State of Oregon
Department of Environmental Quality

Memorandum

To: Mid-Coast Sediment Technical Working Group
From: Peter Bryant - Watershed Management – WQ - HQ

Subject: Documentation of methods and results for identifying watershed characteristics and initial statistical model development.

Purpose

This document describes the methods used to complete the source assessment and linkage analyses for the Mid Coast TMDL that addresses CWA Section 303(d) listings for sedimentation impairments and biological impairments attributed to fine sediment loading to streams and rivers. DEQ presented this information to the Sediment Technical Working Group (TWG) on January 14, 2015.

We (Oregon Department of Environmental Quality) solicit TWG member feedback on the specific methods used to develop the statistical model as explained in this document. Please submit your review by February 20th, 2015. There are four questions we want you to consider in your review of this document:

1. Are there additional considerations we should evaluate for the methods used to select variables and formulate models?
2. Are there additional or alternative model performance criteria we should use for the model selection?
3. Are the variables in the selected model reasonable with respect to your understanding of the systems?
4. Is there additional discussion or are there references that could be added to the document that would help make these methods more understandable?

If you want to highlight previous work, please include a complete citation of the work and provide a short explanation of how this cited work relates to the specific part of the methods used to develop the statistical model. Previous work that you raise and don’t provide the citation and explanation for will be given a lower priority than citations that are complete and include an explanation.

Following review of comments, we will evaluate and finalize the analyses. The final source analyses results will be evaluated for utility in refining point source waste load allocations, nonpoint source load allocations, and other key components of the sediment TMDLs. We appreciate your efforts to improve this document and request that if you intend to send this document to parties outside the TWG that you let the facilitator know so they can be informed of the general process of our work.

General Approach

Sediment dynamics in streams can be modeled using mechanistic models of sediment budgets or with statistical models that relate factors influencing the generation, transport and deposition of sediment to in-stream sediment measurements (USEPA 1999). We employed a statistical model due to the lack of the specific data needed to parameterize a mechanistic model and the prohibitive cost and timeliness of collecting those data. The statistical model, however, could be developed using data readily available and easily compiled through Geographical Information System (GIS) analysis. Furthermore, a statistical model can be applied at scales required for TMDL development while still capturing our conceptual understanding of sediment loading to streams and rivers based on published literature. Because we are analyzing in-stream characteristics, we decided to use a framework for incorporating stream network flow relationships developed by Ver Hoef and Peterson (2010). The statistical model relates watershed characteristics to a measure of in-stream sediment by taking into account spatial relationships among sampling points on the stream network. The measure of in-stream sediment we are using is the fine sediment score (FSS) for aquatic macroinvertebrates. The FSS is a biological index of individual macroinvertebrate taxon responses to the abundance of fine sediment (Huff et al 2006). We used FSS because the impairments addressed
were based on biological criteria. In addition, we have macroinvertebrate data collected at a spatial scale appropriate for this analysis.

The source analysis steps were:

1. Organize the data,
2. Evaluate relationships among characteristics,
3. Select method and formulate models,
4. Model selection,
5. Analysis and interpretation of results.

These steps correspond to the following sections in this document and are displayed in Figure 1. The first section, Organize the data, details the identification of several natural (non-anthropogenic) and human (anthropogenic) watershed characteristics influencing the generation, transport and deposition of sediment through TWG input and their calculation using Geographical Information System (GIS) tools. The second section, Evaluate relationships among characteristics, details the use of the random forest machine learning method used to identify watershed characteristics that had strong relationships with FSS and to identify and remove correlated characteristics. The third section, Select method and formulate models, provides the approaches used to determine which combinations of variables were used to formulate models based on statistical properties and ecological functional relationships. The fourth step, Model selection, details the specific methods and measures used to determine the best model. The fifth step, Analysis and interpretation of results, details model checking methods used to verify model assumptions and to validate the model.

Figure 1. Flow chart detailing the steps in the source assessment and linkage analyses for the Mid Coast Sediment TMDL. Steps 1 through 4 are represented here including; 1) Organize the data, 2) Evaluate relationships among characteristics, 3) Select method and formulate models and 4) Model selection. Step 5, not represented in the chart, is the Analysis and interpretation of results.

1. Organize the data
Through feedback solicited from the Sediment TWG, we identified a list of natural (non-anthropogenic) and human (anthropogenic) watershed characteristics that may influence the generation, transport, or deposition of fine sediments (Appendix A). We had sufficient macroinvertebrate data to calculate FSS at 564 sites, with a total of 783 samples across Oregon watersheds in the level 3 Coast Range Ecoregion. The sampled area includes most of the western coastal watersheds of Oregon draining to the Pacific Ocean and the lower Columbia River (Figure 2; Appendix A). We used GIS to organize the TWG-identified watershed characteristic data and to calculate values for each macroinvertebrate sample.

In order to evaluate the scale at which natural or anthropogenic watershed characteristics had the greatest influence on FSS, we calculated values at the individual site (e.g. precipitation), sampled reach (e.g. reach slope), reach streamside area (RSA), reach contributing area (RCA), accumulated streamside area (ARSA) and the accumulated reach contributing area (ARCA; Figure 3). The sampled reach was defined as 150 meters or 40 times the wetted width, which ever was longest. The RSA was 45 meters from both sides of the stream from the site to the top of the catchment. The RCA was the watershed area within the catchment upstream of the sample location. The ARSA was 45 meters from both sides of the stream along the entire stream network upstream of the sample location. The ARCA was the watershed area in the entire network upstream of the sample location. Most characteristics were expressed as percentages (e.g., percent of erodible lithology in RCA) to standardize measures across different spatial scales. Individual characteristics were selected to represent measurements of a general process or characteristic such as precipitation or lithology.

**Figure 2.** Sampling sites in western Oregon included in the construction of the spatial stream network model to relate landscape variables to fine sediment score (FSS) for the Mid Coast Sediment TMDL source assessment. Bold outlines are the Oregon Water Resources Department administrative boundaries. Light grey outlines are National Hydrography Dataset fourth field HUC boundaries.
Figure 3. Multiple scales of influence where watershed characteristics were calculated. a) Site/Reach: Bottom of sampled reach/reach extent. b) Reach Streamside Area (RSA). c) Reach Contributing Area (RCA). d) Accumulated Reach Streamside Area (ARSA). e) Accumulated Reach Contributing Area (ARCA).

2. Relationships among characteristics

We selected 112 variables to consider for inclusion in the FSS statistical model (refer to Appendix A for a complete list of variables). In order to make this list manageable for statistical model development, we used the random forest (RF) method to select the most important variables explaining variability in FSS scores across sites (Breiman 2001). The advantages of the RF method are it does not require transformation of the variables (e.g., log transformation), and provides a measure of variable importance in explaining the response variable (Cutler et al. 2007, De’ath and Fabricius 2000). This method has been used to identify landscape variables affecting response of aquatic macroinvertebrate communities to landscape stressors (Carlisle 2009, Cutler et al. 2007, De’ath and Fabricius 2000, Murphy et al. 2010, Prasad et al. 2006). The random forest method assembles multiple regression tree models and averages these individual models to estimate an optimal classification scheme. It also withholds a portion of the data set to validate the classification scheme and to assess individual importance of explanatory variables, calculated as the decrease in model accuracy when an individual variable is removed from the model (Cutler et al. 2007).

The RF method requires a specified number of variables from which to randomly select and build regression trees (mtry) and a specified number of trees to grow per forest (ntree). According to Geneur et al. (2010), the optimal values of mtry and ntree to promote the importance of true variables and to stabilize the results are 1/3 the number of variables (p/3) and 2000 trees, respectively. We developed 50 individual random forest models using the randomForest package in the R statistical software (R Development Core Team 2012) in order to calculate a distribution of variable importance values (Geneur et al. 2010). We sorted the variables by the median of the importance measures from the 50 runs. After examining the rankings we selected the top 50% of the ranked variables for consideration in the next model development step (Figure 4).
Figure 4. Plot of the 112 variables included in the 50 runs of random forest (mtry = p/3, ntree = 2000). The x-axis is the percent increase in mean squared error (MSE) and is a measure of variable importance affecting the accuracy of the model when that variable is removed. The y-axis are the individual variables, refer to Appendix A for a complete list of variables. The box shows the variables in the top 50% of ranked variables selected for correlation evaluation.

An important consideration in random forest modeling is that strong correlations among explanatory variables can decrease variable importance scores for set of variables describing similar characteristics. This happens because variables describing a common characteristic remain in the classification tree when an individual variable is removed. Model accuracy is thus preserved by the correlated (redundant) characteristics, artificially decreasing the importance score of the removed variable (Geneur et al 2010, Toloşi and Lengauer 2011).

To address possible reduction in importance due to correlation between explanatory variables in the top 50%, we evaluated the correlation between each variable within a conceptual group. A conceptual group is the group of variables identified that represent measurements of a larger process or characteristic. The conceptual groups identified within the watershed characteristics included precipitation, disturbance, lithology and soils, land use, landslide susceptibility, stream attributes and location. Refer to Appendix A for a list of variables and their assigned conceptual group. Correlations were evaluated using Pearson’s product moment correlation (Figure 5). If two or more variables were correlated within a conceptual group we selected the variable with the highest variable importance score. For example, precipitation variables were all correlated but the rainfall in the 3 years prior to the sample (sum_1095_days) had the highest importance value. We therefore selected sum_1095_days to represent “precipitation” in the subsequent model selection process (Figure 5). After removing significantly correlated variables, we arrived at a list of 14 variables to consider in the subsequent model selection process (Figure 6).
Figure 5. Correlation plots of a select number of variables. In the lower triangle, x and y axes are in the units of the measurement (mm). The diagonal are histograms of each variable. The upper triangle displays the p-value and the Pearson product moment correlation ($r$). When two or more variables were correlated, the variable with the greatest variable importance was selected. These variables represent the precipitation conceptual group. Refer to Appendix A for a complete description of each variable.

Figure 6. Variables remaining for consideration in the model selection ranked by importance as evaluated using random forest method after correlated variables have been removed. The x-axis is the percent increase in mean squared error (MSE) and is a measure of variable importance affecting the accuracy of the model when that variable is removed. The y-axis is the variable names. A complete description of each variable can be found in Appendix A.
3. Select method and formulate models

Ver Hoef and Peterson (2010) developed a framework for incorporating correlations due to spatial proximity into general linear models (GLMs) used for stream network modeling. This framework, the spatial statistical model on a stream network (SSN), incorporates spatial autocorrelation errors between sites in the context of hydrologic distance as opposed to typical geostatistical modeling methods which use Euclidean distance (Figure 7). The GLM component of the framework is a statistical linear model of the relationship between a response variable (FSS) and a set of explanatory variables (watershed characteristics). Examples where this framework has been successfully applied include macroinvertebrate indices in northeastern Australia (Frieden et al 2014), stream temperature in the Pacific Northwest (Isaak et al. 2011), stream temperature in the John Day River basin (Ruesch et al. 2012), and water chemistry in southeastern Australia (Ver Hoef and Peterson 2010).

**Figure 7.** Image from Peterson and Ver Hoef (2010) illustrating spatial relationships incorporated in the SSN framework for considering spatial autocorrelation between sampling locations.

The SSN uses a mixed model approach that incorporates spatial autocorrelation models in three potential moving average relationships; tail-up (upstream effect only; Figure 7a and 7b), tail-down (downstream effect only; Figure 7b and 7d) and Euclidean (straight line distance with bi-directional effects and without consideration of landscape structure; Ver Hoef and Peterson 2010). The four autocorrelation functions that can be used with each autocorrelation relationship include exponential, linear with sill, spherical and MARIAH (Ver Hoef et al 2006). The distribution of these functions in relation to hydrologic distance can be examined in Figure 8. Selection of an appropriate autocorrelation relationship can be done through the model selection process or based on prior knowledge of the autocorrelation relationship in the system measured (Garreta et al 2009). Ver Hoef and Peterson developed the Spatial Tools for the Analysis of River Systems (STARS Version 2.0.1) toolset (Peterson and Ver Hoef 2014) in ESRI ArcGIS and the Spatial Stream Network (SSN) package (Ver Hoef et al 2014) for R statistical software (R Development Core Team 2012) to implement this model framework.
Figure 8. Autocorrelation functions available for inclusion in the SSN construction. The final model used Exponential tail up, Exponential tail down and Exponential Euclidean.

We used the STARS toolset (Peterson and Ver Hoef 2014) and the SSN package (Ver Hoef et al 2014) to develop the SSN network model for the watersheds in the level 3 Coast Range Ecoregion where we have macroinvertebrate sampling data. The STARS package in ArcGIS allowed us to calculate the various watershed characteristics at the multiple scales of influence. The STARS toolset then creates the spatial context for the SSN package to import into R and perform the general linear modeling. Prior to inclusion in the model, explanatory and response variables were scaled from 0 to 1 by setting each variable’s minimum value to 0 and maximum value to 1. This had the effect of decreasing the influence of variables with a larger scale of measurement over variables with a smaller scale of measurement.

We used two approaches for model formulation using the set of 14 variables identified in the previous step. In one approach we used a model selection process using statistical properties to determine the variables to include in the model for consideration. In the other approach, we examined and selected variable combinations to populate models based on ecological processes described in the literature.

For the statistical properties approach we used a two-step process to model formulation and selection (Peterson and Ver Hoef 2010). The first step was to hold the autocorrelation function constant and use a backward deletion method to select explanatory variables. The second step was to hold the list of variables selected in the first step constant and evaluate different combinations of autocorrelation functions. For the first step, we started with all 14 variables identified in the random forest method and ran the model multiple times removing the variable with the highest p-value each time. The autocorrelation functions used for each of the candidate models in the first step were exponential tail-up, exponential tail-down and exponential Euclidean due to the simplicity of the calculation and general familiarity with the shape of the function. For the second step, we ran 11 combinations of tail-up, tail-down and Euclidean autocorrelation functions using the model with the fewest variables from the first step. All three autocorrelation relationships were included to represent a wide range of potential autocorrelations and to avoid the assumption of a specific spatial structure (Peterson and Ver Hoef 2010). To select the best performing combination of autocorrelation functions we used the minimum root mean squared percent error (RMSPE). There was no meaningful difference in RMSPE between the different combinations of autocorrelation functions in the context of FSS values (maximum difference of 0.002 FSS units). Given this consistency we will use the Exponential tail-up,
Exponential tail-down and Exponential Euclidean autocorrelation functions for simplicity of calculation and general familiarity with the shape of this function.

For the second approach to model formulation, we considered ecological functional relationships to determine which watershed characteristics to include in each candidate model. We started with the list of 14 variables and built candidate models based on three groupings including natural factors only, human factors only and combination natural and human factors. In this process we considered hypothesized interactions. We formulated natural factor only models to consider that natural factors alone could explain the variation in FSS (e.g. precipitation, lithology and soil content). We included human factor only models to investigate if these factors alone are drivers (e.g. streamside disturbance and road crossings). We formulated combination natural and human factor models to consider combinations and interactions between land use and landscape characteristics (e.g. interactions between precipitation and disturbance alone as well as lithology). A full list of the ecological models included in the analysis can be found in Appendix A. We used the exponential family of autocorrelation functions for each formulated model as determined by the second step of the statistical approach that evaluated the appropriate autocorrelation function.

4. Model selection

We pooled the models formulated from the Statistical properties and the Ecological processes approaches to select the best model overall (Figure 9). The best model was selected by comparing Akaike Information Criteria (AIC) and RMSPE (Akaike 1974). AIC is a metric that helps identify the most parsimonious model by balancing predictive power and the number of model parameters. The principle of parsimony refers to the idea that the best model is the simplest model that is able to adequately explain the process being modeled (Crawley 2013).

![Flow chart detailing the two approaches employed for model formulation, considering Ecological processes as well as relying on Statistical properties. We pooled the models from each approach and compared them to select the best overall model. We used AIC and RMSPE for model selection criteria.](image)

There were four models with very close AIC and RMSPE values (Table 1). The maximum difference in AIC between the models was 1.8 units. Differences in AIC scores between two models less than 2 suggest that both models are equally plausible given the data (Burnham and Anderson 2002). Given this, we look to the RMSPE to determine the best model. Two of the four models had the same low RMSPE so we selected the model with the fewest variables in the interest of parsimony.
Table 1. The four models with the lowest AIC. The * indicates the selected model considering AIC, RMSPE and the number of model parameters. Given AIC are within 2 and RMSPE are close for these models we selected the model with the fewest number of parameters. The values for RMSPE are in scaled log transformed FSS units. Refer to Table 2 for a description of the selected model variables and Appendix A for descriptions of all the variables.

<table>
<thead>
<tr>
<th>Formula</th>
<th>AIC</th>
<th>RMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>sum_1095_days + PDISRSA_1YR + POWNRCA_PRI + PALITHERODRCA + PAACLAYRCA + PASILTRCA + DAPOPRCA2010</td>
<td>-1226.24</td>
<td>0.103099</td>
</tr>
<tr>
<td>sum_1095_days + STRMPWR + PDISRSA_1YR + POWNRCA_PRI + PALITHERODRCA + PAACLAYRCA + PASILTRCA + DAPOPRCA2010</td>
<td>-1225.45</td>
<td>0.103336</td>
</tr>
<tr>
<td>*sum_1095_days + PDISRSA_1YR + POWNRCA_PRI + PALITHERODRCA + PASILTRCA + DAPOPRCA2010</td>
<td>-1225.34</td>
<td>0.103099</td>
</tr>
<tr>
<td>sum_1095_days + STRMPWR + PDISRSA_1YR + POWNRCA_PRI + PALITHERODRCA + PAACLAYRCA + PASILTRCA + DAPOPRCA2010 + PAOWNRCA_AGR</td>
<td>-1224.42</td>
<td>0.103398</td>
</tr>
</tbody>
</table>

The selected model contained 6 variables and the log transformed response variable. We used a log transformed response variable due to non-constant variance and to use the Gaussian error distribution given the variance is greater than the mean per the methods described in Crawley (2013). The log transformation was completed and the results scaled from 0 to 1 to be consistent with the scaling of the other variables. The variables included in the final model are detailed below in Table 2.

Table 2. Watershed characteristics included in the final model. All watershed characteristics were transformed to a scale from 0 to 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Scale of influence</th>
</tr>
</thead>
<tbody>
<tr>
<td>sum_1095_days</td>
<td>Sum of rainfall (in mm) over a 1095 day period prior to the sample date</td>
<td>Site</td>
</tr>
<tr>
<td>PALITHERODRCA</td>
<td>Percent of Accumulated Reach Contributing Area with erodible classified lithology</td>
<td>ARCA</td>
</tr>
<tr>
<td>PDISRSA_1YR</td>
<td>Percent land disturbed over a 1 year period prior to the sample date within the Reach Streamside Area</td>
<td>RSA</td>
</tr>
<tr>
<td>PASILTRCA</td>
<td>Percent of weighted average percent silt (0.002 to 0.05mm) between 6 – 36 inches within the Accumulated Reach Contributing Area</td>
<td>ARCA</td>
</tr>
<tr>
<td>DAPOPRCA2010</td>
<td>Estimated human population density (count per square km) in the Accumulated Reach Contributing Area in the year 2010</td>
<td>ARCA</td>
</tr>
<tr>
<td>POWNRCA_PRI</td>
<td>Percent of reach contributing area with private forest land</td>
<td>RCA</td>
</tr>
</tbody>
</table>

The formula with parameter coefficients for the selected model is:

\[ Y = 0.44297 \times \text{sum}_1095\_days + 0.31639 \times \text{PDISRSA}_1\_YR + 0.08801 \times \text{POWNRCA}_PRI + 0.05862 \times \text{POWNRCA}_PRI + 0.11911 \times \text{PASILTRCA} + 0.19232 \times \text{DAPOPRCA2010} + Z + \varepsilon \]

where \( Y \) = scaled log transformed FSS, the variable names are described in Table 2 and \( Z \) contains the spatially autocorrelated random variables and \( \varepsilon \) is the random effect. The spatially autocorrelated random variables, \( Z \), are defined by:

\[ Z = \sigma^2 z \times R \]

where \( \sigma^2 \) is the within site variability and \( R \) are the autocorrelation values from the exponential tail up, exponential tail down and exponential Euclidean autocorrelation functions.
5. Analysis and interpretation of results

To check the selected model for validity of assumptions and overall prediction accuracy we evaluated the significance of the model, examined the distribution of the residuals, performed a leave one out cross validation (LOOCV) procedure and compared the selected model to a non-spatial model.

To evaluate the significance of the model we used a likelihood ratio test to compare the selected best model to the null model. A null model fits a model without explanatory variables allowing us to test the effect of the explanatory variables. The $\chi^2$ test statistic was 164.4 with a $p$-value <0.001. These values allow us to reject the null hypothesis that there is no difference between the selected model and a null model.

To check the model assumption of normally distributed errors, we plotted the distribution of the residuals for the model. We observed an approximate normal distribution (Figure 10 top left). We also plotted a normal quantile-quantile plot of the standardized residuals against the theoretical quantiles assuming that the data were normally distributed. We observed a close fit supporting the assumption of normally distributed errors (Figure 10 bottom right). To check the assumption of constant variance in the model, we plotted the residuals and the standardized residuals against the fitted values. We did not observe a pattern which supports the assumption of constant variance (Figure 10 top right and bottom left).

![Figure 10](image)

**Figure 10.** Plots investigating the distribution of residuals from the final model. Top left is a histogram of the residuals. Top right are the raw residuals plotted against the predictions. Bottom left are the standardized residuals plotted against the predictions. Bottom right is the normal quantile-quantile plot of the standardized residuals against the theoretical quantiles if the data were to have a normal distribution.

For model validation, we used a leave one out cross validation (LOOCV) procedure. In this procedure, each response value is excluded one at a time and the final model predicts the removed value (Ver Hoef et al 2014). We plot the LOOCV predictions against the observed data as well as the standard error of the predictions (Figure 11). We observed that the predictions were approximately linear and that standard errors were characteristic of a normal distribution.
Figure 11. Leave one out cross validation (LOOCV) procedure predictions for log10 FSS and their standard errors plotted against observed log10 FSS. The predictions are distributed along a 1:1 line and the standard errors display a pattern indicative of normal errors.

Lastly, we checked to see if a non-spatial model would be as good at estimating the FSS as the model with spatial autocorrelation included. We fit a model without autocorrelation functions using the same variable set of the selected model. Compared to the selected model with autocorrelation included, the non-spatial model performed significantly worse based on AIC scores. The non-spatial model AIC was -1161.879 while the spatial model AIC was -1225.34.

The selected model contains a combination of natural (non-anthropogenic) and human (anthropogenic) factors. Interestingly, we arrived at nearly the same set of explanatory variables when comparing the selected model to the best ecological approach model. The only difference is the best ecological approach model did not have the variable for population density in the accumulated reach contributing area (DAPOPRCA_2010). This convergence of models supports the selection of the model from the statistical approach in addition to its improved model performance measures. The statistical properties approach model had the lowest AIC by 13 units when compared to the best ecological approach model. The statistical properties approach model also had the lowest RMSPE. The selected model will allow for the evaluation of the effects of different management actions on FSS while accounting for background settings.
References


