza and Filzmoser, 2009; Hawkins *et al.*, 2003; Cramer, 1993). In addition, the significance of each PLS model was estimated through response permutation testing (Eriksson *et al.*, 2006).

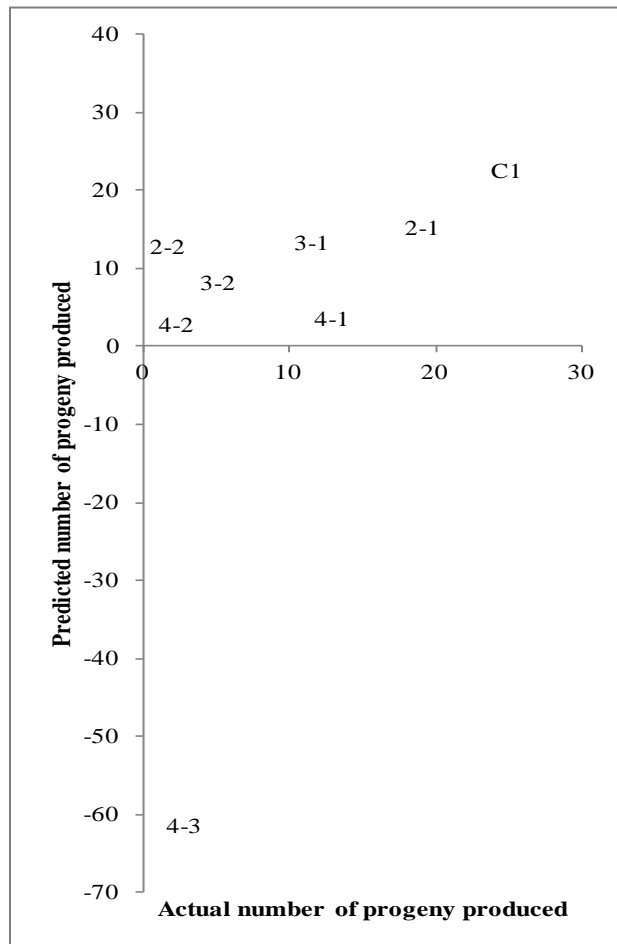
Statistical significance was assessed at  $\alpha \leq 0.05$ . Means were reported as the mean value  $\pm$  standard error. Multivariate statistics (PCA and PLS) and permutation tests were performed in R (R Development Core Team, 2009) using the Chemometrics package (Filzmoser and Varmuza, 2010).

## 5.3 RESULTS AND DISCUSSION

### 5.3.1 PLS Predictions of the Number of Earthworm Progeny Produced

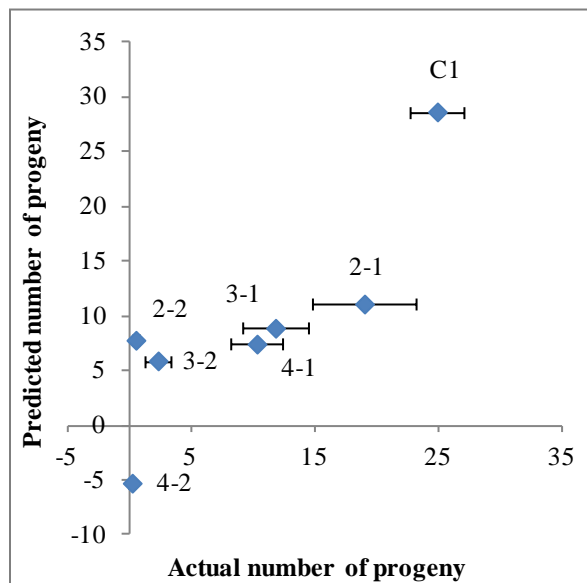
An initial data analysis using the complete dataset (all soils) of Study 3 indicated that no statistically significant model could be constructed between the soil properties and number of earthworm progeny produced ( $p=0.61$  for 'best' model based on 'leave one class out cross-validation' (LOCOCV) and permutation testing). However, a review of the cross-validated model predictions showed that only the values from soil 4-3 were poorly predicted (**Figure 5.1**) when soil 4-3 was left out of the model development. A close examination of the physical and chemical properties of soil 4-3 revealed that the profile of PHC contamination in soil 4-3 differed dramatically from that for the other soils. Specifically, the proportion of total contamination from F1 and F2 were several orders-of-magnitude higher for soil 4-3. Negative values are possible with any regression model; the fact that values are negative tells us that 1) the model fits poorly (seems to be the case with low R-square); 2) there is insufficient data support for one set of

variables (F1 and F2 fractions are higher for this location and study 3 soils were co-contaminated with metals); 3) a model or procedure exists that lacks robustness; 4) there are missing important variables; or 5) some combination of all of the above.



**Figure 5.1: Actual number of progeny produced plotted against predictions for each class based on ‘leave one class out’ cross validation.**  
Error bars represent standard error of the mean. Labels indicate the soil treatment represented.

It appears that in the absence of other soils with similar characteristics in the model, it was not possible to accurately predict the number of earthworm progeny that would be produced in soil 4-3 using the results from the other soils. Therefore, construction of the model was completed with soil 4-3 excluded. In this case, it was possible to construct a statistically significant model ( $p < 0.0025$ , based on LOCOCV and permutation testing) with a reasonably high goodness of fit (indicated by  $Q^2Y = 0.42$ ). The LOCOCV cross-validated predictions for this model are shown in **Figure 5.2**.



**Figure 5.2:** Actual number of progeny produced plotted against predictions for each class based on 'leave one class out' cross validation. Error bars represent standard error of the mean. Labels indicate the soil treatment represented.

Since cross-validation and permutation testing suggested that exclusion of soil 4-3 for the progeny production endpoint allowed for the construction of a meaningful model, aspects of this model were then explored further. Specifically, variable importance in projection (VIP) values were calculated in order to determine which variables contributed most to the predictive model. Variables with top 20 VIP values are shown in **Table 5.1**.

**Table 5.1:** VIP values for PLS model predicting earthworm number of progeny using 'Study 3' results with soil 4-3 excluded.

Variable	VIP
Total Carbon	1.52
Total Nitrogen	1.49
Soil Moisture	1.37
<b>Total Xylenes</b>	<b>1.33</b>
Salt Magnesium	1.32
Sodium Adsorption Ratio	1.30
Salt Chloride	1.26
<b>Total PHC</b>	<b>1.24</b>
Hydrometer Silt	1.24
Salt Calcium	1.23
CCME Coarse	1.23
Ammonia	1.23
Total Sulphur	1.23
<b>F4G</b>	<b>1.19</b>
<b>Toluene</b>	<b>1.17</b>
<b>F4</b>	<b>1.17</b>

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

PLS Approach: Partial Least Squares Regression

September 9, 2013

**Table 5.1: VIP values for PLS model predicting earthworm number of progeny using 'Study 3' results with soil 4-3 excluded.**

Variable	VIP
<b>Arsenic</b>	<b>1.13</b>
Hydrometer Clay	1.12
pH	1.09
Salt Sodium	1.09

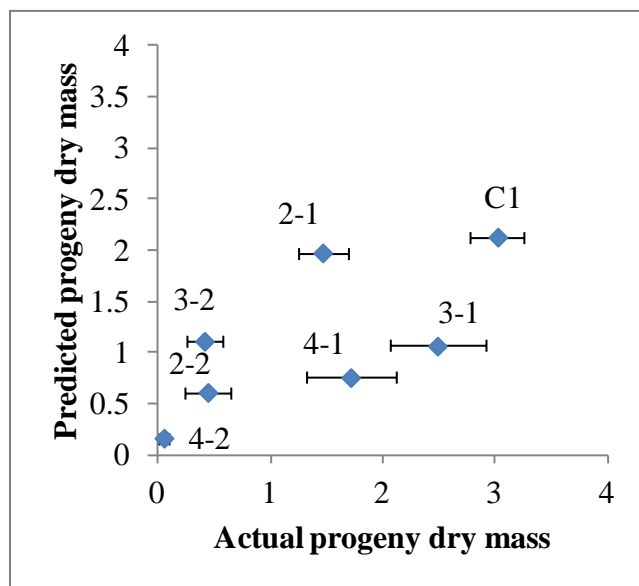
Bold values denote values associated with soil contaminants

VIP – variable importance in projection

It is interesting to note that of the top 20 VIP variables in the model, the majority of them are physical properties of the soil rather than contaminant values which are in bold. This is the same conclusion reported for the DRAMA approach.

### 5.3.2 PLS Predictions of the Dry Mass of Earthworm Progeny

The same analysis was repeated using dry mass of earthworm progeny as the predicted variable. As above, soil 4-3 was identified as an outlier, and the analysis was repeated with soil 4-3 results excluded. This model was less significant and had a lower 'goodness of fit' in comparison to the above model ( $Q^2Y=0.25$ ,  $p=0.03$ , see LOCOCV predictions in **Figure 5.3**).



**Figure 5.3: Actual progeny dry mass produced plotted against predictions for each class based on 'leave one class out' cross validation. Error bars represent standard error of the mean. Labels indicate the soil treatment represented.**

VIP values for the progeny dry mass model are provided in **Table 5.2**. Once again, the majority of the most important predictive variables appear to be soil physical properties. The contaminant values are highlighted in bold script.

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

PLS Approach: Partial Least Squares Regression

September 9, 2013

**Table 5.2: VIP values for PLS model predicting earthworm progeny dry mass using 'Study 3' results with soil 4-3 excluded.**

Variable	VIP
Conductivity	1.48
Salt Calcium	1.42
Total Nitrogen	1.42
Soil Moisture	1.41
Electrical Conductivity	1.41
Salt Sulphate	1.40
Total Sulphur	1.39
Gravel	1.31
<b>F1</b>	<b>1.23</b>
Elemental Sulphur	1.21
Salt Magnesium	1.19
Hydrometer Sand	1.14
Hydrometer Clay	1.11
<b>F4</b>	<b>1.09</b>
Hydrometer Silt	1.09
<b>Cyclodextrin Fraction 2</b>	<b>1.08</b>
<b>F4G</b>	<b>1.08</b>
<b>F3</b>	<b>1.07</b>
<b>Nickel</b>	<b>1.06</b>
Clay	1.06

Bold values denote values associated with soil contaminants

VIP – variable importance in projection

**5.3.3 PLS Prediction of Organism Responses**

Following the initial 'proof of concept' presented above for the earthworm toxicity test results, PLS regressions were also used to determine the strength of the relationships between multivariate soil property information and the remaining soil ecotoxicity endpoints (earthworm, collembola, and plant endpoints) using the results from Study 3 only. Model parameters for each of the fitted models are presented in **Table 5.3**. Graphs of the 'actual vs. predicted' results for the cross-validated models are presented in **Table 5.4** (earthworm), **Table 5.5** (collembola), and **Table 5.6** (plants).

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

PLS Approach: Partial Least Squares Regression

September 9, 2013

Table 5.3: Parameters of fitted PLS models using multivariate soil property information\*

Y-variable	No. of components	Q <sup>2</sup> Y	R <sup>2</sup> X	R <sup>2</sup> Y	P	Sign. Diff.
<b>EARTHWORM ENDPOINTS</b>						
Number of progeny produced	1	<0	53.8	49.4	0.61	
Dry mass of progeny (mg)	1	<0	53.9	46.04	0.61	
Collembola endpoints						
<b>Adult Survival (%)</b>	<b>1</b>	<b>36.7</b>	<b>57.0</b>	<b>60.3</b>	<b>0.0075</b>	<b>**</b>
<b>Number of progeny produced</b>	<b>1</b>	<b>32.4</b>	<b>56.9</b>	<b>39.7</b>	<b>0.005</b>	<b>**</b>
<b>PLANT ENDPOINTS</b>						
BA_emergence	1	<0	14.5	28.5	0.36	
BA_root dry mass	1	<0	53.4	48.6	0.085	
<b>BA_root length</b>	<b>1</b>	<b>27.8</b>	<b>53.0</b>	<b>63.3</b>	<b>0.01</b>	<b>*</b>
<b>BA_shoot dry mass</b>	<b>1</b>	<b>26.5</b>	<b>52.3</b>	<b>68.0</b>	<b>0.01</b>	<b>*</b>
<b>BA_shoot length</b>	<b>10</b>	<b>24.5</b>	<b>100</b>	<b>93.84</b>	<b>0.01</b>	<b>*</b>
NWG_emergence	1	<0	53.9	33.2	0.055	
<b>NWG_root dry mass</b>	<b>1</b>	<b>7.26</b>	<b>53.5</b>	<b>57.1</b>	<b>0.0275</b>	<b>*</b>
<b>NWG_root length</b>	<b>1</b>	<b>26.0</b>	<b>53.9</b>	<b>54.6</b>	<b>0.0125</b>	<b>*</b>
<b>NWG_shoot dry mass</b>	<b>1</b>	<b>56.7</b>	<b>53.9</b>	<b>61.9</b>	<b>0.0025</b>	<b>**</b>
<b>NWG_shoot length</b>	<b>1</b>	<b>48.6</b>	<b>52.2</b>	<b>66.5</b>	<b>0.0025</b>	<b>**</b>
RC_emergence	1	<0	52.5	60.9	0.48	
<b>RC_root dry mass</b>	<b>1</b>	<b>79.7</b>	<b>54.2</b>	<b>85.0</b>	<b>&lt;0.0025</b>	<b>**</b>
<b>RC_root length</b>	<b>3</b>	<b>80.5</b>	<b>81.0</b>	<b>90.1</b>	<b>&lt;0.0025</b>	<b>**</b>
<b>RC_shoot dry mass</b>	<b>1</b>	<b>76.0</b>	<b>54.1</b>	<b>86.3</b>	<b>0.005</b>	<b>**</b>
<b>RC_shoot length</b>	<b>1</b>	<b>66.2</b>	<b>53.7</b>	<b>85.5</b>	<b>0.0025</b>	<b>**</b>

\* Including physical properties and contaminant concentrations as the 'X' matrix, and individual traditional ecotoxicity responses (from earthworm, collembola, or plant toxicity tests) as the 'Y' (response) matrix. Asterisks indicate the statistical significance of models (based on 400-fold permutation testing; \* p < 0.05, \*\* p < 0.01). Highlighted values indicate significant differences at p ≤ 0.05.

BA – barley; NWG – northern wheatgrass; RC – red clover

**Table 5.4: Average predictions of y-values ( $\hat{y}_i$ ) for earthworm endpoints\***

Y-variable	Cross-validated predictions	No. of Components	Q <sup>2</sup> Y	R <sup>2</sup> X	R <sup>2</sup> Y	P
<b>EARTHWORM ENDPOINTS</b>						
Number of progeny produced		1	<0	53.8	49.4	0.61

\* Given the multivariate soil property profile for soil by the PLS model derived during the leave one class out cross-validation (LOCOCV) procedure with soil i omitted for PLS models with optimized number of components. Error bars represent the standard error of the mean of the observed data. Model characteristics from Table 5.3 are also provided for information purposes.



## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

PLS Approach: Partial Least Squares Regression

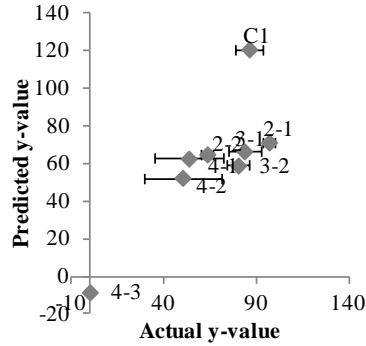
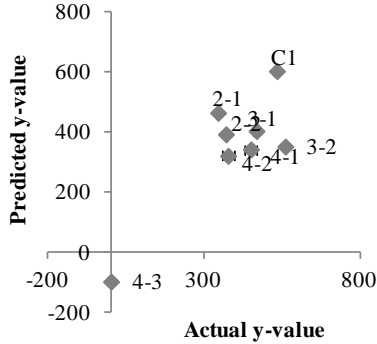
September 9, 2013

Table 5.4: Average predictions of y-values ( $\hat{y}_i$ ) for earthworm endpoints\*

Y-variable	Cross-validated predictions	No. of Components	Q <sup>2</sup> Y	R <sup>2</sup> X	R <sup>2</sup> Y	P
Dry mass of progeny (mg)		1	<0	53.9	46.04	0.61

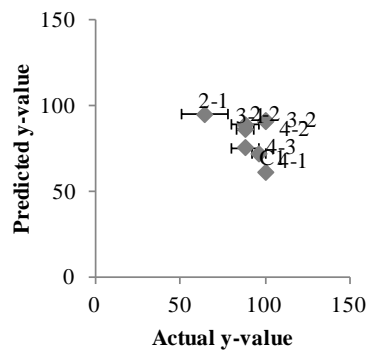
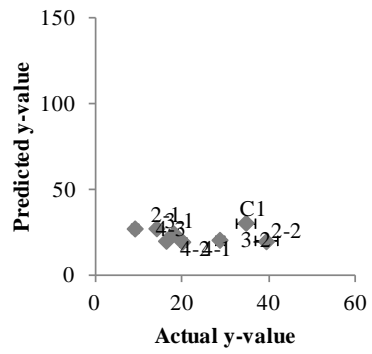
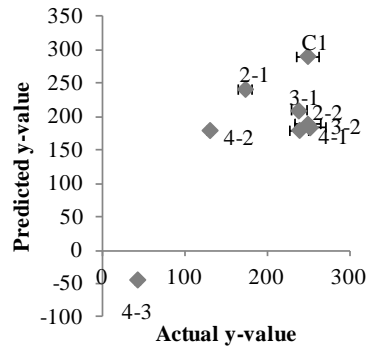
\* Given the multivariate soil property profile for soil by the PLS model derived during the leave one class out cross-validation (LOCOCV) procedure with soil *i* omitted for PLS models with optimized number of components. Error bars represent the standard error of the mean of the observed data. Model characteristics from **Table 5.3** are also provided for information purposes.

**Table 5.5: Average predictions of y-values ( $\hat{y}_i$ ) for springtail endpoints\***

Y-variable	Cross-validated predictions	No. of components	Q <sup>2</sup> Y	R <sup>2</sup> X	R <sup>2</sup> Y	P
<b>COLLEMBOLA ENDPOINTS</b>						
Adult Survival (%)		1	36.7	57.0	60.3	0.0075
Number of progeny produced		1	32.4	56.9	39.7	0.005

\* Given the multivariate soil property profile for soil by the PLS model derived during the leave one class out cross-validation (LOCOCV) procedure with soil i omitted for PLS models with optimized number of components. Error bars represent the standard error of the mean of the observed data. Model characteristics from **Table 5.3** are also provided for information purposes.

**Table 5.6: Average predictions of y-values ( $\hat{y}_i$ ) for plant endpoints\***

Y-variable	Cross-validated predictions	No. of components	Q <sup>2</sup> Y	R <sup>2</sup> X	R <sup>2</sup> Y	P
<b>PLANT ENDPOINTS</b>						
BA_emergence		1	<0	14.5	28.5	0.36
BA_root dry mass		1	<0	53.4	48.6	0.085
BA_root length		1	27.8	53.0	63.3	0.01

\* Given the multivariate soil property profile for soil by the PLS model derived during the leave one class out cross-validation (LOCOCV) procedure with soil i omitted for PLS models with optimized number of components. Error bars represent the standard error of the mean of the observed data. Model characteristics from **Table 5.3** are also provided for information purposes.

# ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

PLS Approach: Partial Least Squares Regression

September 9, 2013

**Table 5.6: Average predictions of y-values ( $\hat{y}_i$ ) for plant endpoints\***

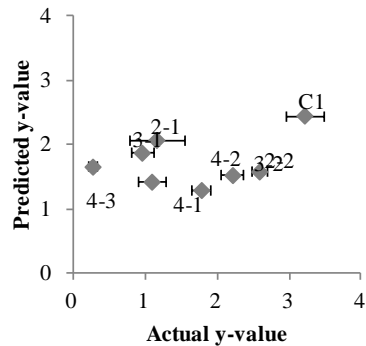
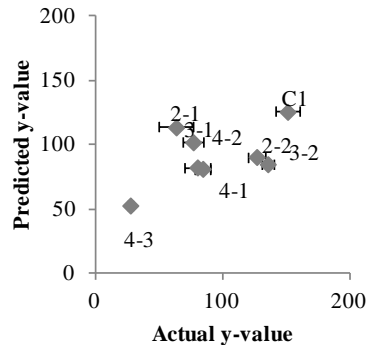
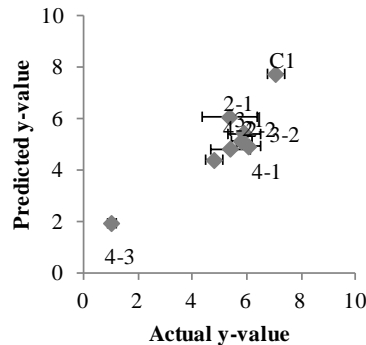
Y-variable	Cross-validated predictions	No. of components	Q <sup>2</sup> Y	R <sup>2</sup> X	R <sup>2</sup> Y	P
BA_shoot dry mass		1	26.5	52.3	68.0	0.01
BA_shoot length		10	24.5	100	93.84	0.01
NWG_emergence		1	<0	53.9	33.2	0.055

# ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

PLS Approach: Partial Least Squares Regression

September 9, 2013

**Table 5.6: Average predictions of y-values ( $\hat{y}_i$ ) for plant endpoints\***

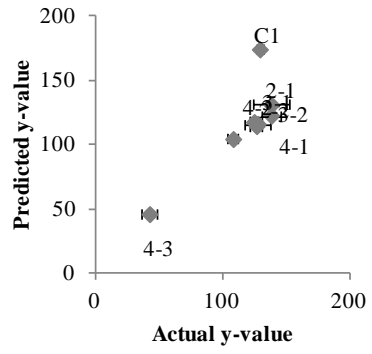
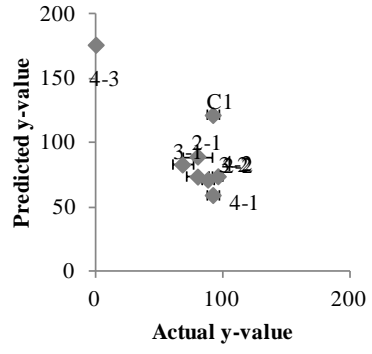
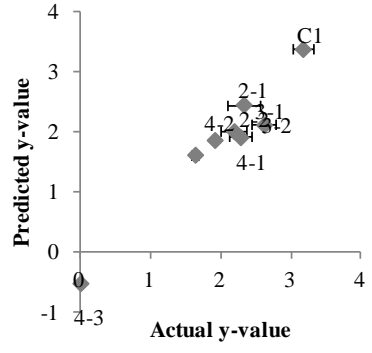
Y-variable	Cross-validated predictions	No. of components	Q <sup>2</sup> Y	R <sup>2</sup> X	R <sup>2</sup> Y	P
NWG_root dry mass		1	7.26	53.5	57.1	0.0275
NWG_root length		1	26.0	53.9	54.6	0.0125
NWG_shoot dry mass		1	56.7	53.9	61.9	0.0025

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

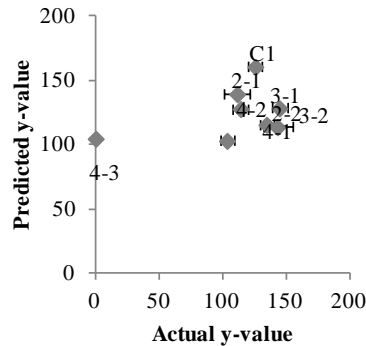
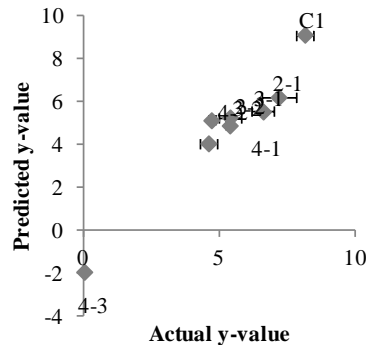
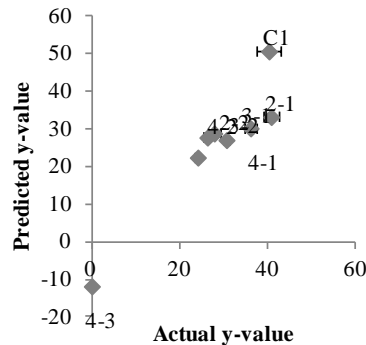
PLS Approach: Partial Least Squares Regression

September 9, 2013

Table 5.6: Average predictions of y-values ( $\hat{y}_i$ ) for plant endpoints\*

Y-variable	Cross-validated predictions	No. of components	Q <sup>2</sup> Y	R <sup>2</sup> X	R <sup>2</sup> Y	P
NWG_shoot length		1	48.6	52.2	66.5	0.0025
RC_emergence		1	<0	52.5	60.9	0.48
RC_root dry mass		1	79.7	54.2	85.0	<0.0025

**Table 5.6: Average predictions of y-values ( $\hat{y}_i$ ) for plant endpoints\***

Y-variable	Cross-validated predictions	No. of components	Q <sup>2</sup> Y	R <sup>2</sup> X	R <sup>2</sup> Y	P
RC_root length		3	80.5	81.0	90.1	<0.0025
RC_shoot dry mass		1	76.0	54.1	86.3	0.005
RC_shoot length		1	66.2	53.7	85.5	0.0025

\* Given the multivariate soil property profile for soil by the PLS model derived during the leave one class out cross-validation (LOCOCV) procedure with soil i omitted for PLS models with optimized number of components. Error bars represent the standard error of the mean of the observed data. Model characteristics from **Table 5.3** are also provided for information purposes.

BA – Barley; NWG – Northern Wheatgrass; RC – Red Clover

Significant models were created for several endpoints; however, for a number of models the significance of the model appears to reflect mainly a discrimination between soil 4-3 and the other soils (e.g., see the ‘actual vs. predicted’ graphs for Collembola number of progeny). Also,

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

PLS Approach: Partial Least Squares Regression

September 9, 2013

the  $Q^2Y$  value, which provides a measure of the “goodness of prediction” for the model (similar to  $R^2$  in a univariate scenario) is generally fairly low, even for the significant models.

The best models (relatively high  $Q^2Y$  values, significant p-values, and best ‘actual vs. predicted’ figures) are for the red clover root dry mass, shoot dry mass, and shoot length. Therefore, the top 20 VIP values, which indicate the variables that contribute most strongly to these models, for each of these endpoints are provided in **Table 5.7**, **Table 5.8** and **Table 5.9**. Again, of the top 20 most important variables in these models, several of them are non-contaminant properties. It is interesting to note that elemental sulphur is the most important soil variable for all three of these endpoints. However, as was noted previously for the earthworm endpoints, measures of soil PHC contamination are also important in all three models.

**Table 5.7: Top 20 VIP values for PLS model predicting RC root dry mass using 'Study 3' results (full dataset)**

Variable	VIP
Elemental_Sulphur	1.39
Molybdenum	1.39
Chromium	1.35
Total_PHC	1.35
Salt_Chloride	1.34
Boron_Saturated_Paste	1.34
F3	1.34
Salt_Sodium	1.34
FineSand	1.33
Boron_Hot_Water_Soluble	1.33
F4	1.33
Total_Sulphur	1.32
Cyclodextrin_Fraction_2_	1.31
OrgCarbon	1.31
F2	1.30
Salt_Potassium	1.30
F4G	1.29
Total_Carbon	1.29
Tin	1.24
Saturation	1.24

Highlighted values are the potential contaminants of concern



**ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS**

PLS Approach: Partial Least Squares Regression

September 9, 2013

**Table 5.8: Top 20 VIP values for PLS model predicting RC shoot dry mass using 'Study 3' results (full dataset)**

Variable	VIP
Elemental_Sulphur	1.37
Molybdenum	1.35
F3	1.35
Boron_Saturated_Paste	1.34
Total_Sulphur	1.33
Salt_Sodium	1.33
F4	1.32
Salt_Potassium	1.32
Boron_Hot_Water_Soluble	1.32
Salt_Chloride	1.31
Total_PHC	1.31
FineSand	1.30
F2	1.29
Cyclodextrin_Fraction_2_	1.29
Chromium	1.28
F4G	1.27
OrgCarbon	1.27
Total_Carbon	1.26
Salt_Magnesium	1.26
Saturation	1.23

Highlighted values are the potential contaminants of concern

**Table 5.9: Top 20 VIP values for PLS model predicting RC shoot length using 'Study 3' results (full dataset)**

Variable	VIP
Elemental_Sulphur	1.41
F3	1.39
Boron_Saturated_Paste	1.38
Salt_Sodium	1.37
Total_Sulphur	1.36
F2	1.36
Salt_Chloride	1.35
Molybdenum	1.35
Cyclodextrin_Fraction_2_	1.35
Total_PHC	1.34
F4	1.34
Salt_Potassium	1.34
Salt_Magnesium	1.30
FineSand	1.29
Boron_Hot_Water_Soluble	1.29
Saturation	1.29
Total_Xylenes	1.28
F1	1.28
F4G	1.28
Organic Carbon	1.27

Highlighted values are the potential contaminants of concern

## 5.4 CONCLUSIONS

This preliminary analysis to investigate the use of PLS approach for SSRO development has demonstrated that it is possible to link multivariate soil properties to certain ecotoxicity endpoints. However, the analyses also highlights at this time that the predictive power of these models is likely to be inadequate for soils with soil properties that vary substantially from the soils used to build the initial model. This could be a function of the small sample size that might be overcome by increasing the number of site soils in the model building exercise and an opportunity to establish model selection criteria. Therefore, there are two possible avenues going forward: 1) further analysis using additional data for a variety of soil types; and/or 2) further assessment of the ability of the “better” models (e.g. red clover root dry mass, red clover shoot dry mass, and red clover shoot length) to predict toxicity endpoints in new soil samples. Alternatively, a model averaging approach might be investigated to improve the utility and application of the models to other soil types. The draft report provided by Dr. Melissa Whitfield-Aslund is available in **Appendix B**.

## 6.0 SEM Approach: Structural Equation Modeling

---

### 6.1 INTRODUCTION AND RATIONALE

Structural equation modeling (SEM) is a very different statistical approach from the analysis of variance and regression statistics familiar to most scientists. SEM is derived from the methods of path analysis developed by Sewall Wright in the 1920s (Wright, 1921). Modern SEM has been widely used in the social and behavioural sciences. Recently, thanks in large measure to the advocacy of ecologists (Shipley 2000; Pugeseck *et al.*, 2003; Grace and Bollen, 2005; Grace, 2006; Grace, 2008; Grace *et al.*, 2010; Lamb *et al.*, 2011; Grace *et al.*, 2012), SEM has become commonly applied in the natural sciences.

Structural Equation Modeling (SEM) is a potential solution for many of the problems encountered in the analysis of field toxicological data. The inter-correlated environmental variables that are so problematic in the current methods used to develop SSROs are readily incorporated in a SEM framework (Grace 2006, Kline 2011, Lamb *et al.* 2011). Further, SEM provides a natural way to incorporate data for multiple species and endpoints from toxicity tests into a single analysis through use of a latent variable (a general concept that is indirectly measured through observation of correlated variables). In toxicity testing, the multiple species and endpoints are effectively indirect measures of the formally unmeasured concept “toxicity” and hence ideal for analysis as a latent variable. Finally, SEM can incorporate measurement error (Grace 2006, Kline 2011, Lamb *et al.* 2011) and thus account for variability in replicate toxicity tests on the same samples.

SEMs have two components, the measurement model and the structural model. The structural model consists of the paths between variables, while the measurement model consists of a latent variable and its associated observed indicator variable(s).

A latent variable represents a concept or quantity that has not been measured directly, but is rather indicated indirectly through one or more observed variables (e.g., measured value) presumed to be highly correlated with the latent variable (**Figure 6.1**). Toxicity is a classic example of a general concept measured in practice by proxy variables (e.g., calculated EC/ICp). A latent variable can be used to estimate the general (unmeasured) species response across a range of toxicant concentrations as indicated by four observed variables (**Figure 6.1**). This latent variable model lets the toxicologist make a clear distinction between the measured proxy and the concept of interest.

Latent variable modeling has a second advantage in that it allows estimates of measurement error to be incorporated into the model. In the case of multiple indicator latent variables such as the example shown in **Figure 6.1**, measurement error is implicitly included as imperfect correlations among the indicator variables. In the case of single indicator latents, the modeler can explicitly state the measurement error by setting the error variance of the observed variable. Measurement error is rarely explicitly considered, yet is nearly always present to some extent in

data. In the model fitting process unacknowledged measurement error can cause problems in the estimation of path coefficients. For example, if measurement error is present in an explanatory variable, the residual error variance will contain both prediction error and measurement error, and as a result the true strength of the relationship between the response and explanatory variables will be underestimated. This underestimation of the true strength of the relationship can cause a downward bias in both the unstandardized and standardized estimates of path coefficients in the structural model.

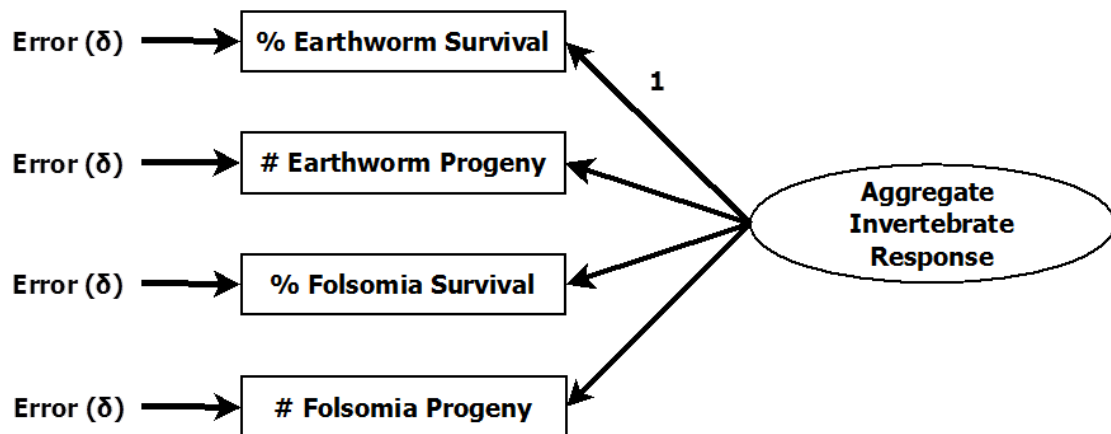


Figure 6.1: An example of a latent variable with multiple indicators.

**Note:** Observed variables are conventionally indicated by rectangles while the unmeasured latent variable is indicated by an ellipse. The arrow extends from the latent variable to the observed variable to indicate that the observed variable is conceptually viewed as having been caused by the latent variable and unmeasured error  $\delta$ . In this case two invertebrate species' responses to a concentration gradient have been measured using four endpoint indicators. The path from earthworm survival is set to 1 indicating that the latent variable is scaled in units of % earthworm survival. Latent variables can alternatively be scaled by setting their variance to equal 1 which scales the latent in units of standard deviation.

The structural (path) model describes the causal relationships among the variables in a model. The structural model consists of either the paths between latent variables or, in an observed variable model, direct relationships among observed variables. Exogenous variables are ones that are causes of other variables in the model, while endogenous variables are caused by at least one other variable. In conventional SEM symbolism a single headed arrows indicates a causal relationship such that a change in the variable at the tail of the arrow results in a predictable change in the variable at the head. A double-headed arrow indicates an unresolved relationship between two variables; typically such a variable is used when the two variables are linked by a causal agent not represented by a variable in the model. As described in **Appendix C**, establishing the initial structural model is a key step in the fitting of a structural equation model.

Testing a theory-based structural model against data allows powerful tests of causality otherwise unavailable with observational (non-experimental) data (Shipley, 2000). The underpinning principle that correlation, linked *a priori* to knowledge of the causal relationships

among variables, could be applied to interpret the strength and relative importance of those causal relationships.

In an SEM study the initial path model represents a causal hypothesis in the form of a set of paths representing causal relationships among the variables in the model. The initial path model implies that a pattern to the variances and covariances between variables then can be tested against the actual variances and covariances in the observed data. The development of the initial model is a crucial step in SEM. This model is typically formulated by the researcher based on past experience or theoretical knowledge. An initial model with adequate fit to the data represents a powerful confirmatory test of the knowledge and theory used to construct the model. In cases where the initial model does not adequately fit the data, SEM can be used in an exploratory mode where modification indices suggest new paths that could be added to the model to improve fit. This exploratory or “data snooping” mode can be nearly as useful as a confirmatory test, since the addition of paths to a model has, in some cases, revealed novel biological hypotheses.

The measurement model specifies how a latent (unobserved) variable is related to one or more observed variables (**Figure 6.1**). A latent variable represents a concept or quantity that has not been measured directly, but is rather indicated indirectly through one or more observed variables presumed to be causally linked to the latent variable. Paths are directed from the latent variable to the observed variable(s) as the observed variables are assumed to be caused by the latent variable. The observed indicators of a latent variable must represent a logically and causally coherent quantity (**Table 6.1**). For example, four different endpoint measurements made on test species X are reasonable indicators of (i.e., are caused by) the latent concept “test species X performance”, but a highly correlated variable such as soil fertility might not be a reasonable indicator. Highly correlated variables that are not reasonable indicators of a particular latent, likely represent additional independent variables that should be connected through the structural model.

Multiple indicators should generally be used to specify latent variables. Two indicator latents can be used, if necessary, but can lead to model fitting issues. Three or more observed indicator variables are preferred as this ensures both the generality of the latent concept and clearly separates the measurement error inherent in any single variable from the general concept of interest. In cases where the concept of interest is tightly linked to the observed variable (i.e. contamination level as indicated by toxin concentration) a single indicator latent can be used. Measurement error can be specified for a latent variable with a single indicator by fixing the error variance  $\delta$  associated with the observed variable. The error variance is calculated as:  $\delta_x = (1 - \lambda^2_x) \times \text{VAR}_x$ ; where  $\delta_x$  is the error variance for observed variable x;  $\lambda^2_x$  is the reliability or the average correlation between repeated measures of x; and  $\text{VAR}_x$  is the variance of the observations of x used in the model. Data from pilot studies or replicate measures can be used to estimate the average correlation between values of x.

Model identification must be considered in the development of the initial model. The sample size is defined by the number of experimental units as this is the number of independent observations used in the calculation of the variance-covariance matrix (**Table 6.1**). Successfully fitting a model requires that the number of estimated parameters be, at a minimum, equal to the number of knowns (t-rule). In practice, the fit of a model can only be evaluated if the number of knowns is greater than the number of parameters to be estimated. The number of elements in the variance covariance matrix is calculated as  $n(n+1)/2$  where  $n$  is the number of observed variables. Satisfying the t-rule is a minimum condition for model identification; however, additional rules must be satisfied in models that contain latent variables (Bollen, 1989; Shipley, 2000; Grace, 2006; Kline, 2011). These additional rules are described in more detail in **Appendix C**.

The number of individual samples remains important in an SEM, even though it does not impact the number of elements in the variance co-variance matrix. Small sample sizes may lead to bias or inaccuracy in the variance-covariance matrix, and hence, both unreliable parameter estimates and tests of model fit (Shipley, 2000; Kline, 2011). Recommendations for sufficient sample size vary widely (Grace, 2006; Kline, 2011). A small sample size relative to the number of variables may lead to a covariance matrix that is not positive definite and hence difficulties in model fitting (Wothke, 1993). Monte-Carlo simulations can be used to estimate minimum sample sizes (Muthén and Muthén, 2002), and bootstrap methods for evaluating model fit can be applied when samples sizes are low (Grace, 2006).

Model fitting is typically carried out using maximum likelihood methods, a family of methods where starting parameter estimates are iteratively improved to increase the fit between model and data (Bollen, 1989; Kline, 2011). There are alternative likelihood-based fitting methods (Muthén and Muthén, 2010; Bollen, 1989). In cases where sample size is small or where distributional assumptions may be violated, bootstrapping may be an effective technique (Grace, 2006; Kline, 2011). Bootstrapping works by randomly resampling a data set with replacement, a large number of times; the model is then fit to each randomized dataset and the distributions of parameter standard errors are used to produce robust standard error estimates (Rodgers, 1999; Kline, 2011). Alternatively, a modified maximum likelihood estimator such as the MLM estimator (maximum likelihood estimates of parameters with standard errors and mean-adjusted  $\chi^2$ ) that is robust to distributional problems may be used (Satorra and Bentler, 1994; Muthén and Muthén, 2010).

**CAPP 09-913-50 ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS**

SEM Approach: Structural Equation Modeling

September 9, 2013

**Table 6.1: Example Variance – Covariance Matrix**

	EA_PR	EA_PR_DM	FC_SURV	FC_PR	BA_SH_L	BA_RO_L	BA_SH_M	BA_RO_M	NWG_SH_L	NWG_RO_L	NWG_SH_M	NWG_RO_M	ALF_SH_L	ALF_RO_L	ALF_SH_M	ALF_RO_M
EA_PR	2.152															
EA_PR_DM	1.487	1.135														
FC_SURV	1.151	0.824	1.109													
FC_PROG	3.505	2.497	2.914	8.755												
BA_SH_L	1.882	1.332	1.750	4.695	4.195											
BA_RO_L	2.197	1.554	1.933	5.320	4.078	4.187										
BA_SH_M	1.887	1.333	1.639	4.495	3.556	3.611	3.130									
BA_RO_M	1.608	1.148	1.410	3.874	2.901	3.019	2.601	2.207								
NWG_SH_L	2.330	1.645	2.078	5.676	4.068	4.344	3.719	3.181	4.730							
NWG_RO_L	2.266	1.590	1.935	5.416	3.698	3.990	3.413	2.921	4.328	4.000						
NWG_SH_M	1.545	1.102	1.196	3.449	2.105	2.338	1.999	1.715	2.516	2.368	1.503					
NWG_RO_M	0.950	0.661	0.697	2.094	1.221	1.369	1.170	1.003	1.468	1.400	0.886	0.547				
ALF_SH_L	1.996	1.417	1.713	4.769	3.254	3.512	3.002	2.572	3.807	3.511	2.081	1.222	3.094			
ALF_RO_L	2.567	1.815	2.140	5.988	3.981	4.325	3.696	3.167	4.679	4.333	2.600	1.530	3.817	4.736		
ALF_SH_M	1.702	1.204	1.324	3.817	2.321	2.577	2.198	1.883	2.768	2.602	1.623	0.967	2.295	2.869	1.794	
ALF_RO_M	1.265	0.885	0.941	2.784	1.599	1.796	1.531	1.314	1.923	1.826	1.160	0.700	1.608	2.029	1.288	0.957

NOTE:

All variables were ln+1 transformed prior to calculating this matrix

ABBREVIATIONS:

AGG. RESP. = Aggregate Response; EA = *Eisenia andrei*; PR = progeny production; FC = *Folsomia candida*; DM = dry mass; SURV = survival; BA = barley; SH = shoot; L = length; M = mass; RO = root; Alf = alfalfa; NWG = northern wheatgrass



Structural equation model fit is evaluated by comparing the model-implied variance-covariance matrix to the observed variance-covariance matrix. The most common test of model fit is a  $\chi^2$  test with a null hypothesis of adequate fit and an alternative hypothesis of inadequate fit. The degrees of freedom for the  $\chi^2$  test are the number of elements in the variance-covariance matrix minus the number of parameters fit in the model. A satisfactory model should have a non-significant  $\chi^2$  test, indicating that no important paths have been omitted from the model. The  $\chi^2$  test is the most common measure of model fit, but there are a wide range of alternatives available, particularly for models containing large numbers of variables (Bollen and Long, 1993; Schermelleh-Engel *et al.*, 2003; Kline, 2011). **Table 6.2** shows several alternative fit indices including the Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Standardized Root Mean Square Residual (SRMR) and Akaike's information criterion (AIC). When fitting a model, one should report both the  $\chi^2$  test results and several of the approximate fit indices. In cases where the  $\chi^2$  is significant but the other fit indices indicate reasonable model fit, the reasons for poor model fit should be carefully assessed and the rationale provided for any changes in model specification made to improve fit.

**Table 6.2: Commonly Used Structural Equation Model Fit Indices<sup>1</sup>**

Fit Index	Criteria for Good Fit	Description
$\chi^2$	$p \geq 0.05$	Test for discrepancies between the observed and model-implied covariance matrices. Non-significant tests indicate an adequate model.
Comparative Fit Index (CFI)	$\geq 0.90$	Measures the relative improvement in model fit over a baseline model. Ranges from 0 to 1 with 1 indicating perfect fit.
Root Mean Square Error of Approximation (RMSEA)	$\leq 0.05$	Measures "badness of fit" where 0 is the best fit.
Standardized Root Mean Square Residual (SRMR)	$\leq 0.08$	Mean absolute correlation residual, or an estimate of the mean differences between observed and predicted correlations.

<sup>1</sup> See Kline (2011) for detailed descriptions

Once overall model fit has been evaluated, it is necessary to examine the strength of individual paths within the model. The  $\chi^2$  test is effective at detecting important missing paths, but a non-significant p-value does not indicate that all of the paths included in the model are important. Unstandardized path coefficients are divided by their standard error to produce a t-statistic (referred to a Critical Ratio (CR) statistic in some software) to test whether a particular path coefficient is significantly different from zero. In cases where a path is not significantly different from zero, a decision must be made to retain or remove those paths. Removing a non-significant structural path that had a strong theoretical justification for inclusion in the initial model is a statement by the researcher that the path is now (theoretically) expected to be unimportant (Grace, 2006). Alternately, one should accept that the (retained) non-significant path is simply non-significant in the context of that particular study. Exploration of why the context of a particular study may lead to a non-significant expected path may be a fruitful research direction. Non-significant paths in the measurement model linking a latent variable and an observed indicator variable, however, are an indication that the latent concept is not a stable



one. In that case, it may be best to reconsider the number and nature of the latent variables included in the model.

Composite variables (Grace and Bollen, 2008) are an important advanced structural equation modeling technique. SEMs are currently very limited in their ability to incorporate non-linear relationships (Grace, 2006; Grace and Bollen, 2008), though recent advances may change this (Grace *et al.*, 2012). Composite variables differ from latent variables in that latent variables represent unobserved variables that are causes of their indicators, while composites are a summary with zero variance representing the collective influence of other variables. Composites have a variety of uses (Grace, 2006), but the most important for the purposes of this project is the ability to model non-linear relationships. Many non-linear relationships can be linearized to some degree through transformation and then directly entered into a standard SEM; however, this is not possible with hump-shaped and other curvilinear relationships that are best described by polynomials. Composite variables provide a method for incorporating polynomial relationships into an SEM with the composite variable typically indicated by a variable and the squared variable. The framework and rationale for the SEM approach (**Subsection 6.1**) is described in greater detail in **Appendix C**. The subsections that follow in this section will detail the material and methods used to apply the SEM approach to existing toxicological data and develop model(s) to describe the relationship between exposure concentrations, pedological properties and biological responses. Greater detail for all aspects of this approach can be found in **Appendix C**.

## 6.2 MATERIALS AND METHODS

The utility of structural equation modeling for assessing toxicological response data was examined in a three stage approach. First, a confirmatory factor analysis (CFA) was conducted with the goal of a) combining multiple endpoint measures within each test species into a series of species level latent variables, and b) combining those species level latent variables into a single second order latent variable representing cross species responses. Second, we contrast the second order latent variable approach with a confirmatory factor analysis where all of the species endpoints were used as direct indicators of a single aggregate species response latent. Third, we developed structural equation models to link species responses to the experimentally manipulated contaminant levels. In all cases, we caution that the models are fit here to small sample sizes relative to the complexity of the models; therefore, interpretation of the model results was done with caution.

### 6.2.1 Confirmatory Factor Analysis and Aggregation of Multiple Endpoints

The methods used for this stage are described in detail in **Appendix C**. To summarize, a conceptual model was constructed using the results of a toxicity assessment of aged and weathered F2 in coarse- and fine-textured soils (spiked multi-concentration tests with two soil types). A confirmatory factor analysis (CFA) was conducted with the goal of a) combining multiple endpoint measures taken within each test species into a series of species level latent

variables, and b) combining those species level latent variables into a single second order latent variable representing cross-species responses.

The first step in the development of a Structural Equation Model is the development of the measurement model used to estimate latent variables. This process is often referred to as “confirmatory factor analysis”. Typically, a number of latent variables with separate indicators would be specified; the curved arrows among the latent variables indicate that they are expected to co-vary, but their relationships are not as of yet subject to direct analysis. The measurement model would then be fit and modified if necessary.

The covariance matrix used in this analysis is shown in Table 6.1. In this case all endpoint variables were transformed using the natural-logarithm +1. Inspection of these data indicated non-linear relationships among the endpoint variables that were linearized by the log+1 transformation. In addition, the transformations served to bring the variances of each variable into a similar range. In cases where transformation is not required or where transformation results in very large or small variances, it may be necessary to achieve model convergence to re-scale variables by multiplying or dividing by factors of 10.

Three alternative models (Model A, B, and C) could be used to combine multiple endpoints into a single latent variable representing species responses to hydrocarbon contamination (**Figure 6.2**). The use of the different models with their corresponding strengths and weaknesses are discussed in detail in the report comprising **Appendix C**. Model C was selected as the most appropriate model for the following reasons: it described the data well; the model fit was better than the alternative models (**Table 6.3**); and, the model had no variables with negative residual variance (**Table 6.4**).

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

SEM Approach: Structural Equation Modeling

September 9, 2013



Figure 6.2: Initial measurement models for Study 4.

BA is Barley, Alf is Alfalfa, NWG is Northern Wheatgrass, EA is *Eisenia andrei*, FC is *Folsomia candida*, and No. is number.

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

SEM Approach: Structural Equation Modeling

September 9, 2013

**Table 6.3: Measurement Model Results for Study 4**

Model Descriptor	Model A	Model B	Model C
No. of model parameters	59	56	68
Deg. Freedom	93	96	84
$\chi^2$	433.22; p<0.0001	477.58; p<0.0001	245.35; p<0.0001
CFI	0.832	0.812	0.921
RMSEA	0.328; 95% CI 0.297-0.360	0.342; 95% CI 0.312-0.373	0.238; 95% CI 0.203-0.273
SRMR	0.047	0.046	0.033
AIC*	362.24	410.31	175.83

\* note that AIC (Akaike's Information Criterion) is an appropriate method for model comparison in this case since all three models contained the same observed variables; CFI – comparative fit index; RMSEA – root mean square error of approximation; SRMR – standardized root mean square residual

**Table 6.4: Full results for Model C including unstandardized path coefficients (col. 2), standard error of the unstandardized coefficients (col. 3), ratio of the unstandardized estimate and standard error (col. 4), test of path significance (col. 5), and standardized path coefficient estimates (col. 6) for Model C**

Path	Unstd. Est.	Std Err	Est. / Std Err	P-Value	Std. Est.
AGG. RESP. BY EA_PR	1.165	0.089	13.137	<0.001	0.806
AGG. RESP. BY EA_PR_DM	0.824	0.068	12.057	<0.001	0.786
AGG. RESP. BY FC_SURV	0.968	0.072	13.420	<0.001	0.933
AGG. RESP. BY FC_PR	2.720	0.155	17.542	<0.001	0.933
AGG. RESP. BY BA_SH_L	1.801	0.289	6.241	<0.001	0.893
AGG. RESP. BY BA_RO_L	1.959	0.213	9.176	<0.001	0.972
AGG. RESP. BY BA_SH_M	1.674	0.196	8.533	<0.001	0.961
AGG. RESP. BY BA_RO_M	1.434	0.144	9.967	<0.001	0.980
AGG. RESP. BY NWG_SH_L	2.116	0.200	10.592	<0.001	0.987
AGG. RESP. BY NWG_RO_L	1.963	0.172	11.428	<0.001	0.996
AGG. RESP. BY NWG_SH_M	1.183	0.087	13.641	<0.001	0.980
AGG. RESP. BY NWG_RO_M	0.698	0.049	14.333	<0.001	0.958
AGG. RESP. BY ALF_SH_L	1.727	0.151	11.46	<0.001	0.997
AGG. RESP. BY ALF_RO_L	2.14	0.178	12.024	<0.001	0.998
AGG. RESP. BY ALF_SH_M	1.3	0.092	14.133	<0.001	0.985
AGG. RESP. BY ALF_RO_M	0.917	0.061	14.995	<0.001	0.951
EA_PR WITH EA_PR_DM	0.483	0.033	14.584	<0.001	0.870
FC_SURV WITH FC_PR	0.196	0.043	4.542	<0.001	0.500
BA_SH_L WITH BA_RO_L	0.431	0.129	3.349	0.001	0.992
BA_SH_L WITH BA_SH_M	0.437	0.132	3.301	0.001	0.990
BA_SH_L WITH BA_RO_M	0.233	0.062	3.753	<0.001	0.878
BA_RO_L WITH BA_SH_M	0.225	0.07	3.221	0.001	0.974
BA_RO_L WITH BA_RO_M	0.122	0.033	3.679	<0.001	0.874
BA_SH_M WITH BA_RO_M	0.123	0.034	3.652	<0.001	0.871
NWG_SH_L WITH NWG_RO_L	0.049	0.002	26.026	<0.001	0.829
NWG_SH_L WITH NWG_SH_M	-0.061	0.003	23.102	<0.001	-0.737

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

SEM Approach: Structural Equation Modeling

September 9, 2013

**Table 6.4:** Full results for Model C including unstandardized path coefficients (col. 2), standard error of the unstandardized coefficients (col. 3), ratio of the unstandardized estimate and standard error (col. 4), test of path significance (col. 5), and standardized path coefficient estimates (col. 6) for Model C

Path	Unstd. Est.	Std Err	Est. / Std Err	P-Value	Std. Est.
NWG_SH_L WITH NWG_RO_M	-0.051	0.002	22.796	<0.001	-0.727
NWG_RO_L WITH NWG_SH_M	-0.023	0.002	10.209	<0.001	-0.545
NWG_RO_L WITH NWG_RO_M	-0.011	0.002	5.065	<0.001	-0.302
NWG_SH_M WITH NWG_RO_M	0.035	0.003	11.962	<0.001	0.682
ALF_SH_L WITH ALF_RO_L	0.008	0.002	5.114	<0.001	0.465
ALF_SH_L WITH ALF_SH_M	-0.019	0.001	14.447	<0.001	-0.621
ALF_SH_L WITH ALF_RO_M	-0.023	0.002	9.372	<0.001	-0.553
ALF_RO_L WITH ALF_SH_M	0.001	0.002	0.490	0.624	0.043
ALF_RO_L WITH ALF_RO_M	0.007	0.004	1.716	0.086	0.178
ALF_SH_M WITH ALF_RO_M	0.058	0.004	13.918	<0.001	0.868

## ABBREVIATIONS:

AGG. RESP. = Aggregate Response; EA = *Eisenia andrei*; PR = progeny production; FC = *Folsomia candida*; DM = dry mass; SURV = survival; BA = barley; SH = shoot; L = length; M = mass; RO = root; Alf = alfalfa; NWG = northern wheatgrass

We contrasted the second order latent variable approach with a confirmatory factor analysis where all of the species endpoints (**Table 6.4**) were used as direct indicators of a single aggregate species response latent. Structural equation models were used to link species responses to the experimentally manipulated contaminant levels. In all cases, the models were fitted here to small sample sizes relative to the complexity of the models; therefore, interpretation of the model results was done with caution.

In all cases, the lethal/sublethal endpoint paradox was addressed by setting reproduction values to zero in treatments where all organisms died. Models A and B used first-order latent variables to combine individual species endpoints into species-specific responses, and a second-order latent variable to combine the first order latent variables together into an aggregate cross-species response. In model A, root and shoot measurements were assumed to contribute together to a single response for each plant species; undirected correlations were included between the two shoot measurements and the two root measurements for each species because those measurements were made on the same plant parts. In Model B, the root and shoot responses for each species are modeled separately to account for the hypothesis that roots and shoots may respond differently to toxicant exposure. Model C combined all of the individual endpoints, regardless of species, directly into a single aggregate cross-species response. Model C included undirected correlations between measures made on a particular endpoint for a species, because all of those measures were for the same individual organism.

Even though model C was clearly the better model, there were indicators that the overall fit was poor (e.g. significant chi-square test, CFI value less than 0.95, and RMSEA value with confidence limits greater than 0.10) and not the result of important missing paths. More than likely, the significant chi-square test resulted from “noisy” data rather than poor overall model fit

which was attributable to the large number of observed variables in the model relative to the sample size. The chi-square test is sensitive to the number of elements in the variance-covariance matrix; the increased test power associated with a large matrix can result in significant tests that detect biologically insignificant lack of fit (Grace, 2006). This problem with the chi-square tests led to the development of alternative methods of assessing model fit (Table 6.3). Inspection of the standardized and unstandardized path coefficients and  $R^2$  values indicated that the model combined the multiple endpoints into a single variable in an acceptable manner. The path coefficients for all of the endpoints are highly significant (**Table 6.4**). The  $R^2$  values for the observed variables ranged from a low of 0.617 to a high of 0.997, with only five of the 16 observed variables having an  $R^2$  less than 0.9. Finally, plots of the raw endpoint data against the aggregate species response variable demonstrated that the aggregate response variable captures the overall responses of all of the endpoints (Figure 3; **Appendix C**). Given the low sample size (35) relative to the number of model parameters (68), the modeling results must be considered preliminary. Therefore, model C was retained as the measurement model aggregating the endpoint measures into a single composite endpoint.

### 6.2.2 Estimation of IC25 and IC50 Values from a Latent Variable

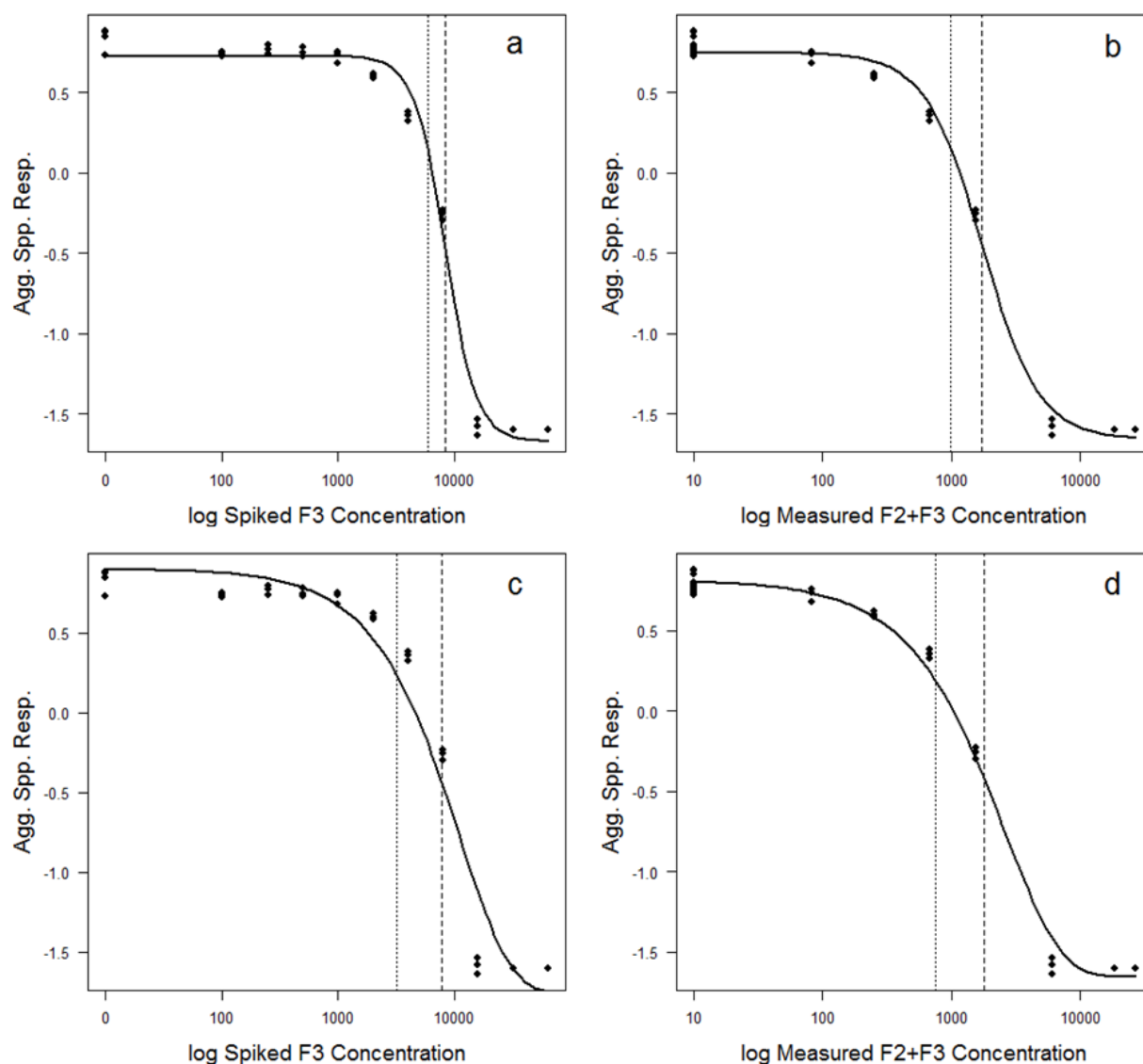
The estimated values for the aggregate species response developed in model C above was used to estimate IC25 and IC50 values. Standard non-linear regression procedures were applied to the data to describe (Stephenson *et al.*, 2000, Environment Canada, 2005) the relationship between the aggregate species response variable predicted in Model C and both spiked F2 concentrations and observed F2 and F3 concentrations. Two models (logistic and exponential) were fit to these data using the “drm” function in the “drc” library in the R 2.14 package (Ritz and Streibig, 2005; R Development Core Team, 2011) and used to estimate the IC25 and IC50 values using the “drc” library function ED.

A logistic model fit better than an exponential model for both the spiked data (logistic AIC = -48.27; exponential AIC = -9.19) and the observed contaminant concentration (logistic AIC = -74.6; exponential AIC = -68.92) (**Figure 6.3**). The lower AIC values for the models with observed concentration as an explanatory variable indicate that observed concentrations were a better predictor of toxicity in this case. Estimated IC25 and IC50 values for each model are listed in **Table 6.5**.

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

SEM Approach: Structural Equation Modeling

September 9, 2013



**Figure 6.3:** Logistic (panels a and b) and exponential (panels c and d) models showing the relationship between aggregate species responses and spiked (a and b) and observed (c and d) contaminant levels.  
Dotted lines show IC25 and dashed lines IC50 values

**Table 6.5:** Estimated IC25 and IC50 values and 95% confidence intervals (CI) for the four nonlinear models fit.

Model	IC25	95% CI	IC50	95% CI
Spiked Sample Logistic	5978 ± 374	5214, 6741	8377 ± 291	7783, 8971
Spiked Sample Exponential	3205 ± 312	2569, 3841	7723 ± 751	6190, 9255
Observed Logistic	993 ± 49	894, 1093	1736 ± 67	1600, 872
Observed Exponential	746 ± 36	673, 819	1798 ± 86	1622, 1973

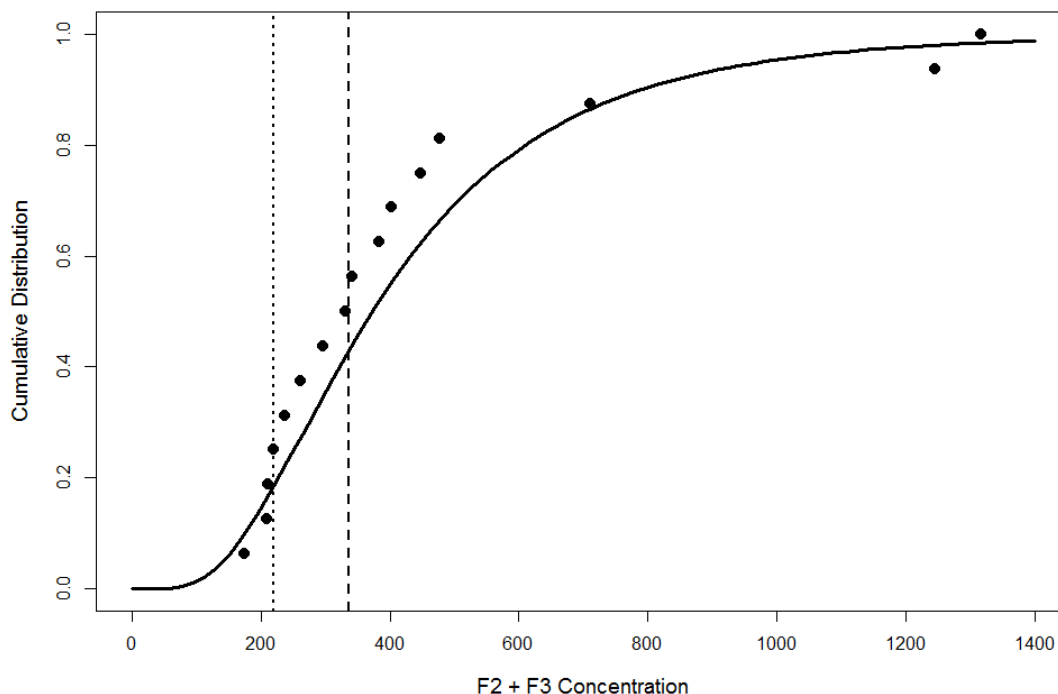
Note: Values are estimates ± 1 standard error



A species sensitivity distribution (SSD) was developed based on the 16 individual species endpoints used in measurement model C for comparison with the results above. IC25 values from this analysis were used to develop a species sensitivity distribution for these species. A standard non-linear regression analysis (Stephenson *et al.*, 2000; Environment Canada, 2005) was used to describe the relationship between the individual endpoints and the observed F2 and F3 concentrations. The observed F2 and F3 concentrations in the analysis above indicated a much stronger relationship between that variable and species responses than the individual nominal values. Two models (logistic and exponential) were fit to these data using the “drm” function in the “drc” library in the R 2.14 package (Ritz and Streibig, 2005; R Development Core Team, 2011). The best model for each endpoint (Table 6; **Appendix C**) was used to estimate the IC25 and IC50 values for each endpoint using the “drm” function (Table 7; **Appendix C**). All acronyms refer to the function names called in the r-scripts.

The estimated IC25 was selected from the best model for each endpoint (Table 7; **Appendix C**), and a species sensitivity distribution was calculated using the “fitdist” function from the “fitdistrplus” library in the R 2.14 package (Delignette-Muller *et al.*, 2010; R Development Core Team, 2011). The cumulative distribution was modeled using log-normal, exponential, and gamma distributions resulting in AIC values of 218.05, 225.07, and 222.05, respectively. The log-normal distribution (mean= 5.74 ±0.63 SD) had the lower AIC value. The empirical and fitted cumulative distribution functions are shown in **Figure 6.4**. This curve suggests target contaminant (F2 + F3) concentrations (20th percentile of the cumulative distribution) of 184 mg/kg, values that are somewhat lower than those suggested by the IC25 values for the logistic (95% CI 758 – 958 mg/kg) and exponential (95% CI 522 – 635 mg/kg) generated using the aggregate species response models above.





**Figure 6.4: Cumulative distribution of IC25 values against contaminant concentration.**  
The dashed line is the fitted lognormal distribution, and the dotted and dashed vertical lines indicate concentrations at the 20th and 50th percentiles of the cumulative lognormal distribution.

### 6.2.3 Structural Equation Modeling

Two structural equation models were fit to the measurement data used for Model C. These models included hydrocarbon contamination as predictors of the aggregate species response (**Figure 6.4**). The fit of both models was relatively poor which was attributable to the small sample size. Nevertheless, both models demonstrated a strong, non-linear relationship between contaminant levels and the aggregate species response with  $R^2$  values of 0.568 and 0.895, respectively, for Model D and E (**Figure 6.5**; Table 8, **Appendix C**). Because of the apparent non-linearity, the relationships between the aggregate species response and both spiked contaminant levels (F2 and F3 concentrations) were modeled as quadratic functions using composite variables (Tables 11, 12 and Figure 8; **Appendix C**). Although, the  $R^2$  values increased, the fit of the models was not improved. These structural equation models were of low utility relative to the direct non-linear modeling of the aggregate response variable demonstrated in the previous section. In particular, the opaque relationship between contaminant concentration and the aggregate response in the SEM relative to non-linear regression suggests that confirmatory factor analysis followed by regression is preferable to SEM in this case.

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

SEM Approach: Structural Equation Modeling

September 9, 2013

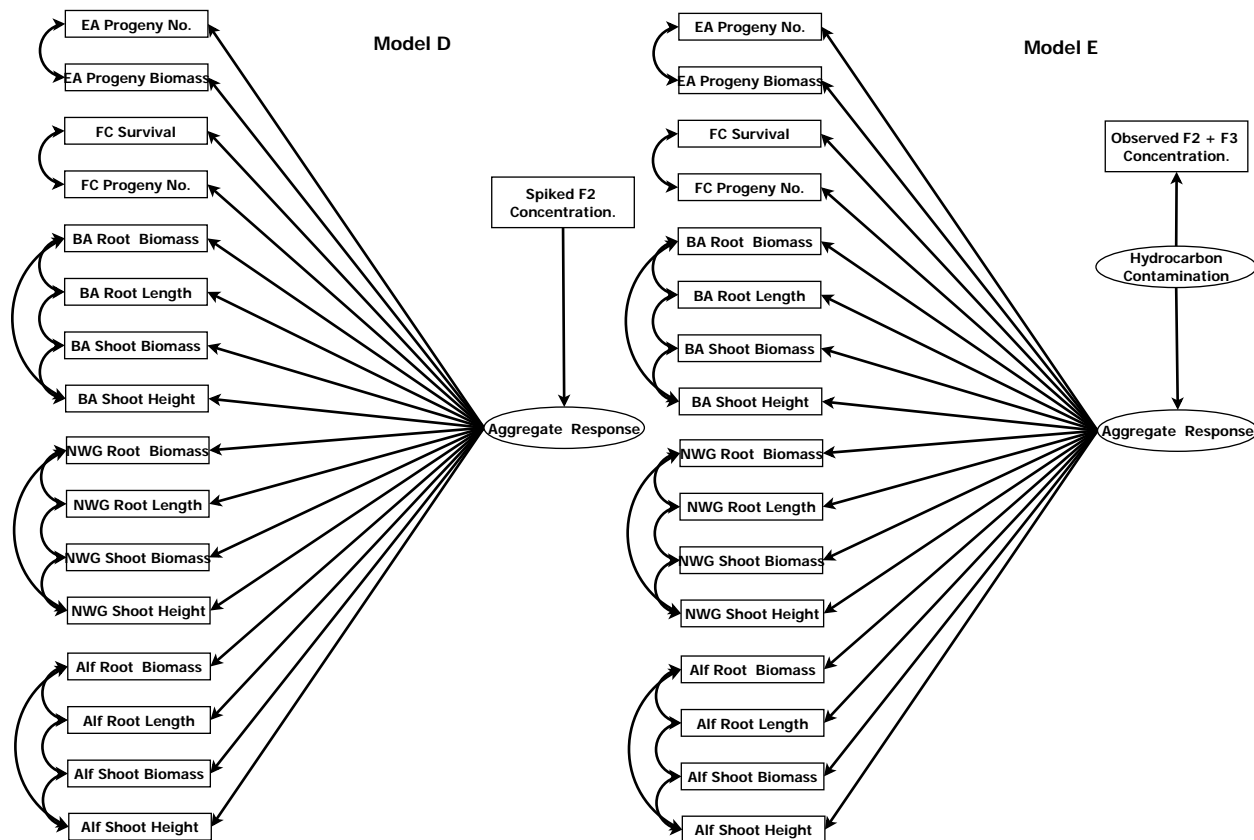


Figure 6.5: Initial structural equation models with hydrocarbon contamination as predictors of the aggregate species response.

## 6.2.4 Modeling of Toxicological Data – Cross Site Field Data

As in the modeling of the spiked sample data, confirmatory factor analyses were conducted to combine multiple endpoints into aggregate latent variables. A series of structural equation models were developed to link measures of both contaminant levels and background environmental conditions to species responses. These models were applied to data for three combined studies with the goal of developing a general predictive model capable of predicting toxicological responses using commonly measured soil characteristics. In all cases, the models were fitted to small sample sizes relative to the complexity of the models; therefore, caution is required in interpreting the model results.

Confirmatory factor analysis was conducted using three possible measurement models (Models H, I, and J; Table 13, **Appendix C**), two with second order latent variables representing toxicity, and one model with a first order latent toxicity variable (Figure 9; **Appendix C**). All of the measurement models had adequate or nearly adequate fit (Table 13; **Appendix C**). The relative strengths and weaknesses of each model are discussed in detail in **Appendix C**. The most appropriate model was the first order latent variable model combining all endpoints into a single toxicity latent (e.g. Model J, **Figure 6.6**, shown below) which incorporated both toxicant

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

SEM Approach: Structural Equation Modeling

September 9, 2013

concentrations and environmental covariates as the basis for predictive modeling. Because the model did not describe the northern wheatgrass data well, Model K was constructed to obviate problems associated with negative residual variances.

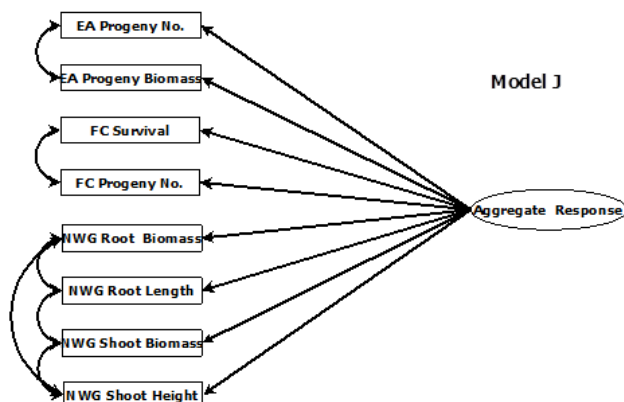


Figure 6.6: Confirmatory factor analysis Model J for the cross-site data.

Standardized and unstandardized path coefficients for Models J and K are shown in **Table 6.6** and the relationships between the aggregate response variable and individual species endpoints are depicted in Figure 10 (**Appendix C**).

As with the spiked sample data, a standard non-linear regression analysis (Stephenson *et al.*, 2000; Environment Canada, 2005a) of the relationship between the aggregate species response variable predicted in the structural equation model (Model K, Table 13, **Appendix C**; **Table 6.6**) and the summed concentrations of the F2, F3, and F4 fractions was conducted. Five analytical regression models (logistic, hormesis, exponential, Weibull, Gompertz) were fit to these data using the “drm” function in the “drc” library in the R 2.14 package (Ritz and Streibig, 2005; R Development Core Team, 2011).

The hormesis model had the best fit (Table 15 and Figure 11; **Appendix C**); however, the fit of these models is not ideal. This is likely the result of varying responses to environmental covariates. Estimated IC25 and IC50 values were not calculated because the poor fit of the models resulted in failure of the “ED” function in the “drc” library.

Table 6.6: Measurement model results for the combined site data

	Model H	Model I	Model J	Model K
Number of model parameters	29	28	32	38
Degrees of Freedom	15	16	12	16
$\chi^2$	24.18; p=0.0620	32.11; p=0.0097	22.07; p=0.0368	32.11; p=0.0097
CFI	0.978	0.962	0.976	0.962
RMSEA	0.073	0.093	0.085	0.093
SRMR	0.044	0.072	0.040	0.072
AIC	1167.15	1176.87	1167.98	1176.87

**Table 6.6: Measurement model results for the combined site data**

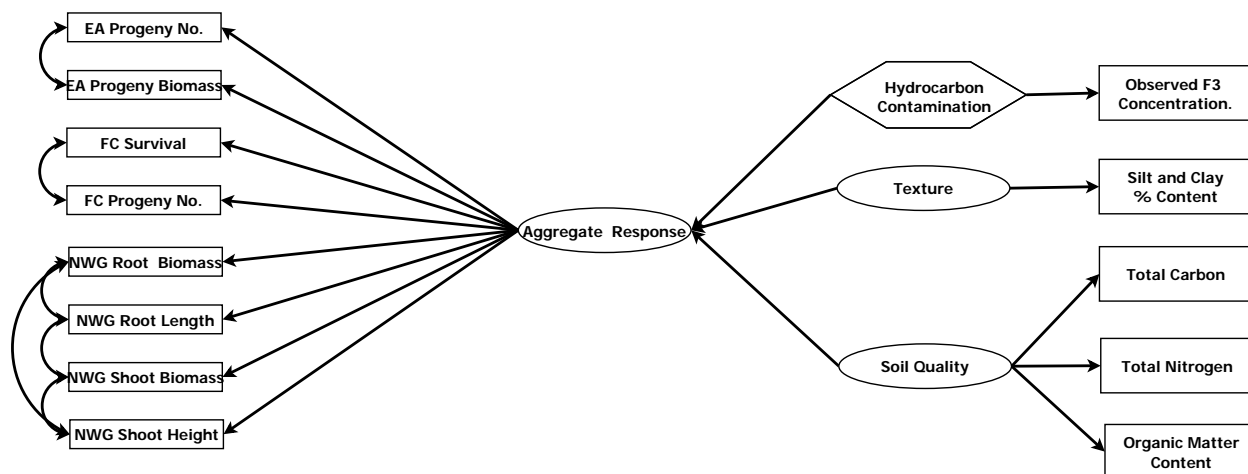
	Model H	Model I	Model J	Model K
--	---------	---------	---------	---------

CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; AIC = Akaike's Information Criteria

### 6.2.5 Models Incorporating Contaminant Levels and Multiple Environmental Predictors

Cross-site measurement models were developed and the measurement models incorporated into structural models containing predictor variables including toxicant concentrations and associated environmental conditions. Because the analyses were limited to the variables measured in common across all three sites, potentially important variables measured at only one site have been ignored.

Several models with different constructs were investigated and found wanting (Model L, M, N; see Tables 16 and 19, **Appendix C** for discussion of their strengths and limitations). Most of the structure of the models was similar to that shown in **Figure 6.7**. All of the models investigated had limitations for various reasons. One of the factors greatly influencing the application of the models was three outlying data points. When these were removed, the fit was substantially improved (e.g., Model N versus Model M; **Table 6.7**), and resulted in a much higher  $R^2$  value (e.g., for the species aggregate response  $R^2 = .951$ ). Despite the improved fit, none of the paths from the species endpoints to the aggregate response variable were significant (**Table 6.8**). Similarly, none of the structural paths from the predictor variables to the aggregate species response were significant (**Table 6.9**). Clearly estimation of the species responses in Models L and M were substantially driven by those data points with the highest contamination level and hence the strongest responses. The sensitivity of these models to a small number of data points highlights the necessity for much larger cross-site data sets to achieve the goal of successful cross-site predictive modeling.



**Figure 6.7: Model L - a structural equation model relating species responses to environmental covariates. The hexagon around hydrocarbon contamination indicates that that variable is a composite incorporating nonlinear biological responses.**

## ALTERNATIVE PROCESS FOR DEVELOPING TIER 2 SSROS

SEM Approach: Structural Equation Modeling

September 9, 2013

**Table 6.7: Results of a cross site models incorporating environmental covariates with (Model M) and without (Model N) three multivariate outlying data points.**

	Model M	Model N
Number of model parameters	52	53
Degrees of Freedom	62	61
$\chi^2$	599.16; p<0.0001	477.35; p<0.0001
CFI	0.731	0.765
RMSEA	0.273	0.246
SRMR	0.194	0.147
AIC	3106.86	2749.75

CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual

**Table 6.8: Measurement model paths (paths between latent and composite variables and their indicators) for Model N**

Path	Unstd. Est.	Std Err	Est. / Std Err	P-Value	Std. Est.
AGG. RESP. BY EA_PR	0.177	0.159	1.112	0.266	0.703
AGG. RESP. BY EA_PR_DM	0.110	0.098	1.116	0.264	0.690
AGG. RESP. BY FC_SURV	0.044	0.042	1.050	0.294	0.280
AGG. RESP. BY FC_PROG	0.071	0.067	1.056	0.291	0.308
AGG. RESP. BY NWG_SH_L	0.058	0.052	1.116	0.264	0.885
AGG. RESP. BY NWG_RO_L	0.028	0.025	1.146	0.252	0.448
AGG. RESP. BY NWG_SH_M	0.075	0.067	1.118	0.264	0.732
AGG. RESP. BY NWG_RO_M	0.041	0.036	1.129	0.259	0.465
Texture BY Fines	1	0	n/a	n/a	0.948
Soil Carbon BY Total_C	1	0	n/a	n/a	0.948
Soil Nitrogen BY Total_N	1	0	n/a	n/a	0.948
Organic Content BY OrgCont	1	0	n/a	n/a	0.948
Contam ON F3	1	0	n/a	n/a	3.097
Contam ON F32	-0.068	0.037	1.859	0.063	-2.224
EA_PR WITH EA_PR_DM	0.189	0.055	3.430	0.001	0.452
FC_SURV WITH FC_PROG	0.521	0.287	1.811	0.070	0.777
NWG_SH_L WITH NWG_RO_L	-0.014	0.003	4.206	<0.001	-0.390
NWG_SH_L WITH NWG_SH_M	0.032	0.005	6.369	<0.001	0.741
NWG_SH_L WITH NWG_RO_M	0	0.005	0.042	0.966	-0.004
NWG_RO_L WITH NWG_SH_M	-0.027	0.006	4.512	<0.001	-0.340
NWG_RO_L WITH NWG_RO_M	0.056	0.006	8.683	<0.001	0.635
NWG_SH_M WITH NWG_RO_M	-0.001	0.012	0.114	0.909	-0.013

NOTE: Single indicator latents and the first variable in a composite variable have unstandardized estimates that are set to one as a requirement of the model fitting process.

ABBREVIATIONS: AGG. RESP. = Aggregate Response; EA = *Eisenia andrei*; PR = earthworm progeny production; PROG = collembolan progeny production; FC = *Folsomia candida*; DM = dry mass; SURV = survival; BA = barley; SH = shoot; L = length; M = mass; RO = root; Alf = alfalfa; NWG = northern wheatgrass

**Table 6.9: Structural model paths in Model N**

Path	Unstd. Est.	Std Err	Est. / Std Err	P-Value	Std. Est.
AGG. RESP. ON Contam	0.977	1.317	0.741	0.459	0.178
AGG. RESP. ON Texture	2.421	2.312	1.047	0.295	0.689
AGG. RESP. ON Soil Carbon	-7.266	6.324	1.149	0.251	-1.030
AGG. RESP. ON Soil Nitrogen	9.746	9.406	1.036	0.300	2.244
AGG. RESP. ON Organic Content	-4.220	4.727	0.893	0.372	-0.938
Soil Carbon ON Organic Content	0.678	0.008	86.817	<0.001	1.063
Soil Nitrogen ON Organic Content	0.981	0.033	29.831	<0.001	0.947

## 6.3 CONCLUSIONS

### 6.3.1 Utility of SEM for Toxicological Data

This section of the report has tried to demonstrate the potential utility of the SEM approach. The major outcome of this investigation was the prospective use of confirmatory factor analysis to aggregate multiple endpoints into a single latent variable that can then be incorporated into standard non-linear modeling procedures to estimate IC25 values. This provides a direct solution to the problem of reconciling divergent ICp estimates from individual endpoints. In particular, the confirmatory factor analysis is uniquely able to identify endpoints that may not be responding in the same manner as the majority (variables with weak and/or nonsignificant loadings on the latent variable). With this approach the toxicologist can determine with confidence whether all of the endpoints are providing equivalent information and, if so, develop a single IC25 estimate from the latent variable using standard procedures.

The overall goal of this project was to develop analytical methods that could incorporate environmental covariates into analyses of toxicological responses and to develop cross-site predictive models that could be used to estimate provisional remediation targets based on readily measured environmental variables. Models L through N (Subsection 6.03.4 and **Appendix C**) linked an aggregate species response variable based on two earthworm endpoints, two collembolan endpoints, and four northern wheatgrass endpoints to toxicant concentrations and measures of soil quality.

These cross-site models are promising, but not ready for implementation in a predictive mode. The models successfully explained the aggregate species responses ( $R^2 > 0.7$ ), but failed many tests of model adequacy (significant  $\chi^2$ , low CFI etc.). A small sample size relative to the complexity of the models is a major impediment to the implementation of these models.

Second, these models were applied to minimal cross-site data (e.g. those collected from three sites). Clearly, the dataset was constrained by too few data and data that were inconsistent in terms of the parameters measured. Data from a much larger number of sites (likely 20+) are required for a scope of inference sufficiently wide for valid cross-site predictive modeling. Also, the non-linear contaminant-aggregate-species relationships are a second important constraint.

Non-linear relationships are expected in toxicological data (Stephenson *et al.*, 2000; Environment Canada, 2005a), but difficult to handle in an SEM framework. Composite variables (Grace and Bollen, 2008) provide a workaround, but do not fully capture the potential range of nonlinear relationships that could reasonably occur. Recent advances in “3rd generation SEM” (Grace *et al.*, 2012) may provide an effective tool for non-linear SEM, but await widespread implementation and acceptance

.

## **7.0 Recommendations**

---

### **7.1 GMR APPROACH**

We could see no advantage to pursuing this approach for deriving SSROs. The alternative approaches proved more interesting and advantageous.

### **7.2 DRAMA APPROACH**

This approach had major advantages in that the model averaging removes the dependence on selecting the “best” fitting model and address the associated uncertainty therein. This approach allows for an objective synthesis of the models and in the process reconciled contradicting interpretations. The DRAMA approach illustrated the importance of non-contaminant soil quality variables (e.g., specific edaphic variables) relative to contaminant variables (e.g., PHC concentrations) with respect to the explaining the variability in the biological response data. The relative “importance” of the non-contaminant variables as explanatory variables might be due to the influence they play on contaminant bioavailability. This approach should be pursued and further developed using a much larger data set and the resulting predictive models developed from the exercise should be verified via designing of a field study to assess the validity of those predictions. The fact that “study” explained less of the variability in the response data than PHCs suggests that the resultant models can be applied effectively across sites.

### **7.3 PLS APPROACH**

The PLS approach was applied to a single study so the results of this analysis should be considered preliminary. Despite that caveat, the approach demonstrated that it is possible to link multivariate soil properties to specific ecotoxicological responses. Although the analyses suggests that the predictive power of these models might be inadequate for soils with properties that vary substantially from those soils from which the initial model(s) was(were) constructed, this could be a function of the small dataset examined. We recommend that the approach, be further developed with a much larger data set and the resulting models assessed through field studies designed to test the predictions from the PLS models. The PLS approach requires the development of model selection criteria in order to identify and apply the most appropriate model(s) for predicting the biological responses expected in light of the existing contaminant levels and pedologic characteristics. We also recommend that the accuracy or veracity of the predictions be tested through a field study.

### **7.4 SEM APPROACH**

From a toxicological perspective, structural equation modeling was the most comprehensive and novel approach assessed in this project. Although a relatively complex approach fraught with constraints and limitations with respect to data quantity and quality, we have identified



further steps required for continued development of this approach. Further steps (Phase 2) in the development of cross-site predictive models should include:

1. Collation of a much larger dataset for evaluation of the cross-site application - a sufficient database will include standard environmental covariates (e.g., texture, organic matter etc.) with a range wide enough to encompass expected soil conditions at any site within the geographic region of interest.
2. Development of SEM models similar to those presented in this report in consideration of the full database.
3. Validation of the model predictive power by fitting the models to subsets of data and comparing predicted versus actual results.
4. Testing and validation of “3rd generation SEM” models using the same data.

A primary barrier in the development of the models in this report was the sampling design for the toxicological dataset that was used. The following modifications (Subsection 7.5) are recommended; these recommendations should improve the statistical value of toxicological data, particularly for cross site comparison for any statistical approach used to examine ecotoxicological data relative to pedological characteristics and contamination profiles.

## **7.5 RECOMMENDED MODIFICATIONS FOR DATA COLLECTION**

### **7.5.1 Choice of Endpoints and Environmental Covariates**

There is extensive literature on the selection of species endpoints in toxicological testing. While there are often important reasons for the selection of particular species, lack of standardization severely hampers cross-site analysis. In this report (Section 6.0) models A-G utilized 16 endpoints collected on 5 separate taxa; however, models H-N utilized only 8 endpoints from 3 separate taxa. Additional endpoints were available in each of the three studies aggregated for models H-N, but could not be used because they were not available from all three studies. This problem is likely to be exacerbated as larger numbers of studies are combined. We recommend that a core list of “default” species be agreed to and tested in all studies. Additional species should be added according to the needs of individual studies. Similarly, a core list of environmental covariates including soil texture, organic matter content, total nitrogen, and total carbon, should be measured in all cases.

### **7.5.2 Sampling Design**

The current sampling designs rely on a small number of field-collected samples and emphasize repeated testing of replicates within those samples. Such a design is not optimal statistically, as it limits the range of environmental covariates, and results in datasets with limited numbers of independent samples. In cases where the number of field-collected samples is severely limited such as in this report, it may be necessary to treat replicates as individual samples to achieve sufficient sample sizes for modeling. The preferred way to deal with replicate subsamples is to

take their mean and thus treat each field-collected soil as an independent sample. This approach, however, requires larger sample sizes than the 8 to 12 independent soil samples available in each of the studies included in this report.

Optimal sampling designs will achieve greater statistical power without substantially increasing the costs of sampling. We suggest that a shift in emphasis from multiple endpoint replicates within each soil sample to a smaller number of replicates across a larger number of soil samples. This, combined with a set of common endpoints as suggested in the section above, would allow greater statistical power for similar effort in the laboratory. Specifically, we recommend:

- A maximum of three endpoint replicates within a given soil sample. Using a mean of 3 replicates in further analysis will reduce the potential impact of an outlier in one of the replicates. Given that 5 to 10 replicates are frequently used, this should allow at least twice as many soil samples to be analyzed, providing a corresponding increase in statistical power. Note that the cost of biological testing is reduced but the analytical costs increase because more soil samples are chemically analyzed.
- Validation of the choice of three replicates as a standard should be done. Specifically, for each standard endpoint the expected range in variance from three replicates should be determined so that practitioners can evaluate whether their sample of three can be considered representative. Laboratory replicates provide information on quality assurance but multiple site soil samples provide information regarding the site, the distribution (magnitude and extent) and effects of the contamination; therefore, it behooves a proponent to collect and analyze more soil samples from the field (true replication) than to increase the number of laboratory replicates (pseudo-replication).

## **7.6 RECOMMENDATIONS FOR FUTURE WORK**

We recommend that development of the SEM and the DRAMA approaches continue with a much larger dataset, as the preliminary analyses conducted for this project has clearly demonstrated the utility of these approaches. Cross-site application was attempted to determine the feasibility of taking these approaches and both appear promising. With models constructed from a much larger database, the adequacy of those models will improve and, as a result, the predictions should prove more useful. Again, once the approach has been further developed and the models developed become robust, then the predictions should be assessed through the implementation of a field study that is designed to test these predictions. Further, the recommendations for modifications to the sampling design would be beneficial regardless of the statistical approach taken and therefore they apply equally to the other approaches as well.

## 8.0 Sign-off

---

This document entitled '*CAPP 09-913-50 Alternative Process for Developing Tier 2 SSROs*' was prepared by Stantec Consulting Ltd. for the account of Petroleum Technology Alliance Canada. The material in it reflects Stantec's best judgment in light of the information available to it at the time of preparation. Any use which a third party makes of this report, or any reliance on or decisions made based on it, are the responsibilities of such third parties. Stantec Consulting Ltd. accepts no responsibility for damages, if any, suffered by any third party as a result of decisions made or actions based on this report.

The following individuals participated in the preparation of this report:



---

**Dr. Gladys Stephenson, Ph.D.**

Environmental Toxicologist  
Stantec Consulting Ltd.  
Guelph, ON



---

**Robin Angell, M.Sc.**

Environmental Scientist  
Stantec Consulting Ltd.  
Guelph, ON



---

**Dr. Eric Lamb, Ph.D.**

Associate Professor  
University of Saskatchewan  
Saskatoon, SK



---

**Dr. Melissa Whitfield-Aslund, Ph.D.**

NSERC Post-doctoral Fellow  
Intrinsic Sciences Inc.  
Mississauga, ON



---

**Dr. Barry Zajdlik, Ph.D. Candidate**

University of Waterloo  
Zajdlik & Associates  
Rockwood, ON

## 9.0 Cited References

---

- Alberta Environment. 2010. Alberta Tier 1 Soil and Groundwater Remediation Guidelines. December 2010 ISBN: 978-0-7785-9949-4 (On-line Edition).
- Alberta Environment. 2007. Tier 2 eco-contact guideline derivation protocol. Edmonton, AB., ISBN: 978-0-7785-6752-3.
- Bailer, A.J., R. Noble, and M. Wheeler. 2005. Model uncertainty and risk estimation for experimental studies of quantal responses. *Risk Analysis*. 25:291-299.
- Barton, K. 2012. MuMIn: Multi-model inference. R package version 1.7.12. <http://CRAN.R-project.org/package=MuMIn>.
- Bates, D., M. Maechler and B. Bolker. 2011. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-41. <http://CRAN.R-project.org/package=lme4>.
- Bollen, K. A. 1989. Structural equations with latent variables. John Wiley & Sons, New York.
- Bollen, K. A. and J. S. Long, editors. 1993. Testing structural equation models. Sage Publications, Newbury Park, CA.
- Buckland, S. T., K. P. Burnham and N. H. Augustin. 1997. Model Selection: An Integral Part of Inference. *Biometrics*. 53:603-618.
- Burnham, K.P. and D.R. Anderson. 2002. Model selection and Multimodel Inference: A Practical Information Theoretic Approach. 2<sup>nd</sup> Ed. Springer, New York. 488pp.
- Byrne, B. M. 2012. Structural Equation Modeling with Mplus. Routledge, New York, NY.
- Canadian Council of Ministers of the Environment. 2006. A protocol for the derivation of environmental and human health soil quality guidelines. Canadian Council of Ministers of the Environment, Winnipeg, MB.
- Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. *J. Roy. Stat. Soc. Ser. A*. 158:419-466.
- Claeskens, G. and N.L. Hjort. 2008. Model Selection and Model Averaging. Cambridge University Press, New York. 312 p.
- Cottingham, K. L., J. T. Lennon, and B. L. Brown. 2005. Knowing when to draw the line: designing more informative ecological experiments. *Frontiers in Ecology and the Environment* 3:145-152.

- Cramer, R. D. 1993. Partial Least Squares (PLS): Its strengths and limitations. *Perspect. Drug Discov. Des.* **2**:269-278.
- Delignette-Muller, M. L., R. Pouillot, J.-B. Denis, and C. Dutang. 2010. *fitdistrplus*: help to fit of parametric distribution to non-censored or censored data.
- Environment Canada. 2005a. Guidance document on statistical methods for environmental toxicity tests. Ottawa, ON.
- Efroymson, R.A., B.E. Sample and M.J. Peterson. 2004. Ecotoxicity test data for total petroleum hydrocarbons in soil: plants and soil-dwelling invertebrates. *Human and Ecological Risk Assessment*: 10(2):207-231.
- Environment Canada. 2004. Biological test method: Tests for toxicity of contaminated soil to earthworms (*Eisenia andrei*, *Eisenia fetida*, or *Lumbricus terrestris*). Report EPS 1/RM/43. Ottawa, ON.
- Environment Canada. 2005b. Biological test method: Test for measuring emergence and growth of terrestrial plants exposed to contaminants in soil. Report EPS 1/RM/45. Ottawa, ON.
- Environment Canada. 2007. Biological test method: Test for measuring survival and reproduction of springtails exposed to contaminants in soil. Report EPS 1/RM/47. Ottawa, ON.
- Eriksson, L., E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström, and S. Wold. 2006. *Multi- and Megavariate Data Analysis Part I Basic Principles and Applications*, 2nd ed., Umetrics, Umeå, Sweden.
- Filzmoser, P. and K. Varmuza. 2010. *Chemometrics: Multivariate Statistical Analysis in Chemometrics. R package version 1.3.7*; <http://CRAN.R-project.org/package=chemometrics>.
- Gough, L. and J. B. Grace. 1999. Effects of environmental change on plant species density: comparing predictions with experiments. *Ecology* 80:882-890.
- Grace, J.B. 2006. Structural equation modeling and natural systems. Cambridge University Press, U.K.
- Grace, J. B. 2008. Structural equation modeling for observational studies. *Journal of Wildlife Management* 72:14-22.
- Grace, J.B. and B.H. Pugeseck. 1997. A structural equation model of plant species richness and its application to a coastal wetland. *American Naturalist* 149:436-460.

- Grace, J.B. and K.A. Bollen. 2005. Interpreting the results from multiple regression and structural equation models. *Bulletin of the Ecological Society of America* 86:283-295.
- Grace, J.B. and K.A. Bollen. 2008. Representing general theoretical concepts in structural equation models: the role of composite variables. *Environmental and Ecological Statistics* 15:191-213.
- Grace, J.B., T.M. Anderson, H. Olff, and S.M. Scheiner. 2010. On the specification of structural equation models for ecological systems. *Ecological Monographs* 80:67-87.
- Grace, J.B., D.R. Schoolmaster, G.R. Guntenspergen, A.M. Little, B.R. Mitchell, K.M. Miller, and E.W. Schweiger. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere* 3:art73.
- Guthery, F.S., L.A. Brennan, M.J. Peterson and J.J. Lusk. 2005. Information theory in wildlife science: critique and viewpoint. *J. Wildlife Management*. 69:457-465.
- Hawkins, D. M., S. C. Basak and D. Mills. 2003. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* 2:579-586.
- Jöreskog, K. G. 1973. A general method for estimating a linear structural equation system. Pages 85-112 in A. S. Goldberger and O. D. Duncan, editors. *Structural equation models in the social sciences*. Seminar Press, New York.
- King, J. R. and D. A. Jackson. 1999. Variable selection in large environmental data sets using principal components analysis. *Environmetrics*. 10:67-77.
- Kline, R. B. 2011. *Principles and practice of structural equation modeling*. 3rd edition. Guilford Press, New York.
- Lamb, E. G., S. J. Shirliffe, and W. E. May. 2011. Structural equation modeling in the plant sciences: an example using yield components in oat. *Canadian Journal of Plant Science* 91:603-619.
- Legendre, P. and L. Legendre. 1998. *Numerical Ecology*, 2nd English Edition. Elsevier, New York. 853pp.
- Magee, L. 1990.  $R^2$  measures based on Wald and likelihood ratio joint significance tests. *Amer. Stat.* 44: 250-253.
- Mazerolle, M.J. 2012. AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c). R package version 1.26. <http://CRAN.R-project.org/package=AICcmodavg>.
- McCullagh, P. and J.A. Nelder. 1989. *Generalized linear models*. Chapman and Hall, 511pp.

- Muthén, L. K. and B. O. Muthén. 2002. How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal* 9:599-620.
- Muthén, L. K. and B. O. Muthén. 2010. *Mplus users guide*. Muthén and Muthén, Los Angeles, CA.
- Nevitt, J. and G. R. Hancock. 2001. Performance of Bootstrapping Approaches to Model Test Statistics and Parameter Standard Error Estimation in Structural Equation Modeling. *Structural Equation Modeling* 8:353-377.
- Ontario Ministry of Environment. 2011. *Soil, Ground water and Sediment Standards for use under Part XV.1 of the Environmental Protection Act*, revised version, April 15, 2011.
- Paton, G.I., K. Killham, H.J. Weitz and K.T. Semple. 2005. Biological tools for the assessment of contaminated land: applied soil ecotoxicology. *Soil Use and Management*. 21:487–499.
- Pinheiro, J.C. and D.M. Bates. 2000. *Mixed Effects Models in S and S-plus*. Springer-Verlag, New York. 528 p.
- Pinheiro, J, D. Bates, S. DebRoy, D. Sarkar and R Development Core Team. 2010. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-97.
- Pugesek, B. H., A. Tomer, and A. von Eye, editors. 2003. *Structural equation modeling: applications in ecological and evolutionary biology*. Cambridge University Press, R Development Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- R Development Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- R Development Core Team. 2011. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- R Development Core Team. 2009. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Renoux, A.Y., B. Zajdlik, G.L. Stephenson, L.J. Moulins. 2012. Risk-based management of site soils contaminated with a mixture of hazardous substances: Methodological approach and case study. *Human and Ecol. Risk Assessment*. (DOI: 10.1080/10807039.2012.691825).



- Ritz, C. and J. C. Streibig. 2005. Bioassay analysis using R. *Journal of Statistical Software* 12:1-18.
- Rodgers, J. L. 1999. The Bootstrap, the Jackknife, and the Randomization Test: A Sampling Taxonomy. *Multivariate Behavioral Research* 34:441-456.
- Satorra, A. and P. M. Bentler. 1994. Corrections to test statistics and standard errors in covariance structure analysis. Pages 399-419 in A. von Eye and C. C. Clogg, editors. *Latent variables analysis: Applications for developmental research*. Sage Publications, Inc, Thousand Oaks, CA, US.
- Schermelleh-Engel, K., H. Moosbrugger, and H. Müller. 2003. Evaluating the fit of structural equation models: tests of significance and descriptive goodness of fit measures. *Methods of Psychological Research Online* 8:23-74.
- Semple, K.T., A.W.J. Morriss and G. I. Paton. 2003. Bioavailability of hydrophobic organic contaminants in soils: fundamental concepts and techniques for analysis. *European Journal of Soil Science*. 54:809–818.
- Shipley, B. 2000. *Cause and correlation in biology*. Cambridge University Press, U.K.
- Stephenson, G. L., N. Koper, G. F. Atkinson, K. R. Solomon, and R. P. Scroggins. 2000. Use of nonlinear regression techniques for describing concentration-response relationships of plant species exposed to contaminated site soils. *Environmental Toxicology and Chemistry* 19:2968-2981.
- Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Stat. Theory and Methods*. A7:12-26.
- Varmuza, K. and P. Filzmoser. 2009. *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press, Boca Raton, FL.
- Vendeginste, B. G. M., D. L. Massart, L. M. C. Buydens, S. de Jong, P. J. Lewi and J. Smeyers-Verbeke. 1998. *Handbook of Chemometrics and Qualimetrics: Part B*, Elsevier, Amsterdam, NL.
- Wang, H, X. Zhang and G. Zou. 2009. Frequentist model averaging estimation: A review. *J. Systems Sci. & Complexity*. 22(4):732-748.
- Wong, DCL, E.Y. Chai K.K. Chu and P.B. Dorn. 1999. Prediction of ecotoxicity of hydrocarbon-contaminated soils using physicochemical parameters. *Env. Toxicol. Chem.* 18: 2611–2621.



Wothke, W. 1993. Nonpositive definite matrices in structural modeling. Pages 256-293 in K. A. Bollen and J. S. Long, editors. Testing structural equation models. Sage publications, Newbury Park, CA.

Wright, S. 1921. Correlation and causation. Journal of Agricultural Research 20:557-585.