

**Modeling and Predicting Median Substrate Size
in Oregon and Washington Streams
Utilizing Geographic Information Systems Data**

Julia J. Smith
December 1, 2005

4025 E. Northern Lights Blvd.
Anchorage, AK 99508
julie.smith@acsalaska.net

Submitted to
Colorado State University
in Partial Fulfillment of the Requirements
for Plan B Master of Science Degree
from the CSU Statistics Department

Abstract

Median substrate size is a statistic that provides information about streambed material. It can be indicative of stream health as it is used to estimate bed load transport capacity, assess macroinvertebrate habitat, and assess the spawning habits for some fish species. In its Environmental and Monitoring and Assessment Program (EMAP), the U.S. Environmental Protection Agency collected data including a measure of median substrate size (LD_{50}) at 485 streams in Oregon and Washington between 1994 and 2004. Using Geographic Information System data compiled at Colorado State University and EMAP data, several models were created to predict LD_{50} using predictors that did not require on-site data collection. The goal was to create a model with a small subset of available variables so that LD_{50} could be predicted without on-site sampling. The categorical nature of LD_{50} and a large set of predictors made this task difficult to accomplish. There were three approaches used in this analysis: stepwise variable selection and multiple regression, classification and regression trees, and a hybrid of multiple regression and classification and regression trees. The hybrid models provided the best predictions at the cost of parsimony.

The Coast Range Ecoregion had a less skewed distribution of LD_{50} than the entire data set. We present a separate analysis for this ecoregion utilizing the same methods. The goal was to find a model that would allow prediction without on-site sampling that could be applied to regions with similar ecosystems. These models had better predictive-abilities than those for the entire data set, and the hybrid models provided the best of these predictions.

Table of Contents

I.	Introduction	1
II.	The Data Available for Analysis	4
	A. The Response Variable: Median Logarithm of the Geometric Mean of Substrate Classes	4
	B. The Predictors	7
	B.1 Land Cover Metrics	8
	B.2 Climatic Metrics	14
	B.3 Geological Metrics	16
	B.4 Geomorphic Metrics	17
	B.5 Ecoregions	19
III.	Methodology and Results Using Oregon and Washington Sites	20
	A. Multiple Regression and Stepwise Variable Selection Using Entire Dataset	20
	A.1 Stepwise Variables Selection from the Set of All Variables	21
	A.1.a Description of analysis	21
	A.1.b Results of Stepwise Variable Selection from the Set of All Variables	22
	A.2 Forward Stepwise Variable Selection within Tiers (Top 4-Tier)	24
	A.2.a Description of Analysis	24
	A.2.b Results for the Top 4-Tier Model	24
	A.3 All Forward Step Predictors for the Geomorphic Tier Included in the Top 4 Model	28
	A.3.a Description of Analysis	28
	A.3.b Results for the Geomorphic plus Top-3 Tier Model	28
	B. Classification and Regression Trees	31
	B.1 Description of Analysis	31
	B.2 Results	33

C.	CART and Top Predictor Hybrid Models	36
C.1	Results for CART Hybrid Top 4-Tier Model	36
C.2	Results for the CART Hybrid Geomorphic plus Top 3-Tier Model	41
D.	A Comparison of Models Using the Oregon and Washington Sites	45
IV.	Coast Range Data and Analysis	46
A.	Description	46
B.	Stepwise Variable Selection for the Coast Range Ecoregion	48
B.1	Top 4-Tier Model Results	48
B.2.	Geomorphic plus Top 3-Tier Model Results	52
C.	CART Using All Variables in the Coast Range Dataset	55
D.	Hybrid CART and Multiple Regression Models for the Coast Range	57
D.1	Top 4-Tier Model Hybrid Results	57
D.2	Geomorphic plus Top 3-Tier Hybrid Model Results	59
E.	Comparison of Models Using the Coast Range Ecoregion Data	63
V.	Conclusions and Future Work	64
VI.	Acknowledgements	66

I. Introduction

With continual development threatening natural resources, there is a need for better understanding of the status of our streams, rivers, and watersheds. Yet, the cost of monitoring watershed systems can be prohibitive. Technological advances have given rise to methods for assessing water quality without sending personnel to the sites of interest. There have been great advances in understanding the relationships between water flow, stream habitat, and stream geology (Allan, 1995). In particular, the size of a stream's substrate is shown to be indicative of its overall health. If it would be possible to estimate median substrate size of a stream (referred to as D_{50} in this paper) from information gathered via satellite, information about a stream's macroinvertebrate habitat, bed load transport, and general fish habitat could be predicted without incurring the costs of visiting the site.

Stream substrate size is one of the most important determinants of macroinvertebrate habitat (USEPA OWOW, 2002), and has both an indirect and direct effect on macroinvertebrate communities (Minshall, 1984). Substrate size affects the oxygen content of water and hence the type of invertebrates and fauna that are able to adapt to these oxygen levels (Eriksen, 1964). Also, substrate size speaks to the stability of the streambed and bed load transport, and invertebrates prefer a more stable environment (Jowett, 2003).

Bed load transport has been estimated using simple equations such as Duboy's tractive force equation and shear stress equations utilizing the specific weight of the water, the mean depth of the water, and the water surface slope (Julien, 1995). However, these predictions are accurate only under uniform stream conditions where the slope is moderate, the predominant substrate size is coarse, and the width to depth ratio is moderate. When the conditions become more variable, substrate size becomes an important predictor for bed load transport (Julien,

1995). There have been several models utilizing D_{50} to calculate initiation of motion and critical dimensionless shear stress (Julien, 1995).

As median substrate size is used to predict bed load transport and invertebrate habitat, it follows that substrate size could also be used to predict fish habitat. The size of substrate affects spawning habits and the invertebrate population affects feeding habits. Thus knowledge of substrate size can lead to knowledge of the fish population in streams. For example, median substrate is too small for Chinook salmon spawning if it is less than 7mm and too large if it is greater than 47 mm (Buffington, Montgomery, and Greenberg 2004).

Although it is an important indicator of stream health, previous attempts to measure median substrate size have involved visiting all sites of interest and sampling substrate in the stream. Two such site-based methods are the zigzag count and Wolman's pebble count. These methods provide sampling estimates and have some measure of uncertainty (Stamp, 2004). More importantly, the cost and time needed to get these estimates could be avoided if median substrate could be estimated without sampling substrate on site.

There have been attempts to predict median substrate size with Shield's critical dimensionless shear stress equation as given by

$$D_{50} = \frac{\tau}{(\rho_s - \rho)gt_{*c}} = \frac{\rho h S}{(\rho_s - \rho)t_{*c}} \quad (1)$$

where τ is the total bank-full shear stress, ρ_s is the density of sediment, ρ is fluid density, g is gravitational acceleration, h is bank-full depth, S is channel slope, and t_{*c} is the critical Shield's stress for movement of D_{50} . However, there are many influences on median substrate size including roughness elements and erosion that affect substrate size and flow. For example, it is

possible that flow resistance associated with wood debris decreases sediment transport (Buffington, Montgomery, and Greenberg 2004). Predicting median substrate from such equations has not been successful on a large scale and the lack of published works in this area confirms the difficulty of the task.

Recent attempts to predict median substrate have utilized airborne imagery of dry bed areas. These techniques essentially use the shadows cast by substrate and thus the depth of color in a photograph to predict size. There has been successful prediction in dry bed areas, but the models are not applicable to wet bed areas due to image distortion that water creates at different depths (Carbonneau, Lane, and Bergeron 2004).

Using Geographic Information Systems (GIS) data collected for 432 streams in Oregon and Washington, the objective of this project is to create a median substrate model that would allow prediction without on-site substrate sampling. The characteristics of the landscape and stream and the relationships between these characteristics and substrate will be utilized to make the predictions. We seek a model that will give accurate predictions at wet bed sites using a small subset of predictors. Many data can be collected, estimated, and analyzed using GIS technology, including metrics that are calculated by combining the relationship between land cover characteristics in a watershed and the effects of that type of land cover based on the distance of its occurrence from the stream outlet. These metrics integrate several aspects of the land cover characteristics and could improve previous attempts to predict median substrate size in wet bed locations. Prediction of median substrate utilizing this technology would provide important information about stream health. Because of the many influences on substrate size and the categorical nature of median substrate distribution, there are challenges to overcome to reach this objective.

II. The Data Available for Analysis

A. The Response Variable: Median Logarithm of the Geometric Mean of Substrate Classes

In its Environmental and Monitoring Assessment Program (EMAP), the U. S. Environmental Protection Agency (EPA) has been collecting data for thousands of watersheds. Between 1994 and 2004, data was collected on site at 485 sites in Oregon and Washington. However, the median logarithm of the geometric mean of substrate class (LD_{50}) was measured at only 432 sites in Washington and Oregon (Figure 1 and Figure 2), so 53 sites were eliminated from the scope of this study. Ninety-four of the remaining sites were visited multiple times; however, since time was not a variable of interest in this study, the response is the earliest LD_{50} measurement at any site.

To measure LD_{50} on site, a meter stick is placed in the water and the substrate at the base of the meter stick is visually estimated for its characteristic diameter (the middle value of its length, width and depth). The crew records a substrate size code and repeats this process at several predetermined places along the stream (USEPA, 1998). The EPA converts these codes to numeric values that are then classified according to substrate type (Table 1). The next calculated value for each sampled substrate is the logarithm (base 10) of the geometric mean of the substrate class boundaries, noting that the EPA created the upper bound for the bedrock class and the lower bound for fines only for the calculation of the geometric mean for those classes (Seeliger, 2005). It is possible for boulders to be larger than 8000 mm and for fine substrate to be smaller than 0.001 mm. LD_{50} , the response variable for this study, is the median of the logarithms of the geometric means at each site. LD_{50} is thus ordinal data, but shall be treated as continuous for the purposes of this study.

TABLE 1: LD_{50} values as classified by substrate type

Size (mm)	Class	Geometric mean	LD_{50} (mm) or Log_{10} of geom.mean
8000-4000	Bedrock	5656.85	3.7527
4000-250	Boulders	1000.00	3.0000
250-64	Cobbles	126.49	2.1020
64-16	Gravel (coarse)	32.00	1.5052
16-2	Gravel (fine)	5.66	0.7526
2-.06	Sand	0.35	-0.4604
.06-.001	Fines	0.00775	-2.1109

Source: USEPA, 1998; Seeliger, 2005

Some recorded LD_{50} values are different than those listed by substrate type, because the calculations for the median require finding an average of two classifications. As a result, there were twelve different LD_{50} values for the Oregon and Washington sites. The distribution of LD_{50} has a left skew (Figure 3). In fact, all sites with an LD_{50} measurement of -2.1109 and the one site with an LD_{50} measurement -1.28565 had values that were classified as outliers as they fall below the boundary defined as one and a half of the interquartile range lower than the third quartile. These lower values were difficult to predict accurately with multiple linear regression and the predictions for these sites were often too high.

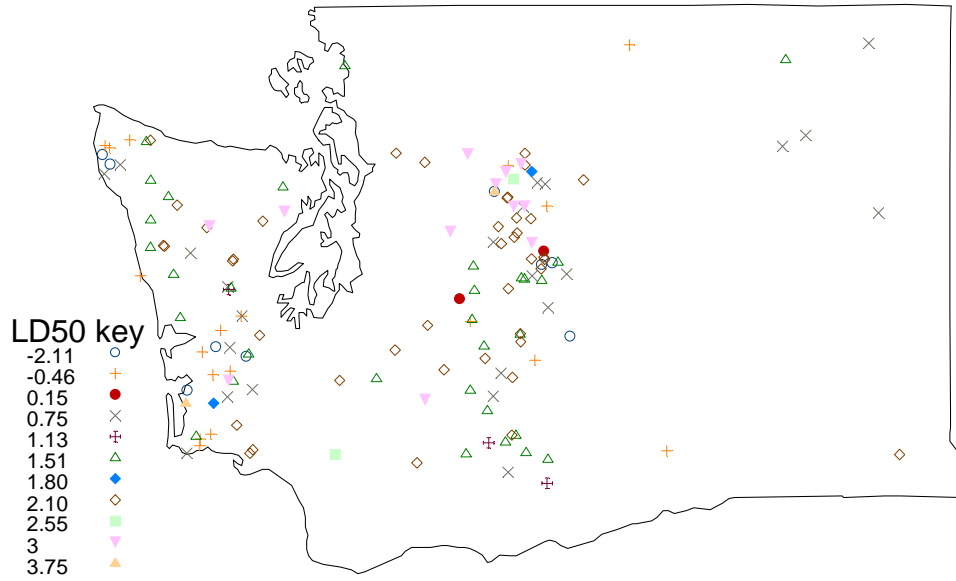


Figure 1: Washington EPA EMAP sites by LD_{50} measurement

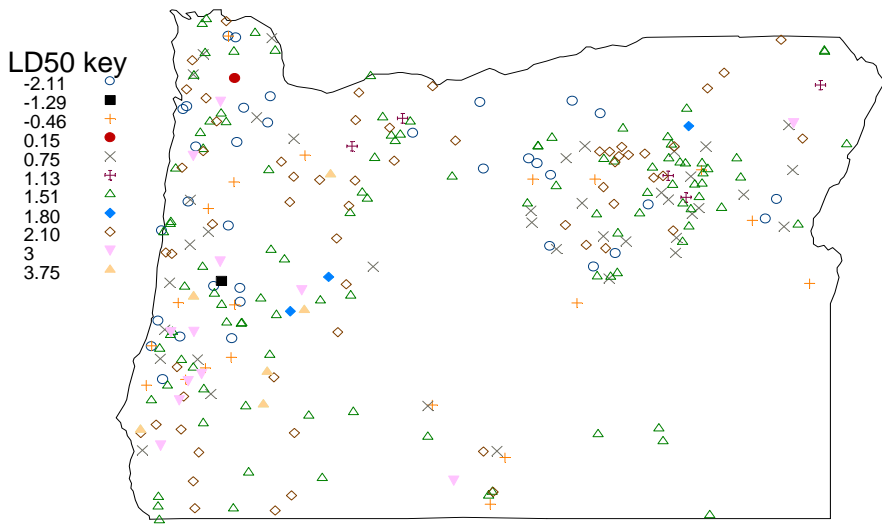


Figure 2: Oregon EPA EMAP sites by LD_{50} measurement

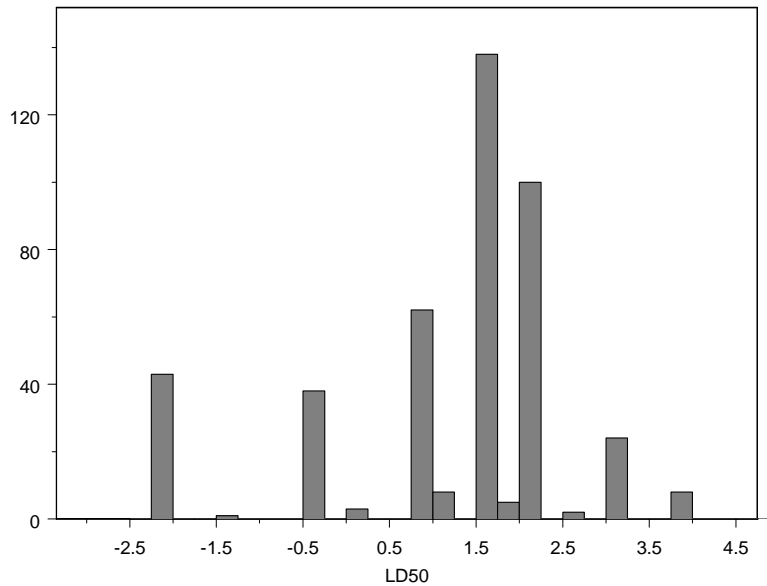


Figure 3: The Distribution of LD_{50} (mm) for 432 EPA EMAP sites

B. The Predictors

There are 1122 predictors available for prediction of the median of logarithm geometric mean of substrate class (LD_{50}). Of these, 412 predictors contained too many missing variables due to complications in the GIS calculations. The remaining 710 predictor variables can be classified into 4 categories: land cover, climatic, geological, and geomorphic (Sanborn and Bledsoe, in press; Cuhaciyar, in prep.) These four categories will be referred to as tiers for the remainder of this paper, indicating that there are four different types of variables. Ideally, we hope to describe how median substrate is affected by each of these types of variables. There were several predictors in the geomorphic tier that had some missing values, but these were not removed as the overarching controls of median substrate are flow energy for sediment transport, flow resistance, and the size and amount of sediment supplied to the outlet. These are

geomorphic influences on median substrate. A final variable, the ecoregion where the site was located, gave information about the ecosystem around the stream network.

B.1 Land Cover Metrics

Land cover metrics describe the land surrounding the channel at each site. Research indicates that the landscape of a watershed affects the median substrate size (Kauffman, Larsen, and Faustini 2004). For example, in the Oregon and Washington sites there is a positive correlation between the percentage of the landscape that is forest and the size of substrate (Figure 4, $r = 0.19$, $p\text{-value} = 3.516 \times 10^{-5}$). Some sites show a relationship that is counterintuitive to the general overall trend. For example, coarse gravel sites ($LD_{50} = 1.505$) were observed over the entire range in the percentage as forest. Such results can make it difficult to predict substrate size for these sites.

The connection between substrate size and landscape is not confined to natural resources. Substrate is easily altered by human activity as development tends to introduce fine substrate into the streambed (Kauffman, Larsen, and Faustini 2004). Thus, many of the predictors in the land cover tier are related to human activity such as the percentage of strip mines and gravel pits and the percentage of highly residential areas.

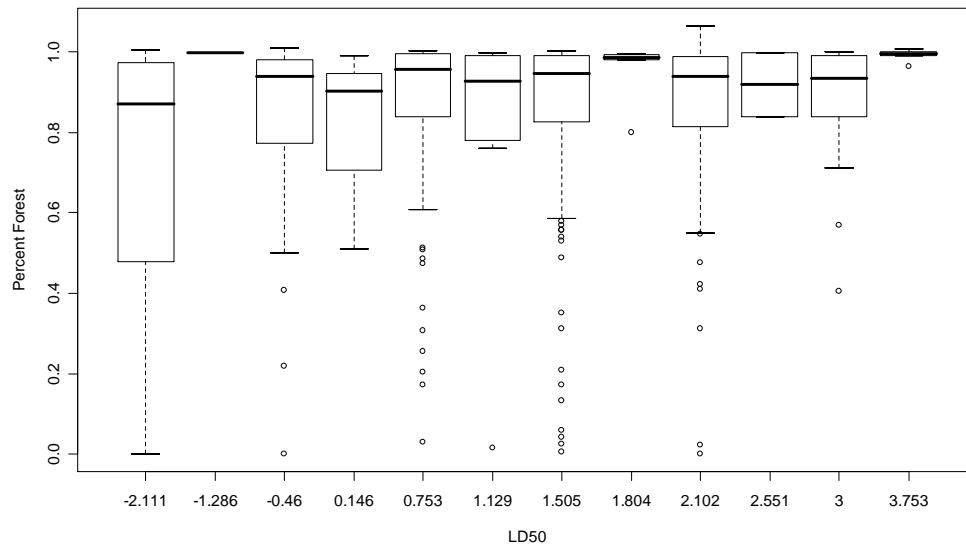


Figure 4: Percentage of the landscape that is forest versus LD_{50}

There are several basic land cover percentage metrics, including watershed percentages for the following; streams, surface water, developed, low-intensity residential, high-intensity residential, commercial, barren, forest, shrub-land, non-natural woody vegetation, grasslands, agricultural, and wetlands. These percentages are for the entire watershed above a sampling site and do not take into account whether the coverage is occurring adjacent to streams or a substantial distance away.

The second set of land cover metrics are distance-weighted land cover areas, expressed as percentages. These predictors are spatially explicit in that they take into account the hydrologic distance between where a particular land cover occurs in the watershed and the sampling location at the watershed outlet. As the hydrologic distance from the outlet increases, the value of a distance-weighted metric decreases. Thus, more weight is given to land covers that occur more often and are close to the outlet. It is possible that the relationship between median substrate

and land cover may depend on the hydrologic distance that the land cover occurs with respect to the stream, and these variables were calculated to investigate this relationship.

The distance-weighting is calculated using the formula

$$\text{Weighted Area}_j = \frac{A_j(e^{-\alpha \bar{d}_j})}{\sum_{i=1}^n A_i(e^{-\alpha \bar{d}_i})} \quad (2)$$

where j represents the land cover type of concern, A_j represents the total area for land cover type j in the watershed, α represents the coefficient of exponential decay, \bar{d}_j represents the average hydrologic distance from the outlet for land cover of type j , and n represents the total number of the land cover types. The distance \bar{d}_j is also referred to as the average flow length, indicating that it is the average of the distances that runoff travels from each parcel of land to the outlet with coverage type j . There are 4 exponential decay coefficients and 22 different land cover types. (Table 2 and 3). Each decay coefficient is used twice: once where the distance is further weighted by the topographic wetness index (Beven and Kirkby, 1979) before it is averaged and once where it is not. Each type of land cover was combined with every combination of the four decay coefficients and weighting or absence of weighting by the topographic wetness index for a total of 176 distance-weighted land coverage metrics.

The exponential decay coefficients determine the importance of hydrologic distance from the outlet that land cover occurs. As the exponential decay coefficient increases, less weight is placed on distance. For example, consider two sites that have an equal percentage of land in the watershed designated as evergreen forest. Further consider that in the first site the average hydrologic distance of evergreens from the outlet is further than that for the second site. If the exponential decay coefficient utilized for distance-weighting is 0.001, there will be a noticeable

difference between the distance-weighted metrics for the two sites. If the coefficient of decay is 0.00001, this difference will be lessened. The inclusion of different exponential decay coefficients allows for the examination of the part played by hydrologic distance in the relationship of LD_{50} and land cover.

The topographic wetness index is an indication of soil moisture content and indicates where surface flow versus subsurface flow may occur. It is calculated using drainage area and slope from digital elevation models. The hydrologic distance from the outlet, as used to calculate the average distance, \overline{d}_j , in the distance-weighting formula can be weighted itself using the topographic wetness index. This is important because the effects of land cover on substrate size may not only depend on the hydrologic distance from the outlet that the land coverage occurs, but also on the ability of run-off to carry fine sediments and debris caused by land coverage to the stream.

There are two sets of generalized distance-weighted metrics that combine similar land cover types. The first set is type A, where metrics are grouped by the tens place (Table 4). For example deciduous, evergreen, and mixed forests were combined into a general forest group. The second set, type B, is identical to type A except that the land cover type “orchards/vineyards/other” is combined with other agricultural cover types. Because of this combination, there is an additional metric created to describe the percentage that is combined agricultural (labeled as percent_8061). These metrics were created to investigate whether the effects of land cover depend not on specific types, but a more general type. For example, it may or may not be important what type of forest is in a watershed, but the fact that there is a forest of any type might be significant.

The third set of metrics in the land cover category include all of the previous land cover types, both with and without distance-weighting, calculated within a given distance, or buffer, around the stream network. Basically, these predictors ignore landscape coverage that occurs outside of the buffer distance. The three buffer distances used are 30 meters, 100 meters, and 300 meters. These predictors further address the issue of the effects of landscape coverage and the distance from the stream that the coverage occurs by utilizing only the area near the stream.

Table 2: Exponential Decay Coefficients

GIS Coefficient Code	Exponential Decay Coefficient (α)
1	0.00001
2	0.0001
3	0.0005
4	0.001
5	0.00001*
6	0.0001*
7	0.0005*
8	0.001*

*Not weighted by topographic wetness index

Table 3: Land cover types

GIS Land Cover Code (i)	Land Cover Type
11	Open water
12	Perennial ice/snow
21	low-intensity residential
22	high-intensity residential
23	commercial/industrial/transportation
31	bare rock/sand/clay
32	quarries/strip mines/gravel pits
33	transitional/changing
41	deciduous
42	evergreen
43	mixed
51	shrub-land
61	orchards/vineyards/other
71	grasslands/herbaceous
72	alpine/tundra
81	pasture/hay
82	row crops
83	small grains
84	fallow
85	urban/recreational grasses
91	woody wetlands
92	emergent herbaceous wetlands

Table 4: Generalized Land Cover Descriptions

GIS Code	Description
20	Developed (21, 22, 23)
30	Barren (31, 32, 33)
40	Forest (41, 42, 43)
70	Grasses (71, 72)
80	Agriculture (81, 82, 83, 84, 85)
90	Wetlands (91, 92)
8061	All Agriculture (61, 81, 82, 83, 84, 85)

B.2 Climatic Metrics

Because climate affects the flow of water and the flow regime, climatic variables have potential to provide information about the size of substrate (Kiffney, Bull, and Feller 2002). For example, there appears to be positive correlation between annual average precipitation and substrate size (Figure 5, $r = 0.199$, $p\text{-value} = 1.56 \times 10^{-6}$). Yet, the relationship between climatic variables and LD_{50} can be difficult to isolate. Often, the correlation is not strong or clear as can be seen in the relationship between LD_{50} and the annual average potential evapotranspiration, which is the amount of water that could evaporate from the soil or be used by plants (Figure 6, $r = -0.046$, $p\text{-value} = 0.342$).

The climatic metrics available from the EPA studies in Oregon and Washington include temperature, solar radiation, precipitation, and potential evapotranspiration. These are measured as monthly averages and annual averages. There are also ratios of precipitation including wettest to driest month, wettest 3 months to driest 3 months, and total annual snowfall to total annual precipitation. The average minimum temperature during November, December, January, February, and March is also included. These measures relate to the severity of winter, on average, for the sites. Finally there is a climatic metric called average aspect which is a measurement of direction of flow in degrees from true north.

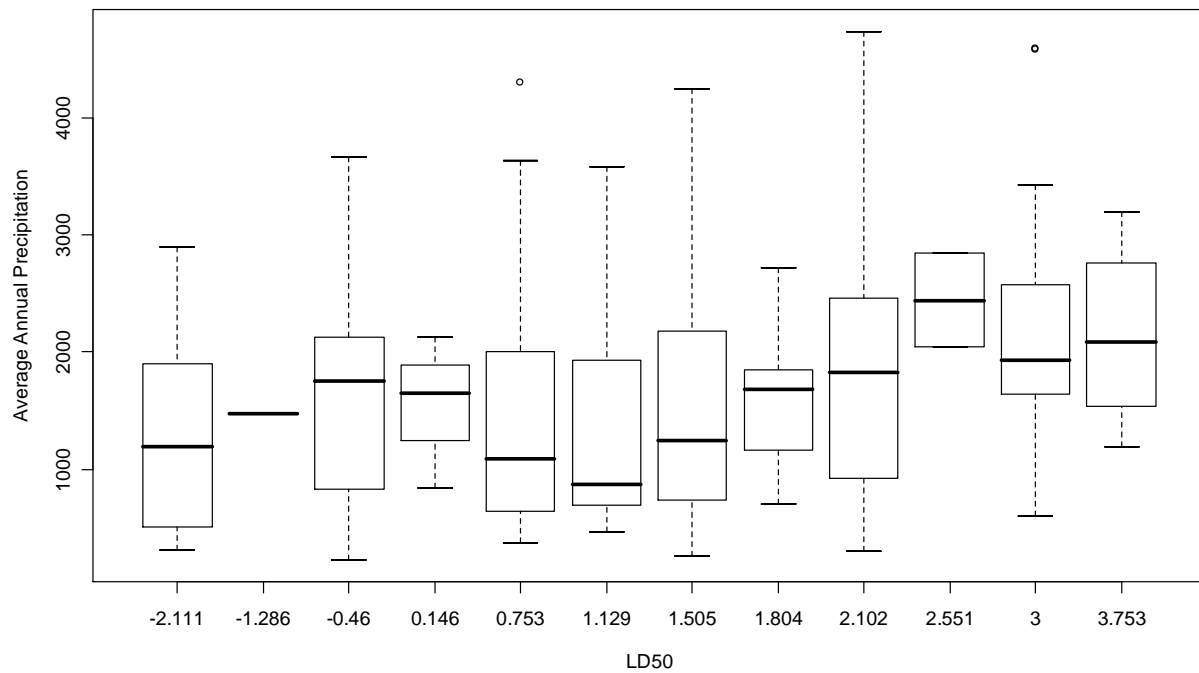


Figure 5: Average annual precipitation (mm) versus LD_{50}

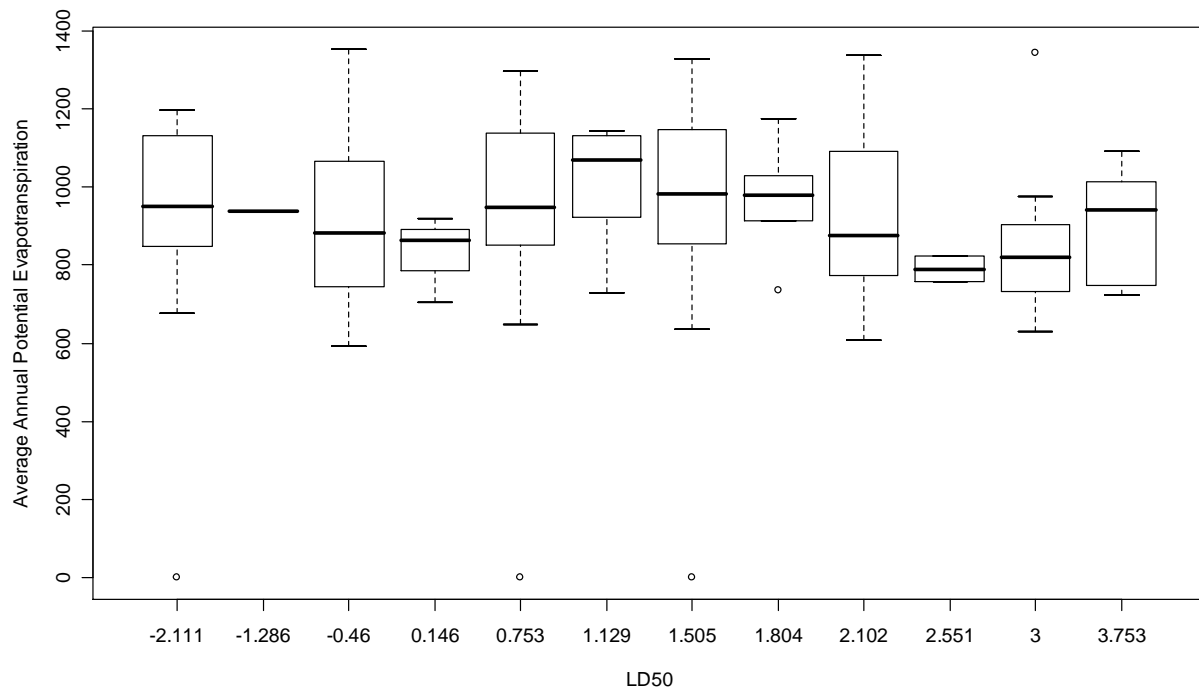


Figure 6: Average annual potential evapotranspiration (mm) versus LD_{50}

B.3 Geological Metrics

The geologic metrics include percentage of the watershed that is underlain by unconsolidated, sedimentary, volcanic, and crystalline geological types. The relationship between LD_{50} and watershed geology constitution show that a higher percentage of unconsolidated type indicates generally a smaller median substrate value (Figure 7, $r = -0.246$, $p\text{-value} = 1.18 \times 10^{-7}$). Additionally in the geology tier, there are percentages of the watershed as unconsolidated fine, unconsolidated coarse, sedimentary fine, sedimentary coarse, coarse sediment producing, extrusive volcanic, unconsolidated, fine-grain soft-sediment producing, fine-grain hard-sediment producing, coarse-grain sediment producing, quaternary landslide deposits, calcareous rocks, and coal.

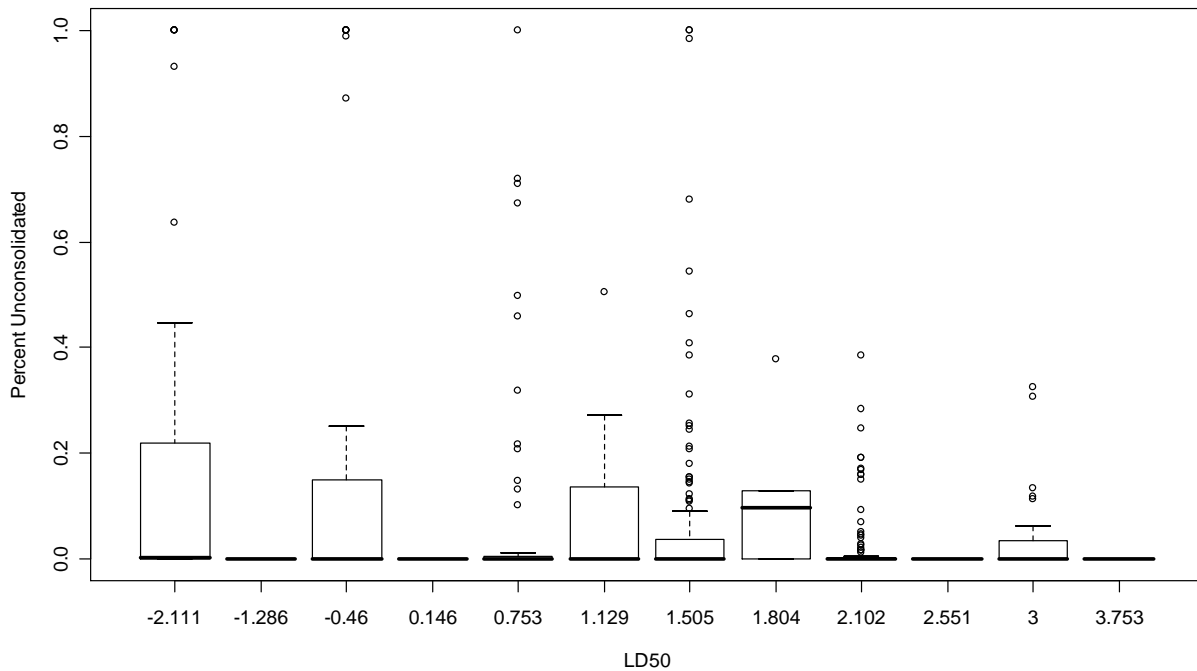


Figure 7: Percentage of watershed as unconsolidated geology type versus LD_{50} measurement

B.4 Geomorphic Metrics

Research indicates a strong relationship between median substrate size and the flow of streams as created by power and slope (Brummer and Montgomery, 2003). This relationship may apply to the Oregon and Washington sites. There is a positive association between LD_{50} and distance-weighted stream power, and there is a positive association between LD_{50} and slope (Figure 8, $r = 0.327$, $p\text{-value} = 2.63 \times 10^{-12}$; Figure 9, $r = 0.214$, $p\text{-value} = 3.78 \times 10^{-6}$). The 42 geomorphic predictors include measures of a stream's sediment transport capacity. The metrics related to the slope include valley entrenchment and its coefficient for each watershed site, the average measure of the width of the valley bottom versus the theoretical width for a sinuous river and its coefficient, the distance to a stream's first tributary, two measures of hillslope connectivity, outlet area, outlet slope, minimum and average elevation, average slope of the outlet, proportion of sites with slope less than 4 and 7 percent, the average slope of the watershed, drainage area, average weighted soil drainage class, average channel slope, average topographic wetness, ratio of the slope to the elongation of the watershed, watershed relief divided by its length, and mean slope within three buffer distances (40-meters, 100-meters, and 300-meters). Also in the geomorphic tier are measures of stream power that are both distance-weighted and not distance-weighted, measures of stream size and the intricacy of a stream's tributary system, percent pool-riffle, percent plane bed, percent step pool, percent cascade, drainage area, drainage density, area-weighted hydrologic group, average area-weighted minimum depth to bedrock, and percentage of watershed that is lakes or other water storage.

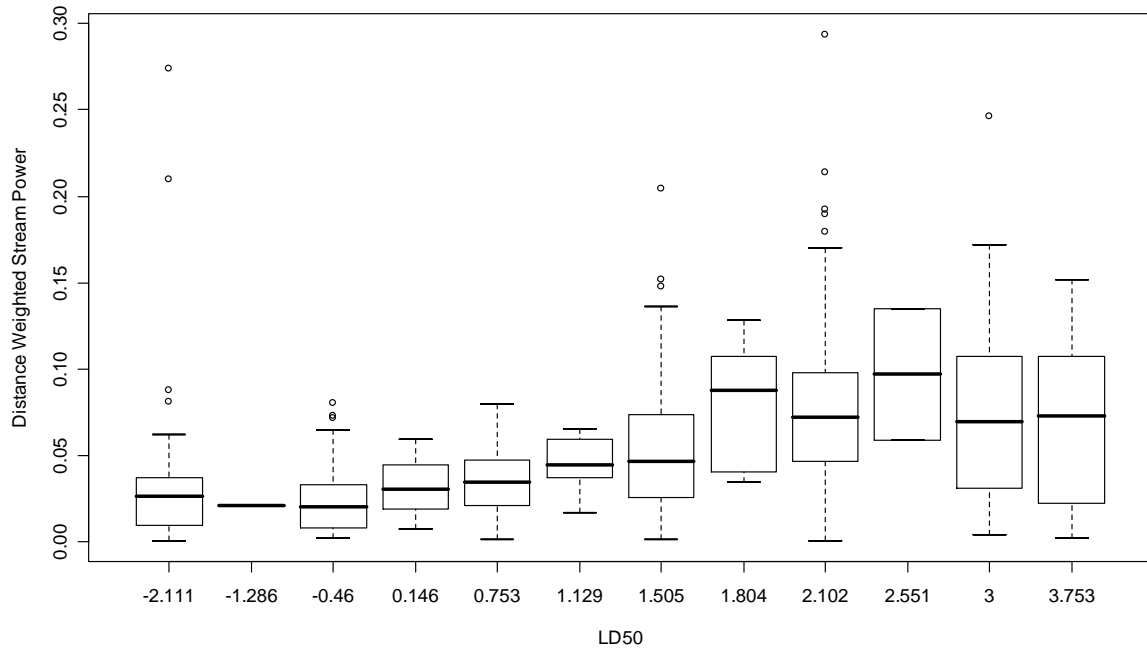


Figure 8: Distance-weighted stream power versus LD_{50} measurement

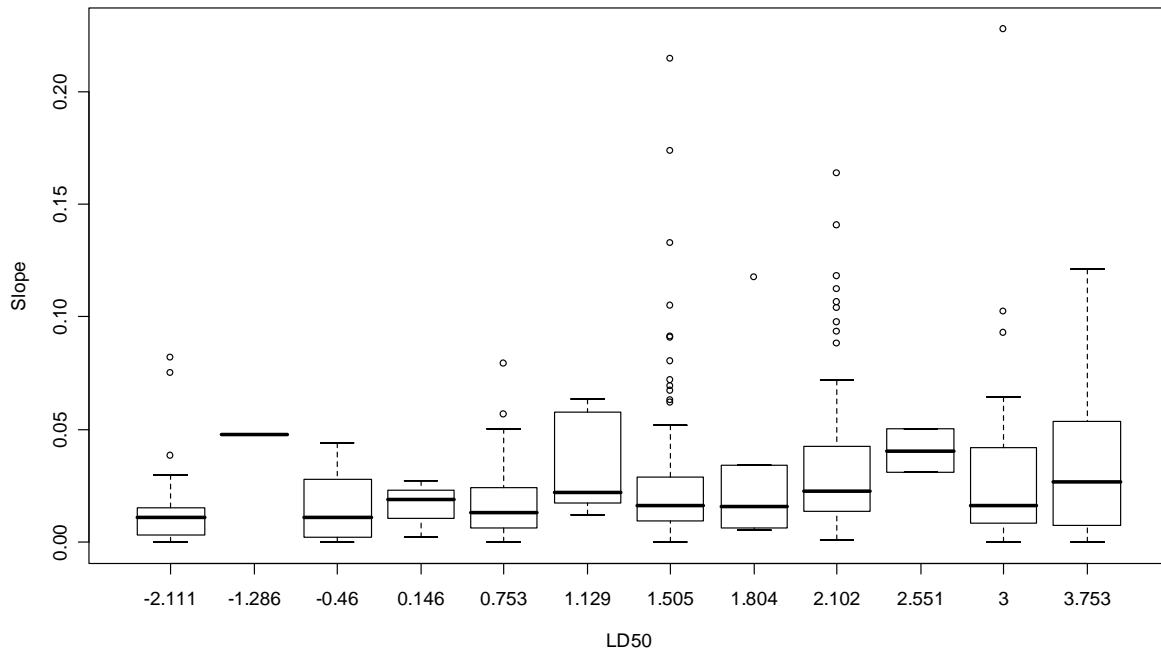


Figure 9: Outlet link mean slope versus LD_{50} measurement

B.5 Ecoregions

The EPA categorizes its EMAP sites into ecoregions, which are areas in which the general ecosystem is similar. There are currently four levels of ecoregions, with level I being the coarsest division and level IV being the most detailed. In this study, level III ecoregions were utilized. The third level ecoregions are based on an analysis of patterns in wildlife, geology, soil, climate, land cover and hydrology (USEPA WED, 2005). Of the 120 different level III ecoregions in the continental United States, 13 of the ecoregions were represented in this study. Because of the similarities in the ecosystems, there is a potential for median substrate size to have a relationship to the ecoregion in which a stream is found. Each ecoregion appears to have a unique distribution of LD_{50} (Figure 10) thus providing another possibility for prediction as will be shown in Section IV in an analysis of the Coast Range Ecoregion.

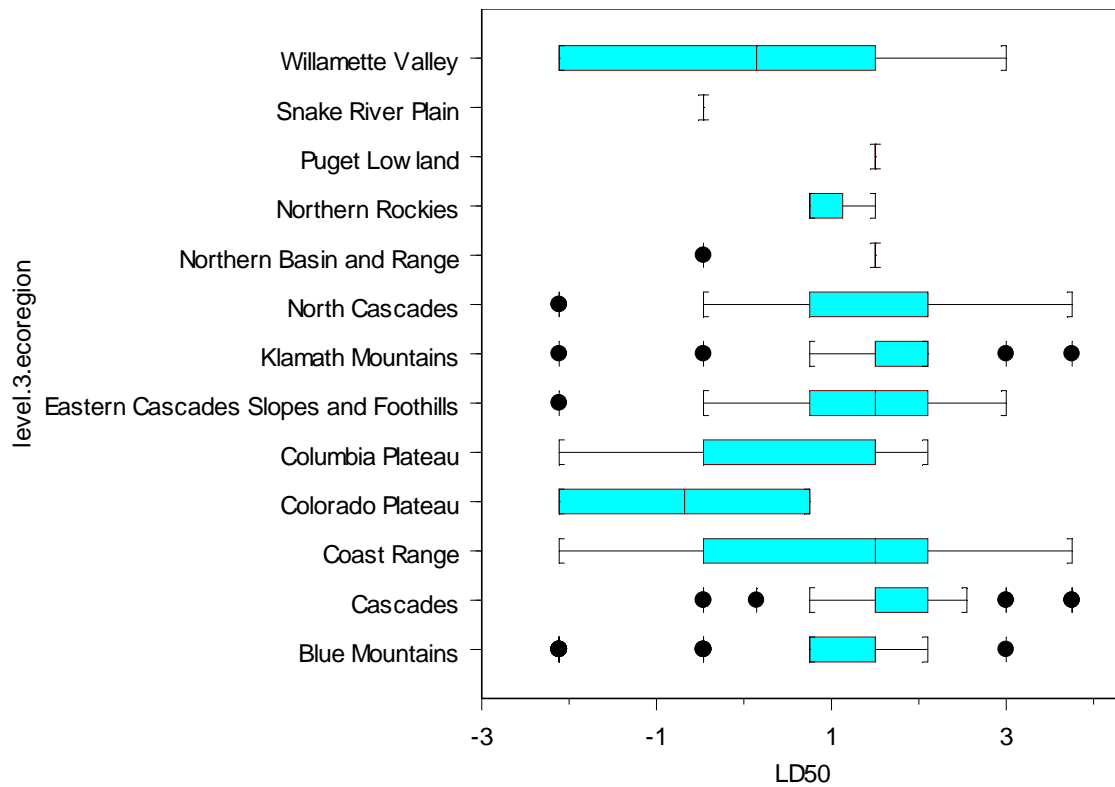


Figure 10: LD_{50} boxplots by Ecoregion

III. Methodology and Results Using Oregon and Washington Sites

We used the predictors described above to produce a model to predict median substrate size. We seek a model that will give accurate predictions of substrate size at unvisited sites. One goal of the analysis is to reduce the number of predictors included in the model to produce a scientifically sensible model.

We considered three potential methods for selecting the variables to predict median substrate: multiple linear regression with stepwise selection of predictors, tree-based modeling, and a hybrid of tree methods and multiple regression methods. All three of these methods were utilized initially on the 432 Oregon and Washington sites. In Section IV, we utilized these same techniques on the Coast Range Ecoregion sites.

A. Multiple Regression and Stepwise Variable Selection Using Entire Dataset

With hundreds of variables to choose from, the main goal of model prediction was variable selection. It is preferred that the model does not have too many variables, as models with many predictors are difficult to interpret and tell us little about the true nature of LD_{50} . In stepwise regression, we used Akaike's Information Criterion (AIC) to select the predictors in the model (Akaike, 1973). The AIC statistic includes a term to indicate how well the data fit the model as well as a penalty term for the number of variables. For a model with N observations and p predictors, the AIC statistic is given by

$$AIC = N \log \left(\frac{RSS}{N} \right) + 2(p+2) \quad (3)$$

where RSS is the sum of squared residuals from the model. In model selection, the model that minimizes AIC is preferred.

We utilized the Predicted Error Sum of Square ($PRESS_p$) criterion to assess the predictive-ability of each model. The $PRESS_p$ criterion for a data set with n observations and p predictors is given by

$$PRESS_p = \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2 \quad (4)$$

where Y_i represents observed response for the i th observation, and $\hat{Y}_{i(i)}$ represents the predicted response for the i th observation from a model fit with the i th site removed (Neter et. al., 1996). Essentially, each of the sites was removed one at a time, a model was fit using the remaining sites and the variables chosen from stepwise regression, and this model was used to predict the omitted site. A small $PRESS_p$ value indicates that the model predicts LD_{50} accurately. The Mean Square Predictive Error ($MSPR$) was calculated by taking the $PRESS_p$ value divided by the number of predictors in the validation set, which is one less than the number of sites available (Neter et. al, 1996). Finally, a statistic that is similar to the coefficient of determination, R^2 , and further indicates the predictive-ability of a model was calculated using the formula given by

$$R^2_{prediction} = 1 - \frac{PRESS_p}{SSTO} \quad (5)$$

where $SSTO$ is the total sum of squared error for the response (Montgomery, Peck, and Vining 2001).

A.1 Stepwise Variables Selection from the Set of All Variables

A.1.a Description of analysis

To reduce the number of variables in the model, we began by performing stepwise variable selection on the set of all variables. As the number of available predictors far surpassed the number of available sites, we performed stepwise predictor selection using subsets of the

variables. There were 571 land cover predictors, far more than in any of the remaining three tiers. As a question of interest in this study was whether buffered or non-buffered land cover metrics would be more effective, the subsets were created separating these two sets of land cover variables. The three subsets are described below.

- Subset 1: Climatic variables (monthly averages over annual averages where possible), non-buffered land cover variables, geology variables and geomorphic variables.
- Subset 2: Climatic variables (annual averages over monthly averages where possible), non-buffered land cover variables, geology variables and geomorphic variables.
- Subset 3: Climatic variables (monthly averages over annual averages where possible), buffered land cover variables, geology variables, and geomorphic variables.

Stepwise regression was performed on each subset using *R* statistical software, specifically the function “stepAIC” from the MASS library (*R Version 2.1.1*, 2005). Starting with the linear model with all predictors, variables were taken out and/or added to find the model with the minimum AIC value.

A.1.b Results of Stepwise Variable Selection from the Set of All Variables

Unfortunately, the models fit using this method had many problems. A large number of variables chosen for the final model made it difficult to interpret the relationships with LD_{50} . The residuals were clearly not normal and it appeared that a transformation of the response would improve the model. The square root of LD_{50} (with a constant added to make all values positive) was the best transformation available. Despite the transformation, there was still a violation of the normality assumptions in each model (Figure 11). The $PRESS_p$ values for each of the three models was more than a trillion, indicating that the model was effective at predicting

LD_{50} for only some sites and grossly overestimated or underestimated other sites. Because of the large $PRESS_p$ values, the predictive R^2 values were negative, indicating that the regression model could explain none of the variability in the predictions (Table 5).

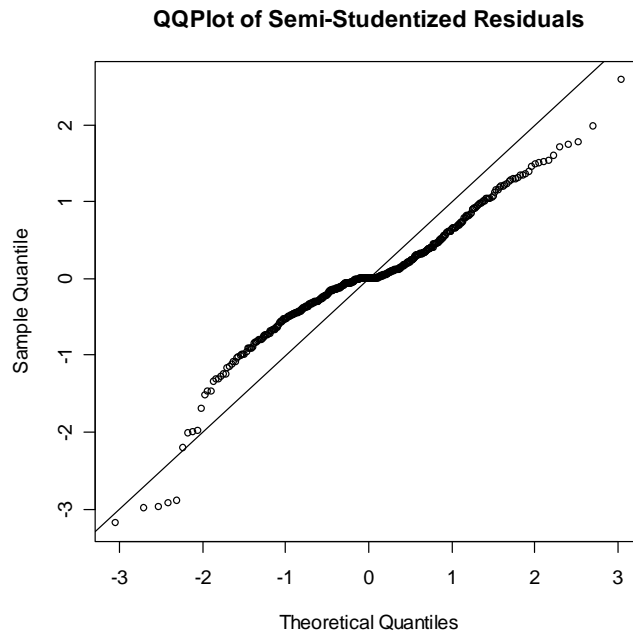


Figure 11: An example of the violation of normality of residuals for the stepwise models (subset 2)

Table 5: Summary statistics of three subset stepwise regression models

Model Subset	Number of Variables in Model	Adjusted R^2	$PRESS_p$	Mean Square Prediction Error	$R^2_{prediction}$
Subset 1	275	0.7573	4.43×10^{13}	1.08×10^{11}	-1.19×10^{12}
Subset 2	203	0.6581	3.03×10^{12}	7.37×10^9	-7.62×10^{10}
Subset 3	346	0.8158	2.50×10^{19}	6.07×10^{16}	-6.67×10^{17}

A.2. Forward Stepwise Variable Selection within Tiers (Top 4-Tier Model)

A.2.a Description of Analysis

In an attempt to create a simple model with fewer predictors, one predictor was to be chosen from each of the tiers including the climatic metrics, land cover metrics, geologic metrics and geomorphic metrics. Forward stepwise selection was performed on each of the four sets of variables utilizing Akaike's Information Criterion and the square root of shifted LD_{50} as the response. The forward direction started with an intercept model and added predictors one at a time with the first variable chosen being that which maximized the reduction of AIC for the intercept model. The next variable chosen maximized the reduction of AIC for the model with both an intercept and the first chosen variable, and this process continued until the AIC value could no longer be reduced. The final model uses the first predictor chosen using forward step regression within each tier, thus reducing the number of variables to four and providing a model that might be scientifically sensible. This model will be referred to as the top 4-tier model.

A.2.b Results for the Top 4-Tier Model

There were some problems and improvements using this tier method. The most obvious problem was a severe drop in adjusted R^2 values with the regression relationship explaining much less of the variation of the transformed LD_{50} . Additionally, there was still a problem with normality of the residuals (Figure 12). Repeating variable selection and regression for the untransformed LD_{50} slightly improved the results, but did not resolve the problem (Figure 13).

In forward stepwise selection processes on both the untransformed and transformed LD_{50} , the same four predictors were selected. The climatic predictor was the average temperature in June (labeled as `avgt_jun`). The land cover predictor was the percent of evergreen forest

distance-weighted with coefficient of exponential decay 0.00001 and not weighted by the topographic wetness index (labeled as b30_r5_142) . The geologic predictor was the percent of the watershed that is unconsolidated geologic type (labeled as uncons). The geomorphic predictor was the mean slope within a 300-meter buffer of the stream (labeled as sbuf_300). Each predictor was significant in both regressions (Table 6 and Table 7).

The $PRESS_p$ criterion was no longer in the trillions, as the prediction errors during cross-validation were more reasonable than those for the all variable step-regression models. While the adjusted R^2 was much lower, this model predicts missing observations nearly as accurately as the regression line fits observed values as indicated by the improved $R^2_{prediction}$ values (Table 8). Although lessened, over-prediction of small LD_{50} values and under-prediction of large LD_{50} was still prevalent (Figure 14). This model indicated that fewer variables improved prediction, but that with the complex nature of LD_{50} , more variables might be necessary for a better fit.

Table 6: Coefficients and statistic for the top 4-tier sqrt($LD_{50} + 5$) model

Coefficient	Coefficient Estimate	Standard Error	t-value	p-value
Intercept	2.397	0.0744	32.214	$< 2 \times 10^{-16}$
Dist-wt % evergreen forest (30m buff)	0.197	0.0577	3.412	7.06×10^{-4}
Mean slope (300m buff)	0.010	0.0017	5.794	1.34×10^{-8}
Percent of watershed as unconsolidated	-0.159	0.0662	-2.400	1.68×10^{-2}
Average temperature in June (°C)	-0.021	0.0042	-4.858	1.67×10^{-6}

Table 7: Coefficients and statistic for the top 4-tier untransformed LD_{50} model

Coefficient	Coefficient Estimate	Standard Error	t-value	p-value
Intercept	0.858	0.3350	2.560	1.08×10^{-2}
Dist-wt % evergreen forest (30m buff)	0.866	0.2598	3.332	9.37×10^{-4}
Mean slope (300m buff)	0.045	0.0075	6.008	4.03×10^{-9}
Percent of watershed as unconsolidated	-0.717	0.2982	-2.404	1.66×10^{-2}
Average temperature in June (°C)	-0.094	0.0191	-4.901	1.36×10^{-6}

Table 8: Top 4-tier model summary statistics

Response	Adjusted R^2	$PRESS_p$ for LD_{50}	Mean Square Prediction Error	$R^2_{prediction}$
square root ($LD_{50} + 5$)	0.2313	647.39	1.502	0.217
LD_{50}	0.2360	641.40	1.488	0.222

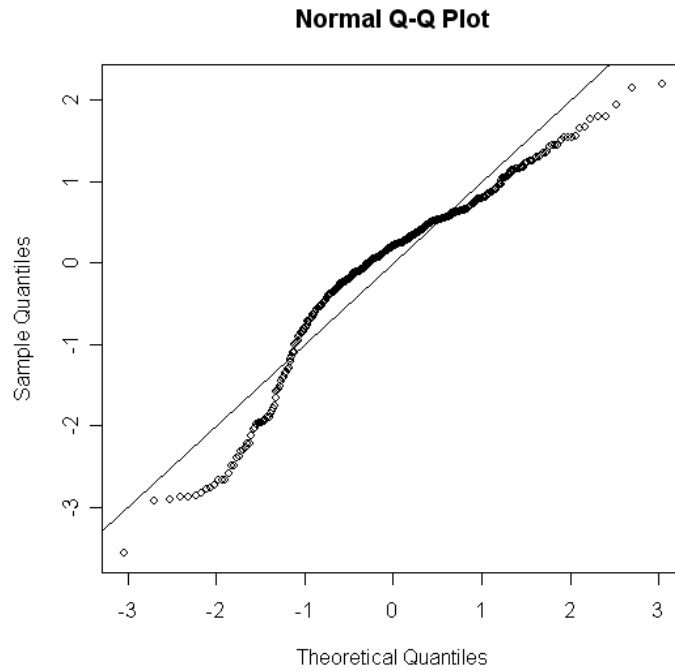


Figure 12: QQ-Plot of semi-studentized residuals for top 4-tier $\sqrt{LD_{50}}$ model

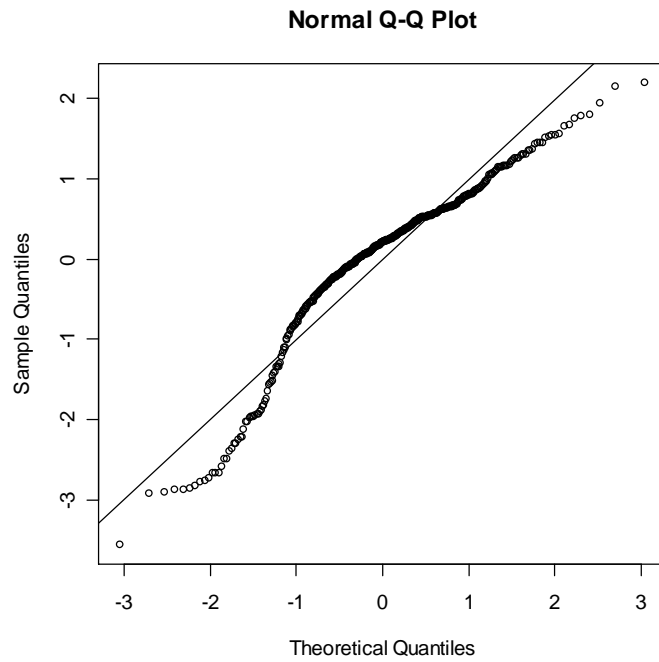


Figure 13: QQ-Plot of semi-studentized residuals for top 4-tier untransformed LD_{50} model

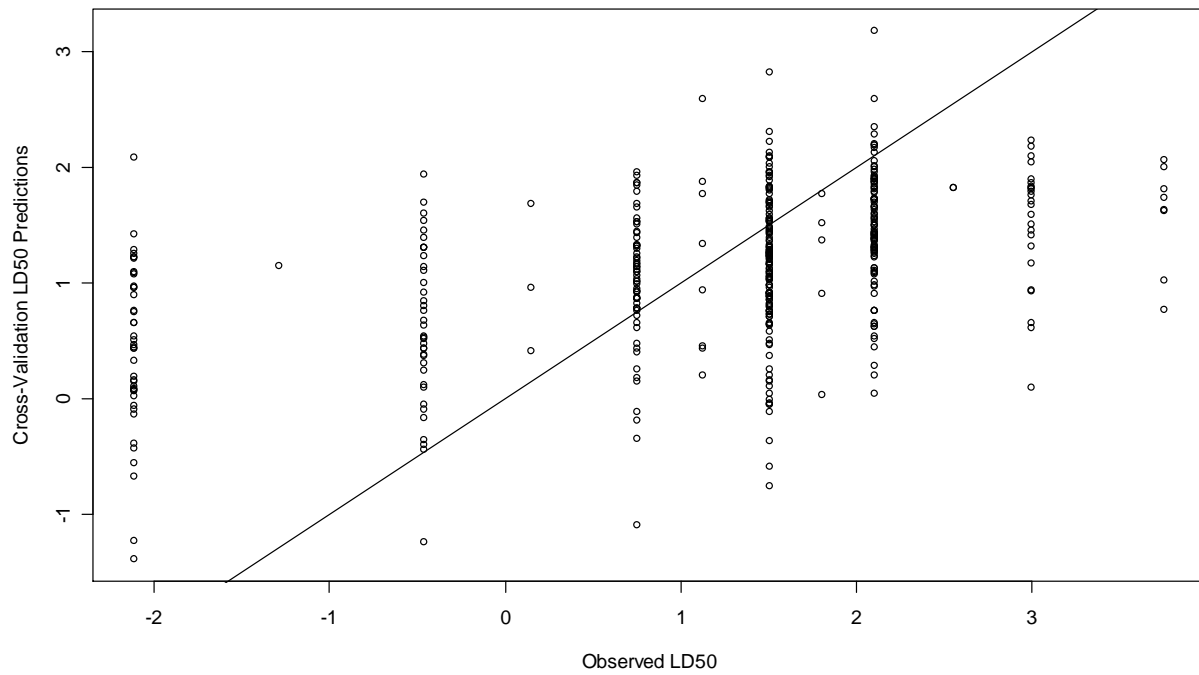


Figure 14: Cross-Validated LD_{50} Predictions versus Observed LD_{50} for top 4-tier untransformed model

A.3 All Forward Step Predictors for the Geomorphic Tier Included in the Top 4 Model

A.3.a Description of Analysis

After examining the forward stepwise variable selection for the geomorphic tier, we decided to include more geomorphic predictors in the top 4-tier model in an attempt to improve the fit and predictive-ability of the model. The reason for utilizing the geomorphic tier was two-fold: the model chosen in forward step variable selection on the geomorphic tier appeared to have greater potential alone than the model with the top predictor from each tier, and past research indicates that geomorphic variables are valuable predictors of median substrate size.

We performed stepwise variable selection again, this time both adding and subtracting variables from the model. It was not our goal to find the top predictor in this tier, but instead to find the best subset of variables, so it was necessary that the direction of stepwise selection was both forward and backward. Because there were several sites with missing geomorphic values, and we felt that having as many geomorphic variables included was important, those sites with missing values were eliminated and the sample size was reduced to 397 sites. The final model was a combination of the subset of geomorphic predictors selected in stepwise variable selection and the top predictor from each of the geologic, climatic, and land cover tiers. The model will be referred to as the geomorphic plus top-3 tier model.

A.3.b Results for the Geomorphic plus Top-3 Tier Model

There were 17 predictors chosen in stepwise variable selection in the geomorphic tier: an average measure of the width of the valley bottom versus the theoretical width for a sinuous stream (labeled as MENTB), a measure of stream power that is the average product of the stream

channel slope near the outlet times the watershed area above the outlet (labeled as link_sa), the mean of the product of the channel slope near the outlet times the watershed area above the outlet raised to the fourth power which is a measure of specific outlet stream power (labeled as link_sa^0.4), a measure of hill connectivity (labeled as CVCON), the ratio of the width of the stream to the width of the floodplain which is a measure of valley entrenchment (labeled as MENTR), the coefficient of the previous metric (labeled as CVENTR), drainage density (labeled as drainden), a measure of the size and complexity of a stream (labeled as shreve), average stream slope (labeled as link_slope), mean slope within a 300-meter buffer of the stream (labeled as sbuff_300), the product of the stream channel slope near outlet and the watershed area above the outlet which is a measure of total stream power (labeled as out_sa), ratio of the slope to the elongation of the watershed (labeled as slp_elon), outlet area (labeled as out_area), average channel slope (labeled as chan_slp), average topographic wetness (labeled as topo_wet), average slope of the watershed (labeled as shed_slp), and the percentage of the stream that is step pool (labeled as pct_SP). We combined these geomorphic predictors with distance-weighted evergreen forest percentage within a 30-meter buffer, percentage of watershed as unconsolidated geologic type, and average temperature in June from the top 4-tier model and performed multiple linear regression (Table 9).

The model had some advantages over the previous models. Adjusted R^2 , $PRESS_p$, $MSPR$, and $R^2_{prediction}$ indicated the fit of the model was an improvement and the predictive-ability of the model was better than all previous models (Table 10). Though improved, there was still a lack of normality of the residuals (Figure 15). Additionally, the problem of under-prediction and over-prediction of large and small LD_{50} measurements was still prevalent (Figure 16). The addition of all top geomorphic variables did create some collinearity problems. When comparing

the top 4-tier model and the geomorphic plus top 3-tier model, the increased number of variables, hence the loss of parsimony, in the latter model is exchanged for higher predictive-ability.

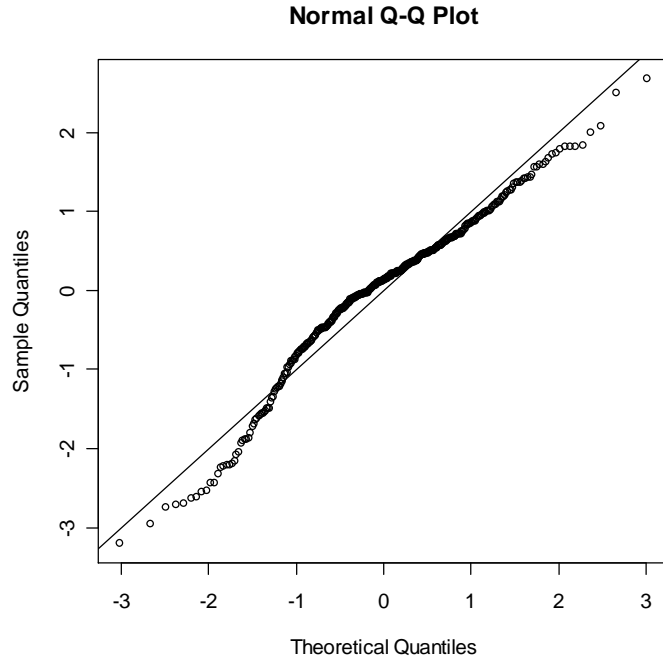


Figure 15: QQ-Plot of semi-studentized residuals for geomorphic plus top 3-tier model

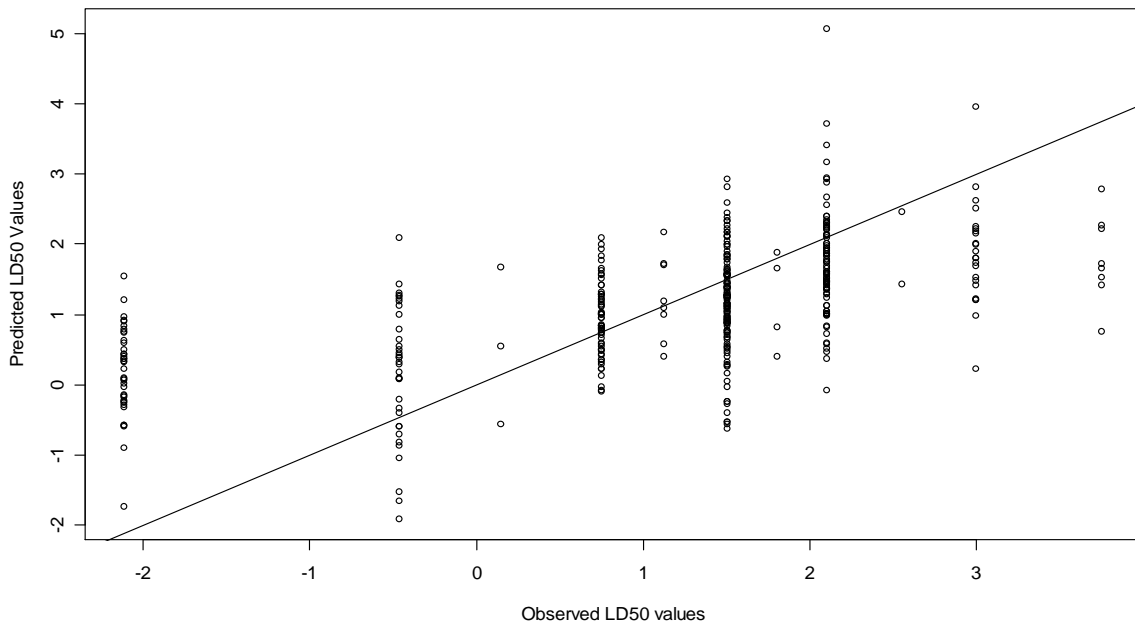


Figure 16: Cross-Validated LD_{50} Predictions versus observed LD_{50} values for geomorphic plus top 3-tier model

Table 9: Coefficients and statistics for the geomorphic plus top 3-tier model

Coefficient	Coefficient Estimate	Standard Error	t-value	p-value
Intercept	-3.814	1.6320	-2.336	2.00×10^{-2}
Valley bottom versus width dvlp river	0.015	0.0074	2.040	4.20×10^{-2}
Stream power (slope x area) ^{0.4}	8.362	2.6460	3.160	1.71×10^{-3}
Stream power (slope x area)	-0.228	0.0988	-2.312	2.13×10^{-2}
Hill connectivity	0.734	0.2387	3.073	2.28×10^{-3}
Valley entrenchment	-0.006	0.0015	-3.853	1.37×10^{-4}
Coefficient of valley	0.443	0.1978	2.241	2.56×10^{-2}
Drainage density (m ⁻¹)	-261.600	75.8100	-3.451	6.23×10^{-4}
Size/complexity of stream	0.003	0.0010	2.811	5.19×10^{-3}
Average stream slope	-6.336	4.5490	-1.393	1.65×10^{-1}
Mean slope (300m buff)	0.019	0.0102	1.881	6.07×10^{-2}
Channel slope x area above the outlet	-0.078	0.0342	-2.268	2.39×10^{-2}
Ratio of slope to elongation	-0.515	0.2112	-2.439	1.52×10^{-2}
Outlet area (km ²)	-0.011	0.0044	-2.465	1.42×10^{-2}
Average channel slope	1.759	1.8140	0.970	3.33×10^{-1}
Average topographic wetness	0.536	0.1770	3.030	2.61×10^{-3}
Average slope in the watershed	0.024	0.0077	3.058	2.39×10^{-3}
% step pool	0.292	0.4895	0.596	5.51×10^{-1}
Average temperature in June (°C)	-0.038	0.0213	-1.788	7.45×10^{-2}
% watershed as unconsolidated	-0.622	0.2972	-2.093	3.70×10^{-2}
Dist-wt % evergreen forest (30m buff)	0.826	0.2639	3.131	1.88×10^{-3}

Table 10: Summary statistics for geomorphic plus top 3-tier model

Response	Adjusted R^2	$PRESS_p$ for LD_{50}	Mean Square Prediction Error	$R^2_{prediction}$
LD_{50}	0.362	504.802	1.274	0.319

B. Classification and Regression Trees

B.1 Description of Analysis

In tree based methods, a binary “yes-no” question is asked about a single predictor variable that allows the response variable to be partitioned into two homogenous subsets. Using recursive partitioning, this process is repeated on each subset of response variables until there are

several homogenous subsets. Each of these subsets is called a terminal node. The prediction of the response is then found by locating the subset that an observation falls into based on the binary questions until a predicted response, the mean of the response in that subset, is given. Using indicator variables, the response of a regression tree with q terminal nodes is given by

$$\hat{y}(x_i) = \sum_{j=1}^q \hat{a}_j 1_{\{x_i \in N_j\}} \quad (6)$$

where $\hat{y}_i(x_i)$ is the predicted response for the i th observation, \hat{a}_j is the mean of the observed response for the observations falling in the j th terminal node, and N_j is the set of observations falling in the j th terminal node (Givens and Hoeting, 2005).

Consider that each variable has at most n possible values, less if there are tied predictor values, and there are p predictors available in the data set. There are at most $p(n - 1)$ possible subsets of observations to determine the split. At each root node of the tree, the next split is chosen by minimizing the residual sum of squared error over all possible $p(n - 1)$ subsets of observations. Therefore, it is possible that a tree with q terminal nodes does not necessarily minimize the residual sum of squared error for all possible trees with q terminal nodes (Givens and Hoeting, 2005). It is also possible for a variable to be chosen as the subject of a binary question more than once within a tree.

To implement Classification and Regression Trees (CART) in R statistical software, the command `rpart` was used (*R Version 2.1.1*, 2005). This command follows the details outlined above, and additionally does some pruning that prevents the final tree from over-fitting the observations. Pruning the terminal nodes of a tree is decided by minimizing the cost-complexity equation

$$R_\alpha = R + \alpha size \quad (7)$$

where R is the residual sum of squared error, α is a value that minimizes error and is computed in cross-validation of the dataset, and $size$ is the number of terminal nodes in the tree. To compute α from a subset of values, R statistical software automatically does cross-validation on pruned trees using 10 subsets of the observations, fitting every pruned tree with each of the nine subsets and predicting the values for the subset that is left out, and then calculating an average residual sum of squared error for the 10 cross-validated models. The α that minimizes the average cross-validation sum of squares over all pruned trees becomes the estimated α , and the smallest tree that comes close to minimizing the cost-complexity equation using this α becomes the final model (Venables and Ripley 1999; Givens and Hoeting 2005).

B.2 Results

CART was performed on the Oregon and Washington sites using all 432 sites and the predictors that were utilized for the top in tier models. There were 15 terminal nodes in the model and 14 variables were utilized in the splits (Table 11). The LD_{50} predictions ranged from -0.1659 to 2.015. The criterion for the splits indicates the importance of the distance-weighted stream power predictor, $DWSP2$, as it was the first variable chosen to create subsets (Figure 17).

There are some values that are severely over-predicted and under-predicted using this regression tree (Figure 18). The range of prediction values does not cover the range of possible LD_{50} values. It was also possible for one of the twelve observed LD_{50} values to be predicted by one of several terminal nodes. For example, sand substrate ($LD_{50} = -0.4604$) appeared in every terminal node. Comparing $PRESS_p$, $MSPR$, and $R^2_{prediction}$ of other models, this model did not have strong predictive-abilities and was not pursued further (Table 12).

Table 11: Predictors utilized in the splits for the CART model

Predictor	Predictor Tier	Description
avgt_jun	Climatic	Average temperature in June (°C)
mint_apr	Climatic	Average minimum monthly temperature in April (°C)
prcp_jan	Climatic	Average precipitation in January (mm)
prcp_may	Climatic	Average precipitation in May (mm)
prcp_sep	Climatic	Average precipitation in September (mm)
snow_jan	Climatic	Average snowfall in January (mm)
min_elev	Geomorphic	Minimum watershed elevation (m)
DWSP2	Geomorphic	Distance-weighted Stream Power
link_sa4	Geomorphic	(Mean of channel slope x watershed area above the outlet) ^{0.4}
MENTR	Geomorphic	Estimated ratio width of stream to width of floodplain (valley entrenchment)
b30_111	Land cover	Percentage of open water within a 30-meter buffer
b100_151	Land cover	Percentage of shrubland within a 100-meter buffer
r8_180_A	Land cover	Percent agricultural (not including orchards/vineyard) distance-weighted (coefficient of decay = 0.001) and not weighted by the topographic wetness grid
b30_r7_130	Land cover	Percentage of barren landscape within a 30-meter buffer, distance-weighted (coefficient of decay = 0.0005) and not weighted by the topographic wetness grid

Table 12: Summary statistics for the CART model

Response	$PRESS_p$ for LD_{50}	Mean Square Prediction Error	$R^2_{prediction}$
LD_{50}	863.921	2.019	-0.0648

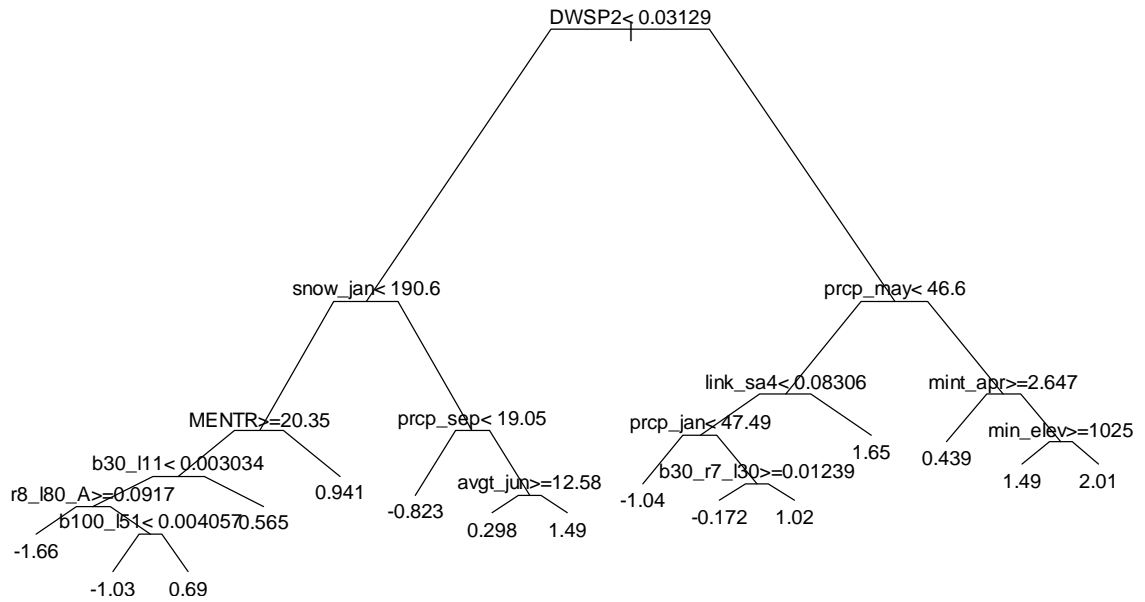


Figure 17: Classification and Regression Tree for LD_{50}

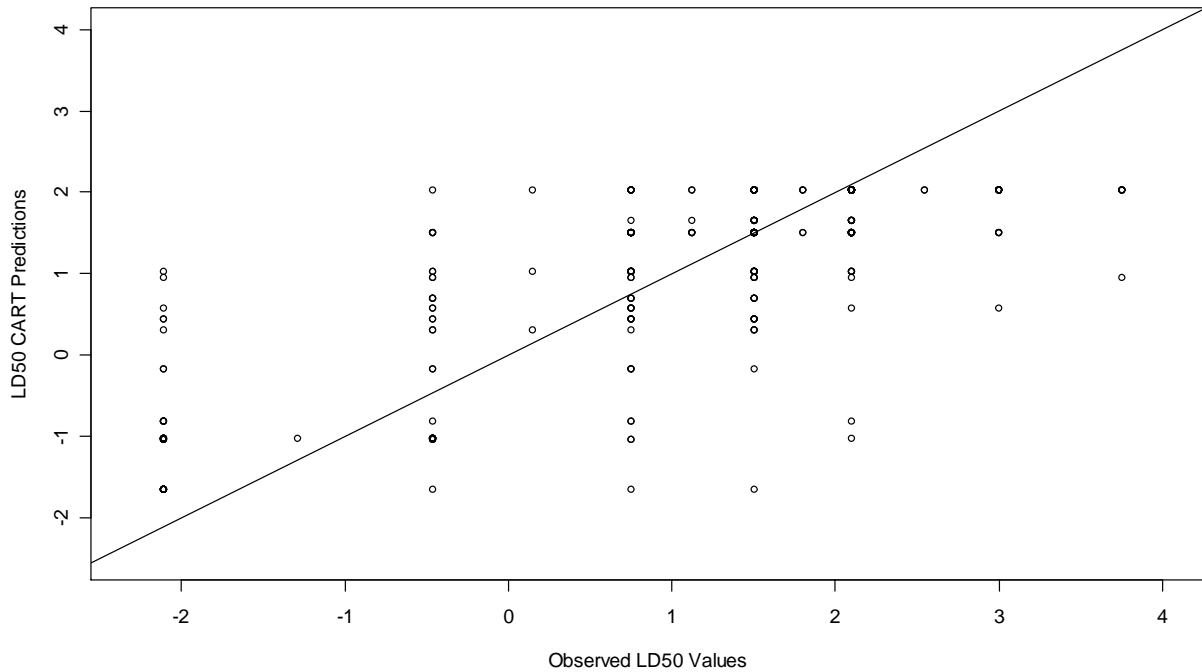


Figure 18: LD_{50} CART Predictions versus Observed LD_{50}

C. CART and Top Predictor Hybrid Models

We used a combination of CART and the top predictor models in an attempt to improve predictions. We performed CART on the residuals of the top in tier models and used the terminal nodes in multiple linear regression. An indicator variable was created for all but the first terminal node, so that there were $q - 1$ variables for a tree with q terminal nodes predicting the residuals. All observations that appear in the first node are coded with “0” for every indicator variable. All observations that appear in the second node are coded with “1” for the first indicator variable, and “0” for the each of the remaining indicator variables. Those observations that appear in the third node are coded with “1” for the second indicator variable and “0” for the remaining. This process continues until each observation is coded according to the terminal node into which it falls. We added the indicator variables to the already existing top predictor models and performed multiple linear regression. These models are hereafter referred to as hybrid models.

C.1. Results for CART Hybrid Top 4-Tier Model

We created two CART hybrid models: one utilizing top four tier predictors for the square root of the shifted LD_{50} and one utilizing the top four tier predictors for the untransformed LD_{50} . When CART was performed on the residuals of the square root model, the percentage of the watershed as unconsolidated geologic type was not utilized in any split. For the untransformed model, all four variables were utilized in the tree. Also, the transformed model had only 11 terminal nodes, whereas the untransformed model had 12. For both models, the indicator variables were named so that a site that fell into the i^{th} terminal node would be coded with “1” for the indicator variable named “node. i ” and “0” for all remaining indicator variables.

Some of the top four predictors were not highly significant when the indicator variables were added (Table 13 and Table 14). The trees for the two models were quite different (Figure 19 and Figure 20), despite the fact that the predictive nature of the hybrid models were not.

Both of these models improved the top in tier prediction models considerably. There were still issues with normality of the residuals and a considerable problem with predicting the large and small values of LD_{50} (Figure 21 and Figure 22). There was improvement in the adjusted R^2 value and the predictive indicators (Table 15). Overall, there was little difference between the two models, but the untransformed hybrid model is preferred for parsimony and slightly higher predictive statistics.

The improvement of the hybrid over the top 4-tier model is in its predictive-ability, and a slightly better fit. Because the residuals are not normal, this model does not match the capability of the top geomorphic plus top-3 tier model.

Table 13: Coefficients and statistics for $\sqrt{LD_{50} + 5}$ top 4-tier hybrid model

Coefficients	Coefficient Estimate	Standard Error	t-value	p-value
Intercept	2.141	0.163	13.167	$< 2 \times 10^{-16}$
Dist-wt % evergreen forest (30m buff)	0.147	0.111	1.321	1.87×10^{-1}
Mean slope (300m buff)	0.009	0.002	5.185	3.38×10^{-7}
% watershed as unconsolidated	-0.247	0.063	-3.957	8.91×10^{-5}
Average temperature in June ($^{\circ}\text{C}$)	-0.022	0.004	-5.289	1.99×10^{-7}
Node.2	-0.038	0.133	-0.287	7.74×10^{-1}
Node.3	0.118	0.139	0.851	3.95×10^{-1}
Node.4	0.139	0.137	1.015	3.11×10^{-1}
Node.5	0.158	0.112	1.416	1.58×10^{-1}
Node.6	0.310	0.100	3.112	1.98×10^{-3}
Node.7	0.362	0.100	3.618	3.34×10^{-4}
Node.8	0.358	0.146	2.449	1.47×10^{-2}
Node.9	0.424	0.112	3.800	1.66×10^{-4}
Node.10	0.449	0.150	2.999	2.87×10^{-3}
Node.11	0.689	0.143	4.808	2.13×10^{-6}

Table 14: Coefficient and statistics for LD_{50} top 4-tier hybrid model

Coefficient	Coefficient Estimate	Standard error	t-value	p-value
Intercept	-0.435	0.756	-0.576	5.65×10^{-1}
Dist-wt % evergreen forest (30m buff)	0.651	0.525	1.239	2.16×10^{-1}
Mean slope (300m buff)	0.049	0.008	6.234	1.12×10^{-9}
% watershed as unconsolidated	-0.669	0.349	-1.918	5.58×10^{-2}
Average temperature in June ($^{\circ}\text{C}$)	-0.098	0.020	-4.928	1.20×10^{-6}
node.2	0.210	0.581	0.362	7.17×10^{-1}
node.3	0.680	0.618	1.100	2.72×10^{-1}
node.4	0.699	0.506	1.380	1.68×10^{-1}
node.5	0.618	0.687	0.900	3.69×10^{-1}
node.6	0.621	0.641	0.970	3.33×10^{-1}
node.7	1.370	0.450	3.043	2.49×10^{-3}
node.8	1.645	0.453	3.630	3.19×10^{-4}
node.9	1.688	0.699	2.413	1.63×10^{-2}
node.10	1.910	0.506	3.777	1.82×10^{-4}
node.11	1.943	0.706	2.753	6.16×10^{-3}
node.12	3.052	0.653	4.677	3.95×10^{-6}

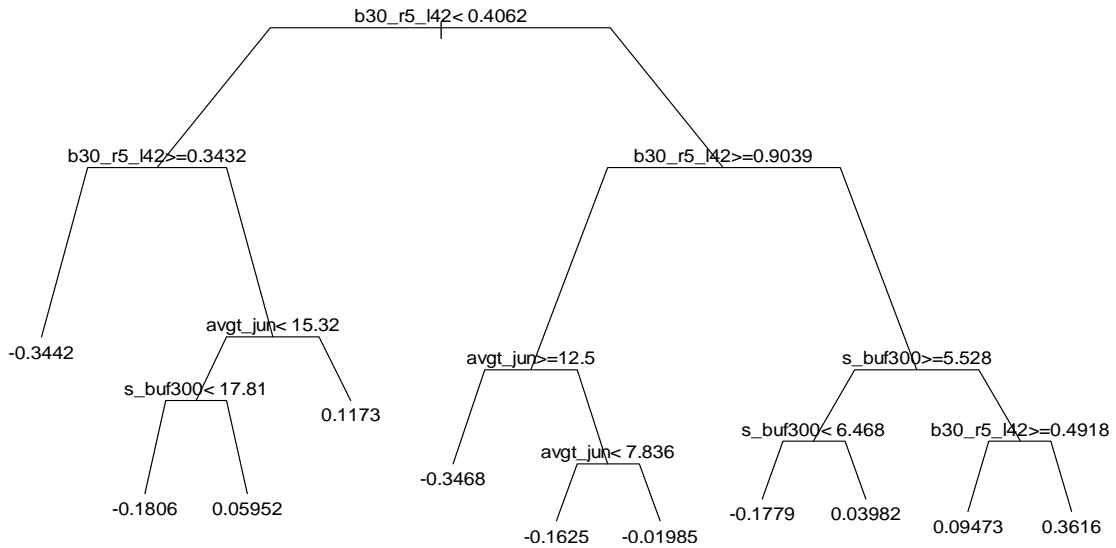


Figure 19: Tree for residuals of the top 4-tier $\sqrt{LD_{50} + 5}$ model

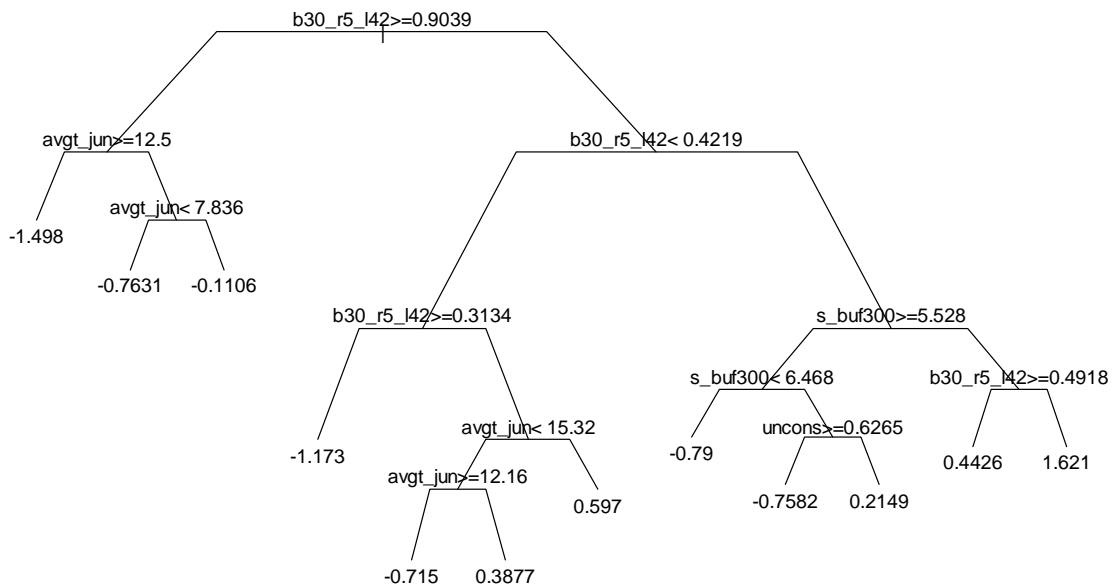


Figure 20: Tree for residuals of the untransformed LD_{50} top 4-tier model

Table 15: Top 4-tier hybrid model summary statistics

Response	Adjusted R^2	$PRESS_p$ for LD_{50}	Mean Square Prediction Error	$R^2_{prediction}$
square root ($LD_{50}+5$)	0.362	559.80	1.299	0.325
LD_{50}	0.363	555.08	1.288	0.327

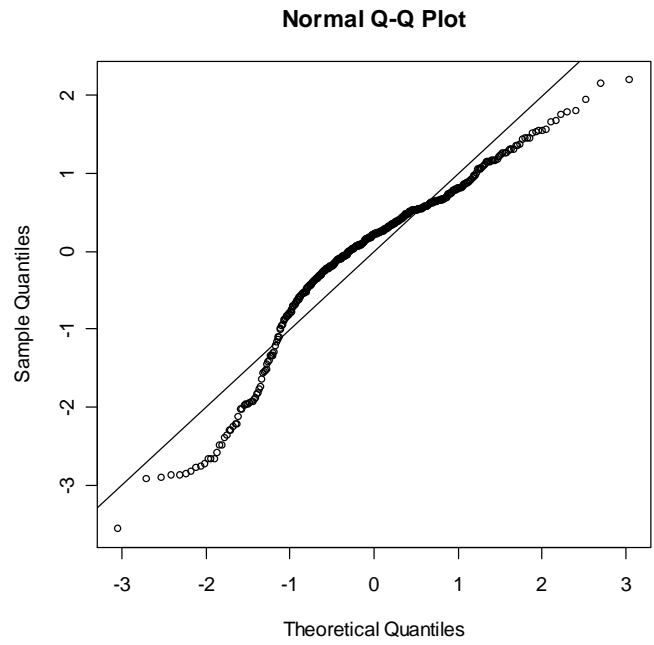


Figure 21: QQ Plot of the semi-studentized residuals for the top4-tier $\text{sqrt}(LD_{50} + 5)$ hybrid model

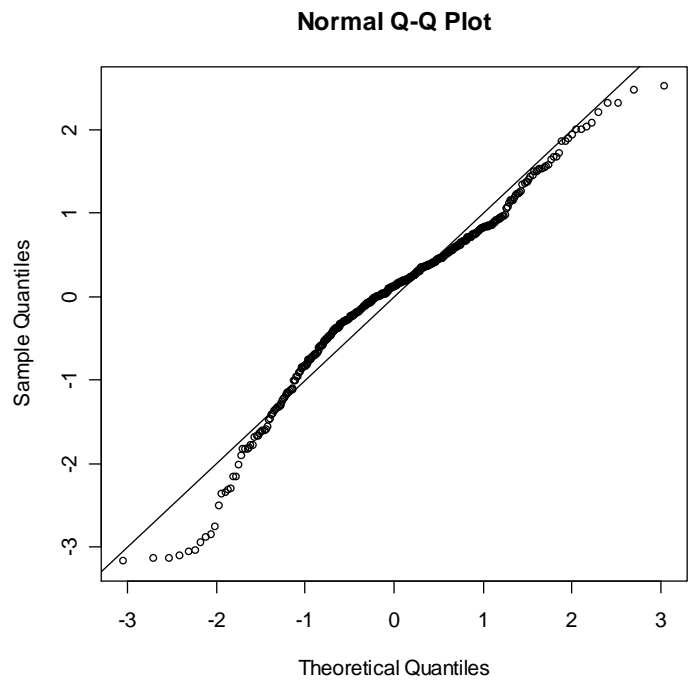


Figure 22: QQ-Plot for the semi-studentized residuals of top4-tier LD_{50} hybrid model

C.2 Results for the CART Hybrid Geomorphic plus Top 3-Tier Model

We performed CART on the residuals of the model with all the variables chosen in stepwise regression on the geomorphic tier and the top predictor from each of the climatic, geologic, and land cover tier. There were 18 terminal nodes in the residual tree utilizing the following 11 variables for the splits: the average temperature in Jun (labeled as *avgt_jun*), distance-weighted percent of evergreen forest within a 30-meter buffer (labeled as *b30_r5_l42*), hill connectivity (labeled as *CVCON*), the coefficient of valley entrenchment (labeled as *CVENTR*), the mean of the product of channel slope and watershed area above the outlet (labeled as *link_sa*), the average stream slope (labeled as *link_slope*), an average measure of the width of the valley bottom versus the theoretical width for a sinuous stream (labeled as *MENTB*), the mean outlet slope times the watershed area above the outlet (labeled as *out_sa*), the average slope of the watershed (labeled as *shed_slp*), the ratio of the slope of the watershed to the elongation of the watershed (labeled as *slp_elon*), and the average topographic wetness (labeled as *topo_wet*) (Figure 23). We created 17 indicator variables with the first node coded with all zeros and performed multiple linear regression on these indicator variables and the variables from the original model (Table 16).

This model outperforms all previous models in predictive-ability (Table 17). Fifty-four percent of the variation in LD_{50} values is explained by the regression model. The $PRESS_p$ criterion is lower and $R^2_{prediction}$ is higher than in any of the previous models. The problems with normality of the residuals and poor prediction of extreme LD_{50} , while greatly improved, have not been completely resolved (Figure 24 and Figure 25). For the models utilizing all sites in the Oregon and Washington, this model is preferred to the previous models if predictive-ability is its

intended use. However, there are many terms and some collinearity issues in the model, and thus other models may be preferred for their simplicity.

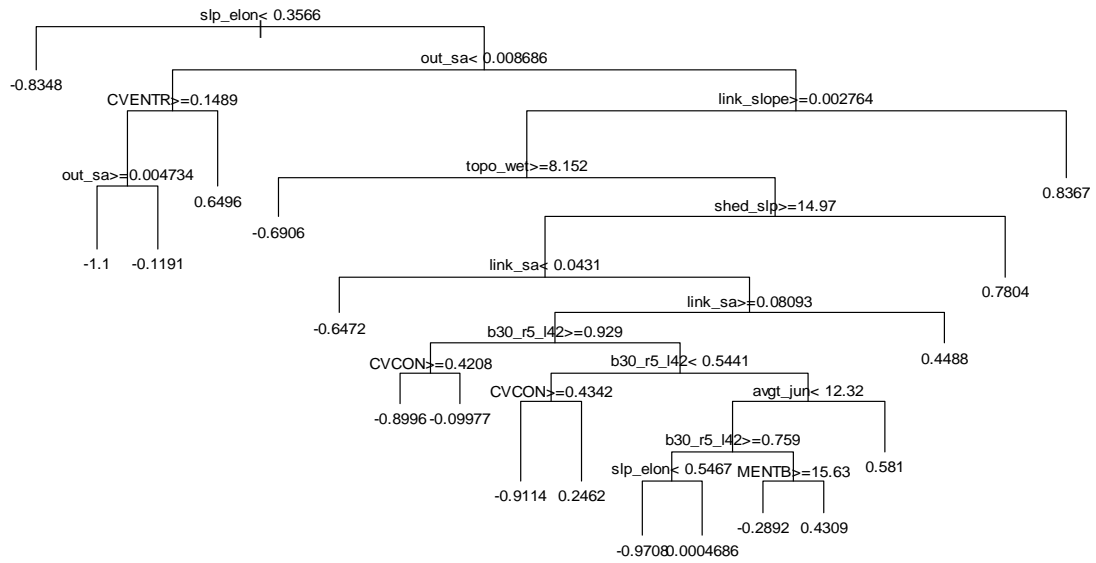


Figure 23: Tree on the residuals of the top geomorphic plus top 3-tier model

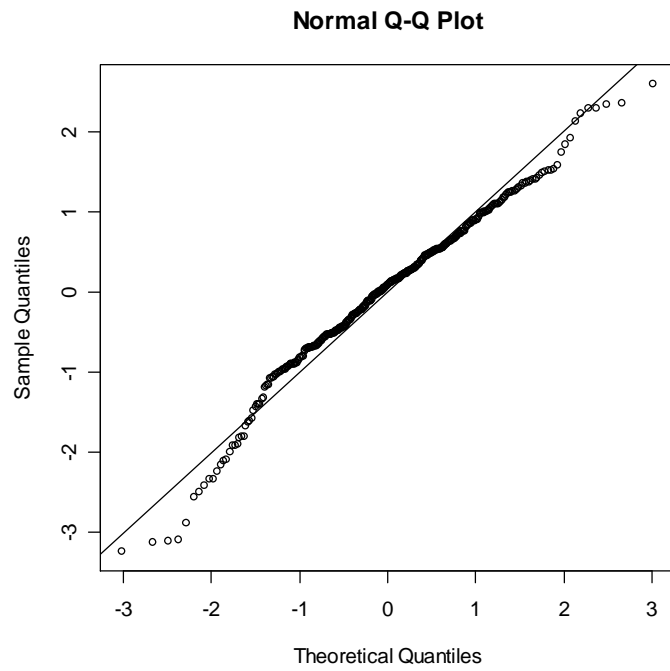


Figure 24: QQ-Plot of the semi-studentized residuals for geomorphic plus top 3-tier hybrid model

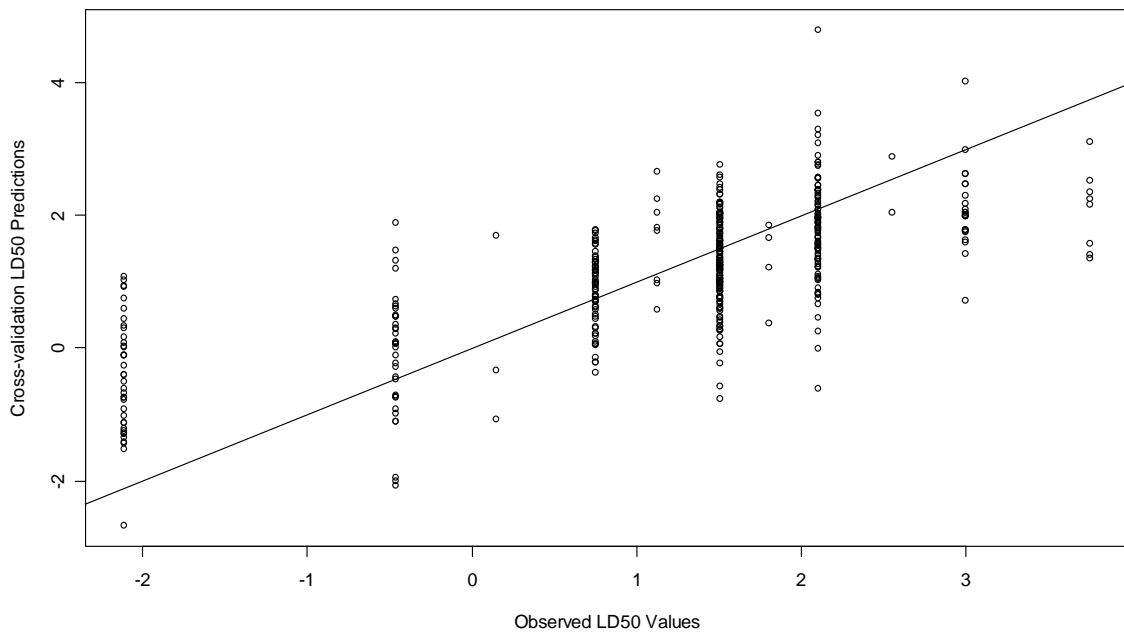


Figure 25: Cross-Validated LD_{50} Predictions versus Observed LD_{50} for geomorphic plus top 3-tier hybrid model

Table 16: Coefficients and statistics for geomorphic plus top 3-tier hybrid model

Coefficient	Coefficient Estimate	Standard Error	t-value	p-value
Intercept	-5.508	1.5140	-3.638	3.150 x 10 ⁻⁰⁴
Valley bottom versus width sinuous stream	0.015	0.0066	2.320	2.092 x 10 ⁻⁰²
Stream power (slope x area) ^{0.4}	6.393	2.5350	2.522	1.212 x 10 ⁻⁰²
Stream power (slope x area)	-0.101	0.0895	-1.123	2.620 x 10 ⁻⁰¹
Hill connectivity	1.081	0.2199	4.915	1.360 x 10 ⁻⁰⁶
Valley entrenchment	-0.006	0.0013	-4.395	1.460 x 10 ⁻⁰⁵
Coefficient of valley entrenchment	0.242	0.1757	1.375	1.700 x 10 ⁻⁰¹
Drainage density (m ⁻¹)	-283.500	65.4700	-4.330	1.940 x 10 ⁻⁰⁵
Size/complexity of stream	0.003	0.0009	3.516	4.950 x 10 ⁻⁰⁴
Average stream slope	-3.968	4.1630	-0.953	3.412 x 10 ⁻⁰¹
Mean slope (300m buff)	0.004	0.0091	3.698	2.510 x 10 ⁻⁰⁴
Channel slope x area above the outlet	-0.092	0.0295	-3.103	2.068 x 10 ⁻⁰³
Ratio of slope to elongation	-0.487	0.1899	-2.565	1.072 x 10 ⁻⁰²
Outlet area (km ²)	-0.013	0.0039	-3.232	1.342 x 10 ⁻⁰³
Average channel slope	0.715	1.5810	0.452	6.513 x 10 ⁻⁰¹
Average topographic wetness	0.577	0.1691	3.410	7.230 x 10 ⁻⁰⁴
Average slope in the watershed	0.024	0.0069	3.512	5.010 x 10 ⁻⁰⁴
% step pool	0.356	0.4372	0.813	4.167 x 10 ⁻⁰¹
Average temperature in June (°C)	-0.042	0.0199	-2.123	3.442 x 10 ⁻⁰²
% watershed as unconsolidated	-0.550	0.2574	-2.135	3.340 x 10 ⁻⁰²
Dist-wt % evergreen forest (30m buff)	0.997	0.2917	3.419	7.020 x 10 ⁻⁰⁴
node.2	0.117	0.3792	0.307	7.586 x 10 ⁻⁰¹
node.3	0.113	0.3430	0.328	7.431 x 10 ⁻⁰¹
node.4	0.096	0.3792	0.254	7.999 x 10 ⁻⁰¹
node.5	0.277	0.3856	0.717	4.736 x 10 ⁻⁰¹
node.6	0.296	0.4136	0.717	4.741 x 10 ⁻⁰¹
node.7	0.468	0.3475	1.347	1.787 x 10 ⁻⁰¹
node.8	0.717	0.3647	1.967	5.000 x 10 ⁻⁰²
node.9	0.985	0.2895	3.403	7.430 x 10 ⁻⁰⁴
node.10	1.017	0.3164	3.214	1.428 x 10 ⁻⁰³
node.11	1.119	0.2750	4.069	5.820 x 10 ⁻⁰⁵
node.12	1.493	0.3206	4.658	4.500 x 10 ⁻⁰⁶
node.13	1.583	0.2868	5.519	6.540 x 10 ⁻⁰⁸
node.14	1.573	0.2890	5.443	9.710 x 10 ⁻⁰⁸
node.15	1.725	0.2881	5.987	5.190 x 10 ⁻⁰⁹
node.16	1.738	0.3656	4.753	2.910 x 10 ⁻⁰⁶
node.17	1.977	0.3329	5.940	6.730 x 10 ⁻⁰⁹
node.18	2.104	0.3196	6.585	1.620 x 10 ⁻¹⁰

Table 17: Summary statistics for geomorphic plus top 3-tier hybrid model

Response	Adjusted R^2	$PRESS_p$ for LD_{50}	Mean Square Prediction Error	$R^2_{prediction}$
LD_{50}	0.540	378.107	0.955	0.490

D. A Comparison of Models Using the Oregon and Washington Sites

In modeling LD_{50} , a balance in the number of variables selected from the large dataset is required. In the all-variable stepwise models, high adjusted R^2 values and low $R^2_{prediction}$ values indicate that the predictive-ability of a model drops drastically with over-fitting of the model. In the top 4-tier models, both a poor fit and poor predictive-ability occur due to the small number of variables. Models utilizing more geomorphic variables in top-in-tier models achieve a balance that partially correct over and under-fitting. These models provide improved fit and improved predictions. The hybrid models provide a unique approach to variable selection by providing information rich terminal node indicator variables. These indicator variables, when added to the original models, provide more explanation of the variation in LD_{50} and provide better predictions (Table 18). There is a loss of parsimony and increase in collinearity in the latter models. Yet, when prediction is the main objective, the extra variables are justifiable.

Table 18: A comparison of models for all Oregon and Washington data

Model	Response	Adjusted R^2	$R^2_{prediction}$	AIC
All variable stepwise subset 1	$\sqrt{LD_{50} + 5}$	0.7573	-1.19×10^{12}	-1470.6
All variable stepwise subset 2	$\sqrt{LD_{50} + 5}$	0.6581	-7.62×10^{10}	-1344.8
All variable stepwise subset 3	$\sqrt{LD_{50} + 5}$	0.8158	-6.67×10^{17}	-1743.2
Top 4-tier	$\sqrt{LD_{50} + 5}$	0.2313	0.217	-1126.9
Top 4-tier	LD_{50}	0.2360	0.222	173.1
Top geomorphic plus top 3-tier	LD_{50}	0.362	0.319	95.2
CART	LD_{50}	NA	-0.065	NA
Top 4-tier hybrid	$\sqrt{LD_{50} + 5}$	0.362	0.325	-1197.7
Top 4-tier hybrid	LD_{50}	0.363	0.327	105.2
Top geomorphic plus top 3-tier hybrid	LD_{50}	0.540	0.490	-19.4

IV. Coast Range Data and Analysis

A. Description

The problems we experienced with over-prediction and under-prediction of bedrock and fine substrate indicate that the skewed distribution over all Oregon and Washington sites made accurate prediction difficult. Unlike other ecoregions, the Coast Range Ecoregion appeared to have an approximately normal distribution of LD_{50} measurements without any values identified as outliers (Figure 26). Within the Coast Range, accurate prediction of LD_{50} might be less difficult than in the entire dataset. There were 134 sites in this ecoregion where the ecosystem has similar characteristics as described by the EPA:

The low mountains of the Coast Range are covered by highly productive, rain-drenched coniferous forests. Sitka spruce and coastal redwood forests originally dominated the fog-shrouded coast, while a mosaic of western red cedar, western hemlock, and seral Douglas-fir blanketed inland areas. Today Douglas-fir plantations are prevalent on the intensively logged and managed landscape (USEPA WED, 2005).

Utilizing only the sites in the Coast Range Ecoregion, our goal is to find a model to predict LD_{50} for regions with similar ecosystems.

To predict LD_{50} for the Coast Range, all previous methods used for the Oregon and Washington sites would be utilized, with the exception of stepwise variable selection over the set of all predictors. Due to the smaller sample size in the Coast Range Ecoregion, step analysis over all variables required several subsets of the variables followed by step analysis performed on a combination of the predictors chosen from the subsets. The lack of predictive success for the set of all Oregon and Washington sites indicated that the top 4-tier method and the top geomorphic plus top 3-tier method were preferable.

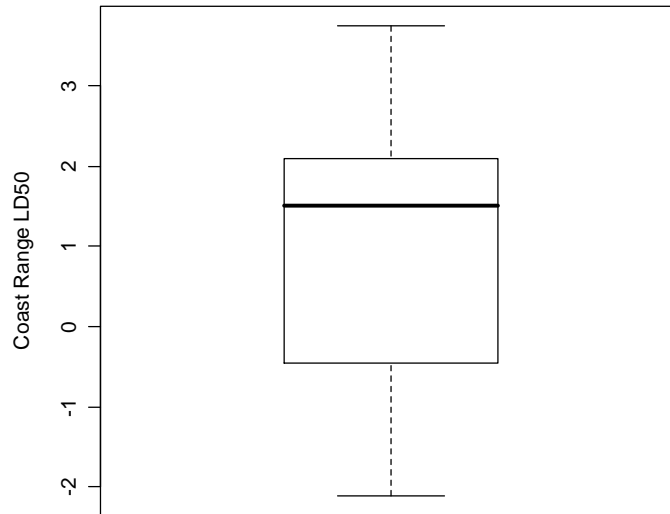


Figure 26: Boxplot of observed LD_{50} values in the Coast Range Ecoregion

As with the set of all Oregon and Washington sites, there were sites in the Coast Region with missing values for important geomorphic variables. After attempting step analysis on the geomorphic variables, there were much better results when missing sites were removed as opposed to missing variables. This left 128 sites of the original 134 Coast Region sites available for the observation set (Figure 27).

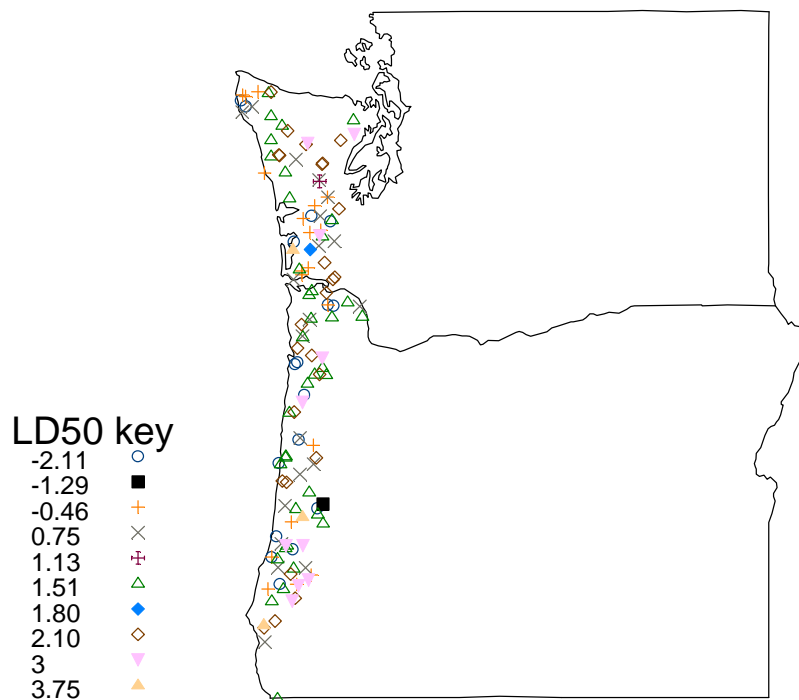


Figure 27: Coast Range Ecoregion Sites

B. Stepwise Variable Selection for the Coast Range Ecoregion

B.1 Top 4-Tier Model Results

We performed forward stepwise variable selection on the climatic, geologic, and geomorphic tiers as before. However, due to the smaller sample size, it was impossible to perform forward step on even two separate subsets of land cover metrics. To select the top predictor in the land cover metrics, we created eight subsets of variables, performed forward

stepwise variable selection on each subset, and, finally, performed forward stepwise variable selection on the subset created by taking the top four predictors from each of the eight processes.

The eight subsets are as follows: the distance-weighted predictors where the distance is weighted further by the topographic wetness index, the distance-weighted metrics with no topographic index weighting, the distance-weighted generalized metrics, the distance-weighted metrics within a 30-meter buffered area, the distance-weighted metrics within a 100-meter buffered area, the distance-weighted metrics within a 300-meter buffered area, the buffered (by 30, 100, or 300 meters) metrics with no distance-weighting, and the buffered (by 30, 100, or 300 meters) generalized metrics. The basic percentage metrics appeared in every subset.

The top predictors chosen for the Coast Range were the average minimum temperature in January (climatic predictor labeled as *mint_jan*), average watershed elevation (geomorphic predictor labeled as *avg_elev*), percentage of watershed underlain by volcanic geologic type (geologic variable labeled as *volcan*), and the percentage of wetlands, distance-weighted with coefficient of decay 0.0001, not weighted by the topographic wetness index, and buffered within 30-meters of the stream (land cover predictor labeled as *b30_r6_190*) (Table 19).

The minimum temperature in January, while chosen first in forward step selection, was not a significant predictor in this model. There is a strong negative correlation between the average watershed elevation and the minimum temperature in January ($r = -0.83$, p -value less than 2.2×10^{-16}). As research indicates that geomorphic variables are important in prediction of LD_{50} , the climatic variable was removed and replaced with the second predictor chosen in forward step selection, the ratio of precipitation in the wettest three months to the driest three months (labeled as *precipr2*). This variable was also correlated to the average watershed elevation ($r = 0.2434$, p -value = 0.0056). The next climatic variable chosen was

average precipitation in November (labeled as *prcp_nov*) and was also correlated to elevation. ($r = 0.3739$, $p\text{-value} = 1.38 \times 10^{-5}$). It was not until the fourth predictor in the forward step selection process that a variable not correlated with elevation was found. That variable was the average aspect, which is the average compass direction of the hillsides in a watershed. As correlation in the hybrid model created from this model was likely to occur, it was important that the first four variables be uncorrelated, and so average aspect was chosen as the climatic predictor for the top 4-tier model.

All five coefficient estimates in the model with average aspect replacing average minimum temperature in January were statistically significant (Table 20). The predictive-ability of this model surpassed the top 4-tier models for the entire data set and the residuals were close to normal (Table 21 and Figure 28). For a parsimonious model, this model does well at modeling and predicting LD_{50} . However, the problem with over-prediction and under-prediction of small and large values is still present (Figure 29). It should be noted that there are only three observations in the Coast Range data set that are classified as having bedrock median substrate, so the sites that are in this category do not have a strong effect on the linear regression model.

Table 19: Coefficients and statistics for the first (rejected) top 4-tier model: Coast Range

Coefficient	Coefficient Estimate	Standard Error	t-value	p-value
Intercept	-0.357	0.2931	-1.219	2.25×10^{-1}
Average minimum temperature in January (°C)	0.008	0.0869	0.097	9.23×10^{-1}
Average watershed elevation (m)	0.003	0.0009	3.422	8.46×10^{-4}
% watershed volcanic geologic type	0.854	0.3165	2.698	7.95×10^{-3}
% wetlands distance-weighted 30m buff	-25.070	10.5200	-2.383	1.87×10^{-2}

Table 20: Coefficients and statistics for the final top 4-tier model for the Coast Range

Coefficient	Coefficient Estimate	Standard Error	t-value	p-value
Intercept	-1.051	0.3660	-2.872	4.80×10^{-3}
Average aspect (°)	0.004	0.0015	2.460	1.53×10^{-2}
Average watershed elevation (m)	0.003	0.0005	6.133	1.08×10^{-8}
% watershed volcanic geologic type	0.888	0.3069	2.894	4.50×10^{-3}
% wetlands distance-weighted 30m buff	-25.840	10.2500	-2.520	1.30×10^{-2}

Table 21: Statistics for top 4-tier Coast Range model

Response	Adjusted R^2	$PRESS_p$ for LD_{50}	Mean Square Prediction Error	$R^2_{prediction}$
LD_{50}	0.384	189.649	1.493	0.362

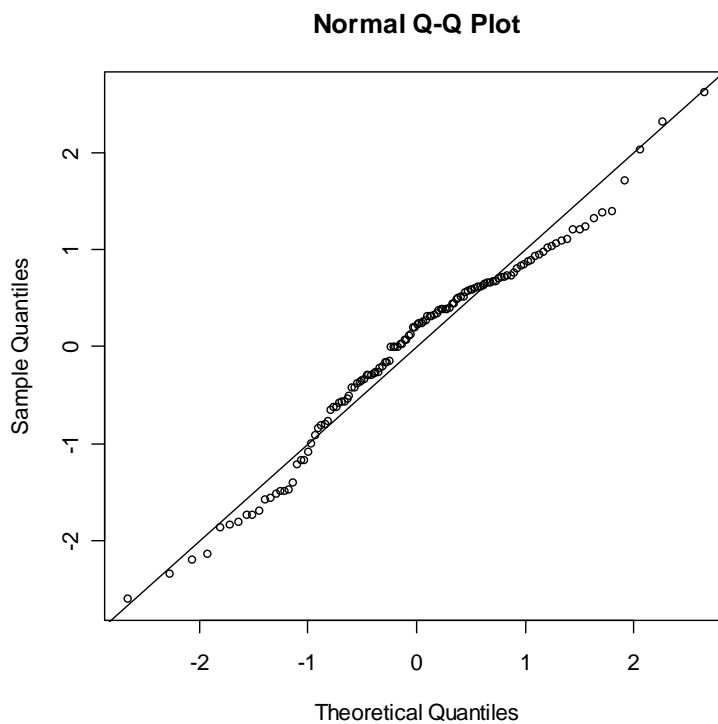


Figure 28: QQ-Plot of semi-studentized residuals for top 4-tier Coast Range model

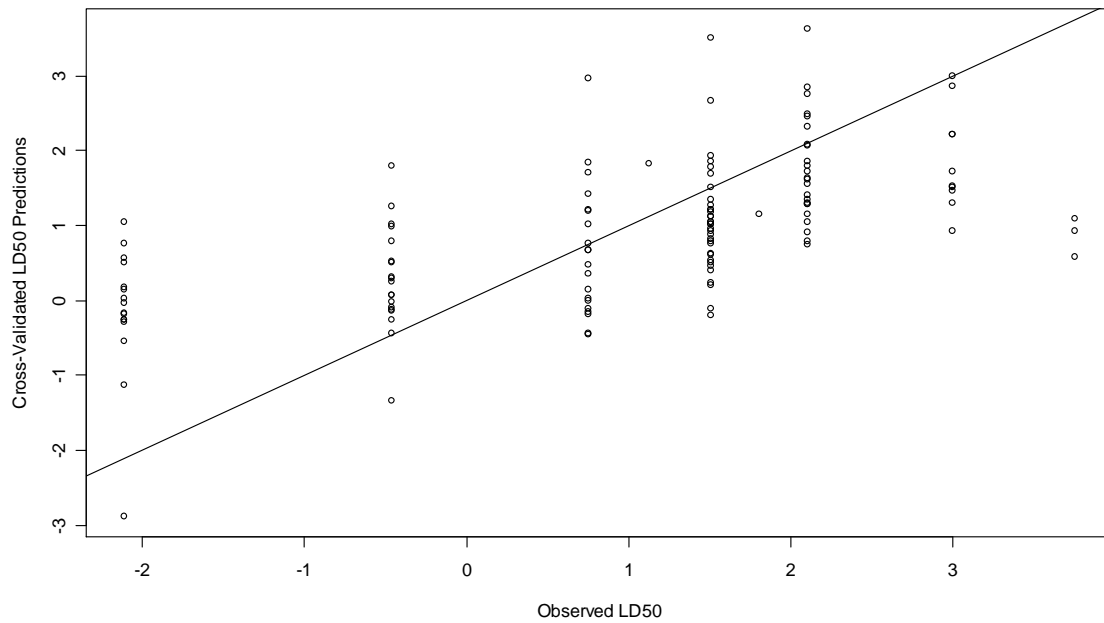


Figure 29: Cross-validated LD_{50} predictions versus observed LD_{50} for the top 4-tier Coast Range model

B.2 Geomorphic plus Top 3-Tier Model Results

We performed forward and backward stepwise variable selection for the geomorphic tiers. Some variables selected via stepwise regression were different than those chosen for all Oregon and Washington sites (Table 22).

Table 22: Top predictors chosen in stepwise variable selection on the geomorphic tier Coast Range

Predictor	Description
avg_elev	Average watershed elevation (m)
drainden	Drainage density
s_buf300	Mean slope within a 300-meter buffer
MENTR	Ratio of width of stream to width of floodplain (valley entrenchment)
CVCON	Coefficient of average connectivity
dist_1	Distance to the first tributary (m)
pct_lt4	Percent of landscape with less than 4% slope
pct_lt7	Percent of landscape with less than 7% slope
strahler	Measure of size and complexity of river
pct_C	Percent of stream as cascade
DWSP2	Distance-weighted stream power
relief_r	Watershed relief divided by its length

These predictors were added to the model with the percentage of the watershed underlain with volcanic geologic type, the aspect ratio, and the distance-weighted percentage of wetlands buffered within 30-meters of the stream (Table 23). This model improved the top 4-tier model in both predictability and fit (Table 24). The residuals appeared normal (Figure 30). Yet, the model still failed to predict the three bedrock observations accurately, and many of the fine-substrate observations were strongly over-predicted (Figure 31). Considering the complexity of predicting LD_{50} , this model does an excellent job of predicting when compared to other models for both the Coast Range and the entire Oregon and Washington sites.

Table 23: Coefficients and Statistics for geomorphic plus top 3-tier Coast Range model

Coefficient	Coefficient Estimate	Standard Error	t-value	p-value
Intercept	-5.686	1.0300	-5.521	2.21×10^{-7}
Average watershed elevation (m)	0.002	0.0007	3.422	8.69×10^{-4}
Measure of size and complexity of river	0.336	0.1130	2.974	3.61×10^{-3}
Mean slope within a 300-meter buffer	0.055	0.0145	3.812	2.26×10^{-4}
Coefficient of average connectivity	0.891	0.3454	2.580	1.12×10^{-2}
Valley Entrenchment	-0.003	0.0020	-1.602	1.12×10^{-1}
Distance to the 1 st tributary (m)	0.002	0.0013	1.725	8.73×10^{-2}
Percent cascade	-22.990	9.9390	-2.313	2.25×10^{-2}
Percent of landscape > 7 % slope	225.800	75.6700	2.985	3.49×10^{-3}
Distance-weighted stream power	31.150	13.3500	2.334	2.14×10^{-2}
Watershed relief divided by its length	1.632	1.5250	1.070	2.87×10^{-1}
Drainage density	-287.500	214.8000	-1.339	1.83×10^{-1}
Percent of landscape > 4% slope	-109.400	79.6500	-1.374	1.72×10^{-1}
% volcanic geologic type	0.349	0.2988	1.167	2.46×10^{-1}
Average aspect (°)	0.002	0.0014	1.518	1.32×10^{-1}
% wetlands distance-weighted 30m buff	-10.960	9.8880	-1.108	2.70×10^{-1}

Table 24: Statistics for geomorphic plus top 3-tier Coast Range model

Response	Adjusted R^2	$PRESS_p$ for LD_{50}	Mean Square Prediction Error	$R^2_{prediction}$
LD_{50}	0.548	150.069	1.182	0.495

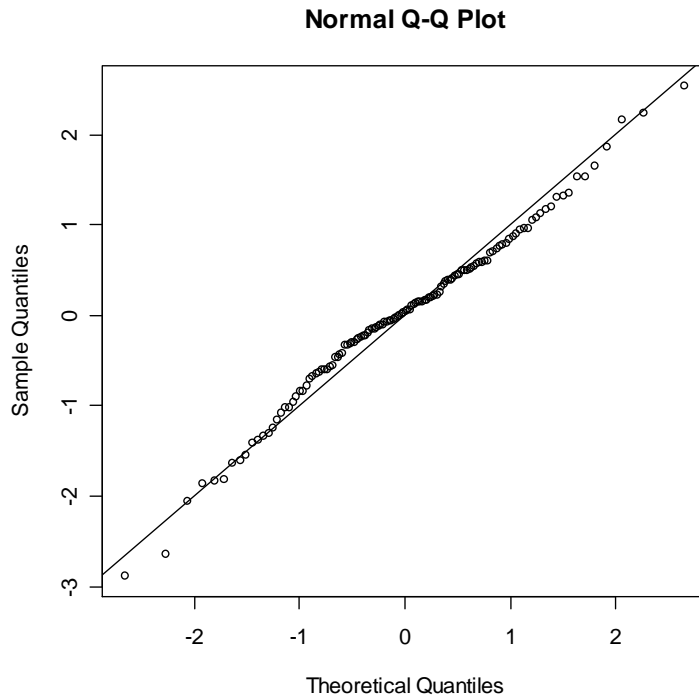


Figure 30: QQ-Plot for geomorphic plus top 3-tier Coast Range model

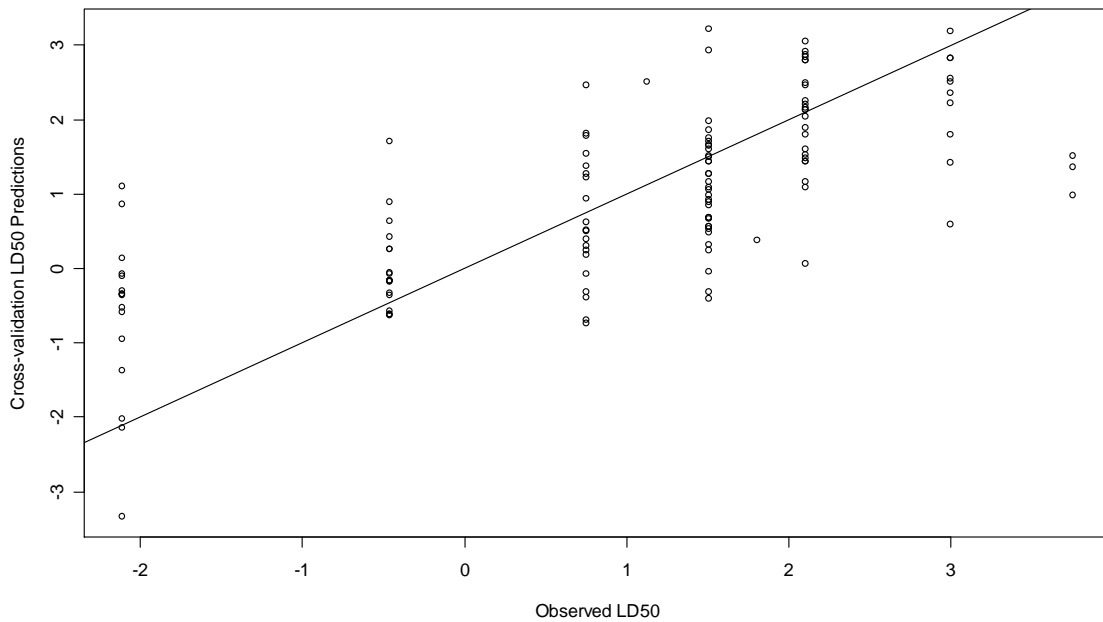


Figure 31: Cross-validated LD_{50} predictions versus observed LD_{50} for geomorphic plus top 3-tier Coast Range model

C. CART Using All Variables in the Coast Range Dataset

We performed CART on LD_{50} using all 134 sites and all variables in the Coast Range dataset. The nine variables utilized in the node splits were average aspect (labeled as aspect), average watershed elevation (labeled as avg_elev), percentage of evergreen forest buffered within 30-meters (labeled as b30_142), average hillside connectivity (labeled as MCON), minimum temperature in December (labeled as mint_dec), outlet area (labeled as out_area), average precipitation in January (labeled as prcp_jan), average precipitation in May (labeled as prcp_may), and distance-weighted percentage of forest ($\alpha = 0.0005$), not weighted by the topographic wetness index (labeled as r7_140_A). The tree was not complex as there were only ten terminal nodes (Figure 32).

Similar to CART performed with all Oregon and Washington sites, this model did not predict LD_{50} values well. Although there was a slight positive association between the predicted values and the observed values, the predictions were not consistently accurate (Figure 33). To see how well this type of model would fit future models, one observation was taken out of the dataset, the regression tree was refit, and the missing observation was predicted (Table 25). While CART improved predictions when used to create a hybrid model, alone it was unstable to the removal of sites and its ability to predict LD_{50} is not strong, even when the model is created on a region where the ecosystem is homogenous.

Table 25: Summary statistics for the CART Coast Range model

Response	$PRESS_p$ for LD_{50}	Mean Square Prediction Error	$R^2_{prediction}$
LD_{50}	286.16	2.152	0.0873

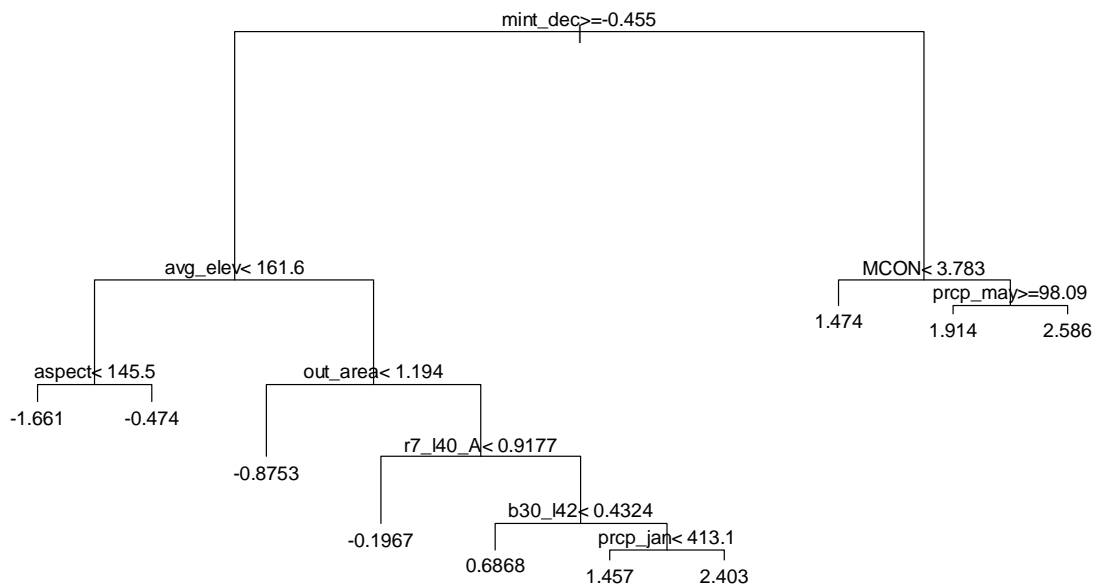


Figure 32: Tree for CART LD_{50} model using all variables for the Coast Range

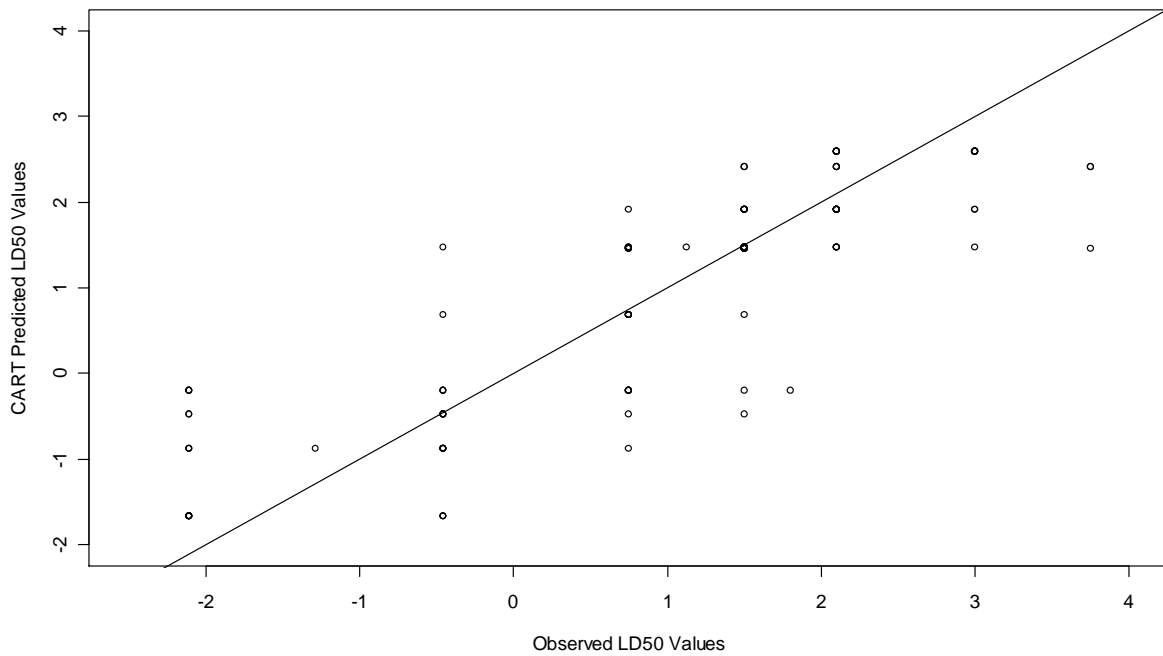


Figure 33: CART predicted LD_{50} values versus observed LD_{50} values

D. Hybrid CART and Multiple Regression Models for the Coast Range

D.1 Top 4-Tier Model Hybrid Results

We performed CART on the residuals of the top 4-tier Coast Range model, and there were nine terminal nodes utilizing average elevation, percentage of the watershed that is underlain with volcanic geologic type, and the distance-weighted percentage of wetlands within a 30-meter buffer (Figure 34). There were some predictors that were not significant when the eight indicator variables were added to the model (Table 26). The model improved the top 4-tier Coast Range model by further improving the normality of the residuals and increasing predictive-ability (Table 27 and Figure 35). There was still over-prediction and under-prediction of bedrock and fine substrate observations (Figure 36). However, for having only twelve predictors, this model performs very well with 50% accuracy in predicting LD_{50} .

Table 26: Coefficients and statistics for top 4-tier hybrid Coast Range model

Coefficient	Coefficient Estimate	Standard Error	t-value	p-value
Intercept	-1.379	0.6334	-2.177	3.15×10^{-2}
Average aspect (°)	0.002	0.0013	1.242	2.17×10^{-1}
Average watershed elevation (m)	0.002	0.0014	1.733	8.58×10^{-2}
% watershed volcanic geologic type	1.078	0.3470	3.106	2.39×10^{-3}
% wetlands distance-weighted 30m buff	-14.029	9.8488	-1.424	1.57×10^{-1}
node.2	0.525	0.8996	0.584	5.61×10^{-1}
node.3	-0.050	0.5439	-0.091	9.27×10^{-1}
node.4	0.334	0.4711	0.710	4.79×10^{-1}
node.5	0.628	0.5526	1.137	2.58×10^{-1}
node.6	1.329	0.5654	2.350	2.05×10^{-2}
node.7	1.334	0.4256	3.135	2.18×10^{-3}
node.8	1.709	0.4918	3.475	7.21×10^{-4}
node.9	1.973	0.4386	4.497	1.65×10^{-5}

Table 27: Statistics for top 4-tier hybrid Coast Range model

Response	Adjusted R^2	$PRESS_p$ for LD_{50}	Mean Square Prediction Error	$R^2_{prediction}$
LD_{50}	0.552	147.516	1.162	0.503

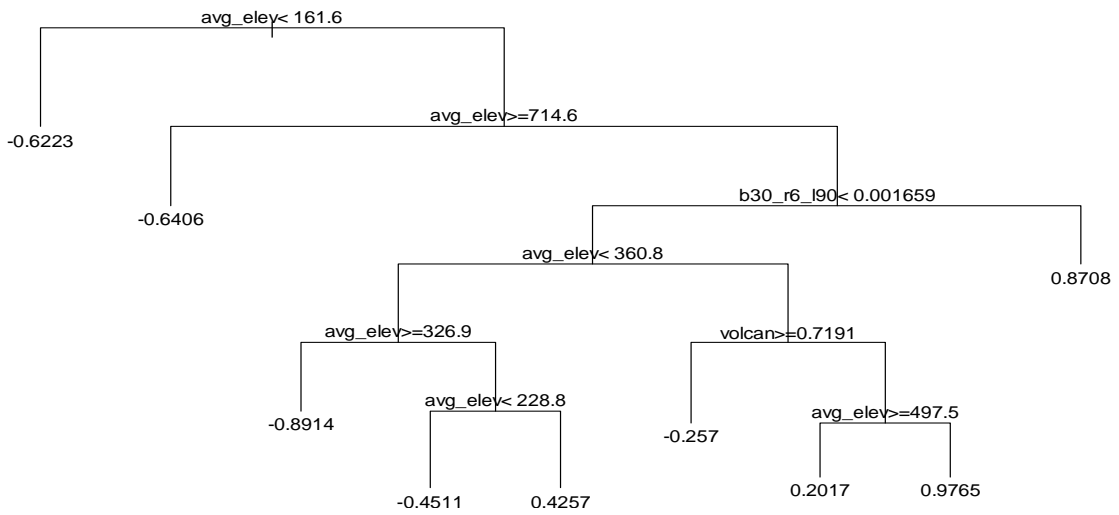


Figure 34: Tree for residuals of top 4-tier Coast Range model

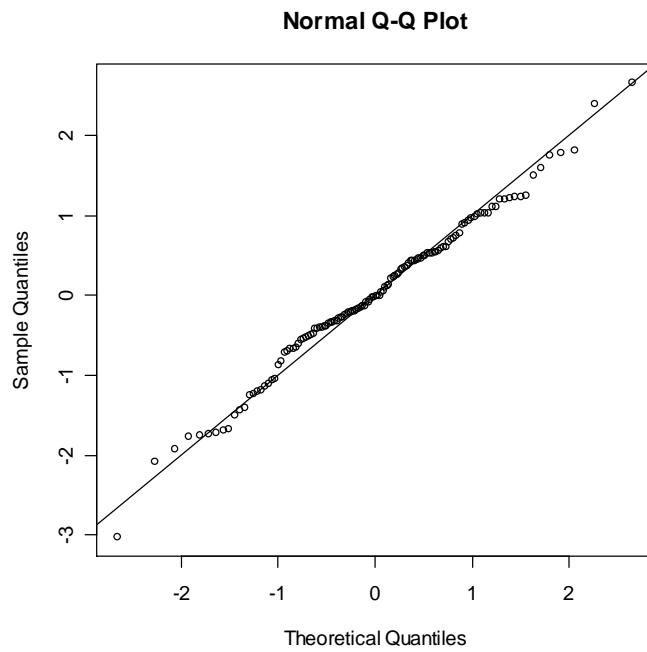


Figure 35: QQ-Plot of semi-studentized residuals for top 4-tier hybrid Coast Range model

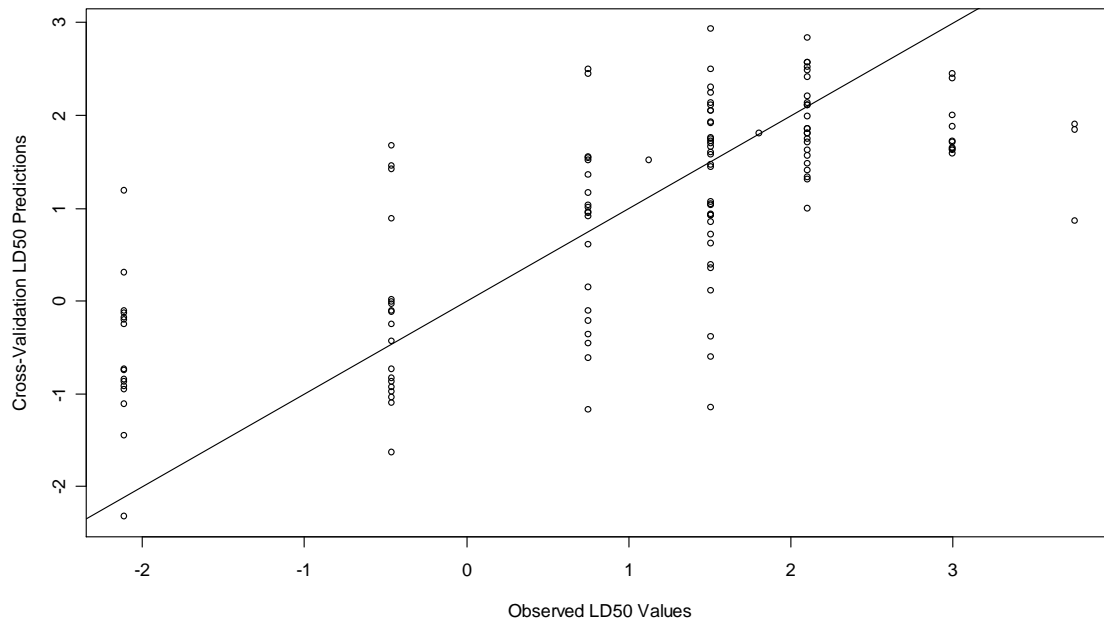


Figure 36: Cross-validated prediction values versus LD_{50} for top 4-tier hybrid model

D.2 Geomorphic plus Top 3-Tier Hybrid Model Results

We fit a classification and regression tree to the residuals of the model containing all geomorphic stepwise variables and the top variable from the climatic, geomorphic, and land cover tiers. The variables chosen for the splits were the coefficient of hill connectivity (labeled as CVCON), distance to the first tributary (labeled as Dist_1), distance-weighted stream power (labeled as DWSP2), ratio of the width of the stream to the width of the floodplain (labeled as MENTR), percent of stream as cascade (labeled as pct_C), percent of landscape with slope less than 7% (labeled as pct_lt7), and watershed relief divided by its length (labeled as relief_r). There were ten terminal nodes in the tree with the first split utilizing the watershed relief divided by its length (Figure 37). We created nine indicator variables and added these to the geomorphic plus top 3-tier model (Table 28).

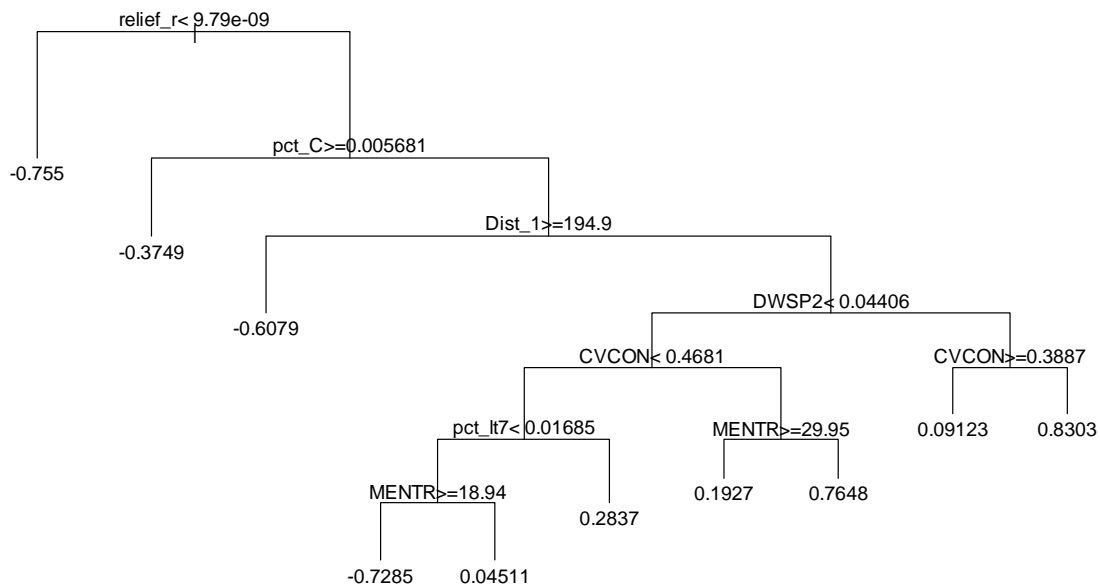


Figure 37: CART on the residuals of the geomorphic plus top 3-tier model

There was collinearity in the model when the CART node indicator variables were added to the model, and thus the residuals were less normal than those for the top 4-tier hybrid model (Figure 38). Yet, this model had 70% of the variation in LD_{50} values explained by the model, notably higher than other models with positive $R^2_{prediction}$ values. The ability of this model to predict omitted observations was the strongest of all previous models with a predictive R^2 of 61.4%. The nine percent difference between $R^2_{prediction}$ and adjusted R^2 may indicate that the large number of variables over-fit the data (Table 29).

The three sites with bedrock LD_{50} values were under-predicted and many of the sites with fine LD_{50} values were over-predicted. However, many of these fine substrate sites were

predicted within a reasonable range, unlike any previous models (Figure 39). For its predictive-ability and fit of the data, this model performs extremely well.

Table 28: Statistics and coefficients for the geomorphic plus top 3-tier hybrid Coast Range model

Coefficient	Coefficient Estimate	Standard Error	t-value	p-value
Intercept	-5.325	0.9067	-5.873	5.29x 10 ⁻⁸
Average watershed elevation (m)	0.003	0.0006	3.965	1.36x 10 ⁻⁴
Measure of size and complexity of river	0.180	0.1070	1.683	9.54 x 10 ⁻²
Mean slope within a 300-meter buffer	0.056	0.0125	4.506	1.75 x 10 ⁻⁵
Coefficient of average connectivity	0.516	0.3425	1.507	1.35 x 10 ⁻¹
Valley Entrenchment	-0.002	0.0017	-1.342	1.83 x 10 ⁻¹
Distance to the 1 st tributary (m)	0.005	0.0013	3.611	4.73 x 10 ⁻⁴
Percent cascade	-9.456	9.0960	-1.040	3.01 x 10 ⁻¹
Percent of landscape > 7 % slope	152.000	64.4800	2.357	2.03 x 10 ⁻²
Distance-weighted stream power	16.200	12.3800	1.308	1.94 x 10 ⁻¹
Watershed relief divided by its length	0.365	1.3510	0.270	7.88 x 10 ⁻¹
Drainage density	-171.900	187.6000	-0.916	3.62 x 10 ⁻¹
Percent of landscape > 4% slope	-56.890	67.6900	-0.840	4.03 x 10 ⁻¹
% volcanic geologic type	0.385	0.2563	1.501	1.36 x 10 ⁻¹
Average aspect (°)	0.002	0.0012	1.450	1.50 x 10 ⁻¹
% wetlands distance-weighted 30m buff	-14.430	8.2200	-1.756	8.21 x 10 ⁻²
node.2	0.225	0.3892	0.578	5.65 x 10 ⁻¹
node.3	0.134	0.5015	0.267	7.90 x 10 ⁻¹
node.4	0.863	0.5438	1.588	1.15 x 10 ⁻¹
node.5	1.251	0.4056	3.083	2.63 x 10 ⁻³
node.6	1.546	0.5140	3.008	3.31 x 10 ⁻³
node.7	1.394	0.3794	3.675	3.80 x 10 ⁻⁴
node.8	1.304	0.3676	3.547	5.89 x 10 ⁻⁴
node.9	2.172	0.4573	4.749	6.63 x 10 ⁻⁶
node.10	2.065	0.4394	4.699	8.10 x 10 ⁻⁶

Table 29: Statistics for the geomorphic plus top 3-tier hybrid Coast Range model

Response	Adjusted R ²	<i>PRESS_p</i> for LD ₅₀	Mean Square Prediction Error	R ² _{prediction}
LD ₅₀	0.700	114.597	0.902	0.614

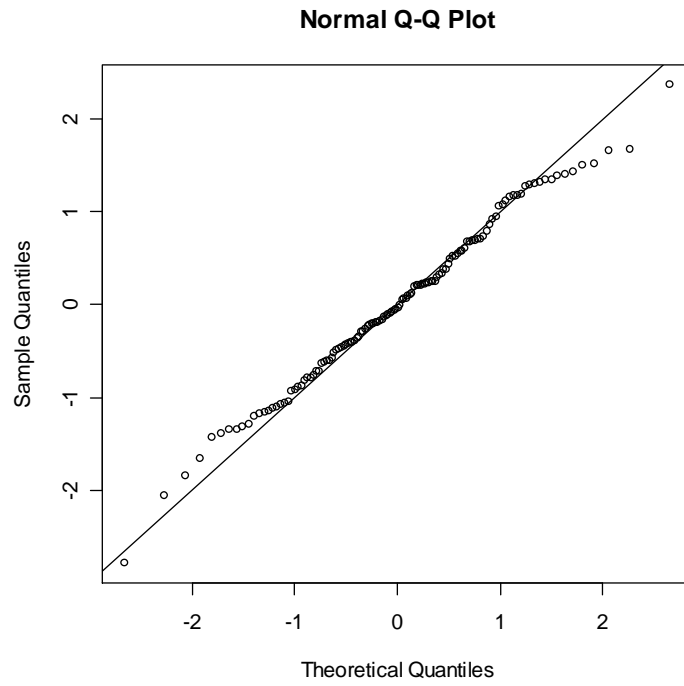


Figure 38: QQ-Plot of semi-studentized residuals for the geomorphic plus top 3-tier hybrid Coast Range

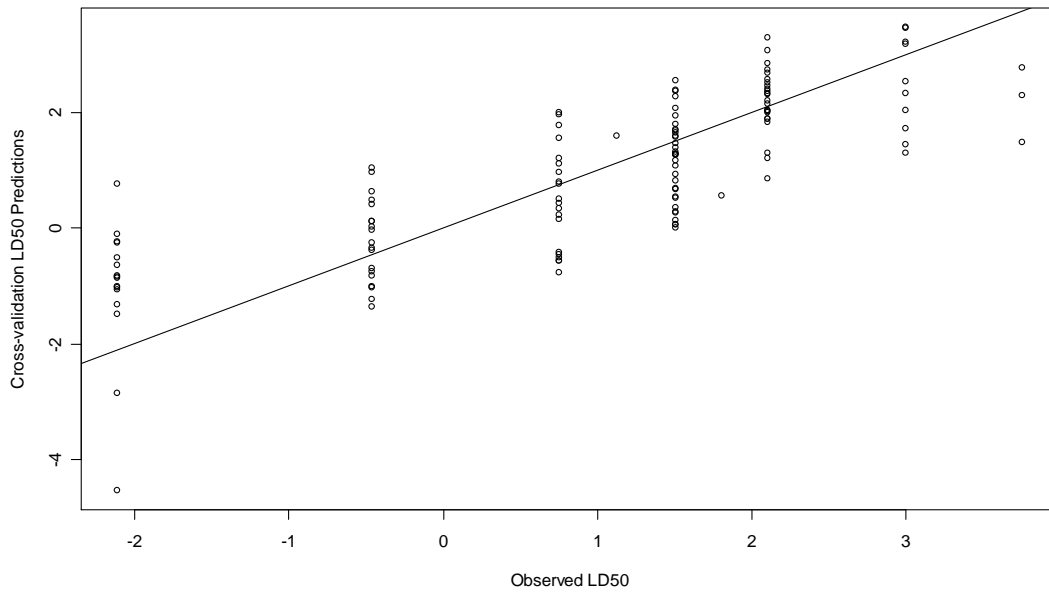


Figure 39: Cross-validation LD_{50} predictions versus observed LD_{50} the geomorphic top 3-tier hybrid Coast Range model

E. Comparison of Models Using the Coast Range Ecoregion Data

In the Coast Range analysis, better prediction and model fit were possible due to the distribution of LD_{50} in this ecoregion. Yet, the need for a balance between over and under-fitting was still necessary. The top 4-tier models, while better performing than those for the entire data set, were under-fit. Adding more geomorphic variables improved both the predictions and the fit of the models. The top 4-tier hybrid Coast model improved upon the fit and predictive-ability of the original top 4-tier Coast model. The geomorphic plus top-3 hybrid model was slightly over-fit as is indicated by the difference between adjusted R^2 and $R^2_{prediction}$ and by the loss of normality in the residuals when compared to the model from which the hybrid originated (Table 30). Yet, the predictive-ability of the hybrid indicates that a similar approach with a more discriminating variable selection process could be stronger performing and better fitting than a non-hybrid model.

Table 30: A comparison of Coast Range models

Model	Adjusted R^2	$R^2_{prediction}$	AIC
Top 4-tier model	0.384	0.362	55.6
Geomorphic plus top-3 model	0.548	0.495	26.0
CART model	NA	0.087	NA
Top 4-tier hybrid model	0.552	0.503	22.3
Geomorphic plus top-3 hybrid model	0.700	0.614	-19.0

V. Conclusions and Future Work

We sought a model with a small subset of variables and strong predictive qualities. Of the three methods approached, the hybrid of CART and multiple regression provided the best predictive models. This approach did not necessary resolve the issues that are inherent in the data, but it improved models in both predictive-ability and fit. The inclusion of more geomorphic predictors also improved the predictive capability of the models. However, this occurred at the expense of a small subset of variables. Choosing only a small set of top predictors provided simple models that lacked predictive-ability. Yet, when analysis was focused on the Coast Range, where the distribution of LD_{50} was less skewed, these models tended to perform reasonably when compared to complex models, considering that the models had only four variables.

The hybrid models improved both the predictive-ability and the fit of the models from which they originated. It is possible that a more discriminating variable selection would increase the fit of the models. Specifically, the geomorphic plus 3-tier hybrid models, while strong prediction models, had some issues of collinearity. Addressing this issue with forward stepwise regression, and eliminating variables with high collinearity to top predictors could solve the issue of over-fitting the model and non-normality of the residuals.

There were mixed results with the land cover metrics when comparing distance-weighting to buffering and the combination of the two. Models utilizing a large number of buffered variables, such as the subset 3 step model, did not perform as well as models with a large number of non-buffered metrics. However, in both top-in-tier variable selections for the entire data set and the Coast Range data, the top predictors chosen from the land cover tier were both buffered

and distance-weighted. Further analyses should be done to determine the differences in the predictive-abilities of these metrics.

These data are difficult to predict accurately. One key issue is the categorical nature of the data itself. LD_{50} , while we have treated it as continuous data, is truly ordinal data. This made it difficult to predict extreme substrate classes. It is possible that treating the data as ordinal and using logistic regression may provide a solution to this problem. However, difficulties may arise using logistic regression as there are twelve categories of the response.

There appears to be spatial correlation in the distribution of LD_{50} . It is possible, as with the Coast Range data that improved predictions can occur when the models focus on regions of similar ecosystems. It is also possible that by utilizing spatial techniques, some difficulties due to the coarse distribution of the data could be resolved.

VI. Acknowledgements

Thanks to:

Curt Seeliger of the EPA, Western Ecology Division, for his explanation of the collection and classification LD_{50} .

Christopher Cuhaciyian of Colorado State University, for explaining the metrics and their complexities, as well as sending data updates.

Keith Olson of Colorado State University, for clarifying several details about the calculation of tributary and power metrics.

Dr. Brian Bledsoe of Colorado State University, for his willingness to serve on my committee and for providing the problem and data access for the analyses in this paper.

Dr. N. Scott Urquhart of Colorado State University, for his willingness to serve on my committee and provide ideas for analysis.

Dr. Jennifer Hoeting of Colorado State University, for her willingness to serve as my advisor, her constant encouragement and guidance through this project, and her countless ideas for new approaches to analysis. This project would not have been possible without her expertise, positive outlook, and constant support.

References

- Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle", *Second International Symposium on Information Theory*, (ed. B. Petrox and F. Caski), pp. 267 – 281, Budapest: Akademiai Kiado.
- Allan, J.D. (1995). *Stream Ecology: Structure and Function of Running Waters*. London; New York: Chapman & Hall.
- Beven, K. and Kirkby, M.J. (1979), "A physically based, variable contributing area model of watershed hydrology, *Hydrological Sciences*, 24(1), 43:69.
- Brummer, C.J. and Montgomery, D.R. (2003), "Downstream Fining in Headwater Channels", *Water Resources Research*, 39, ESG1.1:ESG1.12.
- Buffington, J.M., Montgomery, D.R., and Greenberg, H.M. (2004), "Watershed Scale Availability of Salmonid Spawning Gravel as Influenced by Channel Type and Roughness in Mountain Catchments", *Canadian Journal of Fisheries and Aquatic Sciences*, 61, 2085:2096.
- Carbonneau, P.E., Lane, S.N., and Bergeron, N.E. (2004), "Catchment-scale mapping of surface grain size in gravel bed rivers using airborne digital imagery", *Water Resources Bulletin*, 40, W07202, doi:10.1029/2003WR002759.
- Cuhaciyan, C.O. (in prep), *Developing and Applying Multi-Scale, Geospatially Derived Metrics to Landscape Scale Hydrogeomorphic and Ecological Studies*. Doctoral Dissertation, Colorado State University.
- Eriksen, C. H. (1964), "Benthic Invertebrates and Some Substrate – Current – Oxygen Interrelationships", *Organism-Substrate Relationships in Streams*, Prymatuning Laboratory of Ecology, University of Pittsburgh.
- Givens, G.H. and Hoeting, J.A. (2005). *Computational Statistics*. New Jersey: Wiley.
- Jowett, I. G. (2003), "Hydraulic Constraints on Habitat Suitability for Benthic Invertebrates in Gravel-Bed Rivers", *River Research and Applications*, 19, 495:507.
- Julien, P.Y. (1995). *Erosion and Sedimentation*. Cambridge: Cambridge University Press.
- Kaufmann, P., Larsen, P., and Faustini, J. (2004). "Assessing Relative Bed Stability and Excess Fine Sediments in Streams". Presentation EMAP Symposium in Providence, R.I.
- Kiffney, P.M., Bull, J.P., and Feller, M.C. (2002), "Climatic and hydrologic variability in a coastal watershed of southwestern British Columbia", *Journal of the American Water Resources Association*, 38, 1437:1451.

- Maindonald, J. and Braun, J. (2003). *Data Analysis and Graphics Using R*. United Kingdom: Cambridge University Press.
- Minshall, G.W. (1984). "Aquatic Insect-Substratum Relationships", *The Ecology of Aquatic Insects*, (ed. V.H. Resh and D.M. Rosenberg), pp. 358-400. New York: Praeger.
- Montgomery, D.C., Peck, E.A., and Vining, G. G (2001). *Introduction to Linear Regression Analysis, 3rd edition*. New York: Wiley.
- Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1996). *Applied Linear Statistical Models* (4th edition). Massachusetts: WCB/McGraw-Hill.
- R Development Core Team (2005). *R: A Language and Environment for Statistical Programming Version 2.1.1*. ISBN 3-900051-07-0. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org>
- Sanborn, S.C. and Bledsoe, B.P. (in press), "Predicting streamflow regime metrics for ungauged streams in Colorado, Washington, and Oregon", *Journal of Hydrology*.
- Seeliger, Curt (2005). USEPA Western Ecology Division Contractor. Personal correspondence. August 2005.
- Stamp, J. D. (2004). "Associations between Stream Macroinvertebrates Communities and Surface Substrate Distributions". M.S. Thesis, Ohio University.
- USEPA (1998), EMAP-SW-Streams Field Operations and Methods for Measuring the Ecological Condition of Wadeable Streams Manual, EPA/620/R-94/004F, Section 7, Rev. 4, Sept. 1998, B-10.
- USEPA *Consolidated Assessment and Listing Methodology: Toward a Compendium of Best Practices*. (2002), retrieved October 20, 2005, from USEPA Office of Wetlands, Oceans, and Watersheds (OWOW) Web site: http://www.epa.gov/owow/monitoring/calm/calm_ch8.pdf
- USPEA *Level III Ecoregions* (2005), retrieved November 10, 2005, from USEPA Western Ecology Division (WED) Web site http://www.epa.gov/wed/pages/ecoregions/level_iii.htm
- Venables, W.N. and Ripley, B.D. (1999). *Modern Applied Statistics with S-Plus, 3rd edition*. New York: Springer.