Dissolved Oxygen Spatial Analysis
DEQ-Contract 22012

Technical Progress Report – Phase II
March 7, 2021

Prepared for:

Montana Department of Environmental Quality

Prepared by:

Great Lakes Environmental Center, Inc.
739 Hastings Street
Traverse City, Michigan 49686
Phone: (231) 941-2230
Facsimile: (231) 941-2240

Principal Contact:
Dale A. White
Principal Research Scientist
dwhite@glec.com
Phone: (614) 487-1040
Columbus (Ohio) Laboratory

Table of Contents

Abstract

The provision of an intense five-year monitoring survey (2013-2017) completed by the DEQ Water Quality Standards & Modeling Section meant several key questions could be addressed between downstream response and upstream or upland watershed characteristics. One of the key parameters measured in streams was continuous monitoring of dissolved oxygen (DO) that spanned a period of 2-3 weeks. Classification and Regression Trees (CART) were applied to explore the relationships of watershed stressors and mitigators to a response. The monitoring dataset composed of dissolved oxygen, water chemistry, and aquatic plant metrics for 73 stations located in eastern Montana, extending from the north at tributaries to the Missouri River and southward to the Wyoming state border.

The entire mass of datasets comprised three model structure categories – predictor variables, pure response variables that are affected by stressors or mitigators, and those that may serve a dual role and behave either as a predictor or response variables. Predictors variables included fixed-effects – watershed slope, land use/cover, and oil and gas well presence – and random-effects comprised of several meteorological drought indices. Dual and response variables were built from the point-based monitoring described above. A fourth determinant class were factor variables and included stream category, drainage area, and reference site designation.

Several regression tree models were built and interpreted, namely DO responses of mean delta, maximum delta, average daily minimum, and daily counts of exceeding a delta threshold. Other tree models included aquatic plants as a response, namely microalgae thickness and areal coverage of macrophytes. The depth of each regression tree result was determined by a complexity parameter optimized through k-fold cross-validation. Other model diagnostics were employed including the root-node error, cross-plots of model complexity vs. explained variance (or vs. cross-validation error), and a residuals analysis.

Low levels of watershed disturbance, as seen through natural land cover conditions, and the absence of prolonged drought conditions were the most consistent predictors for optimal DO conditions. Secondary predictors like conductivity (which correlates positively to anthropogenic impacts, with other predictors held constant), nutrient levels, drainage area, and water temperature were also important. Two DO tree models – mean delta and average minimum – offered the most guidance in criteria development. Tree models with plant-based responses behaved differently, as it appears to be that where there are macrophytes and/or filamentous algae, there is often microalgae.

However, by aggregating many decision trees, using methods like bagging, random forests, and boosting, the predictive performance of trees can be substantially improved. These tools could be explored in future work using this same set of developed predictors and robust stream monitoring dataset.

List of Tables

List of Figures

DISSOLVED OXYGEN SPATIAL ANALYSIS

The following Technical Progress Report summarizes completion of Phase II activities which include the development of key stressor and response variables for eastern Montana watersheds and the exploration of statistical relationships among them.

1. Approach

1.a. Stressor vs. Response

Montana DEQ and GLEC discussed several questions that could be answered by the provision of an intense five-year monitoring survey (2013-2017) completed by the DEQ Water Quality Standards & Modeling Section. One of the key parameters measured in streams was continuous monitoring of dissolved oxygen (DO). What is impressionable about these DO surveys is that they often spanned a period of 2-3 weeks. Here hourly observations are used to determine a minimum and maximum for a day and thus produce a daily range or delta; the collection of deltas was then compiled over this multi-week period.

DO is an important component of healthy aquatic systems as most animal organisms require it for maintenance, growth, and reproduction. Other stressor-response variables were part of the DO dataset – several aquatic plant metrics and water chemistry. From the discussions with DEQ and the original request for work, the goal was to explore which factors influence variability in dissolved oxygen, which factors co-occur to produce a specific scenario, and how do other factors respond to stresses or mitigating conditions in their upstream watersheds.

1.b. Statistical Learning – A New Form of Exploring Relationships

GLEC proposed the use of Classification and Regression Trees (or CART) as a methodology for exploring the relationships of watershed stressors and mitigators to a response. Classification trees have a categorical target or response variable whereas regression trees have a continuous response variable. Specifically, the difficulty in uncovering the interactions between stressor variables and their impact on a response variable makes CART an appropriate technique. Nisbet et al. (2009) suggest one of the key advantages of CART is to uncover these complex interdependencies. Faraway (2016) frames tree models as finding interactions – split on one variable and then split on another variable within the partitions of the first variable, an interaction is found between these two variables.

CART methodology was introduced in 1984 (in two California universities) by Breiman, Friedman, Ohlsen, and Stone. The foundational text on this subject was written by these authors (Brieman et al. 1984) though it is quite mathematical in its discussion and offers little practical or application advice. However, several narratives are more practical in their exposition on running and interpreting tree models including advice on running the primary software in R (Therneau and Atkinson 2019), and several useful documents by Milborrow on plotting results (2011), regression surfaces (2018), and residuals (2020).

Helsel (2019) in his presentation "40 Years of Water Quality Statistics" suggest regression trees as one modern approach to examining relationships in water quality. He expounds on the advantages of using regression tree methods, namely:

1. Makes use of machine learning tool to classify data into groups by relating the target variable to cutoffs of explanatory variables.
2. The method is flexible because there are no assumptions of linearity or normality.
3. Data at the 'high end' do not affect relationships at the 'low end'; thus, they are not as restricted as are traditional regression methods.
4. Evaluation of success is done by cross-validation – the percent of correct predictions of categories for the response variables – rather than by p-values.
5. Predictions are made for individual observations rather than the mean of observations (as done in regression).

Hence the key message from Helsel (2019) is that regression trees are non-parametric and not significantly impacted by outliers. Berk (2020) suggests that CART is similar to a stagewise regression with predictors that are indicator variables. Concurrently, tree methodology was developed in machine learning starting in the 1970s (Quinlan 1993).

Pursuing the fundamental questions defined above, GLEC makes use of regression trees, and to some extent on classification trees for answers. The Technical Progress report for Phase II begins with specific narrative on the development of predictor, dual predictor-response, and response variables (Section 2). The dataset description is followed by methods (Section 3), and then a presentation and discussion or interpretation of results (Section 4). The final section (Section 5) concludes with overall success, limitations, and suggestions for future work. An appendix follows Section 5 which discusses a new approach for understanding water quality relationships – Bayesian Network Modeling – and then an inventory of all provided R code and datafiles. All code and datafiles will be uploaded to the Montana DEQ server.

2. Description of Dataset

2a. Study Area

A rich and robust dataset composed of dissolved oxygen, water chemistry, and aquatic plant metrics was sampled by Montana DEQ for the period 2013 to 2017. Sampling stations (73 total) were located in eastern Montana, extending from the north at tributaries to the Missouri River and southward to the Wyoming state border (Figure 2.1). These point-based data assemblages were provided to GLEC for subsequent statistical modeling. Drainage basin boundaries were derived by GLEC from digital elevation models (see Phase I report, p 3). It is within these polygon boundaries that GLEC then developed a set of stressor and mitigator variables that likely bear some relationship to the observed variability in dissolved oxygen (DO), water chemistry, and aquatic plants. Some of the drainage basins extend beyond Montana into small sections of southwestern North Dakota, northwestern South Dakota, and northeastern Wyoming (Figure 2.1).

GLEC characterized the entire mass of datasets into three model structure categories – predictor variables that behave as a stressor or mitigator to the aquatic system, pure response variables that are affected by stressors or mitigators, and those that may serve a dual role and behave either as a predictor or response variable (Figure 2.2). A fourth determinant class that may impact the response are defined as factors (or indicator variables) and include stream category, drainage area, and stream reference site (Figure 2.2). A detailed description of the variables that comprise each of these four classes follow. Appendix B.1 is included as a compact reference (i.e., data dictionary) for all variables used in this study.

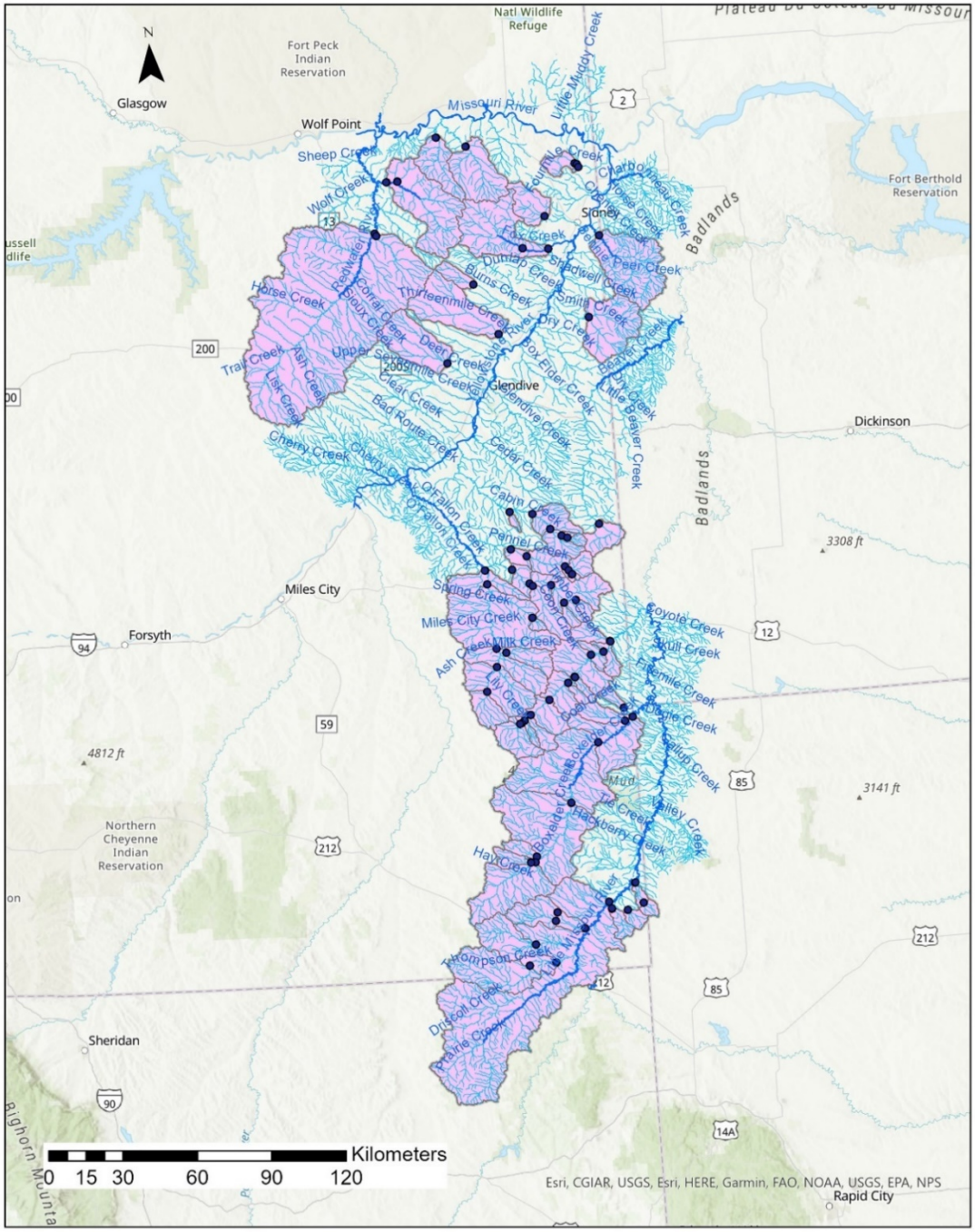


Figure 2.1. Distribution of sampling stations (black dot) and their corresponding drainage basins (pink) used in statistical modeling for this study. Major stream segments in each basin are labeled. Note, several of the stations occur on the same drainage pathway and thus their corresponding drainage basins are nested. Polygon boundaries shown here, in essence, overlap for these nested systems.

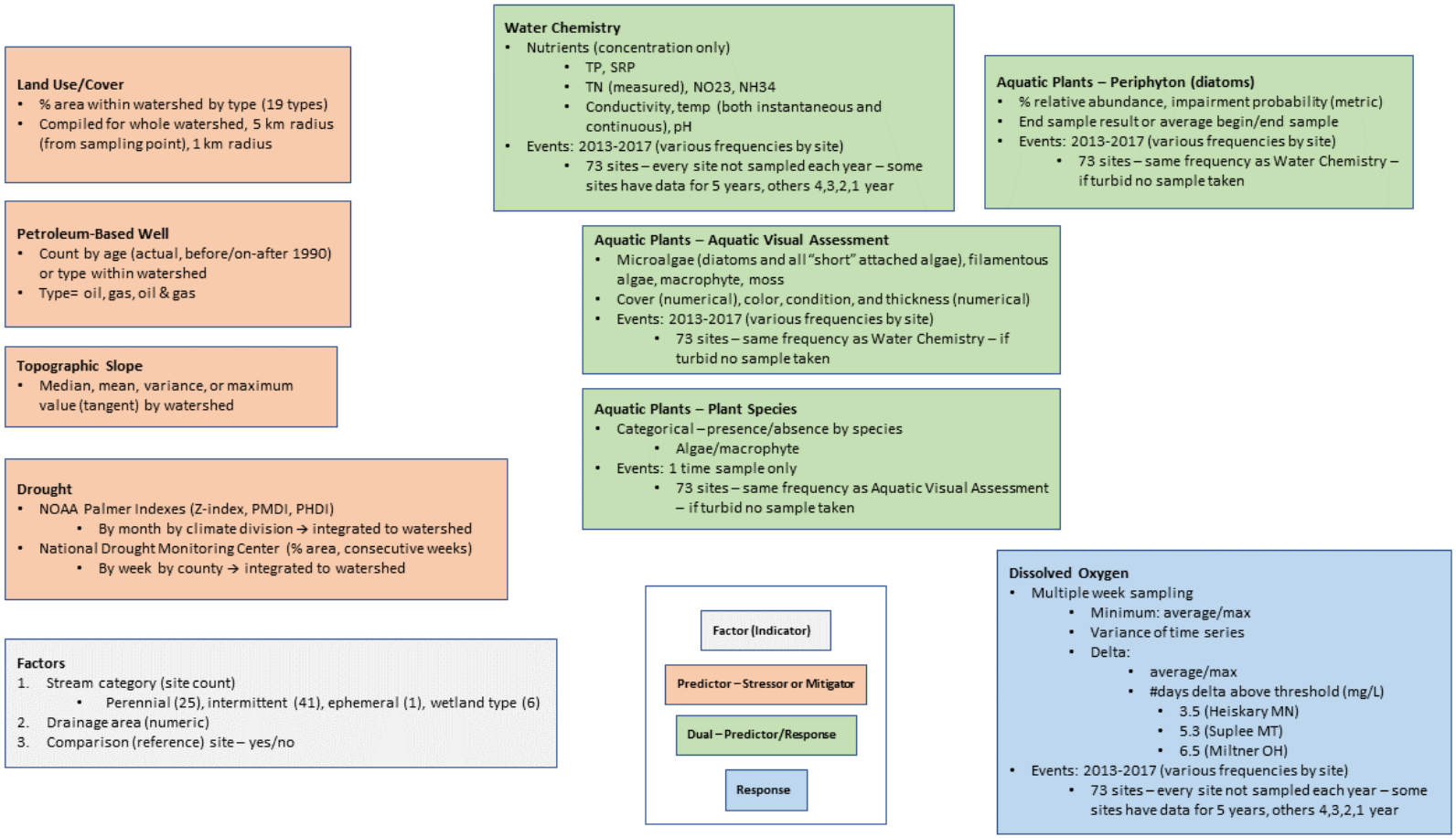


Figure 2.2. Conceptual model of eastern Montana dataset assemblages showing predictor variables that may stress or mitigate the observed response, pure response variables (namely those derived from continuous dissolved oxygen measurements), dual variables that can behave as either a predictor or response, and factors (or indicator variables) that may impact responses.

2.b. Predictor Variables

Predictor variables comprise both stressor and non-stressor (mitigator) variables, with the former behaving as a disturbance that likely causes degradation in a water quality response – e.g. lower dissolved oxygen, enhanced nutrient concentrations that lead to eutrophication, or the deleterious aspects of aquatic plant metabolism and growth. Non-stressor variables support the beneficial aspects of water quality response. The predictor variables developed in this study are a combination of fixed and random effects and exist in these classes: land use/cover (fixed effect), petroleum-based wells (fixed), topographic slope (fixed), and drought (variable effect). The fixed-effect predictors do vary over space (the varied geography of eastern Montana streams and their watersheds) but are fixed in time. The sole random-effect predictor (a series of meteorological drought indices), like the dissolved oxygen response and the dual predictor-response variables, vary *both* in space and time.

2.b.i. Fixed-Effect Predictors (Land Use/Cover, Oil and Gas Wells, and Topographic Slope)

Two sources of land cover and use information were used to build these predictors – the Natural Heritage Program for Montana (NHP) and the National Land Cover Dataset (NLCD) of the US Geological Survey. Both datasets are 2015-2016 vintage with more detailed thematic classification of vegetative land cover provided by NHP; noting that the DO, water chemistry, and plant measurements were collected 2013-2017, with the land cover predictor falling within the midpoint of this time window. More details on the creation of the landcover dataset are shown in the Phase 1 report (p. 5-6). However, a revised approach for computing percent land cover within nested watershed polygons was made with the `Tabulate Area` command in the ESRI ArcGIS Pro application.

There are 57 land cover classes distributed in the eastern Montana study area. Rather than build predictor variables for each individual land use and cover class (e.g., Great Plains Badlands), only the dominant classes were kept. For whole watershed compilations, 98 percent of the watershed area was represented by 14 classes (with a minimum representation of 94 percent). These 14 land cover classes were subsequently aggregated into two major classes – natural [natws][1] and disturbed [distws] – from which predictor variables were built. The natural major class consists of individual classes such as Great Plains Badlands, Great Plains Ponderosa Pine Woodland and Savanna, and Great Plains Wooded Draw and Ravine. The disturbed class consists of individual classes such as Pasture/Hay, Cultivated Crops, and Introduced Upland Vegetation – Annual and Biennial Forbland. Both major classes are exhaustive for each watershed (i.e., natural plus disturbed classes sum to nearly 100 percent[2]) so we would expect to be the relationship to purely inversely linear (Figure 2.3).

[1] All model variables described in Section 2 are identified with brackets [ ]. See Appendix B.1 for definitions of both fixed- and random-effect predictor variables.

[2] Any fragmentary land cover class percentage (< 0.1% of total watershed area) was excluded for efficiency in compilation; hence, in some watersheds the sum of natural plus disturbed land cover is less than 100% of total area.

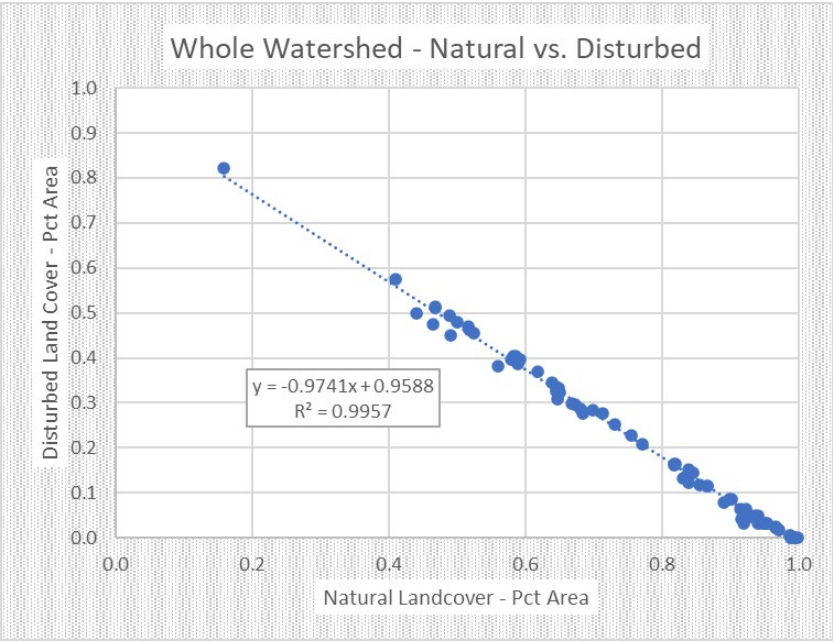


Figure 2.3. Relationship of percent area of natural land cover to that of disturbed land cover within each watershed. Each data point represents the watershed cover for an individual station watershed. A linear equation was fitted to the distribution of points showing a slope nearly equal to -1 and a near complete percent of explained variation.

Besides whole watershed land cover, a second spatial scale, the *near-field* watershed land cover, was developed into a new predictor. The near-field composition is based on a 1000-meter circular buffer taken from a station location as the center that intersects and is within the watershed boundary[3]. In the near-field compilations, 99 percent of the watershed area was represented by 19 land cover classes (with a minimum representation of 92 percent). At the near-field scale both natural [natnf] and disturbed [distnf] predictors were built. Note, the near-field land cover predictors are not necessarily co-linear with the watershed scale predictors (Figure 2.4). In all, four predictors –natural watershed and near-field and disturbed watershed and near-field – were built to represent the mitigation or stress due to the percentage of natural or disturbed land use and cover. The range and average of percent area (natural and disturbed land cover) over the 73 watersheds is shown in Table 2.1.

[3] An improvement in a future rendition of the statistical modeling effort would be to replace the circular buffer with a more precise near-field buffer – represented as the smaller intervening sub-watershed defined by 1000 m traversal of the mainstem stream within each watershed.

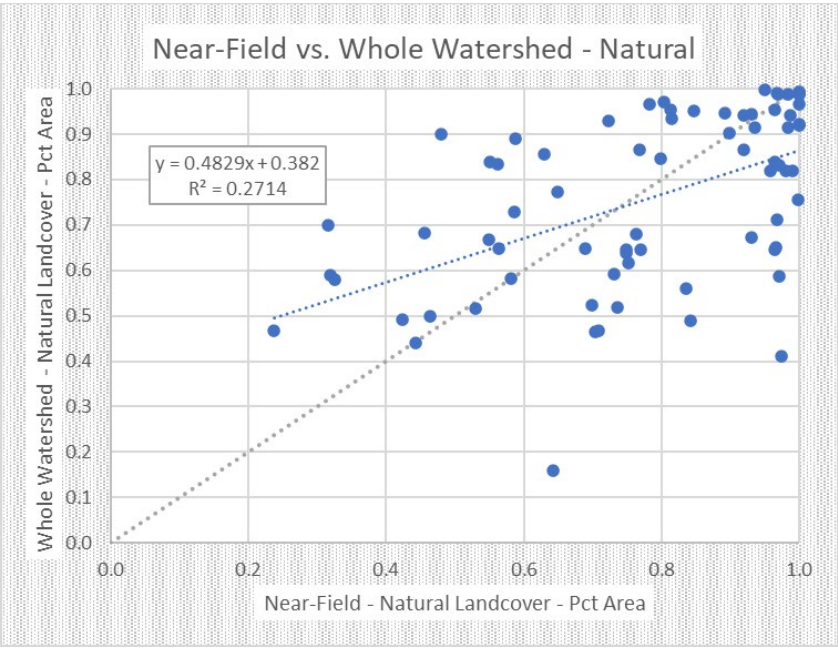


Figure 2.4. Relationship of near-field natural land cover (as a percent) to the same for whole watershed natural land cover. A linear equation (blue) was fit to the distribution of points and 1:1 line (black)is also plotted. As expected, the relationship is a direct one, but not co-linear, with a low R^2 (0.27) and a slope of 0.48. Most agreement occurs in the higher percentages of natural land cover.

Another set of predictors reflects the distribution of oil and gas wells within each watershed of the study area. The well information was processed from a compilation of various master datasets provided by Montana, Wyoming, and North Dakota state agencies. A more detailed description of this compilation is found in the Phase 1 report (p. 8-12). Counts of total number of wells [wells] and number of "old" wells [wellso] (those wells developed before 1990) were made; the `Spatial Join` technique in the ESRI ArcGIS Pro[4] application was used to compile well count by watershed. Montana DEQ suggested that historical wells developed before 1990 may have a greater disturbance to downstream water quality so GLEC accounted for this subcategory. In addition, the distribution of well counts were assigned ranks (0 through 4 for total wells and 0 – 3 for old wells) based on natural breaks in the distribution of counts. The range and average of well counts over the 73 watersheds is shown in Table 2.1.

The final fixed-effect predictor is topographic slope of the watershed[5]. Slope is derived from the topographic (digital elevation model) layer using a neighborhood analysis (3 x 3 cell window).

[4] Both the `Spatial Join` and `Tabulate Area` commands in ESRI ArcGIS Pro work on nested polygons; nested polygons are not honored for similar command names in the ESRI ArcMap application, a predecessor to ArcGIS Pro.
[5] An improvement in a future rendition of the statistical modeling effort would be to compute longitudinal slope of the stream channel, perhaps becoming a more critical predictor variable. Shifting to a higher resolution digital elevation model would be required (e.g., 10 m or 3 m grid resolution) to accurately capture the channel slope. Alternatively, the

Specific details on GLEC's production of this theme are shown in the Phase 1 report (p 5). All slope predictor variables (4 total) are in percent slope (tangent ·100 where tangent = topographic relief / ground distance) and represent summaries of the individual cell estimates within each watershed. Here, the `Tabulate Area` command in the ESRI ArcGIS Pro application was used to compile the average [xslope], median [medslope], maximum [maxslope], and standard deviation [devslope] of slope for each watershed. The range and average of slope over the 73 watersheds is shown in Table 2.1.

Table 2.1. Distributional statistics of fixed-effect predictors on a watershed basis.

Parameter Name (units)	Minimum	Maximum	Average
Drainage area (sq.mi)	1.62	1429.7	196.6
Wells – total (count)	0	235	80.3
Wells – old (count)	0	235	31.5
Slope - average (% slope = tangent·100) (degrees)	1.81 (1.04)	10.1 (5.8)	4.4 (2.5)
Slope - maximum (% slope = tangent·100) (degrees)	17.1 (9.7)	113.2 (48.5)	65.6 (33.3)
Natural land cover – watershed (% area)	15.9	99.8	73.2
Disturbed land cover – watershed (% area)	0	82.2	24.4
Natural land cover – near-field (% area)	23.7	100	75
Disturbed land cover – near-field (% area)	0	76.3	24.2

2.b.ii. Random-Effect Predictor (Drought)

In response to suggestions from Montana DEQ on the importance of drought measures on downstream water quality, GLEC explored several datasets provided by the NOAA National Centers for Environmental Information[6] (NCEI), and then following the suggestions in Heim (2002), explored indices developed by the National Drought Monitoring Center (NDMC). The drought indices are random-effect predictor variables exhibiting variation in space (across eastern Montana) and in time. Drought conditions may vary daily but are computed on a weekly or monthly basis.

The various drought indices that GLEC explored for inclusion in the statistical model are briefly summarized here. Three indices are produced by NOAA (Table 2.2): Z-Index, Palmer Meteorological Drought Index (PMDI), and Palmer Hydrological Drought Index (PHDI)[7]. The

recent availability of the NHDPlus HR (high resolution) hydrographic dataset (from the US Geological Survey) makes slope of channel segment immediately available (from the NHDPlusFlowlineVAA attribute table).

[6] Formerly called the NOAA National Climatic Data Center (NCDC).

[7] Datasets are downloadable from the NOAA at – https://www.ncdc.noaa.gov/temp-and-precip/drought/nadm/indices . NOAA NCEI is currently revamping their web access to drought data and this reference may change in the near future.

PMDI is the most typically used index and often referred simply as the Palmer Drought Index. Each offers a slightly different characterization of drought, with the Z-index being the most dissimilar. The Z-Index is calculated by NOAA from a combination of precipitation, temperature, and soil moisture data, and corresponds to monthly drought conditions with no memory to previous monthly deficits or surpluses. The Z-Index, thus, is a measure of short-term agricultural drought. The PMDI is based on drought-inducing atmospheric circulation patterns. It is a cumulative measure so the PMDI is based on current month and weather patterns of previous months. The PHDI characterizes the hydrological impacts of drought (e.g., reservoir and groundwater levels) and exhibits a longer lag, compared to the PMDI, in drought development and recovery. The time variability for each of three indices is shown in an example for a particular climate division later in this section (Figure 2.7).

The range of values for all three NOAA drought indices varies from < -5.0 (extremely dry) to > +5.0 (extremely wet) and each is compiled by selected stations from the NOAA network or by climate division on a monthly scale (Figure 2.6). After consultation with Montana DEQ, GLEC chose to work with *climate division-based* NOAA indices as these spatial aggregations better represent the distribution of the 73 watersheds in eastern Montana. However, additional drought indices were explored, that add to the NOAA-based indices, and they are discussed in the following paragraphs.

Table 2.2. Summary of drought indices extracted for predictors in the study area for the period 2013-2107. Data sources include NOAA-NCEI (National Centers for Environmental Information) and NDMC (National Drought Mitigation Center).

Index	Source	Time Interval	Spatial Unit
Palmer Z-Index	NOAA-NCEI	Monthly	Climate Division
Palmer Meteorological Drought Index (PMDI)	NOAA-NCEI	Monthly	Climate Division
Palmer Hydrological Drought Index (PHDI)	NOAA-NCEI	Monthly	Climate Division
% Area in Each (Categorical) Drought Level (D0 – D4)	NDMC	Weekly	County
Number of Consecutive Weeks in Each (Categorical) Drought Level (D0 – D4)	NDMC	Weekly	County

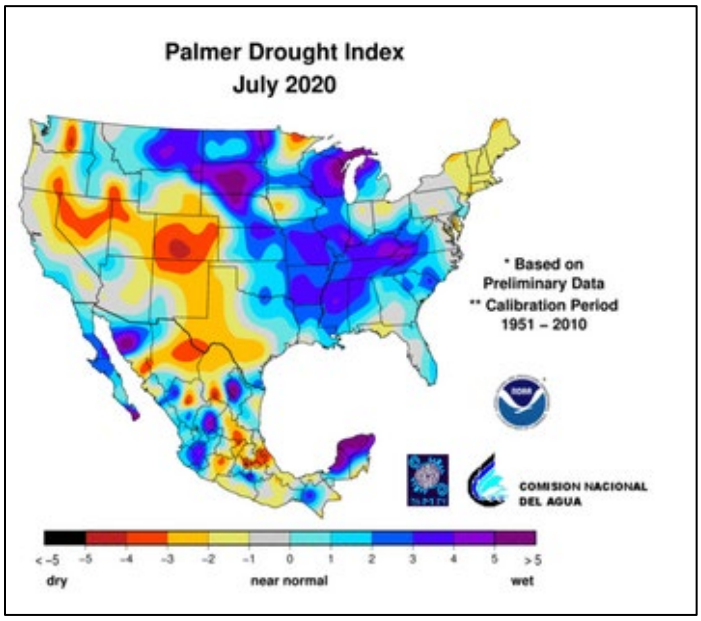


Figure 2.5. Map showing climate division-based Palmer Meteorological Drought Index (PMDI) for the United States and Mexico, and showing extremely wet conditions for eastern Montana for July 2020.

Agencies within NOAA and the US Department of Agricultural (USDA) teamed with the NDMC to produce a weekly US Drought Monitor (DM) product that incorporates climatic data and professional input from all levels (Svoboda 2000). Since no single definition of drought works in all circumstance, the DM authors rely on the analyses of several key indices and ancillary indicators from different agencies to create a final index (Heim 2002). The key parameters(Table 2.3) include the Palmer Drought Index (PMDI), the Crop Moisture Index, soil moisture model percentiles, daily streamflow percentiles, percent of normal precipitation, topsoil moisture (percent short and very short) generated by the USDA, and a satellite-based Vegetation Health Index. The ancillary indicators include the Surface Water Supply Index, the Keetch–Byram Drought Index, the Standardized Precipitation Index, snowpack conditions, reservoir levels, groundwater levels determined from wells, USDA reported crop status, and direct in situ soil moisture measurements.

GLEC extracted two types of drought indicators from the DM[8] – 1) the percent area of a specific drought-type and 2) the number of consecutive weeks of a specific drought-type existing within each county (Table 2.2).

[8] Datasets are downloadable from the US Drought Monitor at – https://droughtmonitor.unl.edu/nadm/Home.aspx

Table 2.3. Key parameters (5) that comprise the US Drought Monitor index D0 through D4 along with possible impacts. Each parameter has a range that corresponds to a particular drought intensity. The NOAA-based PMDI discussed earlier is shown here as the PDSI.

			Ranges				
Category	Description	Possible Impacts	Palmer Drought Severity Index (PDSI)	CPC Soil Moisture Model (Percentiles)	USGS Weekly Streamflow (Percentiles)	Standardized Precipitation Index (SPI)	Objective Drought Indicator Blends (Percentiles)
D0	Abnormally Dry	Going into drought: ▪ short-term dryness slowing planting, growth of crops or pastures Coming out of drought: ▪ some lingering water deficits ▪ pastures or crops not fully recovered	-1.0 to -1.9	21 to 30	21 to 30	-0.5 to -0.7	21 to 30
D1	Moderate Drought	▪ Some damage to crops, pastures ▪ Streams, reservoirs, or wells low, some water shortages developing or imminent ▪ Voluntary water-use restrictions requested	-2.0 to -2.9	11 to 20	11 to 20	-0.8 to -1.2	11 to 20
D2	Severe Drought	▪ Crop or pasture losses likely ▪ Water shortages common ▪ Water restrictions imposed	-3.0 to -3.9	6 to 10	6 to 10	-1.3 to -1.5	6 to 10
D3	Extreme Drought	▪ Major crop/pasture losses ▪ Widespread water shortages or restrictions	-4.0 to -4.9	3 to 5	3 to 5	-1.6 to -1.9	3 to 5
D4	Exceptional Drought	▪ Exceptional and widespread crop/pasture losses ▪ Shortages of water in reservoirs, streams, and wells creating water emergencies	-5.0 or less	0 to 2	0 to 2	-2.0 or less	0 to 2

One would expect drought and water quality to be closely related and Mosley (2015) reviewed this relationship. The obvious impact is reduction in streamflow (discharge) and the corresponding increase in ambient air temperature. Both of these effects propagate into the stream ecosystem. Stream temperature will increase which increases algal metabolism and production and thereby increase the diel range (herein delta) of DO. Further, the saturation level of DO declines with increasing stream temperature and thus reduces the minimum DO each day[9]. With lower streamflow, dilution is reduced causing an increase in stream salinity or the concentration (and thus impact) of any potential stressor from a point source. For example, nutrient concentrations from point sources such as a wastewater discharge or the in-stream presence of livestock promote algal growth and increase delta DO. Reduced streamflow from drought also reduces instream re-aeration. However, drought will also reduce watershed loading to streams, and thus reduce the import of nutrients and sediment (turbidity) from uplands. Other studies that explored the relationship between drought and water quality include Ryberg et al. (2018) using PHDI and total phosphorus load, and Ahmadi and Moradkhani (2018) applying regional ordinal measures of drought to stream DO, temperature, and turbidity response.

Prior to developing predictor variables from NOAA and NDMC data sources, GLEC explored the relationship between the drought indices and the Montana DEQ measured dissolved oxygen and temperature. GLEC found interestingly strong relationships between drought and water quality for

[9] A decline in saturation DO concentration may also decrease the DO daily maximum and thus not affect the diel range. However, increased algal productivity may compensate for the decline in saturation DO and place the system in super-saturated conditions.

Pennel Creek watershed (Montana) (Figures 2.7a-b and 2.8a-c). In the NOAA indices (Figures 2.7a-b), one can observe the smooth time distributions of the PMDI and PHDI relative to the flashier, shorter-term Z-index. PMDI and PHDI practically mimic each other except the PHDI tends to have a longer lag, as expected. In examining DO for Pennel Creek (Figure 2.7a), one observes that DO delta increases and DO minimum decreases with increasing drought severity. The relationship is especially true as drought persists into 2017 and DO delta further increases while DO minimum further decreases. In the earlier part of the study period (2013) when conditions were wetter (positive drought indices), one finds that DO delta is quite small while DO minimums are much higher. In Figure 2.7b, one finds for Pennel Creek that stream temperatures increase measurably with increasing drought (from 2015-2017) whereas temperature was remarkably mitigated in the wetter 2013 period. Note, that temperature measurements shown here were those recorded from the continuous DO monitor sampled by Montana DEQ. As discussed above in the review by Mosley (2015), increasing stream temperatures negatively impacts healthy DO conditions for stream ecosystems.

Similar patterns were found for the NDMC drought indices (Figures 2.8a-c). GLEC introduced a new parameter, as presented by the NDMC, a percent area index integrated across all drought intensity levels (D0-D4) (Table 2.3) and named the Drought Severity and Coverage Index (or DSCI). The DSCI is a weighted sum, for a given week in a given county, expressed as: `DSCI = 1(D0)+2(D1)+3(D2)+4(D3)+5(D4)`, where the higher intensity drought levels are given increasingly higher weight. In Figures 2.8a-b, one observes a similar response of DO delta and minimum to the persistence of drought and increasing intensity. The persistence of drought was further represented by the number of consecutive weeks at a current drought intensity (D0-D4) (Figure 2.8c). One observes that a given area does not experience a higher intensity drought (D3-D4) until some duration of lower intensity drought (D0-D1) exists. As witnessed in Figures 3a-b with percent area of drought, as number of consecutive weeks increases and drought intensity increases, one observes an increase in DO delta and decrease in DO minimum (Figure 2.8c). The Pennel Creek watershed never experienced D4 drought but does experience D3 intensity toward the end of the study period (2017) and it was here one finds exceptionally high DO delta and quite low DO minimum.

Given the promise of drought occurrence in explaining DO variation, GLEC extracted two geo-datasets (polygon features) for the study area that correspond to the indices described above and in Table 2.2 – NOAA climate divisions and county boundaries. Montana DEQ encouraged GLEC to employ readily obtainable drought data; thus, the study area is best represented by 5 NOAA climate divisions (2406, 2407, 3204, 3901, and 4805) and 14 county areas (9 of which are in Montana). GLEC developed nine drought random-effect predictors (Table 2.4).

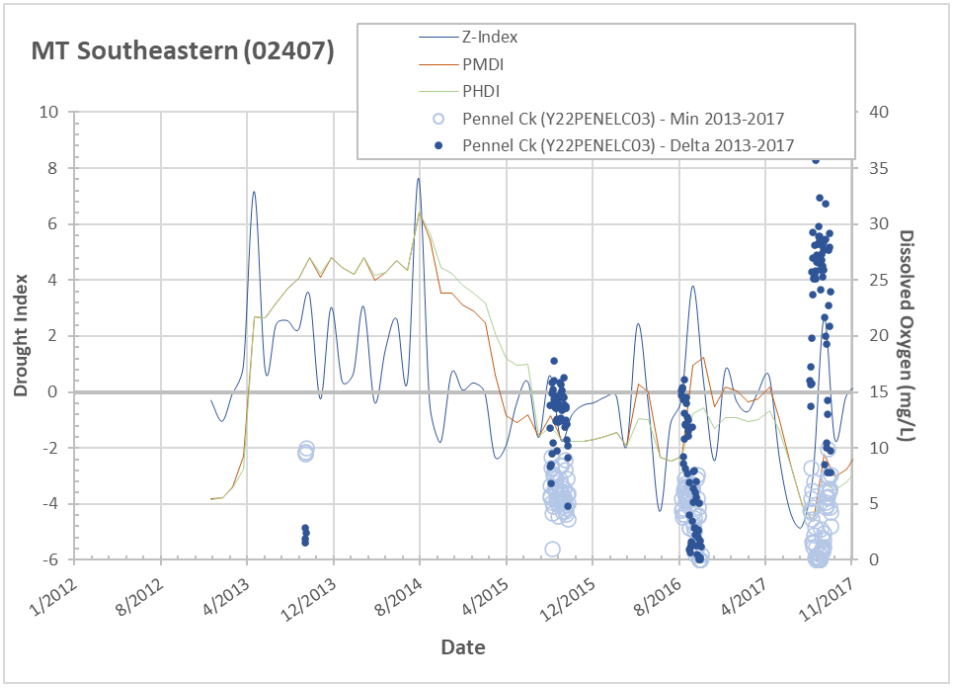
Table 2.4. Random-effect predictor variables to characterize drought intensity and persistence developed for the eastern Montana dissolved oxygen study.

Variable Name	Description	Agency	Time Representation
Zindex	Z-Index	NOAA	Monthly
PMDI	Palmer Meteorological Drought Index	NOAA	Monthly
PHDI	Palmer Hydrological Drought Index	NOAA	Monthly
DSCI	Drought Severity and Cover Index – weighted sum of D0-D4	NDMC	Weekly
DSCIt	transformed (square-root) DSCI	NDMC	Weekly
Dzero	# consecutive weeks at drought severity level D0	NDMC	Weekly
Done	# consecutive weeks at drought severity level D0	NDMC	Weekly
Dtwo	# consecutive weeks at drought severity level D0	NDMC	Weekly
Dthree	# consecutive weeks at drought severity level D3	NDMC	Weekly
Dfour	# consecutive weeks at drought severity level D4	NDMC	Weekly

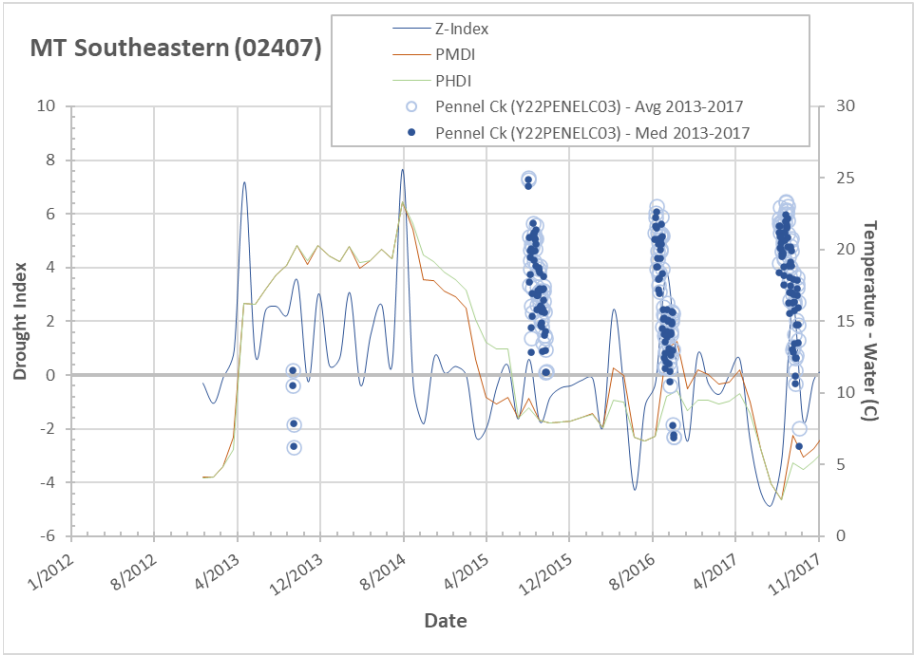
The drought measures shown in Table 2.4 are a true integration of space and time phenomena since they were aggregated over different areal extents and time periods relative to the sampling stations and their associated drainage areas. It was imperative that the spatial extent of the drainage basin, and not the location of the sampling station, be considered when identifying the specific climate division- or county-based drought measure. Thus, each of these drought predictors represent spatial integrations of the original geographical unit (i.e., climate division or county) to the drainage basin polygon. The integrated drought measure was calculated as a weighted sum where the weights represent the percent area of the climate division or county existing within a drainage basin polygon. This integration can be represented, using PHDI as an example predictor, in equation form as:

$$PHDI_L = [\ \%\ area_L \text{ in } CD_a \times PHDI_{CDa}\ ] + [\ \%\ area_L \text{ in } CD_b \times PHDI_{CDb}\ ] + [\ \%\ area_L \text{ in } CD_c \times PHDI_{CDc}\ ]$$

Where $PHDI_L$ is the weighted drought index for a specific watershed L, and CD a, b, and c are three NOAA climate divisions that intersect the boundary of watershed L. Further, $\%\ area_L$ in CD_a is the percent of the watershed L total area in CD_a and so on for CD_b and CD_c. Also note that $\sum\ \%\ area_{L[a,b,c]} = 100$.

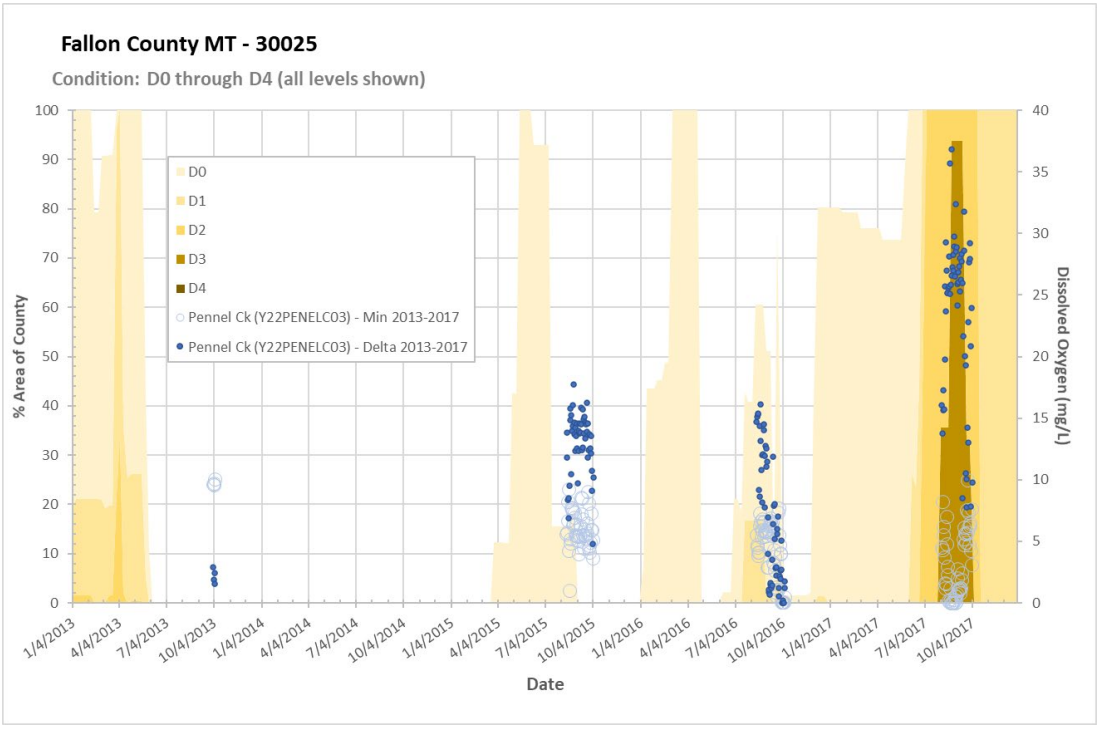


(a)

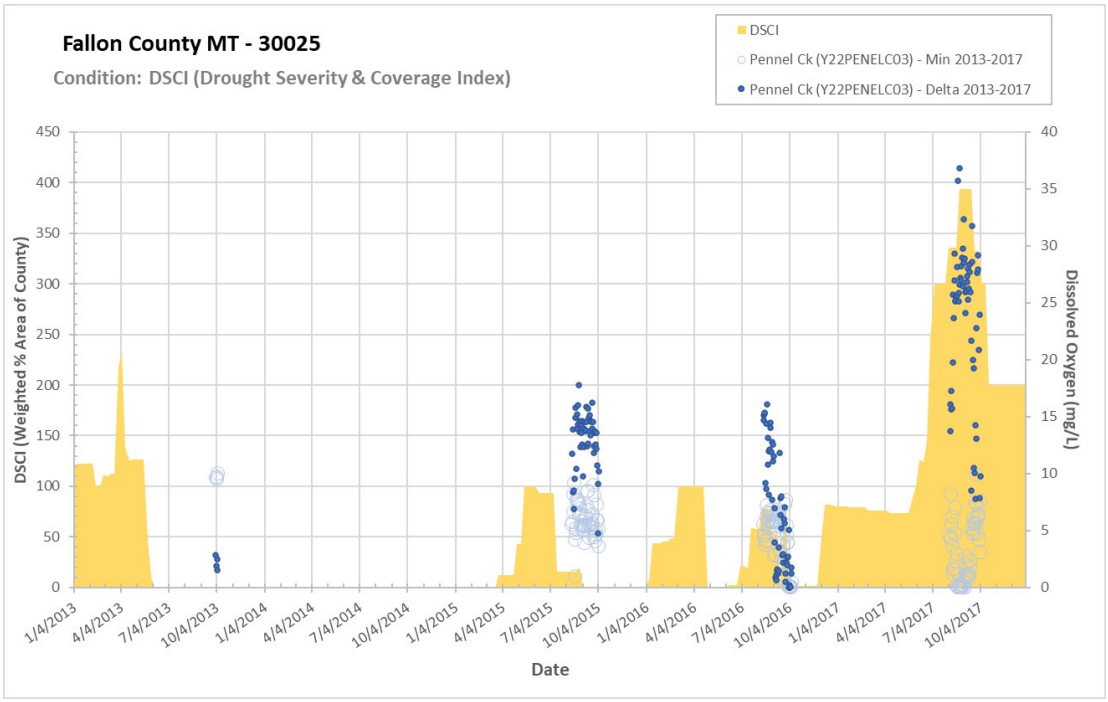


(b)

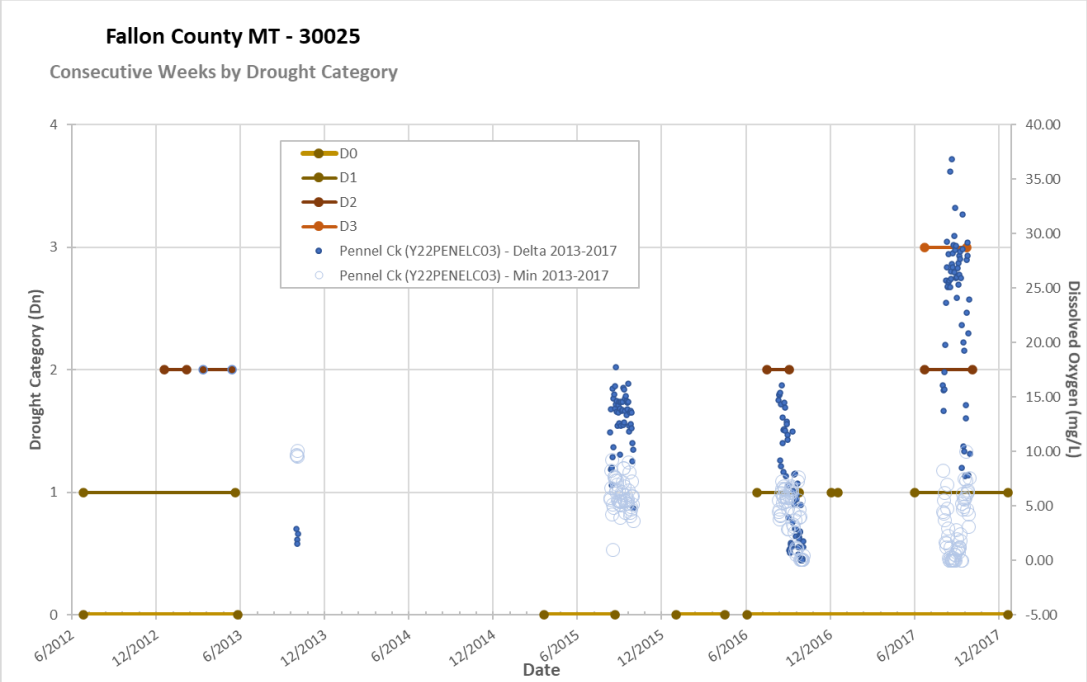
Figure 2.6. Plot showing the time variation of monthly NOAA drought indices (Z-index, PMDI, and PHDI) for the Southeastern MT (02407) climate division along with the daily variation of (a) dissolved oxygen, both minimum (open circles) and delta (solid circles), and (b) stream temperature, both average (open circles) and median (solid circles), for Pennel Creek (Y22PENELC03, P-4) for the 2013-2017 period.


Fallon County MT - 30025
Condition: D0 through D4 (all levels shown)
D0
D1
D2
D3
D4
Pennel Ck (Y22PENELC03) - Min 2013-2017
Pennel Ck (Y22PENELC03) - Delta 2013-2017
% Area of County
Dissolved Oxygen (mg/L)
Date

(a)


Fallon County MT - 30025
Condition: DSCI (Drought Severity & Coverage Index)
DSCI
Pennel Ck (Y22PENELC03) - Min 2013-2017
Pennel Ck (Y22PENELC03) - Delta 2013-2017
DSCI (Weighted % Area of County)
Dissolved Oxygen (mg/L)
Date

(b)



(c)

Figure 2.8. Plot showing the time variation of weekly NDMC drought indices (a) percent area of county in a given drought intensity (D0-D4), (b) weighted percent area of county (DSCI), and (c) number of consecutive weeks in D0-D4 for Fallon County MT (30025) along with the daily variation of dissolved oxygen, both minimum (open circles) and delta (solid circles), for Pennel Creek (Y22PENELC03, P-4) from June 2012 to December 2017.

2.b.iii. Factor Variables

Three factor variables were created or otherwise made available to discriminate major classes of station sampling or watershed variation – stream category [streamcat][10], comparison site [compsite], and drainage area [DA]. Each of these variables is a fixed (time) effect (Figure 2.2). The latter variable is continuous in nature (as opposed to categorical as is typical for a factor variable). The drainage area (sq. mi) was extracted from the drainage basin boundary layer and represents the *total* upstream drainage area from each sampling station to its drainage divide. Stream category assignments reflect one of four types (with number of sites in parenthesis) – perennial (25), intermittent (41), ephemeral (1), and wetland (6) – and were assigned by Montana DEQ prior to this study. The comparison sites (24 of 73 total sites), considered as an *expanded* reference site typing, are based on the absence of any local disturbance issues and a dominance of natural land cover (more than 75 percent by area) in its drainage basin; all based on Montana DEQ direction. The watershed natural land cover percentage [natws] (Section 2.b.i) was used for this computation. Comparison sites also include five sites with the official Montana DEQ *reference site* designation (Suplee et al., 2005). All remaining sites (44 total) were deemed ordinary sites.

[10] See Appendix B.1 for definitions of factor variables.

2.c. Dual Predictor and Response Variables

Dual predictor and response variables affect the DO response in streams and are also affected by the fixed- and random-effect predictor variables described above. These variables include water chemistry (both chemical concentrations and field parameters) and aquatic plant measurements (Figure 2.2). All of the measurements characterized as dual variables were collected by Montana DEQ for the 2013-2017 period. Not every site (of the 73 stations) was sampled in each year of the period of record. Four sites were sampled in each of the five years, while three sites had 4 years of data, nine sites (3 years), 16 sites (2 years), and 41 sites (1 year).

The water chemistry dataset consists of the nutrient parameters – total ammonia [NH3], nitrite and nitrate [Nox][11], total nitrogen (measured *in situ*) [TN], soluble reactive phosphorus (orthophosphate) [OP], and total phosphorus [TP]. All parameters are concentration (mg/L) in their base element (N or P). Four observations had missing values for these parameters. [TN] and [TP] were extended to [TNe] and [Tpe] to include estimates for missing values by substitution with the median of the entire period of record for all stations. While some primitive stream discharge measurements were made by Montana DEQ during the study period, none were reliable or extensive enough to develop corresponding loading values for the nutrient parameters.

Alongside the water chemistry sampling were measurements of aquatic field parameters – pH [pH], specific conductance (µS/cm) [SC], water temperature (instantaneous) (°C), and barometric pressure (mmHg) [BP]. Water temperature was also measured with DO under continuous monitor (discussed in the next section). GLEC chose to compute summary water temperature measures from the continuous monitoring as it would be more integrative of ambient conditions. Hence, two water temperature model variables were developed – the median [medt] and maximum [maxt] of the weekly suite of daily average temperatures.

Analytical results for some of the water chemistry parameters were below their detection limit. To summarize, nitrite and nitrate had 108 non-detect results (36.7% of the total number of observations), total ammonia 66 results (22.4%), soluble reactive phosphorus 28 results (9.9%), field parameters 4 results (2.4%), and total nitrogen 2 results (0.7%). GLEC initially pursued a more advanced treatment of handling non-detect observations by applying regression-on-order statistics (ROS) to impute the missing values for a few sites and a few parameters. Helsel (2012) prescribes robust ROS for imputation in monitoring datasets with a relatively small number of missing observations, as evident in the eastern Montana dataset. Results with ROS showed little improvement compared to the traditional substitution of one-half the detection limit. Hence, GLEC reverted to using this simpler substitution as a time-saving measure, though will make available to Montana DEQ an R script file (containing specific R commands) for completing ROS imputations in other datasets.

[11] See Appendix B.1 for definitions of dual predictor-response variables.

The aquatic plant measurements (Figure 2.2) include Montana DEQ's Aquatic Visual Assessment which is composed of measurements on micro-algae (mainly diatoms plus other "short" attached algae), filamentous algae, macrophytes, and moss. The measurements include thickness, extent of substrate cover, color, and condition (relative age in life cycle). Micro-algae quantification was then used by Montana DEQ to compute their periphyton metrics – relative abundance of nutrient enricher taxa and probability of impairment by stream nutrients. One other aquatic plant measurement includes species identification for macrophytes (15 unique species) and algae (2 unique species). Species identification was not employed as a variable (a factor variable) in the statistical modeling for this study, but in future work their association with DO model predictions could readily be included or at least portrayed in map form. Model variables developed from the Aquatic Visual Assessment dataset include the percent cover of micro-algae as a ranked[12] variable [MARANK], micro-algae thickness (in mm) [MATHICKMM], percent cover of filamentous algae as a rank [FARANK], and percent cover of macrophytes as a rank [MPHYTERANK]. The ranked variables created here yield no loss of information because the Aquatic Visual Assessment, by its nature, was built entirely on discrete (as opposed to continuous) observations. Two remaining dual variables were built from the two periphyton metrics, and describe above – percent relative abundance [RELABUND] and probability of impairment [IMPROB]. Not every site (of the 73 stations) was sampled for aquatic plant information in each year of the period of record. The sampling distribution by year and number of sites follows that of water chemistry explained above.

2.d. Response Variables

The pure response variable and ultimate endpoint in this study was the DO concentration in the stream system (Figure 2.2). DO is an important driver of aquatic health, impacting the growth, reproduction, and maintenance of aquatic organisms, such as fish and macroinvertebrates. Continuous DO measurements (hourly; either with YSI® or miniDOT® instruments) in events spanning 1-3 weeks in length were made by Montana DEQ over the 2013-2017 period, and results summarized daily. The sampling distribution by year and number of sites follows that of water chemistry explained above.

To minimize missingness (presence of a missing value) in the response variable with respect to all predictor and dual variables defined above, **the finest time resolution possible for this study was a weekly aggregation**. The statistical learning techniques employed in this study (Regression Trees; see Methods section) allows missingness in predictor variables but not in response variables. When present in the response variable, the entire observation was removed from the analysis.

Several response variables were developed by GLEC to make use of this rich DO dataset – three were based on diel delta magnitudes, three were based on diel delta exceedance of a threshold, and one was based on the minimum value. GLEC developed a response variable that reflects a count of DO delta threshold exceedances. DO delta thresholds are often used by state environmental

[12] All ranked variables in this document are computed from the range of that particular variable with rank boundaries established from natural breaks in the variable distribution.

agencies as an indicator of nutrient enrichment for a cause of aquatic life use impairment. The response variables were defined as follows (all integrated over a 7-day period or one week)[13]:

1. DO diel delta (mg/L):
    a. average of suite of daily deltas [xdelta]
    b. standard deviation of a suite of daily deltas [sdelta]
    c. maximum of suite of daily deltas [mxdelta]
2. DO minimum (mg/L): average of suite of daily DO minimums [min_avg]
3. Exceedance of a threshold DO delta (count per week):
    a. *Low threshold* [exceedMN]: # days where daily delta exceeds 3.5 mg/L; see Heiskary and Bouchard (2015) - Table 1 for Central River Nutrient Region for Minnesota streams
    b. *Medium threshold* [exceedMT]: # days where daily delta exceeds 5.3 mg/L ; see Suplee and Sada (2016) Table C2-2 and p. C-4 for Montana streams
    c. *High threshold* [exceedOH]: # days where daily delta exceeds 6.5 mg/L; see Miltner (2010) Table 5 for Ohio streams

2.e. Construction of Input Datasets

The final datasets to be used in the statistical learning by regression trees (see Methods section) consist of:

1. **All Predictors** containing 234 observations (records) comprising 73 stations over several weekly events in the 5-year period (2013-2017), 7 DO response variables (Section 2.d), 4 aquatic plant dual (both predictor and response) variables (Section 2.c), and 39 predictor variables (Section 2.b and 2.c).[14]
    a. Each record has an identifier that comprises the station ID, sample year, sample month, and sample *week*.
    b. The fixed-effect predictor variables vary by sampling station watershed but were replicated in time (same for each week).
2. **Expanded Observations** where predictor and dual variables with missingness were eliminated which, in turn, expands the set of DO response observations.
    a. **Weekly**: 762 observations (records), 7 DO response variables, and 24 predictor variables
        i. Each record has an identifier that comprises the station ID, sample year, sample month, and sample *week*.
        ii. The fixed-effect predictor variables vary by sampling station watershed but were replicated in time (same for each week). NOAA monthly drought indices were excluded.

[13] See Appendix B.1 for definitions of response variables.

[14] Water quality variables (chemical concentrations and field parameters) could be employed as a response variable (they are dual variables) but were not used as such in this study due time efficiency. A future study could explore these parameters as an additional set of response variables.

b. **Monthly** (not employed in this study): 318 observations (records), 7 DO response variables, and 24 predictor variables
    i. Each record has an identifier that comprises the station ID, sample year, and sample *month*.
    ii. The fixed-effect predictor variables vary by sampling station watershed but were replicated in time (same for each month). NDMC weekly drought indices were excluded.

3. Methods

3.a. Regression Trees

3.a.i. Introduction

Regression trees are similar to additive models in that they represent a compromise between the linear model and the completely nonparametric approach. Regression trees make use of a recursive partitioning algorithm.

Tree models are also well suited to finding interactions. If the algorithm splits on one predictor variable and then splits on another variable within the partitions of the first variable, one is finding an interaction between these two variables. See Figure 4.1 (in Results) as an example.

The type or method of a tree model is based on the character of the response variable. An advantage of tree models is that many options are available to structure the formulation including responses that are quantitative, count-based, ordinal (factor), and survival.

To grow a tree, the recursive partitioning regression algorithm:

1. Consider all partitions of the region of the predictors into two regions where the division is parallel to one of the axes. A single predictor is partitioned by choosing a point along the range of that predictor to make the split.
2. For each partition, the mean *of the response* in that partition is computed. Then the deviance or residual sum-of-squares (RSS) is computed as RSS(partition) = RSS(part 1)+RSS(part 2). A partition that minimizes the residual sum-of-squares (RSS) is chosen. Deviance (RSS) is equivalent to the concept of node impurity (a node being one of the branching points on a tree) which is the heterogeneity of the distribution at a given node.
3. The partitions are now sub-partitioned in a recursive manner. Sub-partitions within existing partitions are only allowed and *not across* them. This means that the partitioning can be represented using a tree. Further, there is no restriction preventing the splitting of the *same predictor variables* consecutively.

In summary, the tree algorithm *alone* determines which predictor variable becomes the *root node* (the first split) and the value at which the decision (or split) is made to traverse left or right. It is a two-step process. The first step, for each predictor variable, considers all possible binary splits of the predictor values. The *best split for each predictor* is defined as the split that reduces the RSS the most.

With the best split for each predictor determined, the *best split overall* is determined as the second step. That is, the best split for *each predictor is compared* with all of the remaining predictors by the reduction in the error sum-of-squares. The predictor with the largest reduction "wins the competition" (James et al. 2015) – this is the predictor in the pole position or root node. It is the predictor that when properly split leads to the greatest reduction in the RSS compared to the RSS before that partitioning is undertaken.

Regarding missing values, a tree-fitting algorithm can handle them naturally. If a value for some predictor is not available, it could be simply excluded from the criterion. A more complicated approach, and that which was applied here, is to allow a second-choice variable for splitting at a node called a *surrogate split*. Information on the surrogate splits can be obtained by using the `summary` command on the tree object (not shown in this document).

3.a.ii. Model "Tuning" Parameters to Find the Optimal Tree Complexity

GLEC employed `rpart` (for recursive partitioning) within the R application to develop regression trees for the eastern Montana dataset. The algorithm describes how to grow the tree, but determination of optimal size (i.e., complexity) was necessary – too simple a tree results in loss of meaningful relationships between response and predictor, and too complex a tree results in overfitting and introduction of bias, and may be impractical for later implementation of the findings. The default form of `rpart` does restrict the size of the tree, but some intervention was necessary to select the best tree size.

A tree depth may be set prior to model run, termed pre-pruning, through setting of minimum observations before a split may occur, the minimum number of observations in a leaf node (the bucket size), a complexity parameter, and the maximum depth of the tree. All of these pre-pruning parameters in some way are interrelated and, based on frequency of use in the support publications, GLEC chose to work with the complexity parameter (cp). Any split that does not decrease the overall lack of fit by a factor of cp is not attempted. For example, when the model response is quantitative (continuous), the overall R^2 must increase by cp at each step in the tree building. Cp may range from 0 (a deep or saturated tree) to 1 (a shallow tree where no splits are made).

Pruning may also occur following a model run, termed post-pruning, and may achieved by command or interactively with a graphical interface. Post-pruning was not completed in this study but the R command `prp` with `snip=TRUE` will allow for an interactive prune.

3.a.iii Model Diagnostics

Performance of a particular model configuration can be examined through several means – the root-node error, cross-plots of model complexity vs. explained variance or vs. cross-validation error, and a residuals analysis.

The root node error is the percent of correctly sorted records at the first (root) splitting node. It is represented as the total deviance in the particular tree model divided by the number of observations. The value is shown for each model run (see Table 4.3 for an example).

A k-fold cross validation approach was used in this study. The dataset was randomly divided into k roughly equal parts, where k=10 was chosen for this study and as is typically used. Hence, each validation subset is approximately 23 observations (234 total observations/10 folds). GLEC used k-1 (9) parts to predict the cases in the remaining (or 10^{th}) part. The model is run k (10) times, leaving out a different part each time. K-fold cross validation is computationally faster than the alternative "leave (only) one observation out" but its drawback is that the partition is random so that repeating the method will give slightly different numerical results.

The root node error, first mentioned above, can be used to calculate two measures of predictive performance in combination with the relative error (`rel error`) and cross-validation error (`xerror`). For the data used to build the model (the training data), the prediction error rate is equal to the product of the root node error and the relative error. This particular error rate is named the *resubstitution error rate* and it represents the proportion of original observations that were misclassified by various subsets of the original tree. The largest tree (i.e., greatest number of splits) will always yield the lowest resubstitution error rate. However, choosing the tree with the lowest resubstitution rate is not the optimal choice, as this tree will have a bias. Large trees will put random variation in the predictions as they overfit outliers. See `rel error` column in Table 4.3 (in Results) as an example; it is an abbreviated example because all of trees in the Results section have been optimally pruned.

In cross-validation, the prediction error rate is equal to the product of the root node error and the cross-validation error. This particular error rate is named the *cross-validation error rate* and is a more objective measure of predictive accuracy. It is determined by adding up the error across the k- (10) folds. The tree yielding the lowest cross-validated error rate (`xerror`) is selected as the tree that best fits the data. See `xerror` column in Table 4.3 (in Results) as an abbreviated example. Xerror is similar to the PRESS (predicted residual error sum-of-squares) found in linear models. Results for error rates will be further discussed in the next section (Results); a summary of the error rates for each of the model runs is shown in Table 4.1.

There are two additional diagnostic plots to aid in selecting the optimal tree complexity. See Figure 4.4 (in Results) as an example of both plots. The first plot (left-side) is that of explained variance (similar to R^2 in a general linear model) vs. the number of splits (a measure of tree complexity). The R^2 value is simply 1 – `rel error` (named "Apparent") for the original model fit or 1 – `xerror` (named "X Relative") for the cross-validation series. The best tree complexity exists at the maximum value of "X Relative". The value of "Apparent" will always increase with increasing tree complexity because `rel error` always decreases with this complexity (as first noted above); hence, the plot of "Apparent" is not a useful diagnostic other than to compare to "X Relative".

The second plot (right-side) is that of `xerror` plus vertical bars representing its standard deviation vs. number of splits. The standard deviation of `xerror` is named `xstd` and can be found in Table 4.3 (Results) as an example.

A residual analysis is helpful for detecting unusual observations and other potential issues with the data. They can be used both for an overview of the model's performance and to check for outliers.

One can examine each outlier – its nature with respect the value of the response or its predictors – to determine if a particular observation is truly part of the same distribution of the remaining observations (that are themselves not considered outliers). For each model run, the residual diagnostics contains a cumulative distribution vs. residual value, a residual vs. fitted scatter plot with a loess-fitted line, and a quantile-quantile plot. See Figure 4.5 (in Results) as an example.

3.a.iv Model Settings Established for Eastern Montana Streams

Specific to this analysis for eastern Montana streams, three constraints were set for each model – the cost-complexity parameter, method type, and use of surrogates.[15] Initially each model was executed with a very low value of cost-complexity, using a cp close to zero 0 (typically 0.005) to produce a nearly "saturated" tree. See Figure 4.2 (in Results) as an example. The optimal cp value is chosen where the cross-validation error (`xerror` in `rpart` output) is at a minimum. An accepted rule-of-thumb is to select the next simpler model (which is to the left of this minimum on the plot) given that its cross-validation error is lower than the lowest cp plus its standard deviation (defined as the dashed-horizontal line). Jaraway (2016; p 348) offers a similar strategy.

Preliminary model results, of varying tree complexity, were shared with Montana DEQ. After uncovering the key predictor variables and their interactions, as well as recognizing a minimum bucket (leaf-node) size of generally no less than seven observations (the station-dates), DEQ concurred with GLEC on not overfitting trees and requested GLEC place a premium on tree simplicity.

Of the several method types offered by the regression tree software package (`rpart`) only three were applied here (Table 3.1). When the response is quantitative (continuous), `anova` method was chosen, and the most common method in this study (e.g., average DO delta). For count-based response variables, as seen with the number of days per week that a delta threshold is exceeded, the `count` method was chosen. For ordinal responses, such as any ranked variable or any with a continuous-like value but only takes on a small number of discrete values (e.g., MATHICKMM), the `class` method was chosen.

Finally, surrogate variables were used when at least one of the predictor variables were missing for a given observation. Recall, surrogates do not apply to a missing response variable. When a response value is missing, the observation would be deleted (entire record) in a tree model run; this situation occurred in this study only when an aquatic plant variable served as the response.

3.b. Model Formulations to be Tested

Making use of the flexibility of regression trees and their ability to uncover relationships in the data, several model formulations were designed for this study (Table 3.1). GLEC selected a subset of them for actual implementation and in consultation with Montana DEQ as to which model results would likely be of most value.

[15] All custom-defined command lines (or similar scripts) employed in the R framework are included in Appendix B; they have also been saved as digital files and available from GLEC upload (2/2021) to **ePass Montana File Transfer Service**.

Table 3.1. Regression tree model formulations proposed with eastern Montana response-predictor dataset. Formulations implemented in this study are asterisked. The expanded datasets (both weekly and monthly) are included here for reference, though only one model was implemented. Model variable names are shown in brackets [ ].

Response Variable		Predictor Variables Excluded	Method
All Predictors – Weekly (234 observations)			
Dissolved Oxygen	Mean Delta [xdelta]*	None	anova (quantititative response)
	Standard Deviation Delta [sdelta]	None	anova (quantititative response)
	Maximum Delta [mxdelta]*	None	anova (quantititative response)
	Average Minimum [min_avg]*	None	anova (quantititative response)
	Exceedance of delta threshold (Minnesota – low)[16] [exceedMN]	None	Poisson (count response)
	Exceedance of delta threshold (Montana – medium) [exceedMT]*	None	Poisson (count response)
	Exceedance of delta threshold (Ohio – high) [exceedOH]	None	Poisson (count response)
Aquatic Plant	Micro-algae thickness (mm) [MATHICKMM]*	Micro-algae thickness rank [MARANK]	class (factored response)
	Percent cover of macrophytes (rank) [MPHYTERANK]*	None	class (factored response)
	Percent cover of filamentous algae (rank) [FARANK]	None	class (factored response)
Expanded Observations – Weekly (762 observations)			
Dissolved Oxygen	Mean Delta [xdelta]*	· All water chemistry & field parameters · All plant indices · NOAA monthly drought indices	anova (quantititative response)
	… (as in All Predictors –Weekly)	" "	
Expanded Observations – Monthly *(318 observations)*[17]			
Dissolved Oxygen	… (as in All Predictors –Weekly)	· All water chemistry & field parameters · All plant indices · NDMC weekly drought indices	anova (quantititative response)

[16] Future regression tree modeling should compare the change in predictor variable importance among the three delta thresholds (low, medium, high).

[17] Future implementation of regression tree modeling should consider folding this dataset (which contains the NOAA monthly drought indices) into the expanded weekly dataset. Then one simply replicates the monthly-varying drought variables over the 4 weeks in each month (4x replication).

4. Results and Discussion of Regression Tree Model Runs

Results of regression tree models are first discussed with those using the full predictor dataset beginning with four DO-type response models and then two aquatic plant response models. Lastly, one DO model using an expanded observation dataset is discussed. All of these models are identified in Table 4.1 along with their corresponding complexity (cp and number of splits) and associated predictive errors. Complexity, as first described in the Methods section, is a measure of how many branches (or splits) the resulting model tree yields. Predictive errors, also described in the Methods section, are useful for comparing variations of the same model but not between models.

Table 4.1. For each of the model formulations[18], the complexity parameter assigned, resulting number of splits, root node error (total deviance/# observations), and relative and cross validation errors and their corresponding error rates.

Model Name	Complexity Parameter (cp)	# Splits	Root Node Error	Relative Error	Cross-Validation Error	Error Rate	
						Resubstitution	Cross-Validation
Delta Mean DO	0.033	5	17.9	0.653	0.98	11.7	17.5
Delta Maximum DO	0.043 [19]	4	27.0	0.685	0.974	18.5	26.3
Minimum Average DO	0.025	7	6.9	0.565	0.934	3.9	6.4
Delta Exceed-Medium DO[20]	0.039	5	1.24	0.527	0.78	0.7	1.0
Microalgae Thickness[21]	0.069	2	0.453	0.745	0.745	0.3	0.3
Macrophyte % Cover[2,4]	0.0077	12	0.62	0.531	0.841	0.3	0.5
Delta Mean DO Expanded	0.011	16	20.6	0.393	0.579	8.1	11.9

[18] Several other formulations should also be considered in future modeling efforts (e.g., frequency of delta exceedance at lower and higher critical thresholds).

[19] An alternative lowest cross-validation error exists at cp=0.1 with only one split. GLEC chose the smaller cp (0.0077) resulting in a more complex tree to increase meaningfulness of the predictor interpretation for this model run.

[20] Count-based response.

[21] Ordinal (class) response.

4. a. Dissolved Oxygen

4.a.i. Dissolved Oxygen – Mean Delta

The regression tree model with a response of mean weekly DO delta contains five splits and employs five predictor variables (Figure 4.1; Table 4.2). Each of the leaf nodes contain a set of observations (where the total is listed for each node) (Figure 4.1). Each observation was indexed by its station-date identifier and the list of identifiers and their tree decision rules (e.g., as in Table 4.2) have been compiled and made available in electronic files[22] for all model runs shown in Table 4.1.

The primary predictors were watershed land cover (disturbance), low intensity drought (number of consecutive weeks), and conductivity in the water column. In this particular model result, any of the leaf nodes where the mean DO delta is at or above 5.31 mg/L are considered stressed by high DO range and likely negatively impact aquatic organisms[23]. These situations exist where disturbed land in the watershed exceeds approximately 16 percent and low intensity drought is present for more than six consecutive weeks. GLEC suggests, that in these situations, the DO range is increased by lower daily minimums produced by higher water temperatures – lower baseflow from less groundwater recharge (low soil moisture levels and groundwater recharge absent) – and higher daily maximums from reduced light limitation in disturbed land covers. Light limitation can exist from any form of riparian shading, and in the case of eastern Montana, even minimally by the presence of shrubland. With reduced light limitation and a likely increase of nutrient runoff from uplands, photosynthesis from aquatic plants (both micro- and filamentous algae and macrophytes) increases. Further down this same branch of regression tree was the influence of nutrient concentration – in this case total phosphorus at concentrations above 2.22 mg/L – or if not present, when water temperatures are above 20.5 °C.

[22] GLEC upload (2/2021) to **ePass Montana File Transfer Service**.

[23] It is coincidental that the 5.31 mg/L leaf node matches the critical threshold DO range of 5.3 mg/L for Montana streams as prescribed by Suplee and Sada (2016).

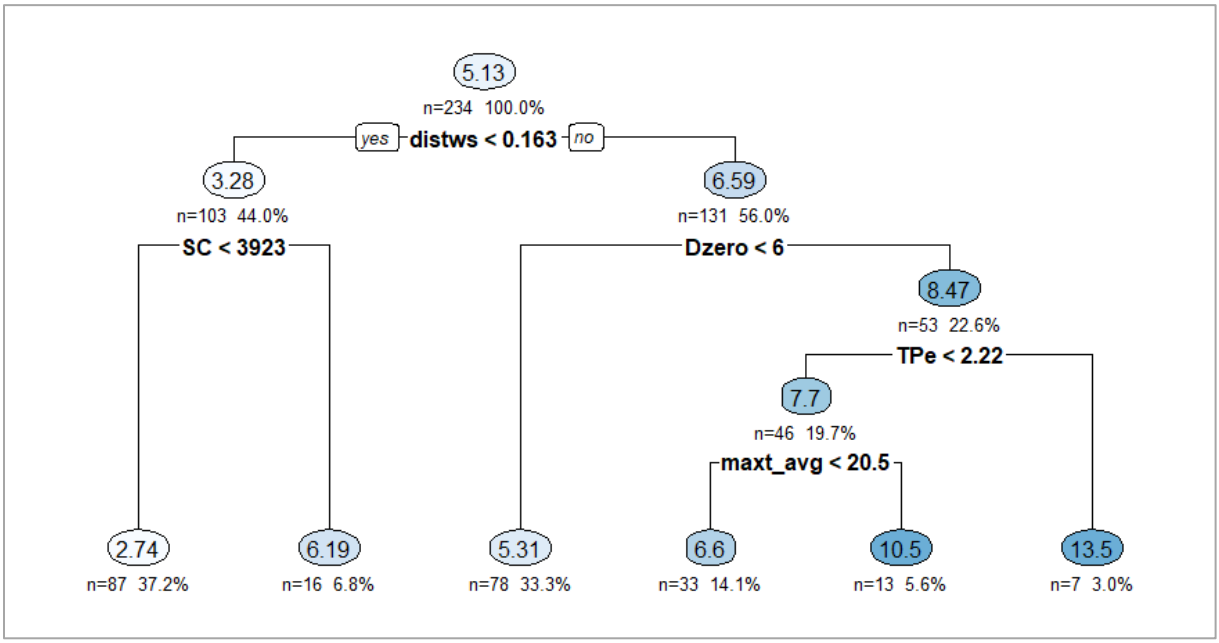


Figure 4.1. Diagram showing the regression tree for a response of **mean weekly DO delta** (mg/L). The predicted value and the number and percentage of total observations are shown for each node. The intensity of the node color is proportional to magnitude of the predicted value. The decision statement to split is located under each node (in bold) – traverse left if the statement is true (yes), otherwise traverse right (no). Branching to the left of "Dzero < 6" represents wetter conditions (i.e., fewer weeks of D_0 drought, whereas its corollary (Dzero > 6) to the right reflects drier conditions.

There is only one leaf node where low mean DO delta is present in this model, and it is extremely low (mean delta of 2.74). Low delta suggests an un-stressed condition, according to the model, where both watershed disturbed land cover and specific conductance are both low (less than 16% and 3923 µS/cm, respectively). Prevalent natural land cover (above 84% of the total watershed area) implies low nutrient runoff (and dissolved solids) and higher levels of riparian shading. Both nutrients and light would be limiting and thus reduce daily DO maximums. Additionally, higher riparian shading and likely improved baseflow from natural land cover would increase daily DO minimums.

Table 4.2. The regression tree model shown above (Figure 4.1) portrayed as a set of decision rules. One set of statements describes a particular leaf node. Primary statement begins with mean predicted response (rounded to one decimal place) and its value (units defined above). Cover percent refers to percent of total observations in the particular node. Secondary statements are decisions about the important predictor variables to reach the leaf node.

```
xdelta is  2.7 with cover 37% when
    distws < 0.16
    SC < 3923

xdelta is  5.3 with cover 33% when
    distws >= 0.16
    Dzero < 6
```

```
xdelta is  6.2 with cover 7% when
    distws < 0.16
    SC >= 3923

xdelta is  6.6 with cover 14% when
    distws >= 0.16
    Dzero >= 6
    TPe < 2.2
    maxt_avg < 20

xdelta is 10.5 with cover 6% when
    distws >= 0.16
    Dzero >= 6
    TPe < 2.2
    maxt_avg >= 20

xdelta is 13.5 with cover 3% when
    distws >= 0.16
    Dzero >= 6
    TPe >= 2.2
```

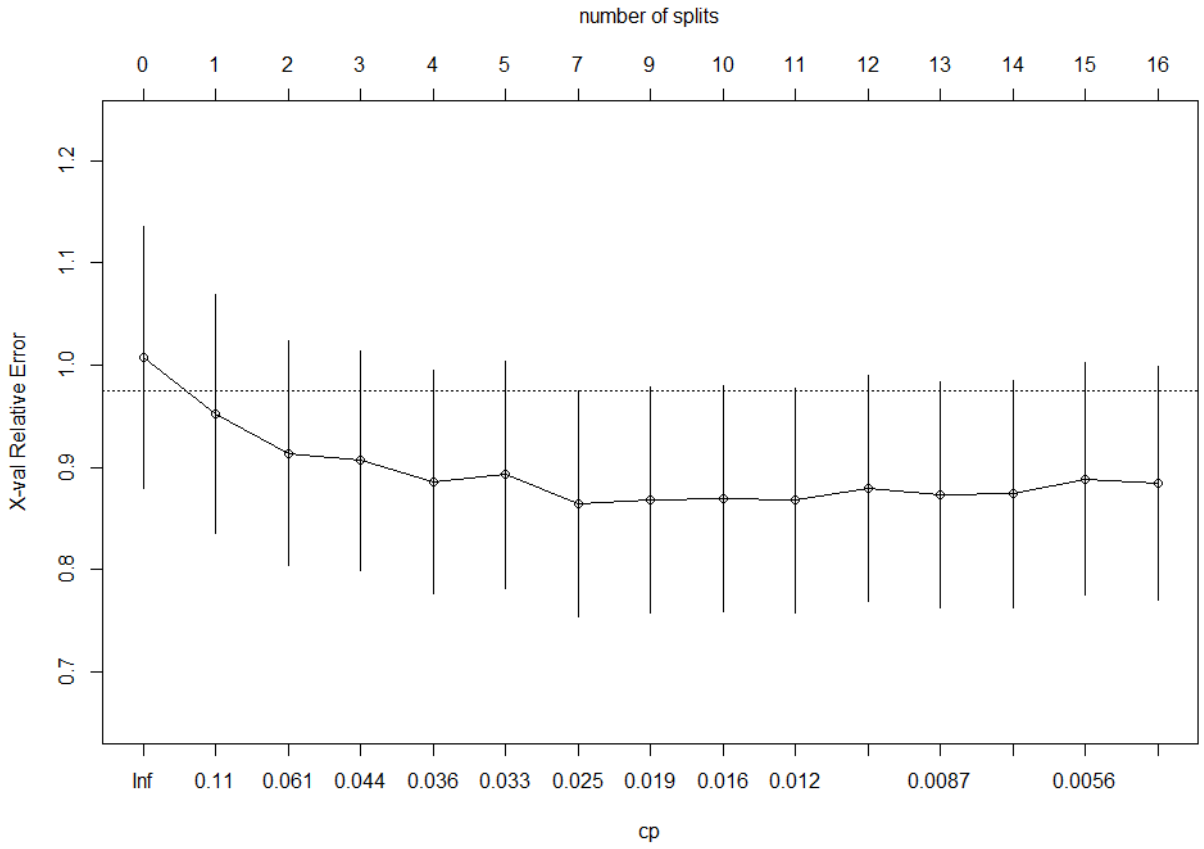


Figure 4.2. Initial plot of cross-validation error (xerror) vs. model complexity (cp) or number of splits in tree (upper x-axis). Error bars for xerror equal to ±1 error standard deviation (xstd). Dashed horizontal line is placed at +1 xstd above the lowest modeled xerror.

As described in the Methods section (Section 2), the initial value of the complexity parameter (cp) was intentionally set very low (typically 0.005) to build a fully saturated tree. Figure 4.2 shows the change in the cross-validation error with respect to changing complexity. The most complex tree is on the right of the plot. The tree was subsequently pruned to a complexity just less than (i.e., at a larger cp) the lowest cross-validation error as long as its own `xerror` is below the dashed horizontal line. From Figure 4.2, the lowest `xerror` was found at a cp = 0.025, which would have resulted in 7 splits, so using the rule-of-thumb discussed in Methods, a tree of complexity nsplits = 5 (upper x-axis) and cp = 0.033 (lower x-axis) was used to build the model shown in Figure 4.1 and Table 4.3.

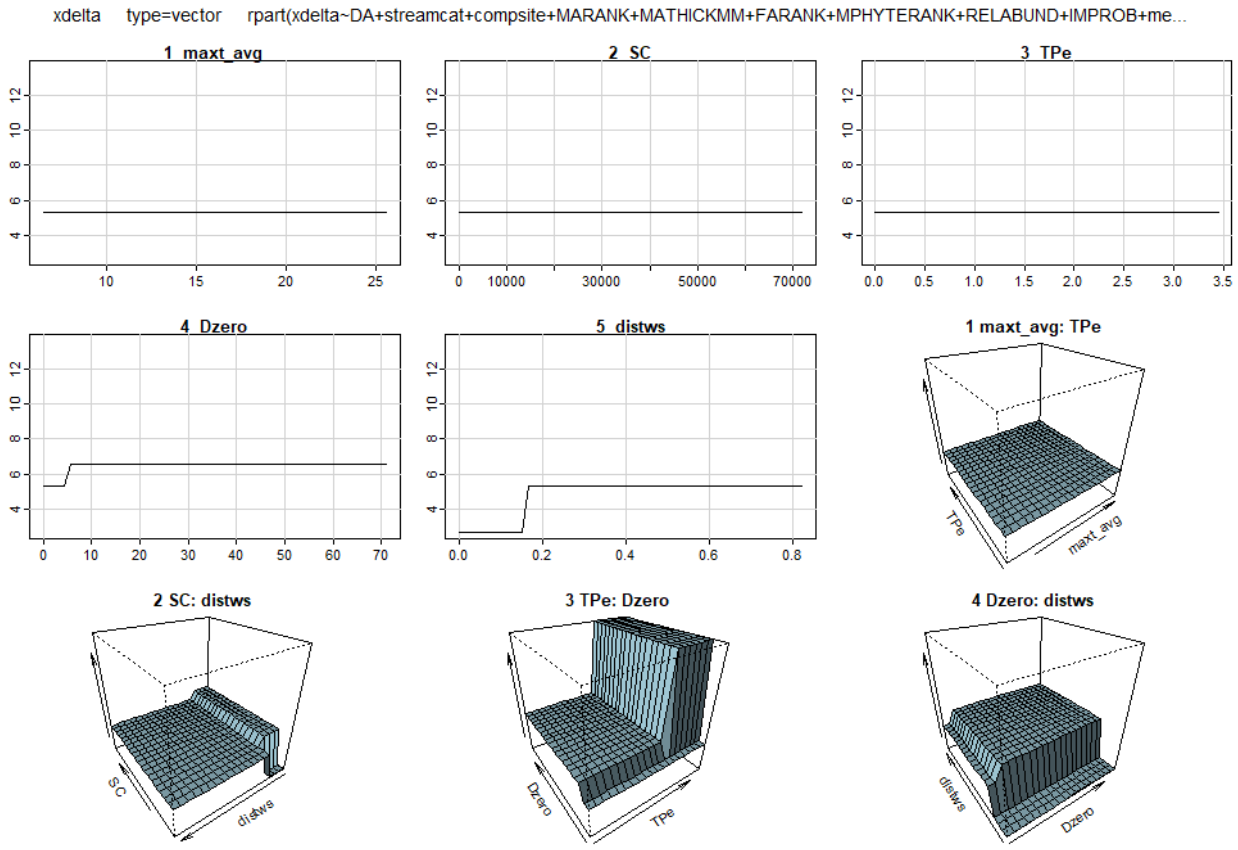


Figure 4.3. Plot of regression tree surface for model formulation of **mean weekly DO delta** (formulation also listed at the top of the diagram). Upper series of plots show relationships between each predictor variable and the response variable. All variables used in the tree are shown. Lower series of plots shows interactions between a pair of predictor variables, with all other variables held

at their median values, and the response variable. The pairs are chosen by which variables appear in the parent-child pairs of the regression tree (see Figure 4.1 or Table 4.2).

To better visualize the interaction between a pair of predictor variables, bivariate plots of predictors vs. the response variable are shown (Figure 4.3, bottom plots). Univariate plots of a single predictor indicate at what thresholds, if any, does the response change with respect to a changing value of a single predictor (Figure 4.3, upper plots). Some predictor variables in these univariate plots show no relationship to the response variable (e.g., SC). However, the predictor becomes important in its interaction with other predictors (e.g., SC and distws). For the mean weekly DO delta tree, one can observe, for example in Figure 4.3 (lower right), the delta response to drought but only when a threshold of disturbed watershed land cover was exceeded (more than approximately 16 percent of total watershed area).

Variables that do not appear in plot were considered background variables. In each plot (Figure 4.3, bottom), the background variables were held fixed at their median values (the medians are calculated from the original data). If a background variable is not continuous but rather a factor, then the most common factor level is used instead of the median. In true partial-dependence plots, the effect of the background variables is averaged. This is computationally complex in regression tree model so the median value is computed instead.

Table 4.3. Initial model formulation, list of predictor variables used in tree construction after model complexity chosen, root node error, and table showing – in order of decreasing cp (increasing tree complexity) – cp value, number of splits, relative error (rel error), cross-validation error (xerror), and standard deviation of cross-validation error (xstd).

Regression tree:

```
rpart(formula = xdelta ~ DA + streamcat + compsite + MARANK +
    MATHICKMM + FARANK + MPHYTERANK + RELABUND + IMPROB + medt +
    maxt_avg + BP + NH3 + Nox + OP + pH + SC + TN + TP + Tpe +
    Tne + Zindex + PMDI + PHDI + DSCI + DSCt + Dzero + Done +
    Dtwo + Dthree + Dfour + natws + distws + natnf + distnf +
    wells + wellso + wellcat + wellocat + maxslope + medslope +
    xslope + devslope, data = do_wk, method = "anova", x = FALSE,
    y = FALSE, control = rpart.control(cp = 0.033, usesurrogate = 2))

Variables actually used in tree construction:
[1] distws   Dzero    maxt_avg SC       Tpe

Root node error: 4185/234 = 17.9

n= 234

      CP nsplit rel error xerror  xstd
1 0.1509      0     1.000   1.02 0.129
2 0.0750      1     0.849   1.05 0.137
3 0.0493      2     0.774   1.04 0.133
4 0.0384      3     0.725   1.01 0.134
5 0.0335      4     0.686   1.00 0.134
6 0.0330      5     0.653   0.98 0.132
```

A plot of explained variance (similar to an R^2 in a general linear model) and model complexity (shown as number of splits) indicates that a 3-split model for mean delta DO was optimal (Figure 4.4, left). The distribution of interest is labeled "X Relative" based on a cross-validation of the original dataset; the distribution labeled "Apparent" is based only on the original data and will always increase with increasing complexity and should generally be ignored. Note, a 3-split model was found in this cross-validation whereas a 5-split model was found to be optimal in the original test when setting a low value of cp (Figure 4.2). In cross-validation, the partition is random so that repeating the method will give slightly different numerical results; a disadvantage of tree modeling. However when looking at the distribution of cross-validation error (`xerror`) and model complexity (Figure 4.4, right), the 3-split model is optimal but either a 4- or 5-split model was acceptable. Note, that the 3-split model is, in effect, visible in the 5-split tree (Figure 4.1) – it is simply the first three splits. Table 4.3 shows the actual values of `xerror` and `rel error`, with the latter equal to 1-R^2.

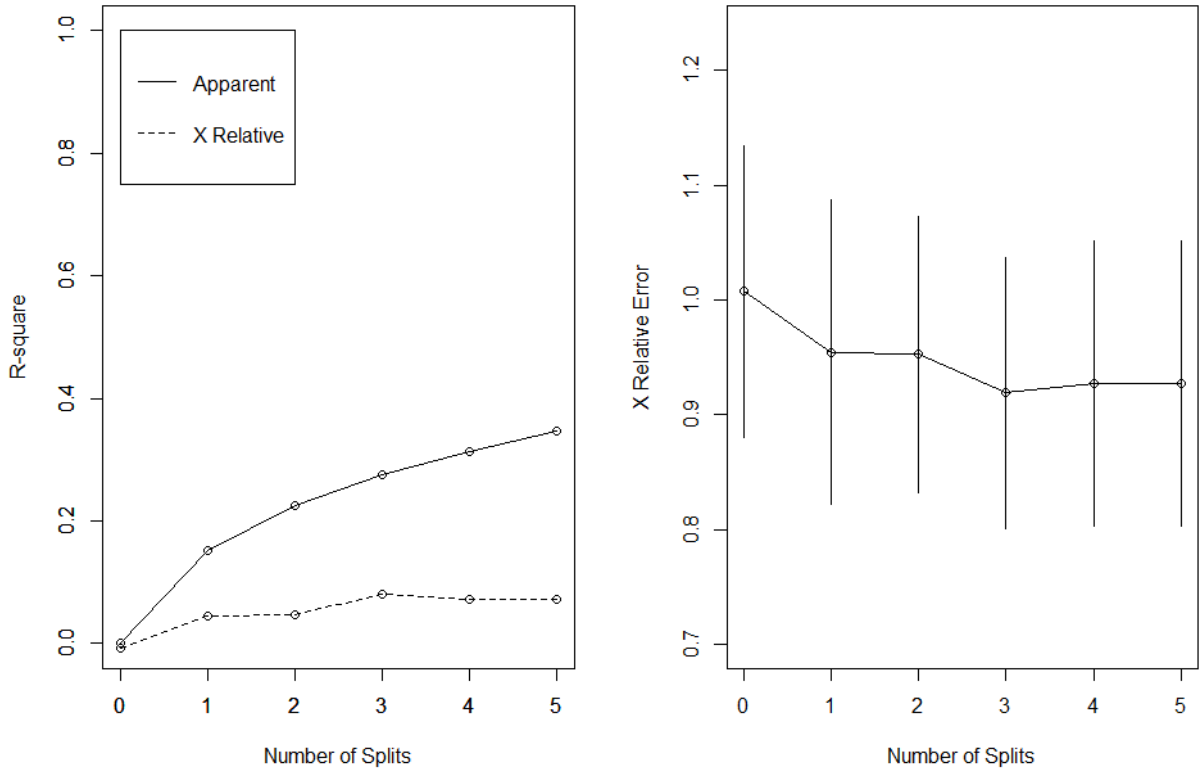


Figure 4.4. (**left**) Diagnostic plot showing explained variance (R^2) vs. the number of splits in tree diagram (a measure of complexity). R^2 (Apparent) = 1 – `rel error` for the original model fit; R^2 (X Relative) = 1 – `xerror` for the cross-validation series. (**right**) Diagnostic plot showing cross-validation relative error (`xerror`) vs. number of splits. Vertical bars represent ±1 standard deviation (`xstd`) of `xerror`.

A residual analysis of the mean weekly DO delta is shown in Figure 4.5. A typical residual vs. model fit of the training data is shown in the upper right. Notice that the fits are quantized which is

characteristic of `rpart` regression tree models. Each vertical line of plotting points corresponds to a leaf of the tree (see mean values of each leaf node in Figure 4.1). Note, a "jitter" of the plotting points was added to reduce overplotting caused by quantization here. The red line is a LOESS line (locally weighted smoothing of the scatter of points) to show the trend of residual magnitudes.

At Red Butte Creek (site 01 – station-ID Y22RDBTC01) on the 3rd week of 9/2013 (see case[24] 62 on Figure 4.5), for example, one observes that the model estimates a moderate average delta DO (xdelta) (the fitted value is 5.31), the observed level was actually much higher – 13.36 mg/L higher or 18.67 mg/L! A likely explanation of this model underprediction was simply a function of aggregating daily measurements into one week. This particular week (the 3rd week of September 2013) had only one daily value (as opposed to a maximum of seven daily values). Hence the mean delta was computed from a mean of one and it happened to be quite large on that particular day. For case 215 (Figure 4.5), a similar data aggregation limitation occurred at Little Beaver Creek (Y27LBVRC12) also on the 3rd week of 9/2013 as above where two samples were in a weekly bin. The third outlier (case 37) is O'Fallon Creek on the 4th week of 8/2017. While there was a daily DO delta for each day in the weekly bin, there was high variance among the daily deltas for that particular week.

From the quantile-quantile (QQ) plot (Figure 4.5, bottom left), this residual is unusual. The QQ plot shown here is a plot of the quantiles of the residuals (y-axis) set against a normal distribution of quantiles (x-axis). A quantile (or percentile) is the fraction (or percent) of points below the given value. The normal quantiles are theoretical and simply shown to identify unusual residuals, which are observations that deviate from the relatively straight pattern along the dashed line. The regression tree model has no assumptions of an underlying theoretical distribution.

Since case 62 (Red Butte Creek) was the largest absolute residual (x-axis), it determines the right bound of the cumulative distribution plot (Figure 4.5, top left). Also, in this same plot, one observes that 50 percent of the observations have a mean delta residual of less than 2 mg/L, and 75 percent of the observations have a residual of less than 3 mg/L. Both delta magnitudes were small indicating that this model captures the test data reasonably well.

[24] Case number shown on the residual plots (red font) refers to the same station-date record for all model runs in this study; thus, peculiar residuals can be discussed in the context of several or all model response variables.

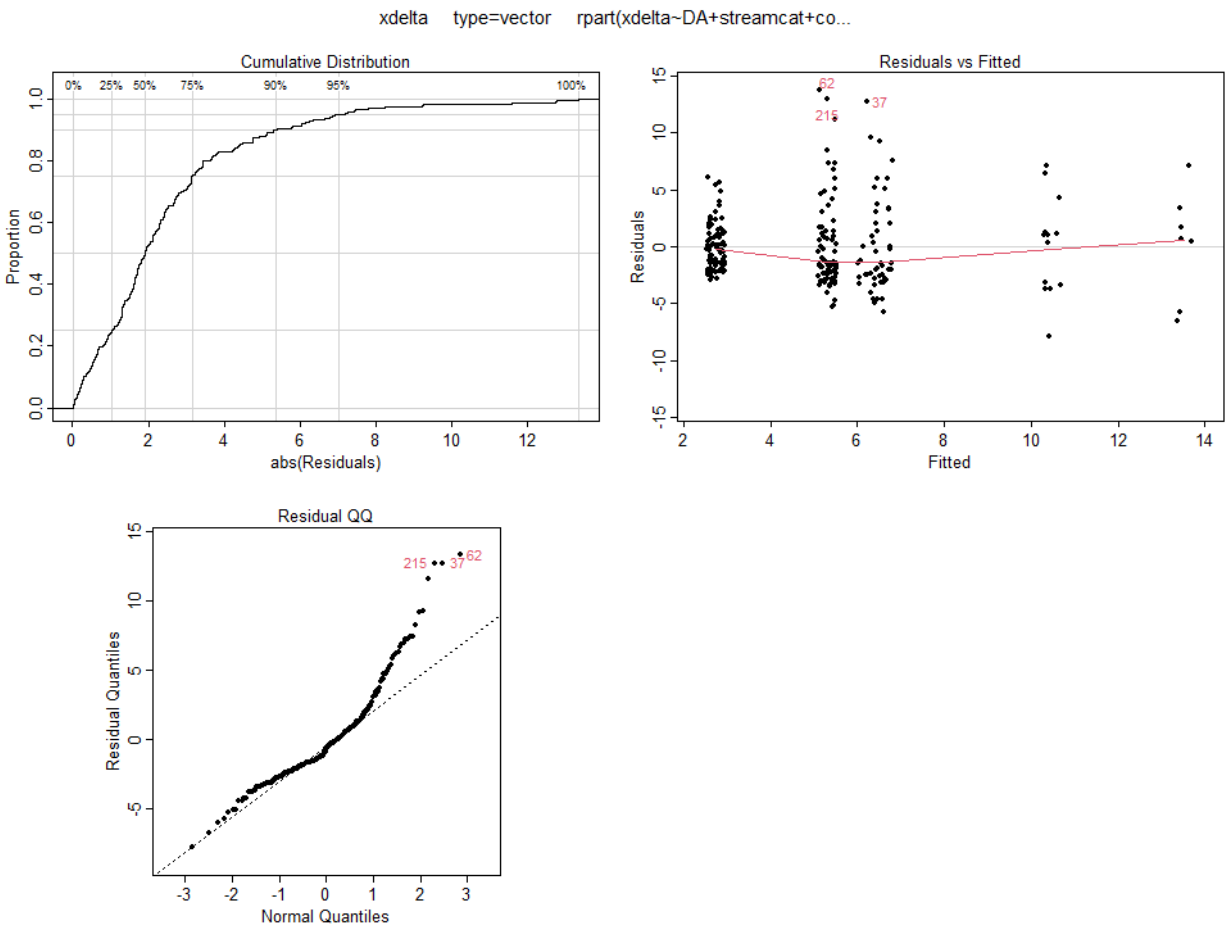


Figure 4.5. Residual analysis showing (**upper left**) cumulative distribution (proportion) of residuals (their absolute value) for entire dataset, (**upper right**) residual magnitude vs. predicted value from regression tree model (red line is loess-fitted value and observations with a high residual are noted by their record number (red font) in the dataset), and (**lower left**) quantile-quantile (QQ) plot of residuals. A positive residual indicates the observed value exceeds the predicted (fitted) value (i.e., the model underpredicts the observed).

4.a.ii. Dissolved Oxygen – Maximum Delta

The regression tree model with a response of maximum weekly DO delta contains four splits and employs four predictor variables (Figure 4.6; Table 4.4). The primary predictors were watershed land cover (disturbance) and meteorological drought (PMDI). Because the leaf node delta value is the average of its membership, the maximum delta values are somewhat subdued, except for the high Nox node (Figure 4.6). Nonetheless, the right half of the tree presents a situation of high disturbed land cover (yet it is a low threshold of 16% or greater disturbed) and meteorological conditions of slightly wet to extreme drought (drier conditions corresponding to PMDI less than +2.54; see also Figure 2.6). The 16% area threshold of disturbed land cover exists in 13 of 73 watersheds. Hence, delta DO declines when watershed conditions were moderately wet or greater and as shown on the left half of the tree. The far-right branch of the tree (maximum delta averaged to 14.4 mg/L)

presents a situation of high watershed disturbed land cover, drainage areas (DA) that exclude the largest in the study area (this represents 53 of the 73 study watersheds), and nitrite-nitrate concentrations above 0.9 mg/L. While nutrient influx is an obvious covariate to high diel DO range, and particularly nitrogen as suggested by Montana DEQ as the limiting nutrient in eastern Montana streams, the moderate to small drainage area is more susceptible to higher maximum deltas.

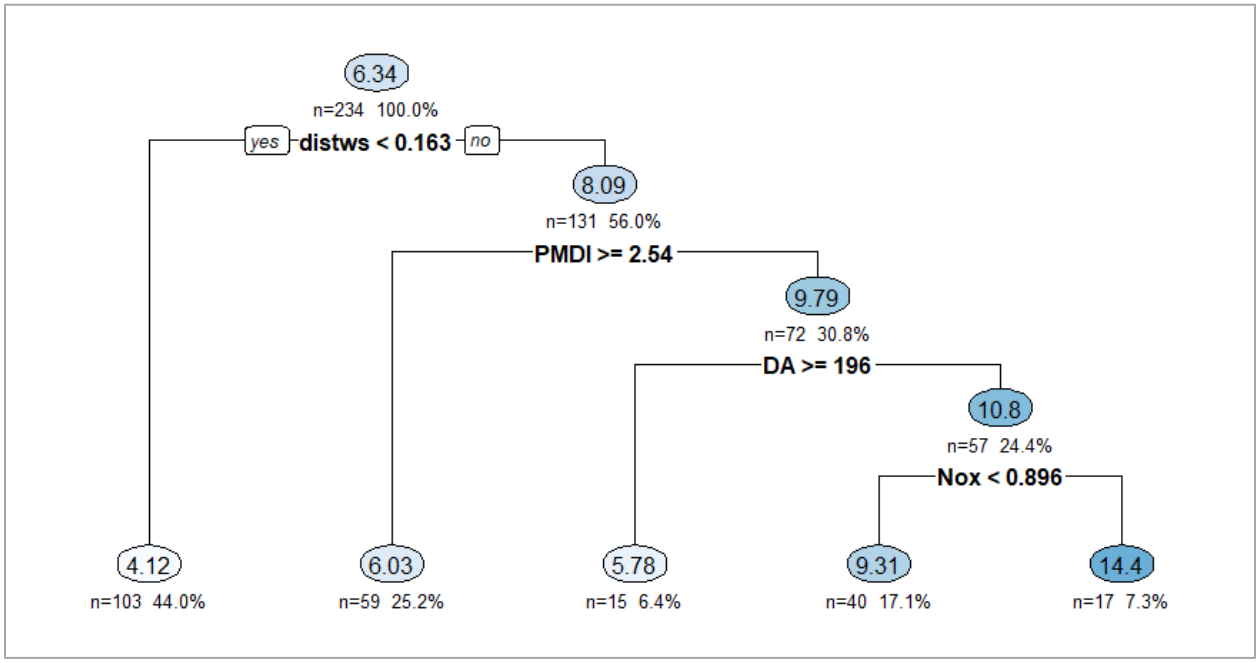


Figure 4.6 Diagram showing the regression tree for a response of **maximum weekly DO delta** (mg/L). The predicted value and the number and percentage of total observations are shown for each node. The intensity of the node color is proportional to magnitude of the predicted value. The decision statement to split is located under each node (in bold) – traverse left if the statement is true (yes), otherwise traverse right (no).

Table 4.4. The regression tree model shown above (Figure 4.6) portrayed as a set of decision rules. One set of statements describes a particular leaf node. Primary statement begins with mean predicted response (rounded to one decimal place) and its value (units defined above). Cover percent refers to percent of total observations in the particular node. Secondary statements are decisions about the important predictor variables to reach the leaf node.

```
mxdelta is  4.1 with cover 44% when
    distws < 0.16

mxdelta is  5.8 with cover 6% when
    distws >= 0.16
    PMDI < 2.5
    DA >= 196

mxdelta is  6.0 with cover 25% when
    distws >= 0.16
    PMDI >= 2.5

mxdelta is  9.3 with cover 17% when
    distws >= 0.16
    PMDI < 2.5
    DA < 196
    Nox < 0.9

mxdelta is 14.4 with cover 7% when
    distws >= 0.16
    PMDI < 2.5
    DA < 196
    Nox >= 0.9
```

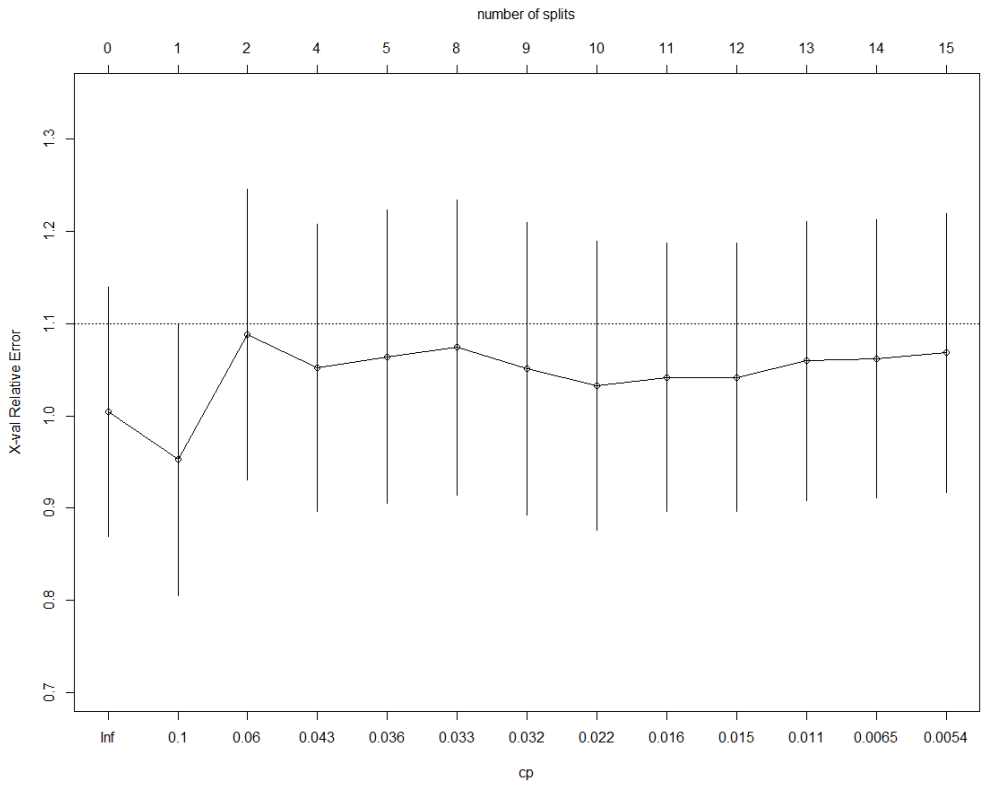


Figure 4.7. Initial plot of cross-validation error (xerror) vs. model complexity (cp) or number of splits in tree (upper x-axis). Error bars for xerror equal to ±1 error standard deviation (xstd). Dashed horizontal line is placed at +1 xstd above the lowest modeled xerror.

Figure 4.7 shows the change in the cross-validation error with respect to changing complexity. The most complex tree is on the right of the plot. The tree is subsequently pruned to a complexity just less than (i.e., at a larger cp) the lowest cross-validation error as long as its own `xerror` is below the dashed horizontal line. From Figure 4.7, the lowest `xerror` was found at a cp = 0.1 with only one split. The resulting tree at that low complexity would be expectedly simple and somewhat uninteresting in terms of predictor interactions. Hence, GLEC pursued interpretation of a tree of more complexity with nsplits = 4 (upper x-axis) and cp = 0.043 (lower x-axis); that model is shown in Figure 4.6 and Table 4.5.

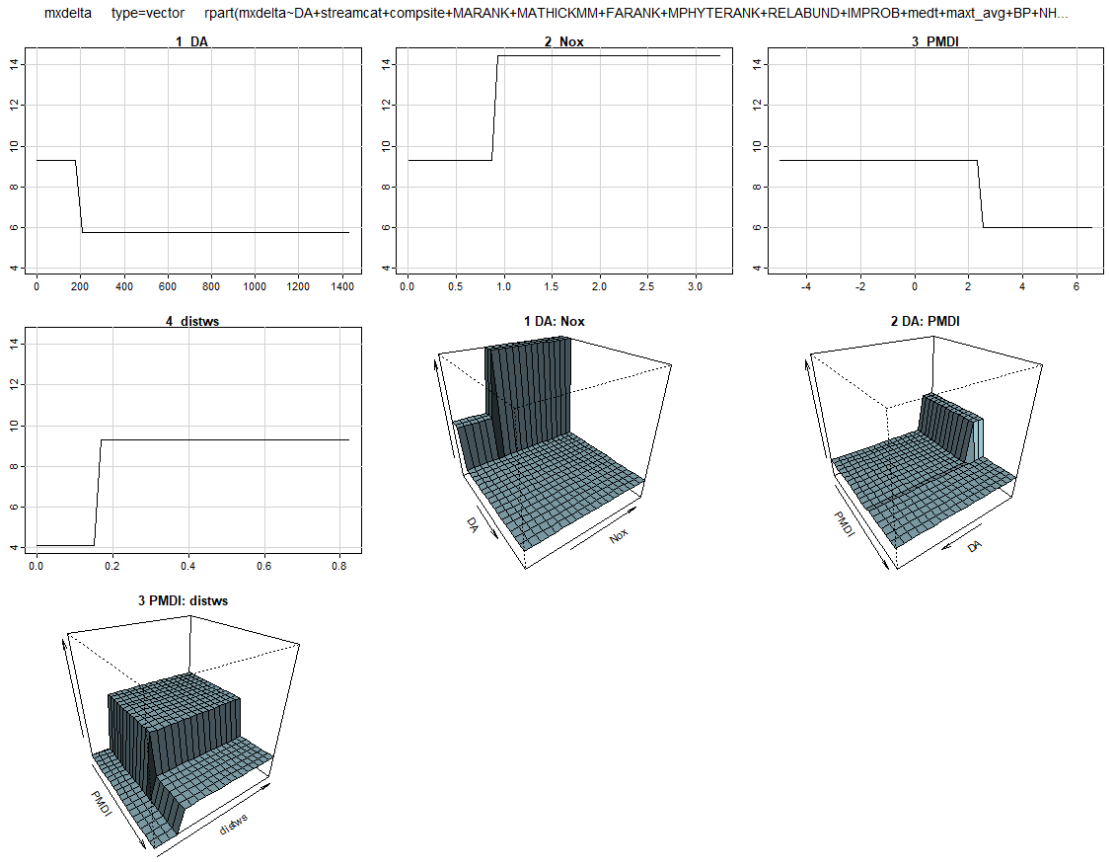


Figure 4.8. Plot of regression tree surface for model formulation of **maximum weekly DO delta** (formulation also listed at the top of the diagram). Upper series of plots show relationships between each predictor variable and the response variable. All variables used in the tree are shown. Lower series of plots shows interactions between a pair of predictor variables, with all other variables held at their median values, and the response variable. The pairs are chosen by which variables appear in the parent-child pairs of the regression tree (see Figure 4.6 or Table 4.4).

Bivariate plots of important predictors vs. the response variable are shown in Figure 4.8 (bottom) and univariate plots of a single predictor and response are shown in the same figure (upper). For the maximum weekly DO delta tree, one can observe, for example in Figure 4.8 (lower left), the delta response from drought only occurs when a threshold of disturbed watershed land cover was exceeded (more than approximately 16 percent of total watershed area).

Table 4.5. Initial model formulation, list of predictor variables used in tree construction after model complexity chosen, root node error, and table showing – in order of decreasing cp (increasing tree complexity) – cp value, number of splits, relative error (rel error), cross-validation error (xerror), and standard deviation of cross-validation error (xstd).

```
Regression tree:

rpart(formula = mxdelta ~ DA + streamcat + compsite + MARANK +
    MATHICKMM + FARANK + MPHYTERANK + RELABUND + IMPROB + medt +
    maxt_avg + BP + NH3 + Nox + OP + pH + SC + TN + TP + TPe +
    TNe + Zindex + PMDI + PHDI + DSCI + DSCt + Dzero + Done +
    Dtwo + Dthree + Dfour + natws + distws + natnf + distnf +
    wells + wellso + wellcat + wellocat + maxslope + medslope +
    xslope + devslope, data = do_wk, method = "anova", x = FALSE,
    y = FALSE, control = rpart.control(cp = 0.043, usesurrogate = 2))

Variables actually used in tree construction:
[1] DA     distws Nox    PMDI

Root node error: 6310/234 = 27

n= 234

      CP nsplit rel error xerror  xstd
1 0.1442      0     1.000  1.005 0.136
2 0.0726      1     0.856  0.953 0.148
3 0.0489      2     0.783  1.004 0.149
4 0.0430      4     0.685  0.974 0.154
```

A plot of explained variance and model complexity (shown as number of splits) indicates that a 1-split model for maximum delta DO was optimal (Figure 4.9, left). But one can observe, after an initial decline, an improvement in explained variance towards a 4-split model. When looking at the distribution of cross-validation error (`xerror`) and model complexity (Figure 4.9, right), the 1-split model provides the best fit to the data but one can then identify a second local minima at a 4-split model. Table 4.5 shows the actual values of `xerror` and `rel error`, with the latter equal to $1\text{-}R^2$.

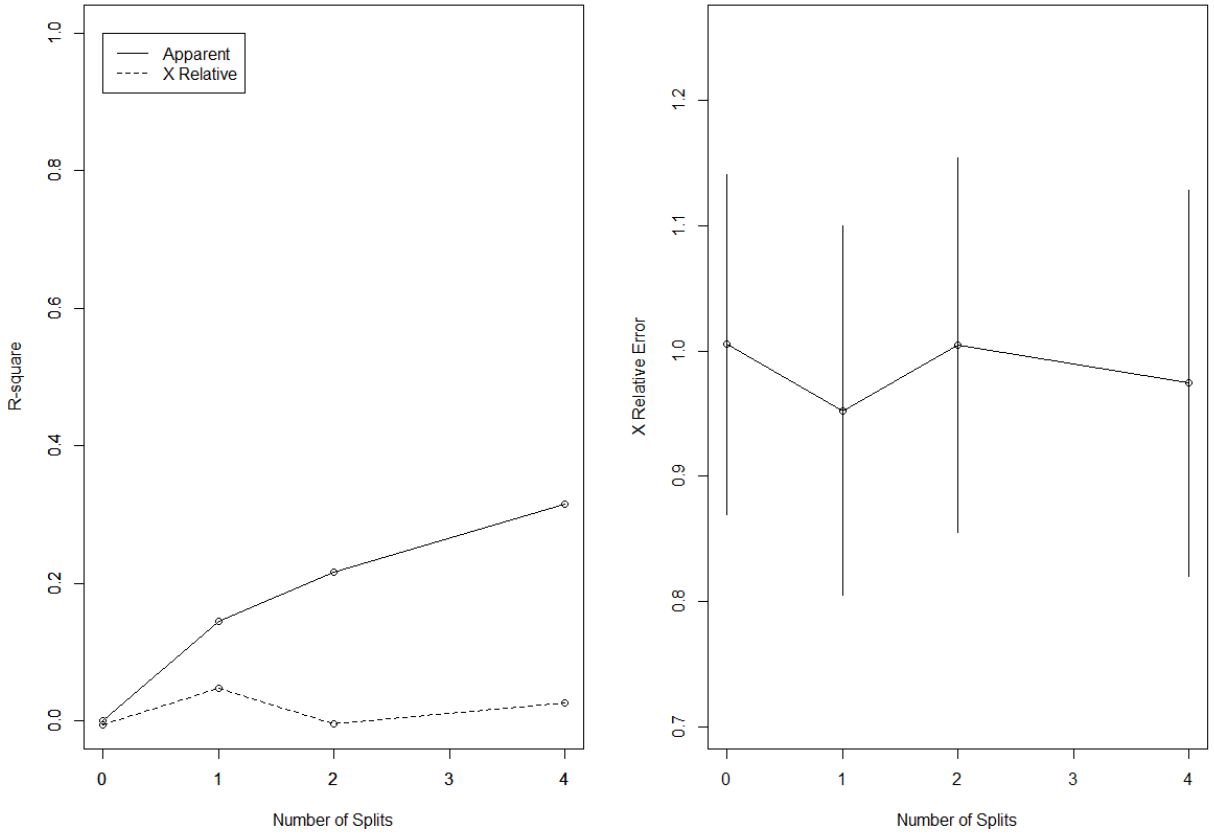


Figure 4.9. (**left**) Diagnostic plot showing explained variance (R2) vs. the number of splits in tree diagram (a measure of complexity). R2 (Apparent) = 1 – rel error for the original model fit; R2 (X Relative) = 1 – xerror for the cross-validation series. (right) Diagnostic plot showing cross-validation relative error (xerror) vs. number of splits. Vertical bars represent ±1 standard deviation (xstd) of xerror.

A residual analysis of the maximum weekly DO delta is shown in Figure 4.10. In case 37 (O'Fallon Creek – Y22OFALC11 on 8/2017 4th week), the model estimates a relatively low maximum delta DO (mxdelta) (the fitted value is 4.12), but the observed level was markedly higher – 18.95 mg/L higher or 23.07 mg/L (Figure 4.10, upper right). Note, case 37 also appeared as an outlier in the model fit for mean weekly delta (see Section 4.a.i). From the QQ plot (bottom left), case 37 is the biggest absolute residual (x-axis) and thus determines the right bound of the cumulative distribution plot (top left of Figure 4.10). Other unusual residuals are cases 74 (Y22SNDSC06 on the 2nd week of 9/2013) and 214 (Y27LBVRC06 on the 1st week of 10/2013), having the same pattern as case 37 but are not further discussed.

Also, in the cumulative distribution plot (Figure 4.10, top left), one observes that 50% of the observations have a maximum delta residual of less than 2.5 mg/L, and 75% of the observations

have a residual of less than 4 mg/L. Both residual magnitudes were low, especially for maximum DO delta, and suggests a good model fit.

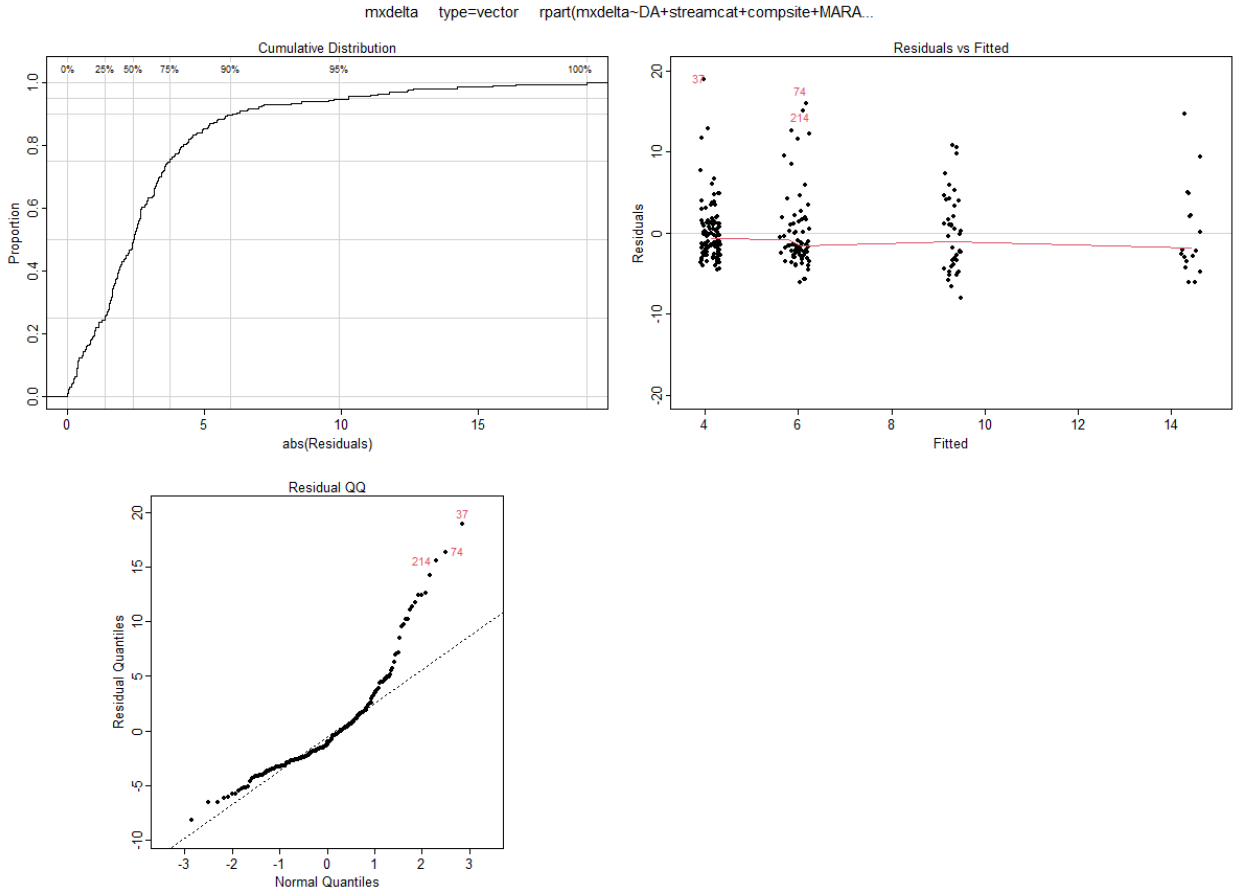


Figure 4.10. Residual analysis showing (**upper left**) cumulative distribution (proportion) of residuals (their absolute value) for entire dataset, (**upper right**) residual magnitude vs. predicted value from regression tree model (red line is loess-fitted value and observations with a high residual are noted by their record number (red font) in the dataset), and (**lower left**) quantile-quantile (QQ) plot of residuals. A positive residual indicates the observed value exceeds the predicted (fitted) value (i.e., the model underpredicts the observed).

4.a.iii. Dissolved Oxygen – Average Minimum

The regression tree model with a response of average weekly DO minimum contains seven splits and employs six predictor variables (Figure 4.11; Table 4.6). The primary predictor for the DO minimum as a response variable is weekly median water temperature (medt), and this would be expected as DO saturation concentration is primarily dependent on temperature. One can observe higher DO minimums (indicating health aquatic systems) when median water temperature is below 15°C. Watershed drainage area (DA) plays the next most important role; DO minimum is highest in only the largest systems (14 of the 73 study watersheds).

Stressed aquatic systems may be indicated by very low weekly average minimum DO. There are three key leaf nodes found on Figure 4.11 – (1) high water temperature in a smaller watershed with low moisture to very dry meteorological conditions (far-left branch), (2) lower water temperatures in a smaller watershed with low pH (middle branch), or (3) with conditions of high pH, low count of older oil/gas wells, and relative abundance of nutrient taxa. The first scenario suggests low flow, warm aquatic systems where minimum DO would be lower particularly during dry conditions. The second scenario might be explained as that which occurs during respiration in non-daylight hours[25]. Here water temperatures would be expected to be lower without radiational heating. The pH tends to follow the DO diel hourly trace (peaking in late afternoon) as CO_2 is consumed during daylight from photosynthesis, but released at night during respiration. The third scenario would be difficult to explain on a physico-chemical basis and may be due to model over-fitting.

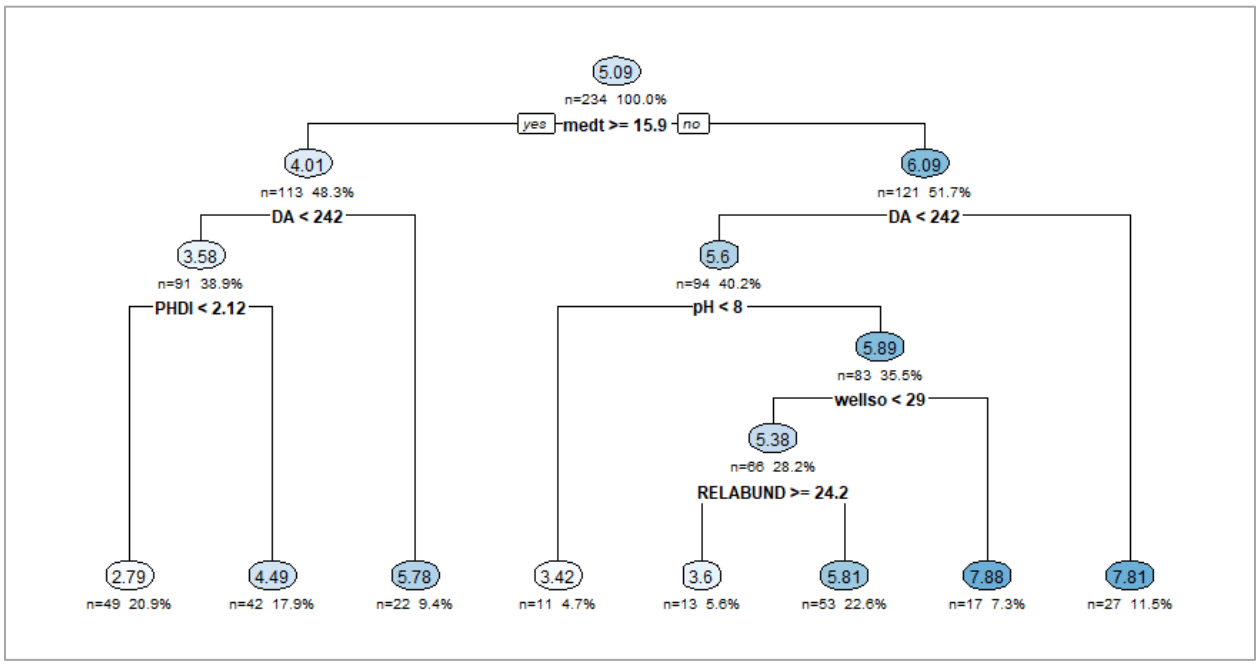


Figure 4.11. Diagram showing the regression tree for a response of **mean weekly DO minimum** (mg/L). The predicted value and the number and percentage of total observations are shown for each node. The intensity of the node color is proportional to magnitude of the predicted value. The decision statement to split is located under each node (in bold) – traverse left if the statement is true (yes), otherwise traverse right (no).

[25] Another possibility, as proposed by Montana DEQ, is that the time-of-day for when the sample was collected might better determine the magnitude of pH. If a sample was collected early in the day (when DO is at or near its diel minimum), then pH would be corresponding low. Conversely, if the sample was collected later in the day (when DO is near its diel maximum), then pH would also be at its diel peak. Further separation of samples by time-of-day might be warranted to better understand this influence.

Table 4.6. The regression tree model shown above (Figure 4.11) portrayed as a set of decision rules. One set of statements describes a particular leaf node. Primary statement begins with mean predicted response (rounded to one decimal place) and its value (units defined above). Cover percent refers to percent of total observations in the particular node. Secondary statements are decisions about the important predictor variables to reach the leaf node.

```
min_avg is 2.8 with cover 21% when
    medt >= 16
    DA < 242
    PHDI < 2.1

min_avg is 3.4 with cover 5% when
    medt < 16
    DA < 242
    pH < 8

min_avg is 3.6 with cover 6% when
    medt < 16
    DA < 242
    pH >= 8
    wellso < 29
    RELABUND >= 24

min_avg is 4.5 with cover 18% when
    medt >= 16
    DA < 242
    PHDI >= 2.1

min_avg is 5.8 with cover 9% when
    medt >= 16
    DA >= 242

min_avg is 5.8 with cover 23% when
    medt < 16
    DA < 242
    pH >= 8
    wellso < 29
    RELABUND < 24

min_avg is 7.8 with cover 12% when
    medt < 16
    DA >= 242

min_avg is 7.9 with cover 7% when
    medt < 16
    DA < 242
    pH >= 8
    wellso >= 29
```

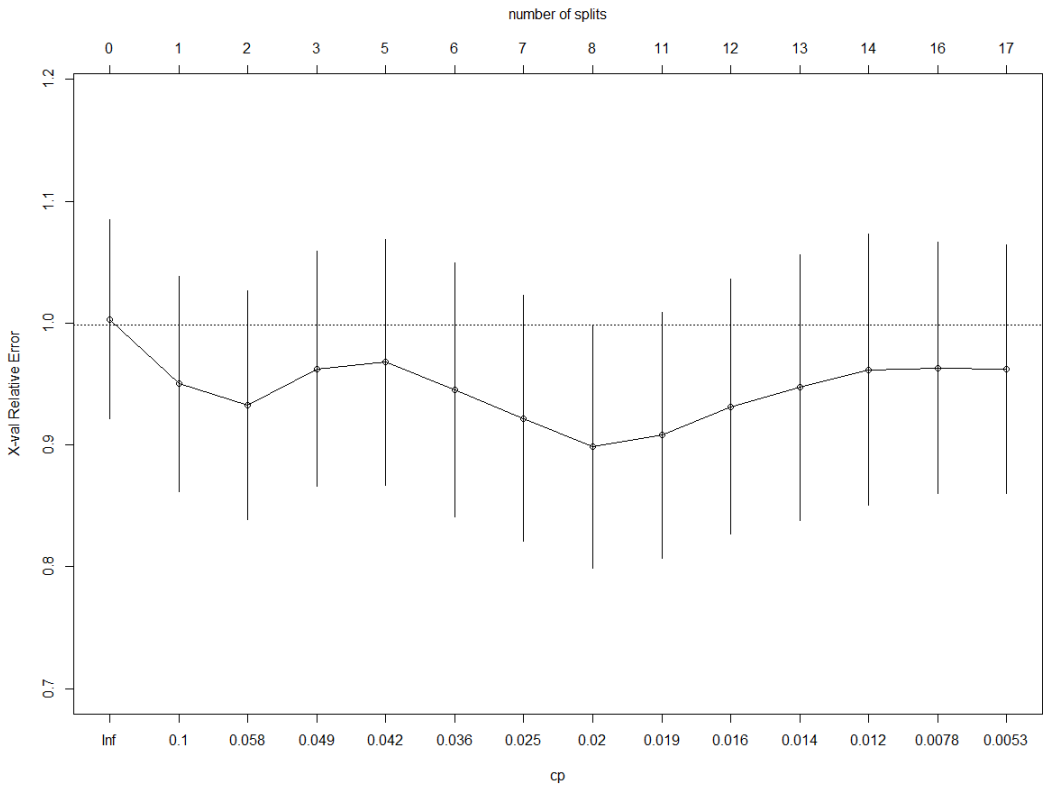


Figure 4.12. Initial plot of cross-validation error (*xerror*) vs. model complexity (cp) or number of splits in tree (upper x-axis). Error bars for *xerror* equal to ±1 error standard deviation (*xstd*). Dashed horizontal line is placed at +1 *xstd* above the lowest modeled *xerror*.

Figure 4.12 shows the change in the cross-validation error with respect to changing complexity. The most complex tree is on the right of the plot. The tree is subsequently pruned to a complexity just less than (i.e., at a larger cp) the lowest cross-validation error as long as its own `xerror` is below the dashed horizontal line. From Figure 4.12, the lowest xerror is obviously found at a cp = 0.02 so a tree of complexity nsplits = 7 (upper x-axis) and cp = 0.025 (lower x-axis) was used to build the model shown in Figure 4.11 and Table 4.7.

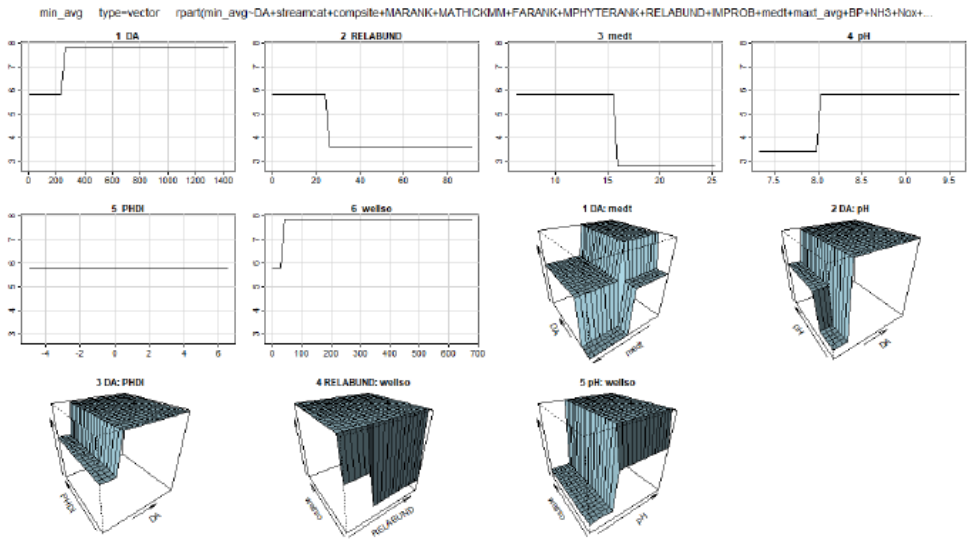


Figure 4.13. Plot of regression tree surface for model formulation of **mean weekly DO minimum** (formulation also listed at the top of the diagram). Upper series of plots show relationships between each predictor variable and the response variable. All variables used in the tree are shown. Lower series of plots shows interactions between a pair of predictor variables, with all other variables held at their median values, and the response variable. The pairs are chosen by which variables appear in the parent-child pairs of the regression tree (see Figure 4.11 or Table 4.6).

Bivariate plots of important predictors vs. the response variable are shown in Figure 4.13 (bottom) and univariate plots of a single predictor and response are shown in the same figure (upper). For the average weekly DO minimum tree, one can observe the minimum DO response from high water temperatures only occurs at smaller drainage areas (Figure 4.13, lower left).

Table 4.7. Initial model formulation, list of predictor variables used in tree construction after model complexity chosen, root node error, and table showing – in order of decreasing cp (increasing tree complexity) – cp value, number of splits, relative error (rel error), cross-validation error (xerror), and standard deviation of cross-validation error (xstd).

```
Regression tree:

rpart(formula = min_avg ~ DA + streamcat + compsite + MARANK +
    MATHICKMM + FARANK + MPHYTERANK + RELABUND + IMPROB + medt +
    maxt_avg + BP + NH3 + Nox + OP + pH + SC + TN + TP + TPe +
    TNe + Zindex + PMDI + PHDI + DSCI + DSCt + Dzero + Done +
    Dtwo + Dthree + Dfour + natws + distws + natnf + distnf +
    wells + wellso + wellcat + wellocat + maxslope + medslope +
    xslope + devslope, data = do_wk, method = "anova", x = FALSE,
    y = FALSE, control = rpart.control(cp = 0.025, usesurrogate = 2))

Variables actually used in tree construction:
[1] DA       medt     pH       PHDI     RELABUND wellso

Root node error: 1614/234 = 6.9

n= 234

      CP nsplit rel error xerror   xstd
1 0.1574      0     1.000  1.020 0.0831
2 0.0634      1     0.843  0.922 0.0861
3 0.0535      2     0.779  0.940 0.0901
4 0.0446      3     0.726  0.890 0.0901
5 0.0404      5     0.637  0.953 0.1097
6 0.0316      6     0.596  0.931 0.1096
7 0.0250      7     0.565  0.934 0.1115
```

A plot of explained variance and model complexity (shown as number of splits) indicates that either a 5- or 7-split model for average DO minimum was optimal (Figure 4.14, left). GLEC chose a 7-split model for this response variable. When looking at the distribution of cross-validation error (xerror) and model complexity (Figure 4.14, right), the 3-split model is optimal. As mentioned in Section 4.a.i, partitioning in cross-validation is random so that repeating the method will give slightly different numerical results. Table 4.7 shows the actual values of `xerror` and `rel error`, with the latter equal to $1\text{-}R^2$.

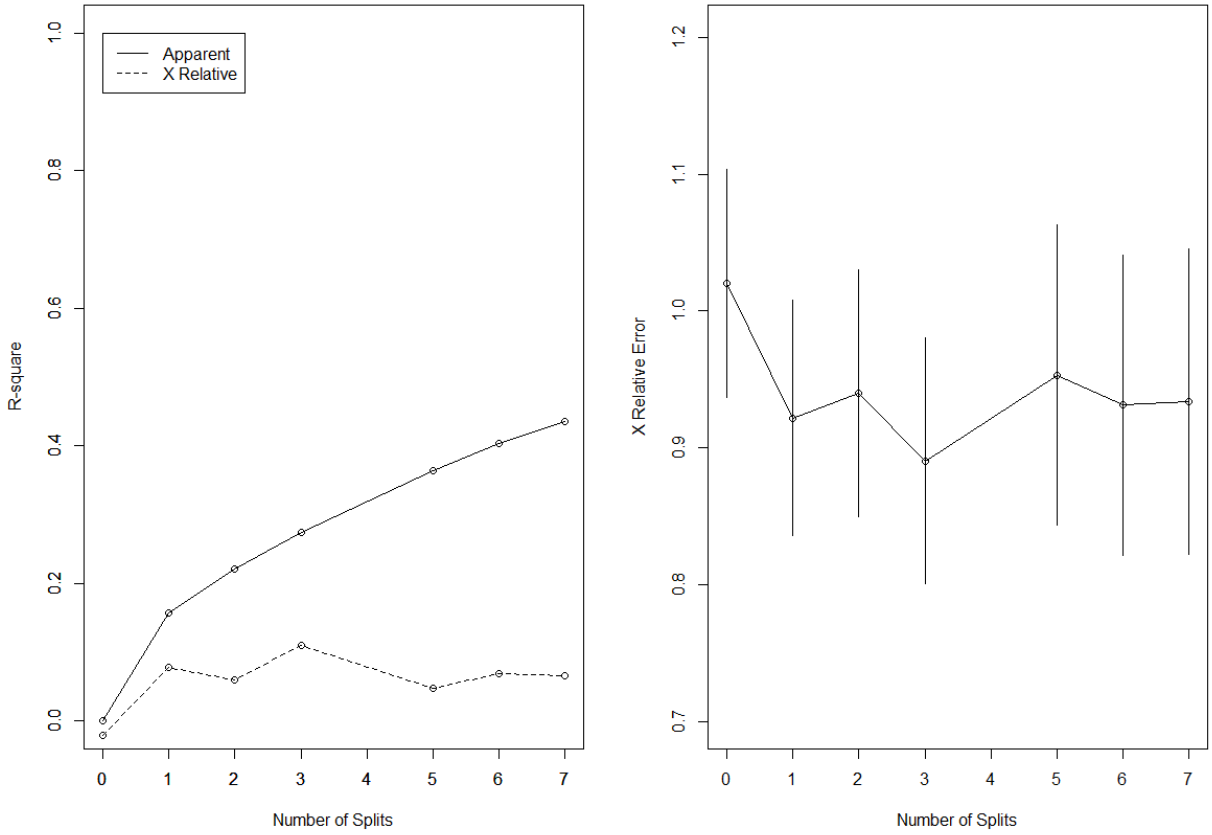


Figure 4.14. (**left**) Diagnostic plot showing explained variance (R^2) vs. the number of splits in tree diagram (a measure of complexity). R^2 (Apparent) = 1 – `rel error` for the original model fit; R^2 (X Relative) = 1 – `xerror` for the cross-validation series. (**right**) Diagnostic plot showing cross-validation relative error (`xerror`) vs. number of splits. Vertical bars represent ±1 standard deviation (`xstd`) of `xerror`.

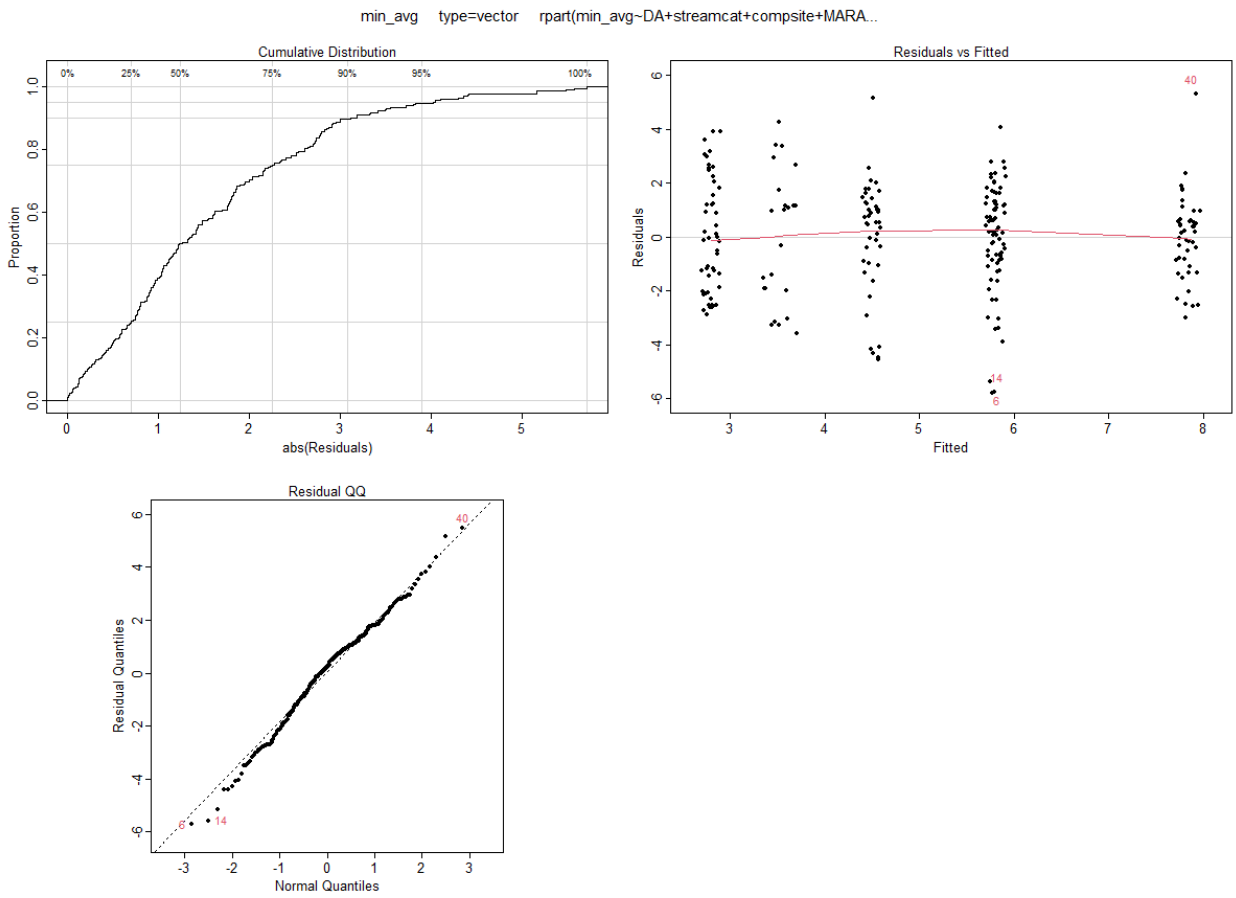


Figure 4.15. Residual analysis showing (**upper left**) cumulative distribution (proportion) of residuals (their absolute value) for entire dataset, (**upper right**) residual magnitude vs. predicted value from regression tree model (red line is loess-fitted value and observations with a high residual are noted by their record number (red font) in the dataset), and (**lower left**) quantile-quantile (QQ) plot of residuals. A positive residual indicates the observed value exceeds the predicted (fitted) value (i.e., the model underpredicts the observed).

The residual analysis of the DO minimum regression tree shows a better-behaved model (Figure 4.15). The scatterplot of residuals vs. fitted (upper-right) is very evenly distributed – the LOESS line nearly follows the zero-residual line, and the QQ scatterplot has little deviation (bottom-left). There are a few outlier residuals – cases 6 and 14 show the model overpredicted the DO minimum. Case 6 is Pasture Creek (M48PSTRC02) on the 2nd week of 10/2016, and case 14 is Charlie Creek East (M51CHLYC04) on the 2nd week of 10/2017. For Charlie Creek East, the explanation is simple – only two sampling days occurred during the 2nd week of 10/2017. When examining the observed data for Pasture Creek during that time frame, every day experienced the same DO minimum of 0.01 mg/L. Note, the maximum DO observed during the same week was only 0.11 mg/L. This pattern of minimum and maximum for Pasture Creek was seen for 25 consecutive days (September-October 2016) and may be suggestive of a major hypoxic event (e.g., a spill of high organic matter into the stream system).

4.a.iv. Dissolved Oxygen – Exceed Delta Threshold (MT) (count)

The first three regression tree models discussed above were built with a quantitative response variable (model method type was anova). In this model with a count of delta threshold exceedance, the Poisson model type is chosen (Table 4.9). A count variable is a rate, and in this case it is the number of threshold exceedances per week (per 7 days).

The regression tree model has a response of number of days per week that a daily delta DO exceeds a critical threshold of 5.3 mg/L. The threshold was derived by Montana DEQ in their guidance for determining wadeable stream nutrient impairment (Suplee and Sada 2016). This count model contains five splits and employs four predictor variables (Figure 4.16; Table 4.8). Like the other DO delta models (average and maximum), there is a primary relationship between disturbed land cover and drought (right-branch). Here as in most of the right branches are a large number of daily exceedances per week. For example, nearly every day of the week experience an exceedance (mean exceedance of 6.84 days) when disturbed land cover in the watershed exceeds 33% of total area and drought is severe (PMDI less than -4.8 considered extreme drought). The same physical explanation as describe in Section 4.a.i can be applied here. In the left branch, where disturbed land cover is low, the minimal presence of macrophytes (0,1 – the two lowest categories) results in the lowest number of exceedences of the 5.3 mg/L threshold in the entire tree, whereas higher densities of macrophytes (2,3,4 – the three highest categories) led to a near doubling of exceedences (2.39/week). This finding is consistent with the observation that macrophyte photosynthesis contributes to DO supersaturation and (therefore) more exceedences of the threshold.

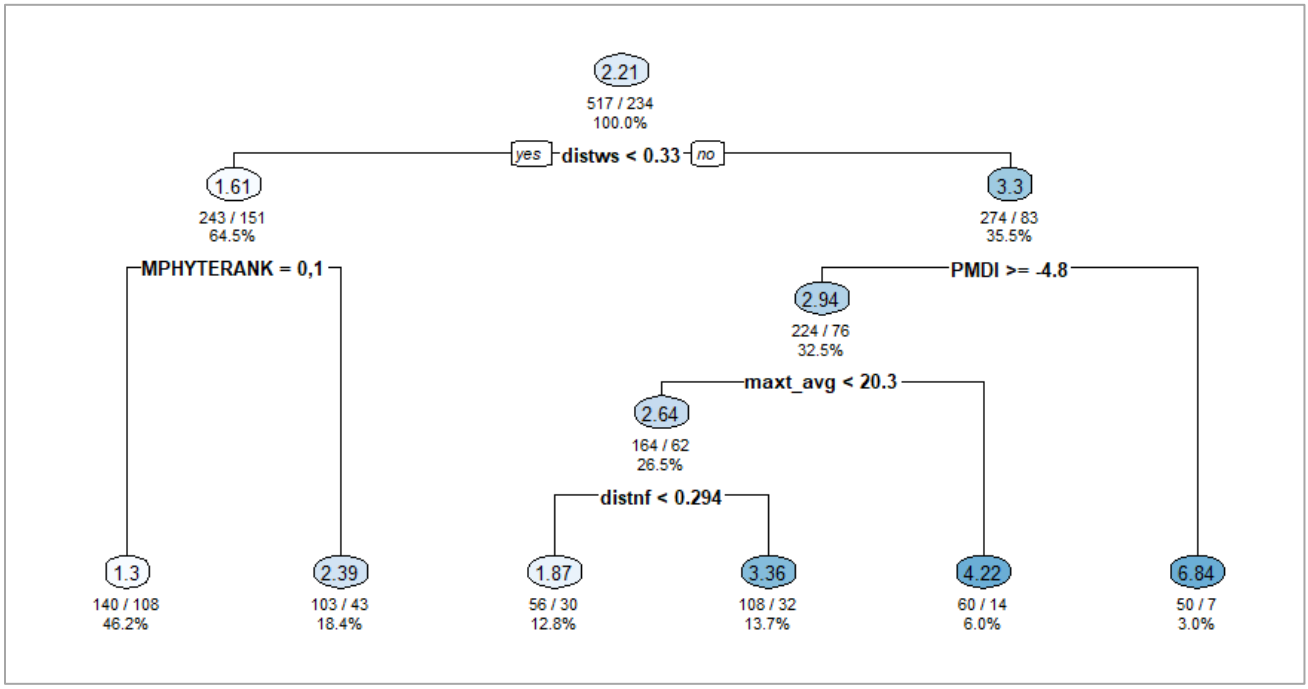


Figure 4.16. Diagram showing the regression tree for a response of number of exceedances per week of a critical threshold (medium-level threshold of 5.3 mg/L). Shown for each node is the predicted value, then a pair separated by "/" listing the total number of events (1 event = 1 day of

exceedance) and the number of observations, and the percentage of total observations. The intensity of the node color is proportional to magnitude of the predicted value. The decision statement to split is located under each node (in bold) – traverse left if the statement is true (yes), otherwise traverse right (no).

Table 4.8. The regression tree model shown above (Figure 4.16) portrayed as a set of decision rules. One set of statements describes a particular leaf node. Primary statement begins with mean predicted response (rounded to one decimal place) and its value (units defined above). Cover percent refers to percent of total observations in the particular node. Secondary statements are decisions about the important predictor variables to reach the leaf node.

```
exceedMT is 1.3 with cover 46% when
    distws < 0.33
    MPHYTERANK is 0 or 1

exceedMT is 1.9 with cover 13% when
    distws >= 0.33
    PMDI >= -4.8
    maxt_avg < 20
    distnf < 0.29

exceedMT is 2.4 with cover 18% when
    distws < 0.33
    MPHYTERANK is 2 or 3 or 4

exceedMT is 3.4 with cover 14% when
    distws >= 0.33
    PMDI >= -4.8
    maxt_avg < 20
    distnf >= 0.29

exceedMT is 4.2 with cover 6% when
    distws >= 0.33
    PMDI >= -4.8
    maxt_avg >= 20

exceedMT is 6.8 with cover 3% when
    distws >= 0.33
    PMDI < -4.8
```

Figure 4.17 shows the change in the cross-validation error with respect to changing complexity. The most complex tree is on the right of the plot. The tree is subsequently pruned to a complexity just less than (i.e., at a larger cp) the lowest cross-validation error as long as its own `xerror` is below the dashed horizontal line. From Figure 4.17, the lowest `xerror` is found at a cp = 0.035 so a tree of complexity nsplits = 5 (upper x-axis) and cp = 0.039 (lower x-axis) was used to build the model shown in Figure 4.16 and Table 4.9.

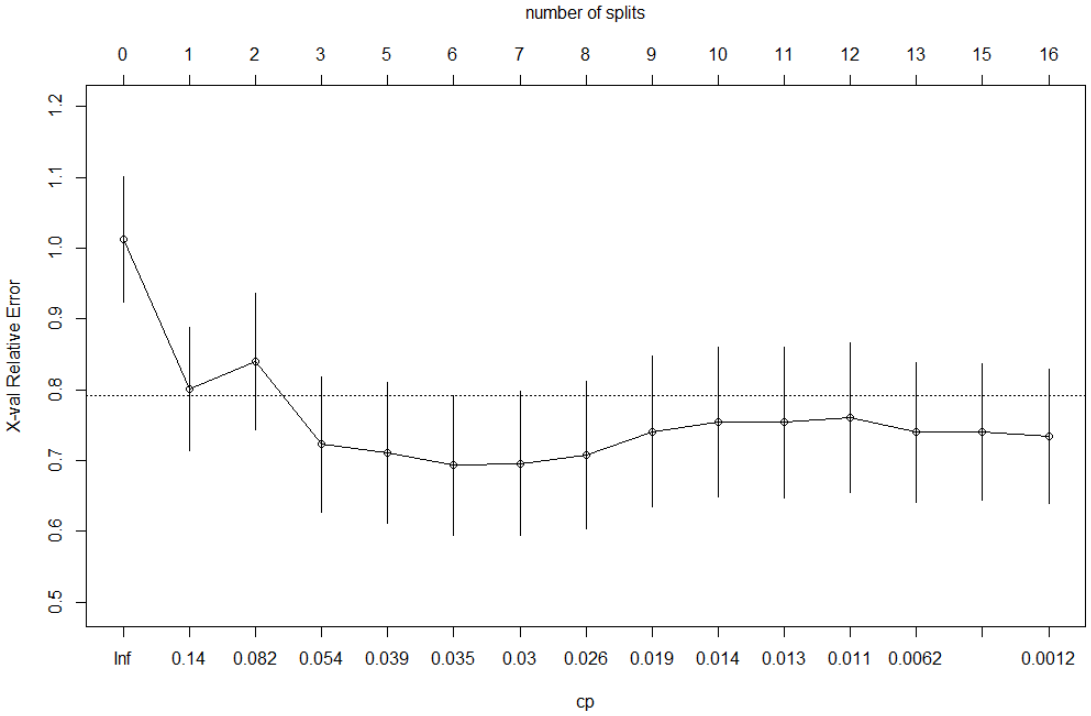


Figure 4.17. Initial plot of cross-validation error (`xerror`) vs. model complexity (cp) or number of splits in tree (upper x-axis). Error bars for `xerror` equal to ±1 error standard deviation (`xstd`). Dashed horizontal line is placed at +1 `xstd` above the lowest modeled `xerror`.

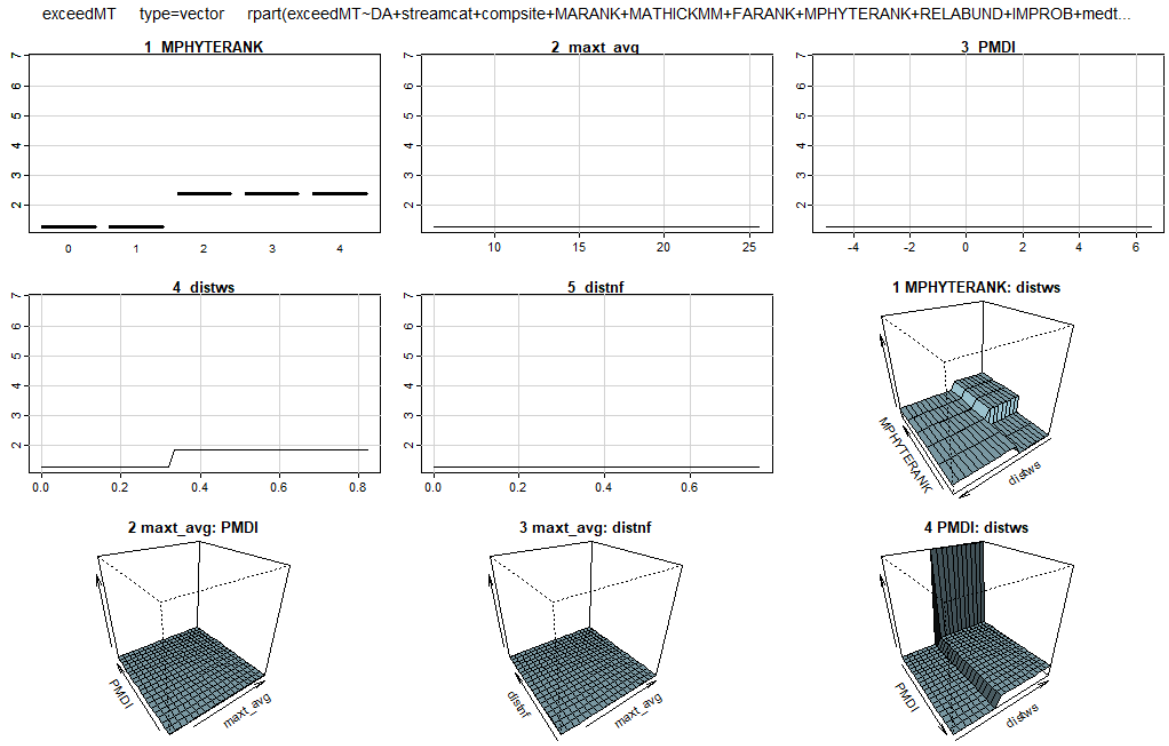


Figure 4.18. Plot of regression tree surface for model formulation of **number of exceedances per week** of a critical threshold (medium threshold of 5.3 mg/L) (formulation also listed at the top of the diagram). Upper series of plots show relationships between each predictor variable and the response variable. All variables used in the tree are shown. Lower series of plots shows interactions between a pair of predictor variables, with all other variables held at their median values, and the response variable. The pairs are chosen by which variables appear in the parent-child pairs of the regression tree (see Figure 4.16 or Table 4.8).

Bivariate plots of important predictors vs. the response variable are shown in Figure 4.18 (bottom) and univariate plots of a single predictor and response are shown in the same figure (upper). For the number of exceedances per week tree, one can observe the highest number of exceedances at very low values of PMDI (extreme drought) and high percent area of disturbed land cover (Figure 4.18, upper right).

Table 4.9. Initial model formulation, list of predictor variables used in tree construction after model complexity chosen, root node error, and table showing – in order of decreasing cp (increasing tree complexity) – cp value, number of splits, relative error (rel error), cross-validation error (xerror), and standard deviation of cross-validation error (xstd).

```
Rates regression tree:

rpart(formula = exceedMT ~ DA + streamcat + compsite + MARANK +
    MATHICKMM + FARANK + MPHYTERANK + RELABUND + IMPROB + medt +
    maxt_avg + BP + NH3 + Nox + OP + pH + SC + TN + TP + TPe +
    TNe + Zindex + PMDI + PHDI + DSCI + DSCt + Dzero + Done +
    Dtwo + Dthree + Dfour + natws + distws + natnf + distnf +
    wells + wellso + wellcat + wellocat + maxslope + medslope +
    xslope + devslope, data = do_wk, method = "poisson", control =
rpart.control(cp = 0.039,
    usesurrogate = 2))

Variables actually used in tree construction:
[1] distnf     distws     maxt_avg   MPHYTERANK PMDI

Root node error: 289/234 = 1.24

n= 234

      CP nsplit rel error xerror   xstd
1 0.2285      0     1.000  1.010 0.0878
2 0.0910      1     0.772  0.804 0.0866
3 0.0740      2     0.681  0.830 0.0962
4 0.0398      3     0.607  0.783 0.0986
5 0.0390      5     0.527  0.780 0.0995
```

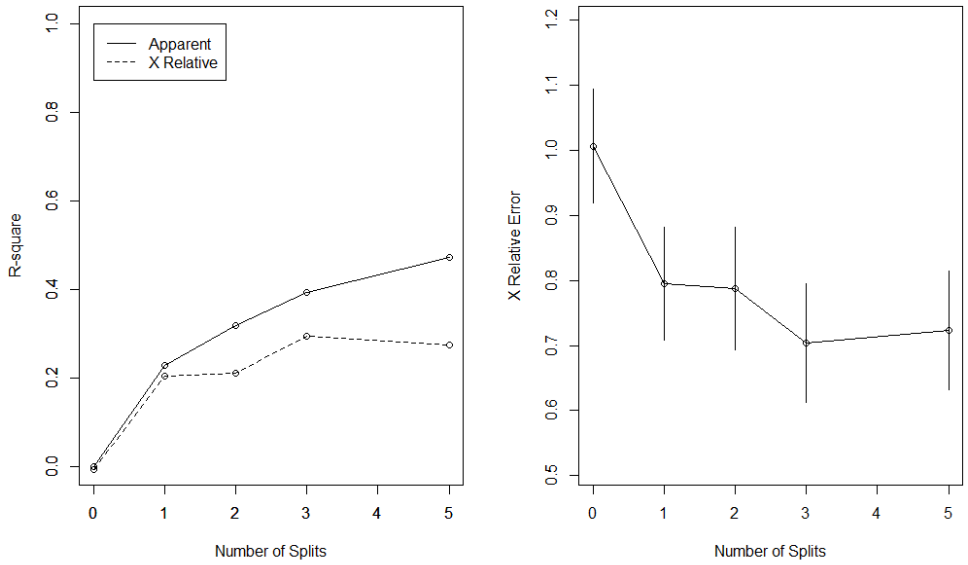


Figure 4.19. (**left**) Diagnostic plot showing explained variance (R^2) vs. the number of splits in tree diagram (a measure of complexity). R^2 (Apparent) = 1 – `rel error` for the original model fit; R^2 (X Relative) = 1 – `xerror` for the cross-validation series. (**right**) Diagnostic plot showing cross-validation relative error (`xerror`) vs. number of splits. Vertical bars represent ±1 standard deviation (`xstd`) of `xerror`.

A plot of explained variance and model complexity (shown as number of splits) indicates that either a 3- or 5-split model for exceedance count is optimal (Figure 4.19, left). GLEC chose a 5-split model for this response variable. When looking at the distribution of cross-validation error (`xerror`) and model complexity (Figure 4.19, right), the 3- or 5-split model are optimal. Table 4.9 shows the actual values of `xerror` and `rel error`, with the latter equal to $1\text{-}R^2$.

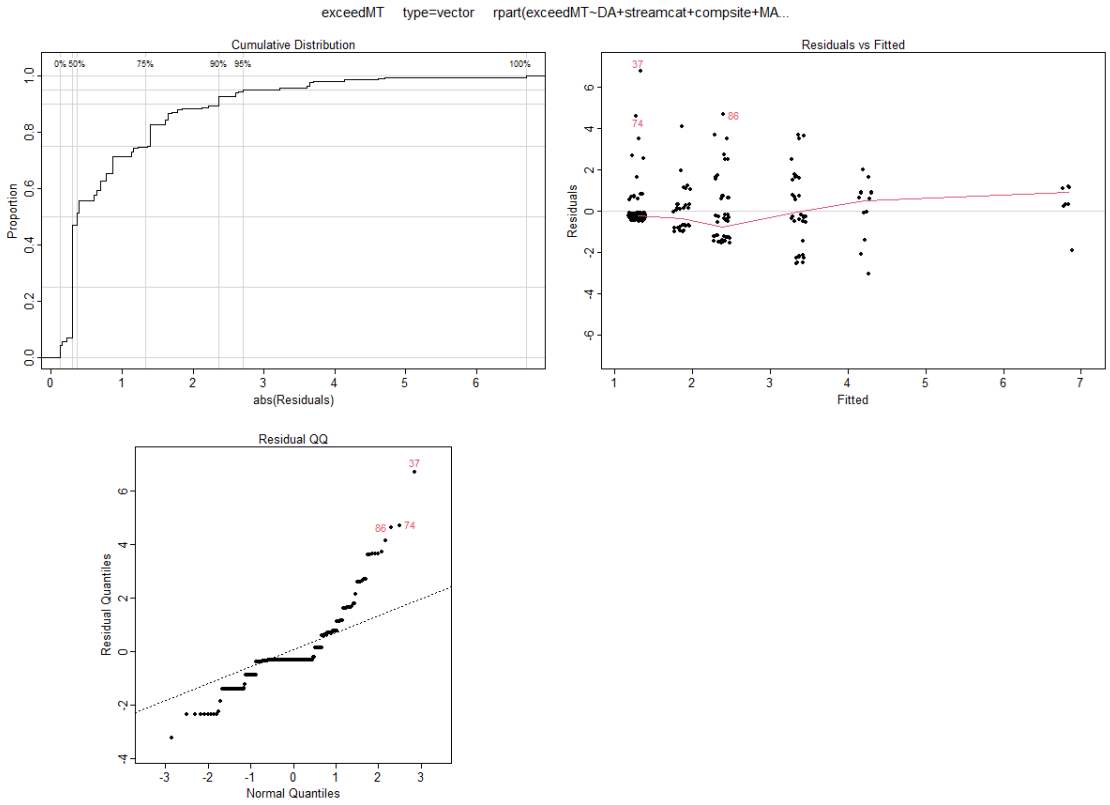


Figure 4.20. Residual analysis showing (**upper left**) cumulative distribution (proportion) of residuals (their absolute value) for entire dataset, (**upper right**) residual magnitude vs. predicted value from regression tree model (red line is loess-fitted value and observations with a high residual are noted by their record number (red font) in the dataset), and (**lower left**) quantile-quantile (QQ) plot of residuals. A positive residual indicates the observed value exceeds the predicted (fitted) value (i.e., the model underpredicts the observed).

The residual analysis of the count exceedance regression tree shows the least-behaved model seen so far in this study (Figure 4.20). The scatterplot of residuals vs. fitted (upper-right) is unevenly distributed, and the QQ scatterplot has high deviation (bottom-left). This model predicts counts so the fitted values are 1 through 7 days (per week). No qualitative explanation, as done for selected outliers in models shown above, was made here.

In the cumulative distribution plot (Figure 4.20, top left), one observes that 50% of the observations have an exceedance residual of less than 0.3 days/week, and 75% of the observations have a residual

of less than 1.25 mg/L. Both residual magnitudes are low and suggests a good model fit for most of the observations in the study dataset.

4.b.i. Aquatic Plant – Microalgae Thickness (rank)

The next two regression tree models use aquatic plants as a response variable (this section and Section 4.b.ii). While the DO delta, exceedance, and minimum constructions are the ultimate response (Figure 2.2), GLEC sought to understand the important predictors for aquatic plant indices collected by Montana DEQ, with themselves behaving as both predictor and response variable.

Both aquatic plant responses are ranked variables (ordinal) and use method "class" in the `rpart` technique (see Tables 4.11 and 4.13); these models should more accurately be called classification trees as opposed to regression trees discussed above. Ordinal responses are discrete and have an order or rank to them.

Here microalgae thickness (in mm) is the response, but as discussed previously (Section 3.a.iv), the aquatic *visual* assessment, by its nature, allows for only five entries in a field observation ("absent" through "very heavy"). The classification tree model contains two splits and employs only two predictor variables (Figure 4.21; Table 4.10). Both predictors are themselves aquatic plant indices – percent cover of both macrophytes and filamentous algae. When macrophyte cover is all but low (rank of 1, 2, 3, or 4), microalgae thickness is moderately higher. In addition, when macrophyte cover was low but filamentous algal cover is anything but absent (i.e., "sparse" through "very heavy"), microalgae thickness is also moderately higher. These findings suggest that conditions which stimulate macrophytes and filamentous algae can also increase microalgae thickness (nutrient enrichment could induce this, for example). One also may observe that when both macrophytes and filamentous algae have low cover, microalgae thickness was also low, all indicating heavily shaded, very turbid, and/or low nutrient-loaded stream systems.

Perhaps because so few observations occur in the higher classes of microalgae thickness (1.8 and 3.0 mm), the tree model was not able to identify other important predictor variables. One would expect that land cover type or nutrient chemistry would play a role in algae growth.

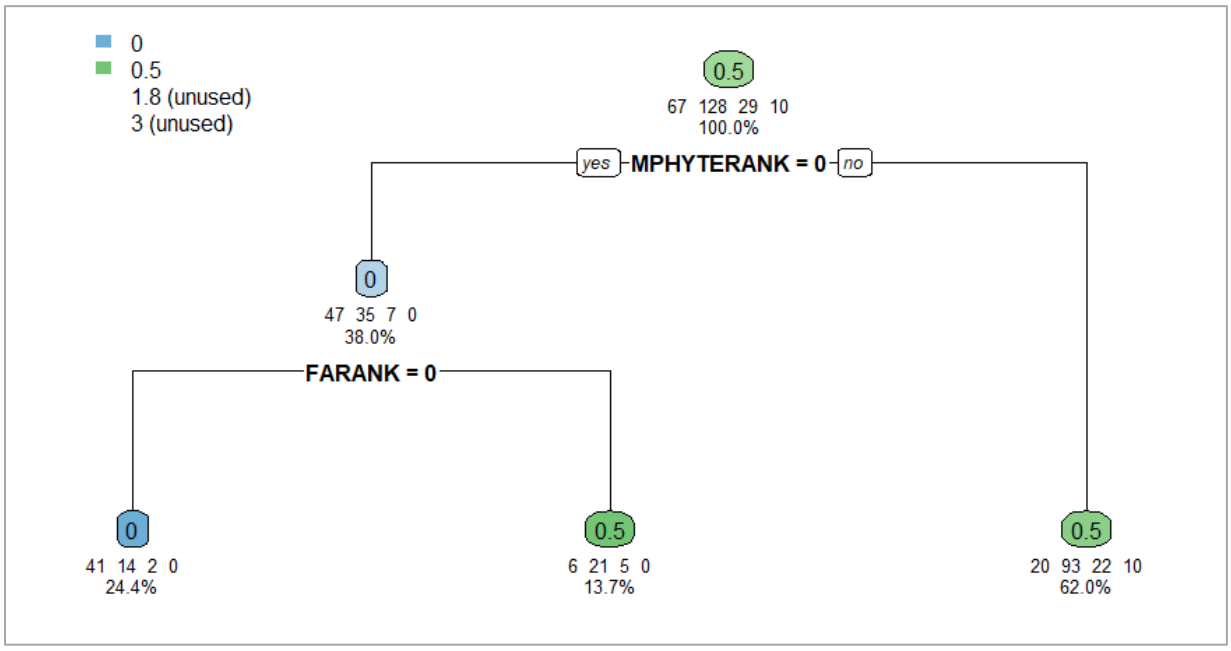


Figure 4.21. Diagram showing the regression tree for a *ranked* response of **microalgae thickness** (mm). The predicted value, the number of observations in each class, and the percentage of total observations are shown for each node. The predicted value is 1 of 4 classes equivalent to 0, 0.5, 1.8, and 3 mm, though the upper two classes were not used in the model because so few observations were available. The intensity of the node color is proportional to magnitude of the predicted value. The decision statement to split is located under each node (in bold) – traverse left if the statement is true (yes), otherwise traverse right (no). MPHYTERANK and FARANK are ranks of percent cover for macrophyte and filamentous algae species, respectively.

Table 4.10. The regression tree model shown above (Figure 4.21) portrayed as a set of decision rules. One set of statements describes a particular leaf node. Primary statement begins with predicted response (class) and its value (units defined above) followed by, in brackets [ ], the percentage of node observations in each class. Cover percent refers to percent of total observations in the particular node. Secondary statements are decisions about the important predictor variables to reach the leaf node.

```
                        0 0.5 1.8   3
MATHICKMM is 0    [.72 .25 .04 .00] with cover 24%
when
    MPHYTERANK is 0
    FARANK is 0

MATHICKMM is 0.5 [.14 .64 .15 .07] with cover 62%
when
    MPHYTERANK is 1 or 2 or 3 or 4

MATHICKMM is 0.5 [.19 .66 .16 .00] with cover 14%
when
    MPHYTERANK is 0
    FARANK is 1 or 2 or 3 or 4
```

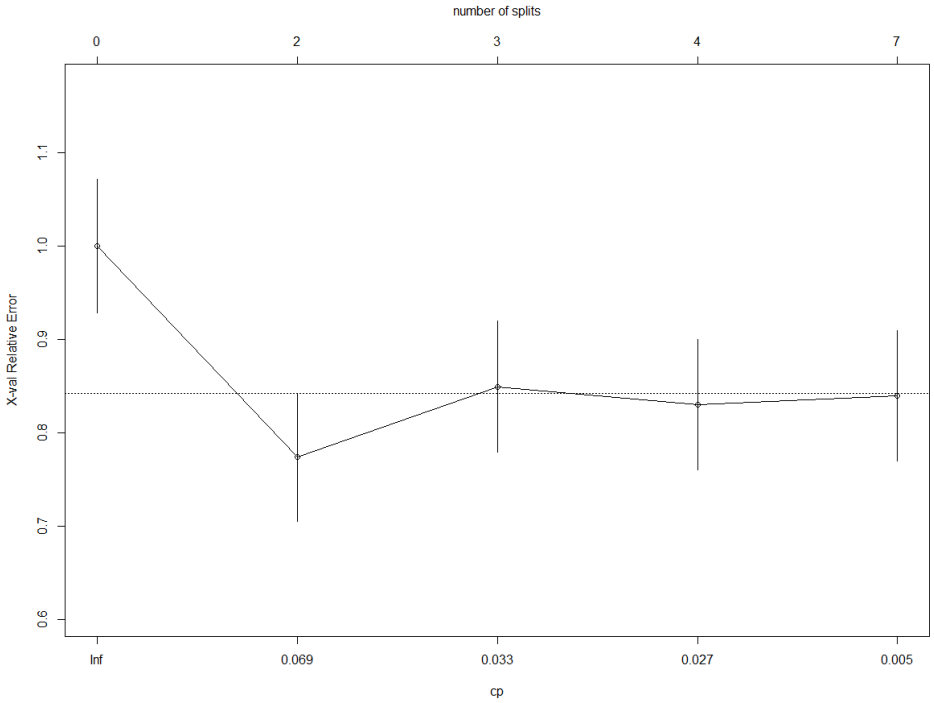


Figure 4.22. Initial plot of cross-validation error (`xerror`) vs. model complexity (cp) or number of splits in tree (upper x-axis). Error bars for `xerror` equal to ±1 error standard deviation (`xstd`). Dashed horizontal line is placed at +1 `xstd` above the lowest modeled `xerror`.

Figure 4.22 shows the change in the cross-validation error with respect to changing complexity. The most complex tree is on the right of the plot. The tree is subsequently pruned to a complexity just less than (i.e., at a larger cp) the lowest cross-validation error as long as its own `xerror` is below the dashed horizontal line. From Figure 4.22, the lowest `xerror` is found at a cp = 0.069 which would suggest running the model at nsplits = 0. To produce some level of predictor interpretation, GLEC subsequently built at model at cp = 0.069 with two splits (Figure 4.21 and Table 4.10).

Bivariate plots of important predictors vs. the response variable are shown in Figure 4.23 (bottom) and univariate plots of a single predictor and response are shown in the same figure (upper). The classification tree is slightly more understandable when shown as a 3-dimensional surface (bottom). Class 2 (rank = 1 or "thin microalgae thickness") is considered the maximum rank in this model.

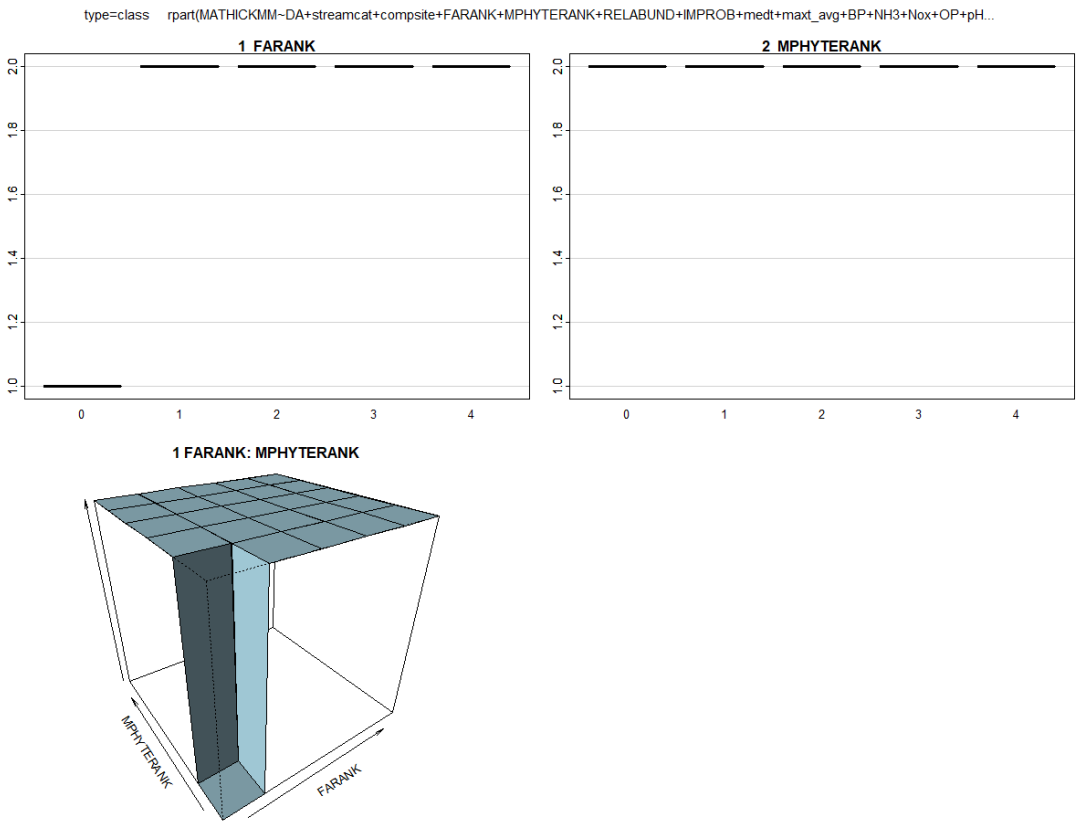


Figure 4.23. Plot of regression tree surface for model formulation of **microalgae thickness** (formulation also listed at the top of the diagram). Upper pair of plots show relationships between each predictor variable and the response variable. Response values are class number where class 1=rank 0 or "absent" and class 2=rank 1 or "thin". All variables used in the tree are shown. Lower plot shows interactions between a pair of predictor variables, with all other variables held at their median values, and the response variable. The pair chosen by which variables appear in the parent-child pairs of the regression tree (see Figure 4.21 or Table 4.10).

Table 4.11. Initial model formulation, list of predictor variables used in tree construction after model complexity chosen, root node error, and table showing – in order of decreasing cp (increasing tree complexity) – cp value, number of splits, relative error (rel error), cross-validation error (xerror), and standard deviation of cross-validation error (xstd).

```
Classification tree:

rpart(formula = MATHICKMM ~ DA + streamcat + compsite + FARANK +
    MPHYTERANK + RELABUND + IMPROB + medt + maxt_avg + BP + NH3 +
    Nox + OP + pH + SC + TN + TP + TPe + TNe + Zindex + PMDI +
    PHDI + DSCI + DSCt + Dzero + Done + Dtwo + Dthree + Dfour +
    natws + distws + natnf + distnf + wells + wellso + wellcat +
    wellocat + maxslope + medslope + xslope + devslope, data = do_wk,
    method = "class", control = rpart.control(cp = 0.069, usesurrogate = 2))

Variables actually used in tree construction:
[1] FARANK     MPHYTERANK

Root node error: 106/234 = 0.453

n= 234

     CP nsplit rel error xerror   xstd
1 0.127      0     1.000  1.000 0.0718
2 0.069      2     0.745  0.745 0.0682
```

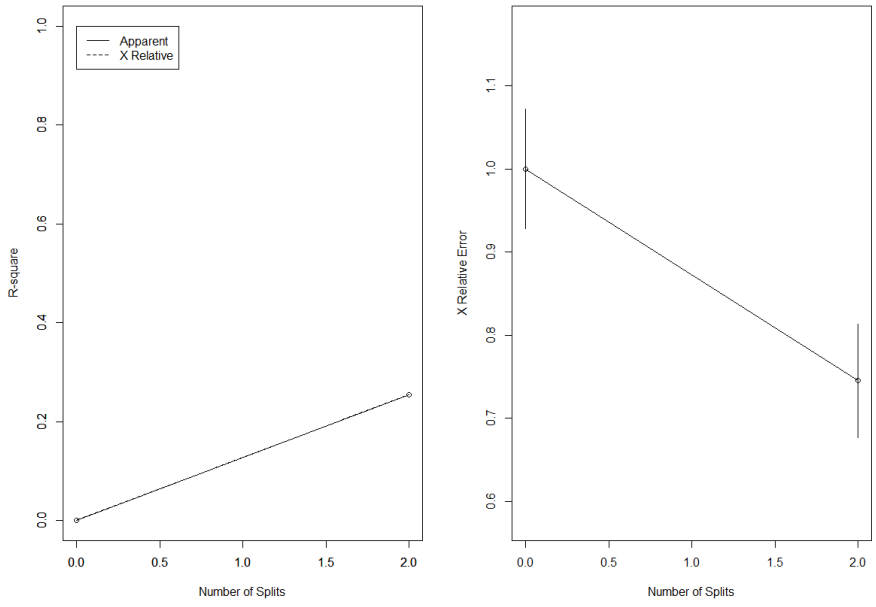


Figure 4.24. (**left**) Diagnostic plot showing explained variance (R^2) vs. the number of splits in tree diagram (a measure of complexity). R^2 (Apparent) = 1 – `rel error` for the original model fit; R^2 (X Relative) = 1 – `xerror` for the cross-validation series. (**right**) Diagnostic plot showing cross-validation relative error (`xerror`) vs. number of splits. Vertical bars represent ±1 standard deviation (`xstd`) of `xerror`.

A plot of explained variance and model complexity (shown as number of splits) indicates that a 2-split model for microalgae thickness is optimal (Figure 4.24, left). When looking at the distribution of cross-validation error (`xerror`) and model complexity (Figure 4.24, right), the 2-split model is an improvement over a no-split model. Table 4.11 shows the actual values of `xerror` and `rel error`, with the latter equal to $1\text{-}R^2$.

The residual analysis of the count exceedance regression tree shows a poor-behaved model (Figure 4.25). The scatterplot of residuals vs. fitted (upper-right) is unevenly distributed and there is a downward trend to the higher class; the QQ scatterplot has high deviation but only at high residual values (bottom-left). This model predicts counts so the fitted values are 1 through 7 days (per week). No qualitative explanation, as done for selected outliers in models shown above, was made here.

In the cumulative distribution plot (Figure 4.25, top left), one observes that 50% of the observations have a class residual of 1.75 mm and considered fairly large; 75% of the observations have a residual of 2.2 mm.

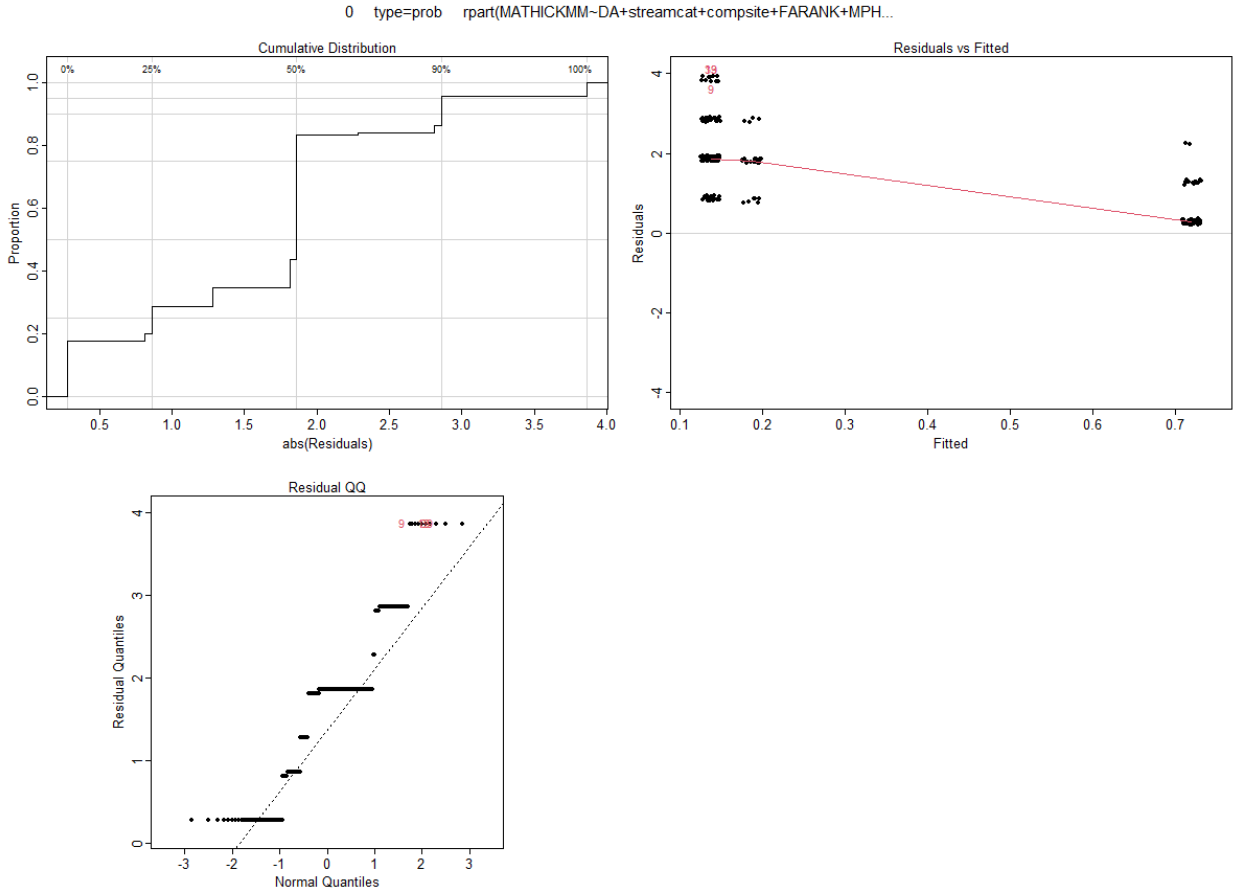


Figure 4.25. Residual analysis showing (**upper left**) cumulative distribution (proportion) of residuals (their absolute value) for entire dataset, (**upper right**) residual magnitude vs. predicted value from regression tree model (red line is loess-fitted value and observations with a high residual are noted by their record number (red font) in the dataset), and (**lower left**) quantile-quantile (QQ) plot of residuals. A positive residual indicates the observed value exceeds the predicted (fitted) value (i.e., the model underpredicts the observed).

4.b.ii. Aquatic Plant – Macrophyte (rank)

The second aquatic plant response variable is the ranked value of macrophyte cover of the stream area. Macrophyte cover was also estimated through Montana DEQ's aquatic visual assessment. As in microalgae thickness, the model is also considered a classification (as opposed to a regression) tree (Table 4.13).

The classification tree model shown here is more complex – it has 12 splits and employs 11 predictor variables (Figure 4.26; Table 4.12). All five ordinal classes ("absent" to "very heavy" cover) have sufficient observations in them to be predicted in this model. The primary predictors are microalgae cover (MARANK), filamentous algae cover (FARANK), and drought intensity (PHDI). One can observe the highest macrophyte cover occurs in one of three paths. One scenario (Figure 4.26, left side) describes conditions where there is no microalgae present, filamentous algae is abundant, and nitrate is < 1.06 mg/L; such conditions have high macrophytes. One explanation as to why nitrate is lower in this node is that the abundant filamentous algae and macrophytes are drawing down the nitrate from the stream water. A second scenario (Figure 4.26, far-right) can be described by high microalgae thickness, long term drought (from low to high intensity), and warmer water temperature. The third scenario (middle of Figure 4.26) also includes high microalgae thickness (the primary predictor), but now wet conditions (lack of drought) for intermittent and perennial streams (streamcat), and increased near-field disturbed land cover.

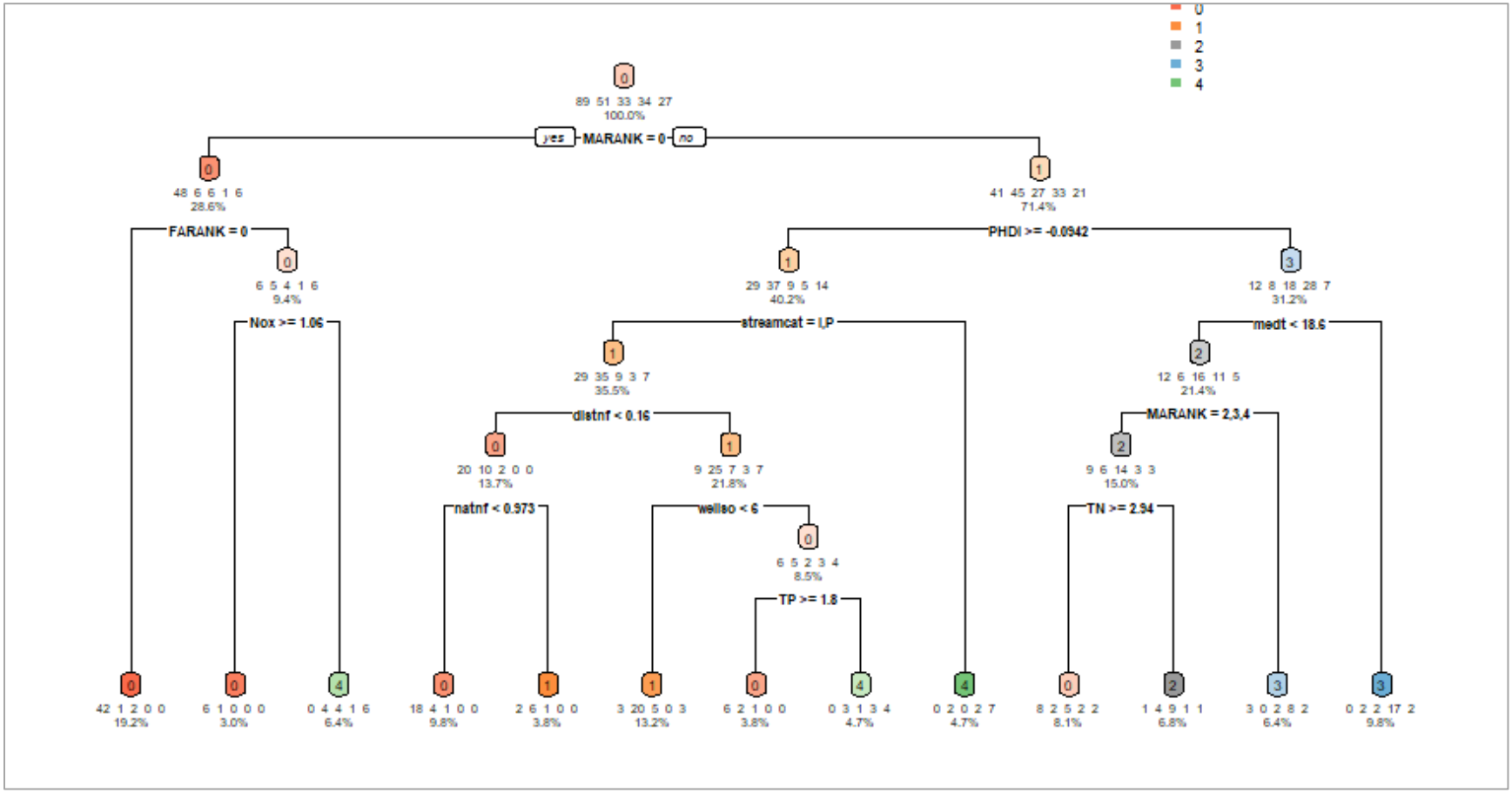


Figure 4.26. Diagram showing the regression tree for a *ranked* response of **macrophytes** (% cover). The predicted value (1 of 5 classes equivalent to "absent" (class 0) through "very heavy" (class 4)), the number of observations in each class, and the percentage of total observations are shown for each node. The intensity of the node color is proportional to the number of observations in the predicted class. The decision statement to split is located under each node (in bold) – traverse left if the statement is true (yes), otherwise traverse right (no).

Table 4.12. (2 parts)**.** The regression tree model shown above (Figure 4.26) portrayed as a set of decision rules. One set of statements describes a particular leaf node. Primary statement begins with predicted response and its rank (explanation shown above) followed by, in brackets [ ], the percentage of node observations in each class. Cover percent refers to percent of total observations in the particular node. Secondary statements are decisions about the important predictor variables to reach the leaf node.

```
              0   1   2   3   4
MPHYTERANK is 0 [.42 .11 .26 .11 .11] with cover 8%
when
    MARANK is 2 or 3 or 4
    PHDI < -0.094
    medt < 19
    TN >= 2.9

MPHYTERANK is 0 [.67 .22 .11 .00 .00] with cover 4%
when
    MARANK is 1 or 2 or 3 or 4
    PHDI >= -0.094
    streamcat is I or P
    distnf >= 0.16
    wellso >= 6
    TP >= 1.8

MPHYTERANK is 0 [.78 .17 .04 .00 .00] with cover 10%
when
    MARANK is 1 or 2 or 3 or 4
    PHDI >= -0.094
    streamcat is I or P
    distnf < 0.16
    natnf < 0.97

MPHYTERANK is 0 [.86 .14 .00 .00 .00] with cover 3%
when
    MARANK is 0
    FARANK is 1 or 2 or 3 or 4
    Nox >= 1.1

MPHYTERANK is 0 [.93 .02 .04 .00 .00] with cover 19%
when
    MARANK is 0
    FARANK is 0

MPHYTERANK is 1 [.10 .65 .16 .00 .10] with cover 13%
when
    MARANK is 1 or 2 or 3 or 4
    PHDI >= -0.094
    streamcat is I or P
    distnf >= 0.16
    wellso < 6

MPHYTERANK is 1 [.22 .67 .11 .00 .00] with cover 4%
when
    MARANK is 1 or 2 or 3 or 4
    PHDI >= -0.094
    streamcat is I or P
    distnf < 0.16
    natnf >= 0.97
```

Table 4.12 (continued)**.**

```
MPHYTERANK is 2 [.06 .25 .56 .06 .06] with cover 7%
when
    MARANK is 2 or 3 or 4
    PHDI < -0.094
    medt < 19
    TN < 2.9

MPHYTERANK is 3 [.20 .00 .13 .53 .13] with cover 6%
when
    MARANK is 0 or 1
    PHDI < -0.094
    medt < 19

MPHYTERANK is 3 [.00 .09 .09 .74 .09] with cover 10%
when
    MARANK is 1 or 2 or 3 or 4
    PHDI < -0.094
    medt >= 19

MPHYTERANK is 4 [.00 .27 .09 .27 .36] with cover 5%
when
    MARANK is 1 or 2 or 3 or 4
    PHDI >= -0.094
    streamcat is I or P
    distnf >= 0.16
    wellso >= 6
    TP < 1.8

MPHYTERANK is 4 [.00 .27 .27 .07 .40] with cover 6%
when
    MARANK is 0
    FARANK is 1 or 2 or 3 or 4
    Nox < 1.1

MPHYTERANK is 4 [.00 .18 .00 .18 .64] with cover 5%
when
    MARANK is 1 or 2 or 3 or 4
    PHDI >= -0.094
    streamcat is W
```

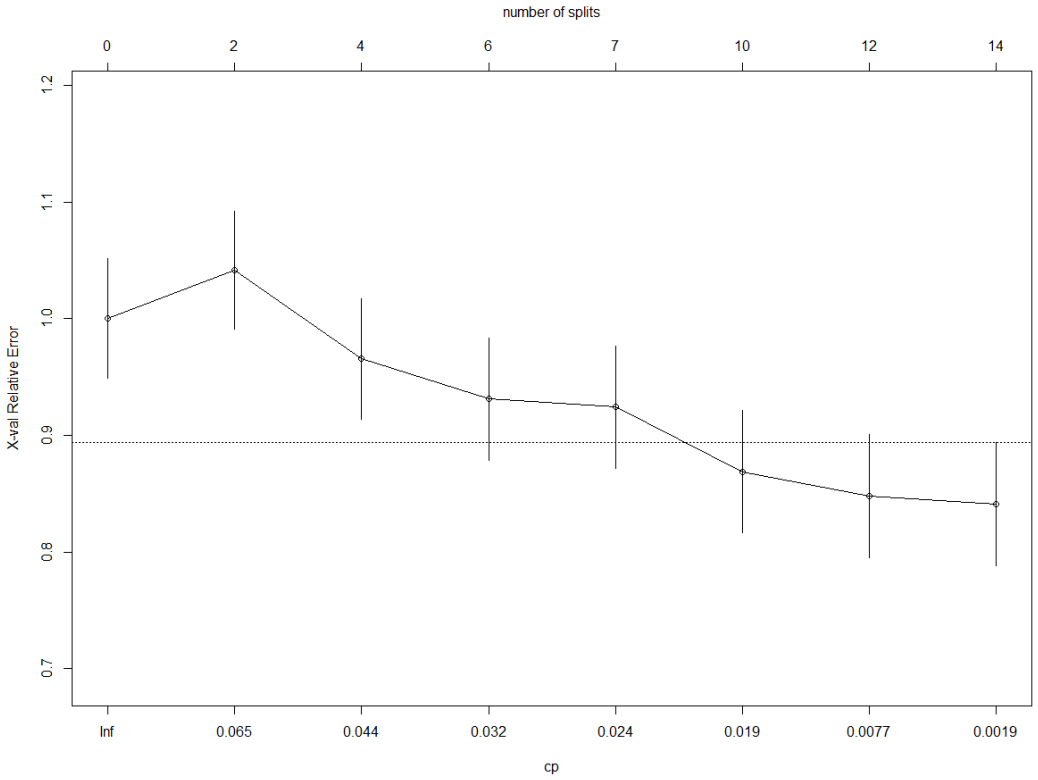


Figure 4.27. Initial plot of cross-validation error (`xerror`) vs. model complexity (cp) or number of splits in tree (upper x-axis). Error bars for `xerror` equal to ±1 error standard deviation (`xstd`). Dashed horizontal line is placed at +1 `xstd` above the lowest modeled `xerror`.

Figure 4.27 shows the change in the cross-validation error with respect to changing complexity. The most complex tree is on the right of the plot. The tree is subsequently pruned to a complexity just less than (i.e., at a larger cp) the lowest cross-validation error as long as its own `xerror` is below the dashed horizontal line. From Figure 4.27, the lowest `xerror` is found at a cp = 0.0019 which would suggest running the model at nsplits = 12 and a cp = 0.0077 (Figure 4.26 and Table 4.12).

Bivariate plots of important predictors vs. the response variable are shown in Figure 4.28 (bottom) and univariate plots of a single predictor and response are shown in the same figure (upper). Response values are a class number where class 1=rank 0 or "absent", class 2=rank 1 or "sparse", through class 5=rank 4 or "very heavy".

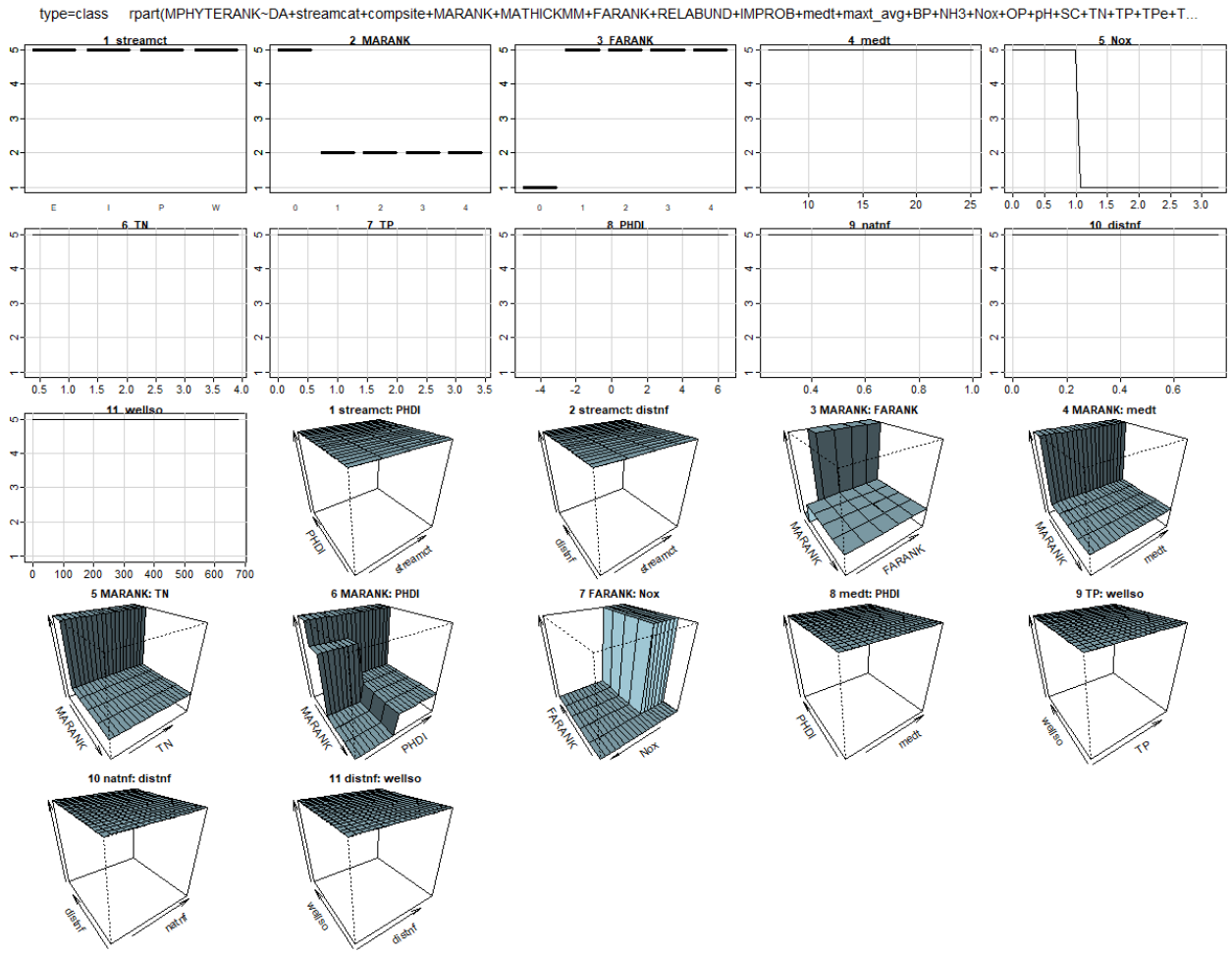


Figure 4.28. Plot of regression tree surface for model formulation of *ranked* response of **macrophytes** (% cover) (formulation also listed at the top of the diagram). Upper series of plots show relationships between each predictor variable and the response variable. Response values are class number where class 1=rank 0 or "absent", class 2=rank 1 or "sparse", through class 5=rank 4 or "very heavy". All variables used in the tree are shown. Lower series of plots shows interactions between a pair of predictor variables, with all other variables held at their median values, and the response variable. The pairs are chosen by which variables appear in the parent-child pairs of the regression tree (see Figure 4.26 or Table 4.12).

Table 4.13. Initial model formulation, list of predictor variables used in tree construction after model complexity chosen, root node error, and table showing – in order of decreasing cp (increasing tree complexity) – cp value, number of splits, relative error (rel error), cross-validation error (xerror), and standard deviation of cross-validation error (xstd).

```
Classification tree:

rpart(formula = MPHYTERANK ~ DA + streamcat + compsite + MARANK +
    MATHICKMM + FARANK + RELABUND + IMPROB + medt + maxt_avg +
    BP + NH3 + Nox + OP + pH + SC + TN + TP + TPe + TNe + Zindex +
    PMDI + PHDI + DSCI + DSCt + Dzero + Done + Dtwo + Dthree +
    Dfour + natws + distws + natnf + distnf + wells + wellso +
    wellcat + wellocat + maxslope + medslope + xslope + devslope,
    data = do_wk, method = "class", control = rpart.control(cp = 0.0077,
        usesurrogate = 2))

Variables actually used in tree construction:
 [1] distnf    FARANK    MARANK    medt      natnf     Nox       PHDI
streamcat TN        TP        wellso

Root node error: 145/234 = 0.62

n= 234

      CP nsplit rel error xerror   xstd
1 0.0828      0     1.000  1.000 0.0512
2 0.0517      2     0.834  1.014 0.0510
3 0.0379      4     0.731  0.979 0.0515
4 0.0276      6     0.655  0.890 0.0525
5 0.0207      7     0.628  0.848 0.0527
6 0.0172     10     0.566  0.855 0.0527
7 0.0077     12     0.531  0.841 0.0527
```

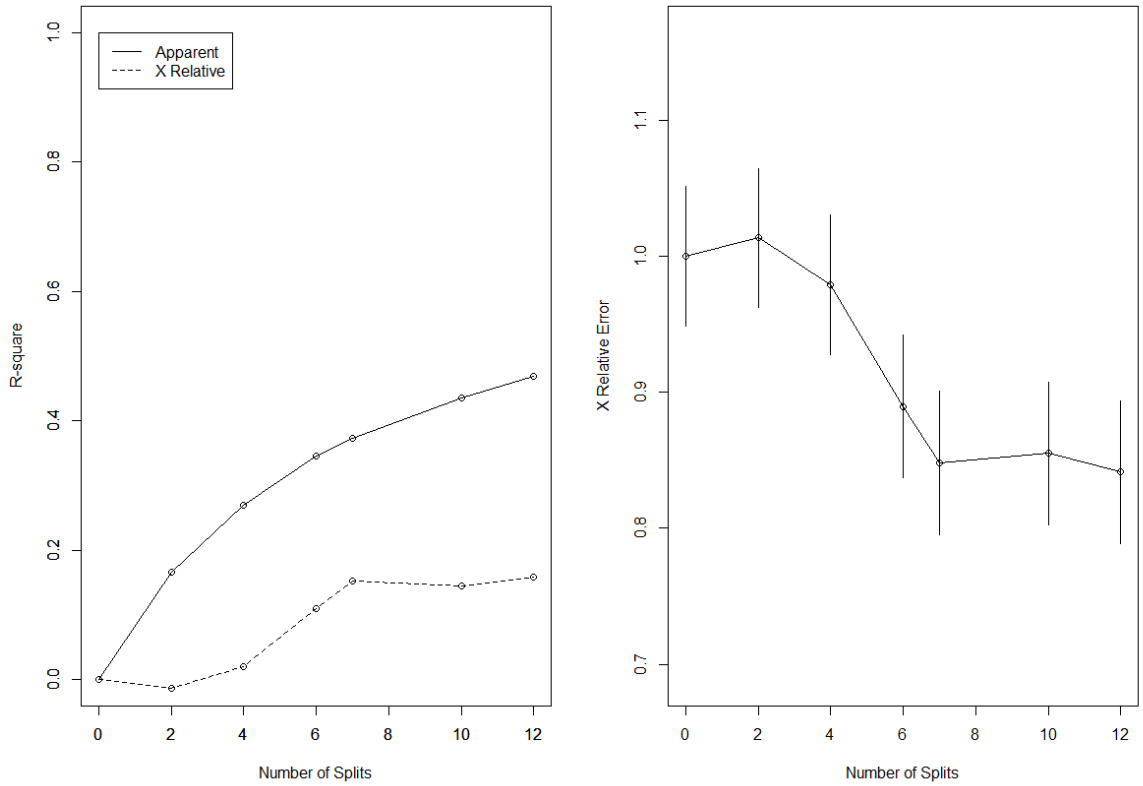


Figure 4.29. (**left**) Diagnostic plot showing explained variance (R^2) vs. the number of splits in tree diagram (a measure of complexity). R^2 (Apparent) = 1 – `rel error` for the original model fit; R^2 (X Relative) = 1 – `xerror` for the cross-validation series. (**right**) Diagnostic plot showing cross-validation relative error (`xerror`) vs. number of splits. Vertical bars represent ±1 standard deviation (`xstd`) of `xerror`.

A plot of explained variance and model complexity (shown as number of splits) indicates that a 10- or 12-split model for macrophyte cover is optimal (Figure 4.29, left). So it is possible to prune this classification tree to a 10-split tree and reduce possible overfitting. When looking at the distribution of cross-validation error (`xerror`) and model complexity (Figure 4.29, right), the 12-split model is optimal though a 10-split model is a close alternative. Table 4.13 shows the actual values of `xerror` and `rel error`, with the latter equal to 1-R^2.

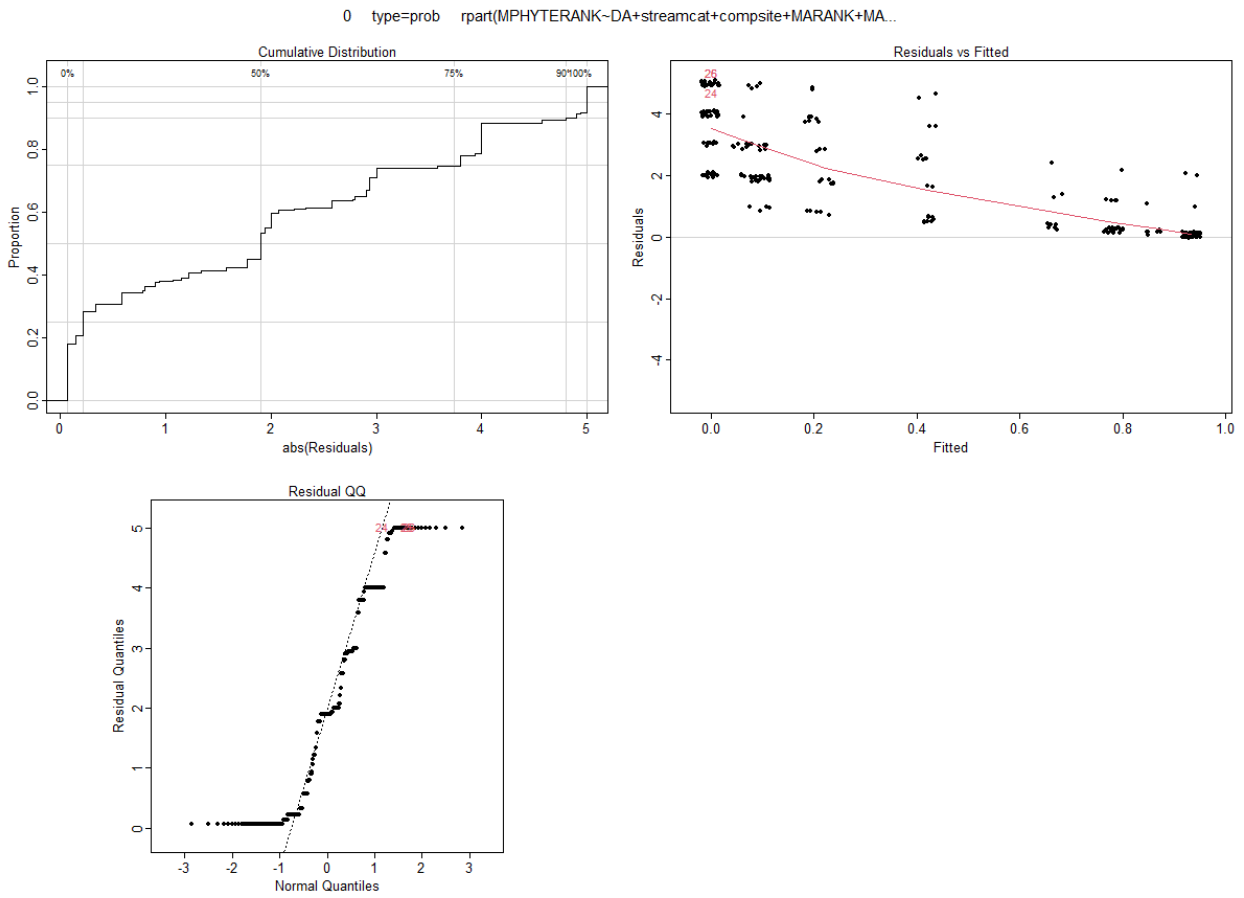


Figure 4.30. Residual analysis showing (**upper left**) cumulative distribution (proportion) of residuals (their absolute value) for entire dataset, (**upper right**) residual magnitude vs. predicted value from regression tree model (red line is loess-fitted value and observations with a high residual are noted by their record number (red font) in the dataset), and (**lower left**) quantile-quantile (QQ) plot of residuals. A positive residual indicates the observed value exceeds the predicted (fitted) value (i.e., the model underpredicts the observed).

The residual analysis of the count exceedance regression tree shows a poorly-behaved model (Figure 4.30) though better-behaved than the previous plant response discussed above (Figure 4.25). The scatterplot of residuals vs. fitted (upper-right) is unevenly distributed and there is a strong downward trend to the higher class of macrophyte cover – the lower-class predictions have higher residuals. The associated QQ scatterplot has high deviation but at both low and at high residual values (bottom-left). No qualitative explanation, as done for selected outliers in models shown above, was made here.

In the cumulative distribution plot (Figure 4.30, top left), one observes that 50% of the observations have a class residual of 1.0; 75% of the observations have a residual of 3.75. Both are considered fairly large residuals.

4.c.i. Dissolved Oxygen – Mean Delta / Expanded Observations (Weekly)

A response of mean weekly DO delta was visited again but now with a larger dataset – 762 observations vs. 234 observations in the model with the same response in Section 4.a. While this model makes use of a larger n, the suite of predictor variables is reduced to generate a longer dataset. The regression tree model shown here is much more complex with 15 splits and employs eight unique predictor variables (Figure 4.31; Table 4.14). As allowed by a regression tree model, a few of these predictor variables occur repeatedly at more than one level – near-field disturbed land cover (3 levels), watershed disturbed land cover (2 levels), and a drought index (2 levels). Some predictor variables appear in this tree that have not appeared in the preceding models, namely two indicator variables – comparison sites and stream category – and also watershed slope.

As in the previous mean delta model, drought and disturbed land cover are the primary predictor variables, though the actual predictor types are slightly different. The drought index selected here was that of the NDMC weighted-percent area of drought intensity. With intense drought over a larger area, mean deltas are at their highest (see right branch of Figure 4.31).

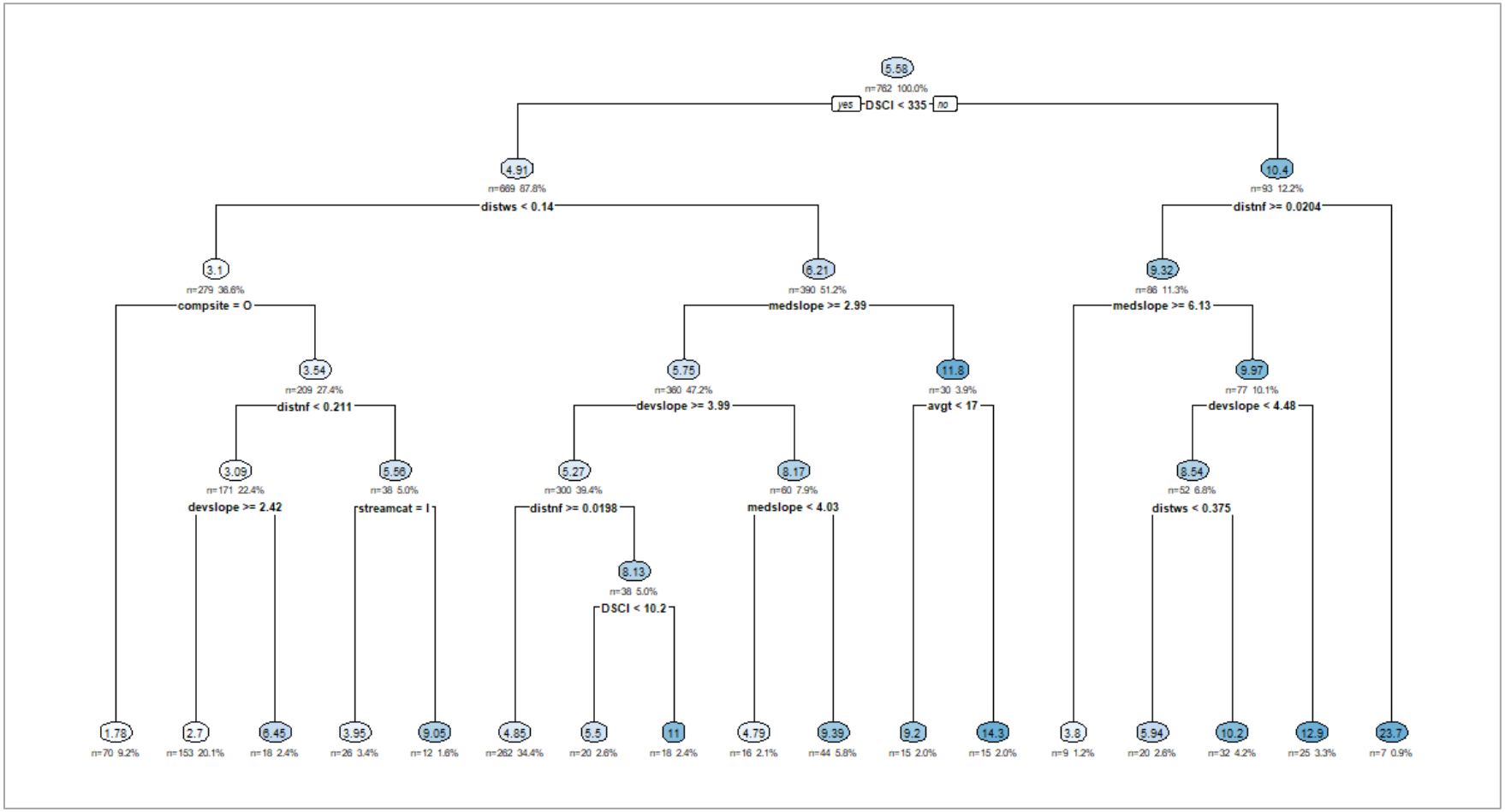


Figure 4.31. Diagram showing the regression tree for a response of **mean weekly DO delta** (mg/L) using the expanded number of observations (n=762) dataset. The predicted value and the number and percentage of total observations are shown for each node. The intensity of the node color is proportional to magnitude of the predicted value. The decision statement to split is located under each node (in bold) – traverse left if the statement is true (yes), otherwise traverse right (no).

Table 4.14. (2 parts)**.** The regression tree model shown above (Figure 4.31) portrayed as a set of decision rules. One set of statements describes a particular leaf node. Primary statement begins with mean predicted response (rounded to one decimal place) and its value (units defined above). Cover percent refers to percent of total observations in the particular node. Secondary statements are decisions about the important predictor variables to reach the leaf node.

```
xdelta is  1.8 with cover 9% when
    DSCI < 335
    distws < 0.14
    compsite is O

xdelta is  2.7 with cover 20% when
    DSCI < 335
    distws < 0.14
    distnf < 0.21
    devslope >= 2.4
    compsite is C or R

xdelta is  3.8 with cover 1% when
    DSCI >= 335
    distnf >= 0.02
    medslope >= 6.1

xdelta is  3.9 with cover 3% when
    DSCI < 335
    distws < 0.14
    distnf >= 0.21
    compsite is C or R
    streamcat is I

xdelta is  4.8 with cover 2% when
    DSCI < 335
    distws >= 0.14
    medslope is 3.0 to 4.0
    devslope < 4.0

xdelta is  4.9 with cover 34% when
    DSCI < 335
    distws >= 0.14
    distnf >= 0.02
    medslope >= 3.0
    devslope >= 4.0

xdelta is  5.5 with cover 3% when
    DSCI < 10
    distws >= 0.14
    distnf < 0.02
    medslope >= 3.0
    devslope >= 4.0

xdelta is  5.9 with cover 3% when
    DSCI >= 335
    distws < 0.37
    distnf >= 0.02
    medslope < 6.1
    devslope < 4.5

xdelta is  6.4 with cover 2% when
    DSCI < 335
    distws < 0.14
    distnf < 0.21
    devslope < 2.4
    compsite is C or R
```

Table 4.14 (continued)**.**

```
xdelta is  9.1 with cover 2% when
    DSCI < 335
    distws < 0.14
    distnf >= 0.21
    compsite is C or R
    streamcat is P

xdelta is  9.2 with cover 2% when
    DSCI < 335
    distws >= 0.14
    medslope < 3.0
    avgt < 17

xdelta is  9.4 with cover 6% when
    DSCI < 335
    distws >= 0.14
    medslope >= 4.0
    devslope < 4.0

xdelta is 10.2 with cover 4% when
    DSCI >= 335
    distws >= 0.37
    distnf >= 0.02
    medslope < 6.1
    devslope < 4.5

xdelta is 11.0 with cover 2% when
    DSCI is 10 to 335
    distws >= 0.14
    distnf < 0.02
    medslope >= 3.0
    devslope >= 4.0

xdelta is 12.9 with cover 3% when
    DSCI >= 335
    distnf >= 0.02
    medslope < 6.1
    devslope >= 4.5

xdelta is 14.3 with cover 2% when
    DSCI < 335
    distws >= 0.14
    medslope < 3.0
    avgt >= 17

xdelta is 23.7 with cover 1% when
    DSCI >= 335
    distnf < 0.02
```

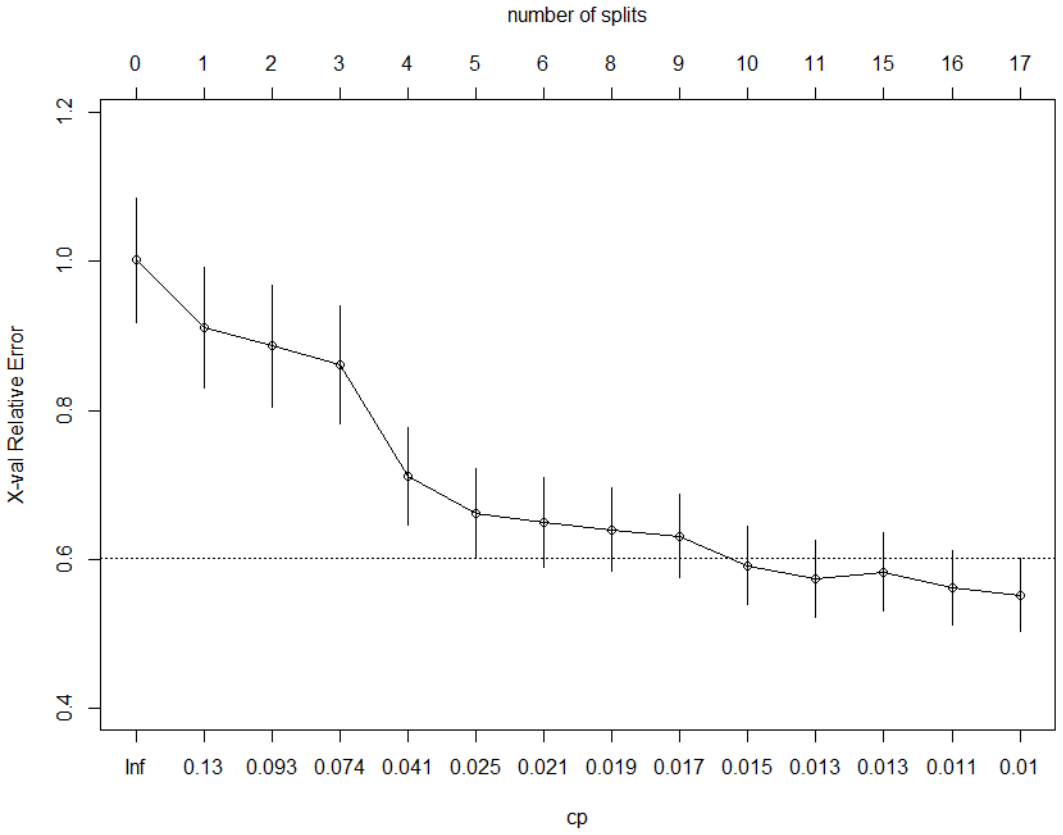


Figure 4.32. Initial plot of cross-validation error (`xerror`) vs. model complexity (cp) or number of splits in tree (upper x-axis). Error bars for `xerror` equal to ±1 error standard deviation (`xstd`). Dashed horizontal line is placed at +1 `xstd` above the lowest modeled `xerror`.

Figure 4.32 shows the change in the cross-validation error with respect to changing complexity. The most complex tree is on the right of the plot. The tree is subsequently pruned to a complexity just less than (i.e., at a larger cp) the lowest cross-validation error as long as its own `xerror` is below the dashed horizontal line. From Figure 4.32, the lowest `xerror` is found at a cp = 0.01, and without any close alternatives, so a tree of complexity nsplits = 16 (upper x-axis) and cp = 0.01 (lower x-axis) was used to build the model shown in Figure 4.31 and Table 4.15.

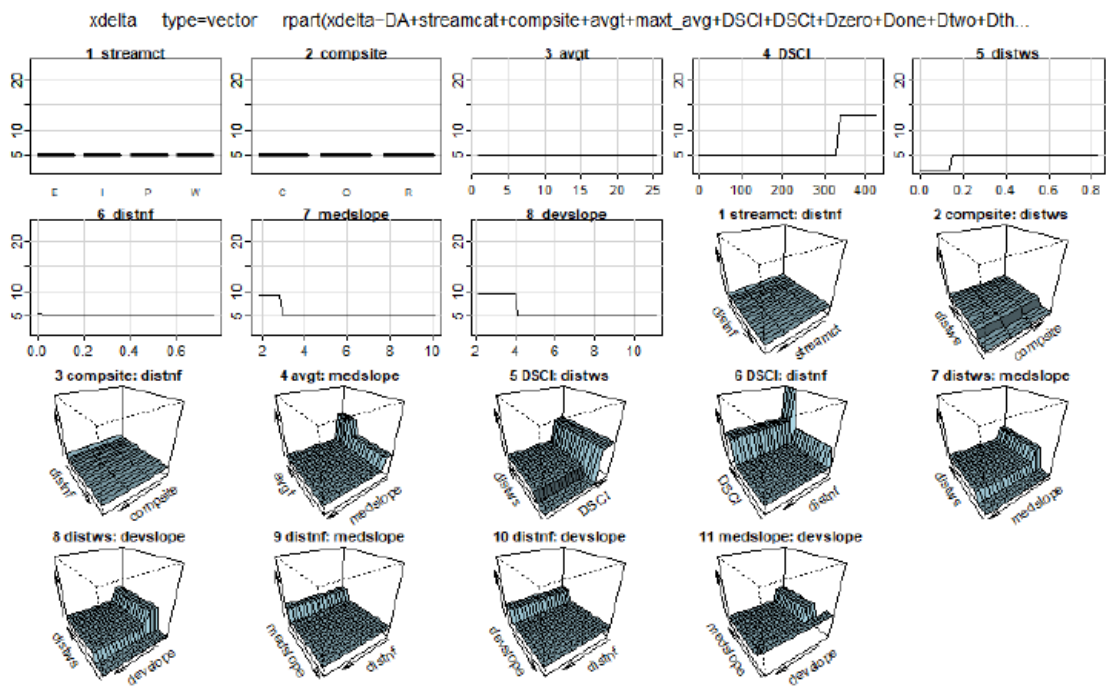


Figure 4.33. Plot of regression tree surface for model formulation of **mean weekly DO delta** using the expanded number of observations (n=762) dataset (formulation also listed at the top of the diagram). Upper series of plots show relationships between each predictor variable and the response variable. All variables used in the tree are shown. Lower series of plots shows interactions between a pair of predictor variables, with all other variables held at their median values, and the response variable. The pairs are chosen by which variables appear in the parent-child pairs of the regression tree (see Figure 4.31 or Table 4.14).

Bivariate plots of important predictors vs. the response variable are shown in Figure 4.33 (bottom) and univariate plots of a single predictor and response are shown in the same figure (upper). For the mean weekly DO delta tree, one can observe increasing mean delta with both increasing drought severity and extent and increasing watershed land disturbance (Figure 4.33, upper middle).

Table 4.15. Initial model formulation, list of predictor variables used in tree construction after model complexity chosen, root node error, and table showing – in order of decreasing cp (increasing tree complexity) – cp value, number of splits, relative error (rel error), cross-validation error (xerror), and standard deviation of cross-validation error (xstd).

```
Regression tree:

rpart(formula = xdelta ~ DA + streamcat + compsite + avgt + maxt_avg +
    DSCI + DSCt + Dzero + Done + Dtwo + Dthree + Dfour + natws +
    distws + natnf + distnf + wells + wellso + wellcat + wellocat +
    maxslope + medslope + xslope + devslope, data = do_exp_wk,
    method = "anova", control = rpart.control(cp = 0.011, usesurrogate = 2))

Variables actually used in tree construction:
[1] avgt      compsite  devslope  distnf    distws    DSCI      medslope
streamcat

Root node error: 15724/762 = 20.6

n= 762

        CP nsplit rel error xerror   xstd
1  0.1567      0     1.000  1.001 0.0838
2  0.1004      1     0.843  0.901 0.0759
3  0.0855      2     0.743  0.885 0.0733
4  0.0635      3     0.657  0.856 0.0718
5  0.0267      4     0.594  0.772 0.0693
6  0.0226      5     0.567  0.676 0.0606
7  0.0201      6     0.545  0.640 0.0550
8  0.0186      8     0.504  0.641 0.0538
9  0.0158      9     0.486  0.650 0.0539
10 0.0140     10     0.470  0.619 0.0496
11 0.0126     11     0.456  0.594 0.0473
12 0.0124     15     0.406  0.592 0.0471
13 0.0110     16     0.393  0.579 0.0471
```

A plot of explained variance (similar to an R^2 in a general linear model) and model complexity (shown as number of splits) indicates that a 16-split model for mean delta DO is optimal (Figure 4.34, left). A similar model complexity was found when examining the distribution of cross-validation error (`xerror`) and model complexity (Figure 4.34, right), the 16-split model has the lowest `xerror` but perhaps there are less complex models at 6-split or 11-split where xerror has stabilized. Table 4.15 shows the actual values of `xerror` and `rel error`, with the latter equal to $1\text{-}R^2$.

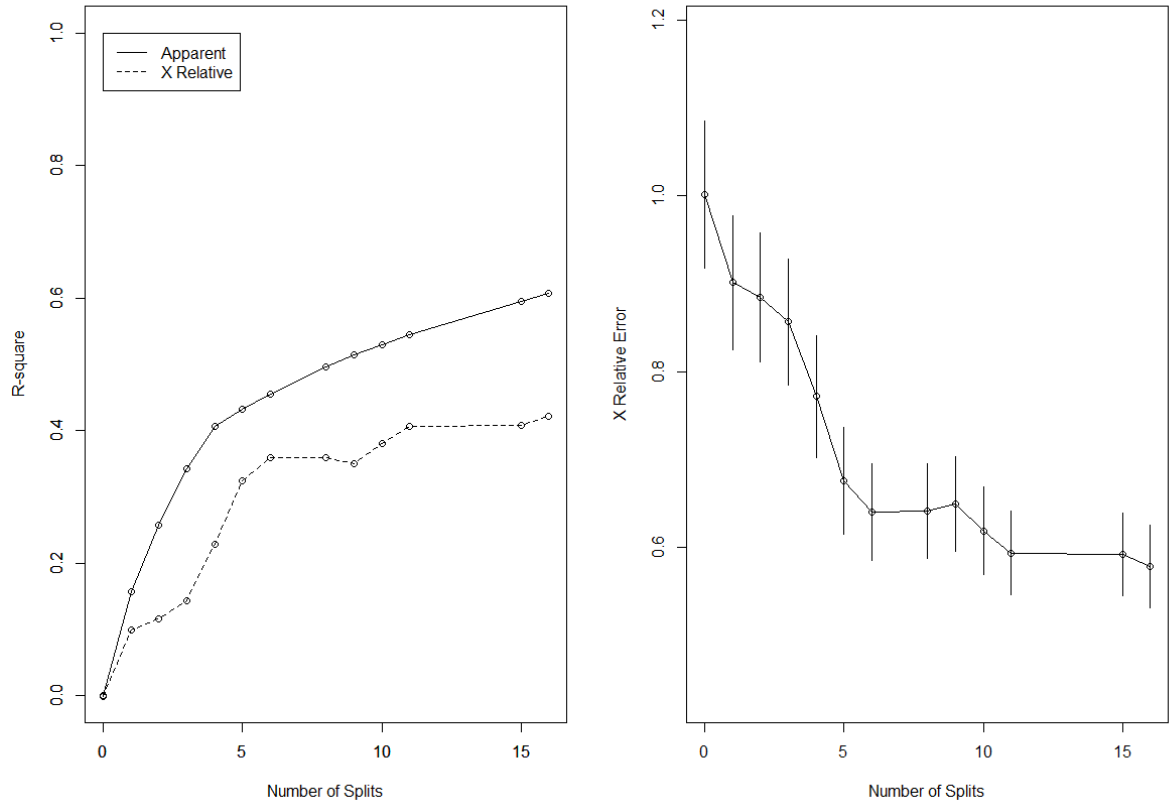


Figure 4.34. (**left**) Diagnostic plot showing explained variance (R^2) vs. the number of splits in tree diagram (a measure of complexity). R^2 (Apparent) = 1 – `rel error` for the original model fit; R^2 (X Relative) = 1 – `xerror` for the cross-validation series. (**right**) Diagnostic plot showing cross-validation relative error (`xerror`) vs. number of splits. Vertical bars represent ±1 standard deviation (`xstd`) of `xerror`.

A residual analysis of the mean weekly DO delta with an expanded observation set is shown in Figure 4.35. The scatterplot of residuals vs. fitted (upper-right) is unevenly distributed with a trend line showing increasing positive residual value with increasing predicted mean delta. The QQ scatterplot has high deviation at its upper tail (bottom-left). No outlier analysis was completed for this particular model, though case 66 and 65 (unnamed tributary to Fourmile Creek: M51FORMT01 during the 2nd and 1st weeks, respectively, of 10/2015) and 236 (Sandstone Creek: Y22SNDSC06 during the 2nd week of 9/2013) have high positive residuals – this model severely underpredicts (10-17 mg/L) for these three cases.

In the cumulative distribution plot (Figure 4.35, top left), one observes that 50% of the observations have a delta residual of less than 1.7 mg/L and 75% of the observations have a residual of less than 2.5 mg/L. Both residual magnitudes are low and suggests a good model fit for most of the observations in the study dataset.

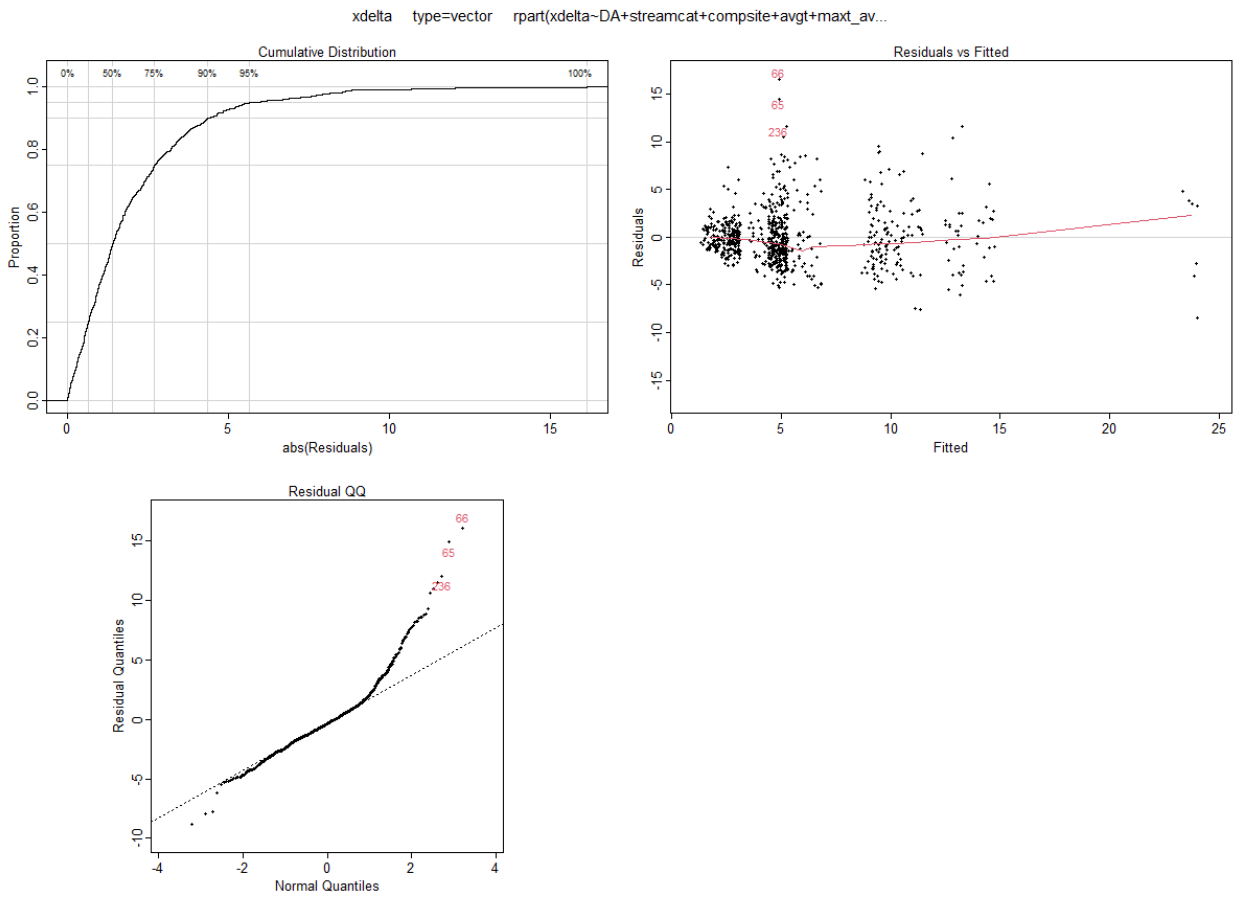


Figure 4.35. Residual analysis showing (**upper left**) cumulative distribution (proportion) of residuals (their absolute value) for entire dataset, (**upper right**) residual magnitude vs. predicted value from regression tree model (red line is loess-fitted value and observations with a high residual are noted by their record number (red font) in the dataset), and (**lower left**) quantile-quantile (QQ) plot of residuals. A positive residual indicates the observed value exceeds the predicted (fitted) value (i.e., the model underpredicts the observed).

5. Conclusions

The analysis of DO response to various fixed- and random-effect predictors presented in this work is a form of statistical learning. It is supervised statistical learning because the models built for predicting the response were based on one or more inputs.

5.a. Overall Findings and Success of Tree Models

Low levels of watershed disturbance and the absence of prolonged drought conditions were the most consistent predictors for optimal DO conditions, expressed as either diel variation or as a minimum. Secondary predictors like conductivity (which correlates positively to anthropogenic impacts, with other predictors held constant), nutrient levels, drainage area, and water temperature were also important.

When comparing all of the models and the information they offer, GLEC suggests that two DO models – mean delta (Figure 4.1) and average minimum (Figure 4.11) – offer the most guidance in criteria development. Both responses are summary measures (per week) and this aggregation is intuitively most stable. One could manage the weekly aggregations with manual adjusting to avoid weekly summaries with only a few days of monitoring. Most of the outliers (high residuals identified in Results and Discussion) were likely caused by these situations.

The one delta exceedance model (Figure 4.16, using the 5.3 mg/L Montana threshold) deployed in this study has a useful role – count of exceedance per week – as it suggests the number of days the aquatic system is stressed by high delta. As found in numerous Ohio-based watershed assessment documents, a primary determinant of the presence of deformities, lesions, and tumors in sampled fish was the frequency of high DO deltas – the higher organisms are stressed by continuous adaptation to changing DO conditions. More work is needed to compare the tree rules (and thus the primary predictors and their boundaries) from the Montana threshold with lower (Minnesota) and higher (Ohio) DO delta threshold values.

Tree models with plant-based responses behaved differently, as the conclusion appears to be that where there are macrophytes and/or filamentous algae, there is often microalgae. The macrophyte (% cover) response model was likely overpredicted or may simply be a poor response variable. The plant-based models were not that useful and perhaps an alternative approach, for example a plant taxonomic model, would be more appropriate.

A final area of continued work is to determine whether the "expanded observations" dataset reveals more information on influential predictors compared to the "all predictors" dataset. Here, expanding from 234 to 762 observations may better exploit the power of statistical learning techniques (including new approaches discussed below). However, in this expanded dataset, there is concurrent loss of some of the random-effect predictor variables, namely the water chemistry and plant-based predictors (e.g., relative abundance of nutrient-enricher taxa), which were not sampled over the entire multi-week DO survey. Given that two fixed-effect predictors – those related to drought and land cover – were prominent, the loss of these random-effects perhaps is mitigated. Future monitoring strategies might expand the frequency of water chemistry monitoring (likely less labor-intensive than monitoring the plant indices) so that all or most of the DO weekly aggregations would be retained.

In this study there is one example, for average delta DO, where a comparison between "all predictors" and "expanded observations" could be made. The expanded observations model was much more complex (16 splits vs. 5 splits) but otherwise was similar to the "all predictors" model. Key predictors of drought and disturbed land cover were present in both models, and the residuals were highly variant in both as well. Additional work on the "expanded observations" dataset should include, along with the NDMC predictors, the NOAA drought predictors – they were excluded because they are monthly in frequency. Once rebuilt, use this larger dataset with the suite of DO response variables.

5.b. Limitations of Regression Tree Modeling

Decision trees for regression and classification have a number of advantages over the more classical approaches of linear regression or multiple linear regression (James et al. 2015). Results of tree models are easier to explain. Decision trees more closely mirror human decision-making than do regression and other classification approaches (e.g., cluster analysis). Trees can be displayed graphically as shown here and are easily interpreted. Trees can easily handle qualitative predictors without the need to create dummy variables.

However, trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches. Additionally, trees can be very non-robust. In other words, a small change in the data can cause a large change in the final estimated tree. Perhaps, the most obvious difficulty is the stability in classes assigned to terminal nodes (Berk 2020).

However, by aggregating many decision trees, using methods like bagging, random forests, and boosting, the predictive performance of trees can be substantially improved. Two of these concepts are introduced in the next section.

5.c. Future Directions – New Approaches

Most of the narrative in this section has been built from the work of James et al. (2015; pp. 316-321) in their textbook on statistical learning.

Bagging and random forests use "trees as building blocks" to construct more powerful prediction models. The decision trees presented in this study do suffer from high variance. If the training dataset is split into two parts at random, and a decision tree is fit to both halves, the subsequent results could be quite different. In contrast, a procedure with low variance will yield similar results if applied repeatedly to distinct data sets. Linear regression tends to have low variance, if the ratio of the number of observations to the number of predictor variables is moderately large. In the "all predictors" dataset employed here, there were 234 observations and 43 predictor variables.

Bootstrap aggregation, or bagging, is a general-purpose procedure for reducing the bagging variance of a statistical learning method; it is particularly useful and frequently used in the context of decision trees. To apply bagging to regression trees, B regression trees are constructed from B bootstrapped training sets and average the resulting predictions. These trees are grown deep (i.e., fully saturated), and are not pruned. Hence each individual tree has high variance, but low bias. Averaging these B trees reduces the variance. Bagging has been demonstrated to give impressive improvements in accuracy by combining together hundreds or even thousands of trees into a single procedure.

Random forests provide an improvement over bagged trees by way of a random small adjustment that removes the correlation from the suite of trees. As in bagging, a forest of decision trees is built from bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of a subset of predictors is chosen as split candidates from the full set of predictors. The split is allowed to use only one of the subset of predictors – the tree algorithm cannot consider a majority of the available predictors.

The approach is useful when a very strong predictor exists in the data, along with a number of other moderately strong predictors. Then in the collection of bagged trees (as for bagging), most or all of the trees will use this strong predictor in the top split. Consequently, all of the bagged trees will look quite similar (and thus be highly correlated) to each other.

When a number of highly correlated trees are averaged, there is not a large reduction in variance compared to averaging many uncorrelated quantities. Bagging may not lead to a substantial reduction in variance over a single tree in this setting. Random forests can overcome this problem by forcing each split to consider only a subset of the predictors.

Connecting these newer approaches back to the present study imply they would be logically the next approach following the current CART analysis of eastern Montana streams. Improvement in predictive result is not automatic as their benefit would vary from study to study. Also, might the potential of a technique be greater than what is observed after it is implemented? Hence, bagging and random forests may be potential improvements with these datasets.

5.d. References Cited

Ahmadi, B. and H. Moradkhani (2018) Revisiting hydrological drought propagation and recovery considering water quantity and quality. *Hydrological Processes*, 33: 1492-1505.

Berk, R.A. (2020) *Statistical Learning from a Regression Perspective*, Chapter 3: Classification and Regression Trees (CART), 3rd Edition, Springer, Switzerland, pp. 157-203.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984) *Classification and Regression Trees*. Wadsworth Books, p. 358.

Faraway, J.J. (2016) *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapter 16: Trees, 2nd Edition, CRC Press LLC, pp. 343-363.

Heim, R.R. (2002) A review of twentieth-century drought indices used in the United States. *Bulletin of the American Meteorological Society*, 83(8): 1149-1166.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2015) *An Introduction to Statistical Learning: with Applications in R*. Chapter 8: Tree Based Methods, [Corrected at 6th printing 2015] ed. New York: Springer; 2015, pp. 303-323.

Heiskary, S.A. and R.W. Bouchard (2015) Development of eutrophication criteria for Minnesota streams and rivers using multiple lines of evidence. *Freshwater Science*, 34(2): 574–592.

Helsel, D. (2019) *Forty Years of Water Quality Statistics: What's Changed, What Hasn't ?* 60 minute webinar at https://PracticalStats.com

Helsel, D. (2012) *Statistics for Censored Environmental Data using Minitab and R*, 2nd edition. Wiley, pp. 79-86.

Milborrow, S. (2011) rpart.plot: Plot rpart models. An enhanced version of plot.rpart, R package: http://www.milbo.org/rpart-plot.

Milborrow, S. (2018 plotmo: Plot a Model's Residuals, Response, and Partial Dependence Plots, R package: https://CRAN.R-project.org/package=plotmo.

Milborrow, S. (2020) Plotting model residuals with plotres.

Miltner, R.J. (2010) A method and rationale for deriving nutrient criteria for small rivers and streams in Ohio. *Environmental Management*, 45(4): 842–855.

Mosley, L.M. (2015) Drought impacts on the water quality of freshwater systems: review and integration. *Earth-Science Reviews*, 140: 203–214

Nisbet, R., J. Elder IV, and G. Miner (2009) *Handbook of Statistical Analysis and Data Mining Applications*. Chapter 11: Classification, Academic Press, pp. 235-258.

Quinlan, J.R. (1993) Chapter 1: FOIL: A Midterm Report in P.B. Brazdil (ed.) *Machine learning, ECML-93: European Conference on Machine Learning*, Vienna, Austria, Springer, pp. 3-20.

Ryberg K.R., J.D. Blomquist, L.A. Sprague, A.J. Sekellick, and J. Keisman (2018) Modeling drivers of phosphorus loads in Chesapeake Bay tributaries and inferences about long-term change. *The Science of the Total Environment*. 616-617: 1423–1430.

Suplee, M.W., Sada de Suplee, R., Feldman, D., and T. Laidlaw (2005) *Identification and Assessment of Montana Reference Streams: A Follow-up and Expansion of the 1992 Benchmark Biology Study*. Helena, MT: Montana Department of Environmental Quality.

Suplee, M.W. and R. Sada (2016) *Assessment Methodology for Determining Wadeable Stream Impairment Due to Excess Nitrogen and Phosphorus Levels*. Helena, MT: Montana Dept. of Environmental Quality, p. C-4 and Table C2-2.

Svoboda, M. (2000) An introduction to the Drought Monitor. *Drought Network News*, 12: 15–20.

Therneau, T.M. and E. Atkinson (2019) *An Introduction to Recursive Partitioning Using the RPART Routines,* Mayo Foundation, 60 pp.

Appendix A

A.1. Bayesian Network Model for Delta DO for Eastern Montana

Author: Robert Miltner, Columbus OH

Why a Bayesian Network?

Bayesian networks (BNs) are particularly well suited to investigating environmental questions centered on conditional states in relation to stated thresholds[26] as they supply answers in terms of modeled distributions[27] or probabilities[28]. Another feature of BNs is that they provide a graphical representation of modeled relationships[2], and they are flexible in allowing expert knowledge and new data to update and inform the model[29]. For those unfamiliar with BNs, one can think of them as a way to query a multiple linear regression model in such way that predictions based on the model integrate the uncertainties carried by the various predictors, thus providing better estimates (i.e., credible intervals) of the uncertainties in model predictions.

Deriving the Network Model

An initial, tentative model was suggested by first inspecting the dataset as to the nature of individual variables (shape, distribution, and missingness), and subsequently making transformations (e.g., log of the drainage area, square roots of percentages and delta dissolved oxygen) or coding variables to categories where appropriate. To pair water chemistry observations with the dissolved oxygen observations, the weekly average of daily delta DO was matched to the chemistry observation occurring in that week (i.e., chemistry observations were not made on all days that sondes were deployed), yielding 234 records. Next, bivariate relationships were examined to identify colinear variables, and variables that correlated with delta DO (Figure A.1). Candidate variables thusly identified were included in a call to the hill climbing (hc) algorithm[30,31] in the bnlearn package (R 4.03) to draw a directed acyclic graph (DAG; Figure A.2). The candidate variables and their descriptions are in Table A.1.

---

[26] Chen, Serena H., and Carmel A. Pollino. "Good practice in Bayesian network modelling." *Environmental Modelling & Software* 37 (2012): 134-145.
[27] Qian, Song S., and Robert J. Miltner. "A continuous variable Bayesian networks model for water quality modeling: A case study of setting nitrogen criterion for small rivers and streams in Ohio, USA." *Environmental Modelling & Software* 69 (2015): 14-22.
[28] Chen, Serena H., and Carmel A. Pollino. "Good practice in Bayesian network modelling." *Environmental Modelling & Software* 37 (2012): 134-145.
[29] Uusitalo, Laura. "Advantages and challenges of Bayesian networks in environmental modelling." *Ecological modelling* 203, no. 3-4 (2007): 312-318.
[30] http://www.cs.cornell.edu/selman/papers/pdf/02.encycl-hillclimbing.pdf
[31] Scutari, Marco, Pietro Auconi, Guido Caldarelli, and Lorenzo Franchi. "Bayesian networks analysis of malocclusion data." *Scientific reports* 7, no. 1 (2017): 1-11.

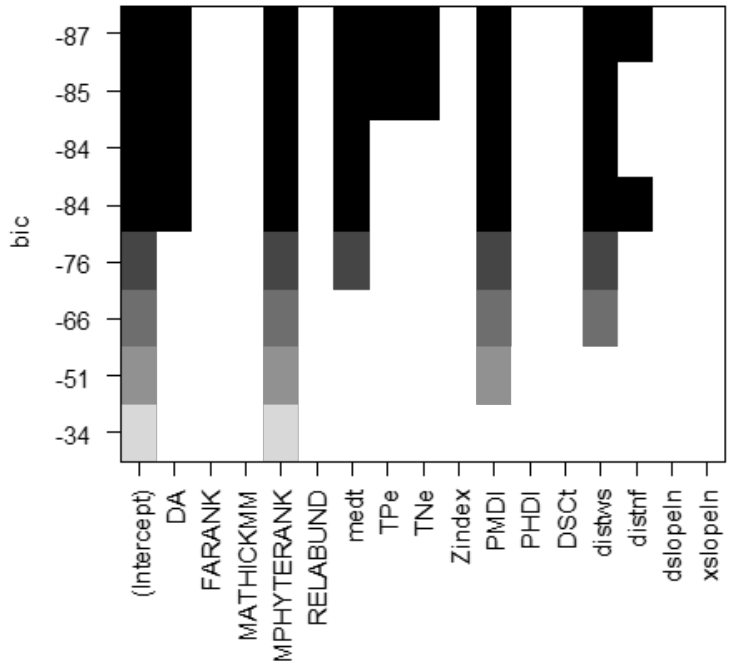


Figure A-1. Variables screened through all subsets regression as predictors of delta DO. The same variables were repeatedly selected. Note that although PMDI was repeated selected, it is colinear with PHDI. Both were iteratively screened by developing separate DAGs and examining the resulting information content scores. PHDI resulted in a slightly lower (i.e., better) score when included in the BN; however, the PMDI was retained given that it formed splits in the regression trees. Similarly, natural land use/cover (not included in the figure) and disturbed land use/cover were inversely related, and therefore essentially equivalent, but disturbed land use/cover provided a better network score.

Table A-1. Candidate variables included in the hill climbing search for a directed acyclic graph (DAG).

wellocat - old wells were categorized (0,1,2 & 3) as 0, 1-16, 20-42, and 79-679 wells per watershed based on inspection of a CDF plot
distws - disturbed land, watershed
Dthree - measure of cumulative drought intensity – D3 level
Dzero - measure of cumulative drought intensity – D0 level
PMDI - measure from Palmer drought index
TNe - total nitrogen, the "e" denotes that 4 values were imputed (using the median of all values)
TPe - total phosphorus, 4 values imputed with the median
medt - median weekly temperature
MPHYTERANK - percent macrophyte cover categorized (0, 1, 2, 3 & 4) as none or unobserved, 5%, 25%, 58% and 88%
DA - watershed drainage area
sqd - square root of the average weekly dissolved oxygen range

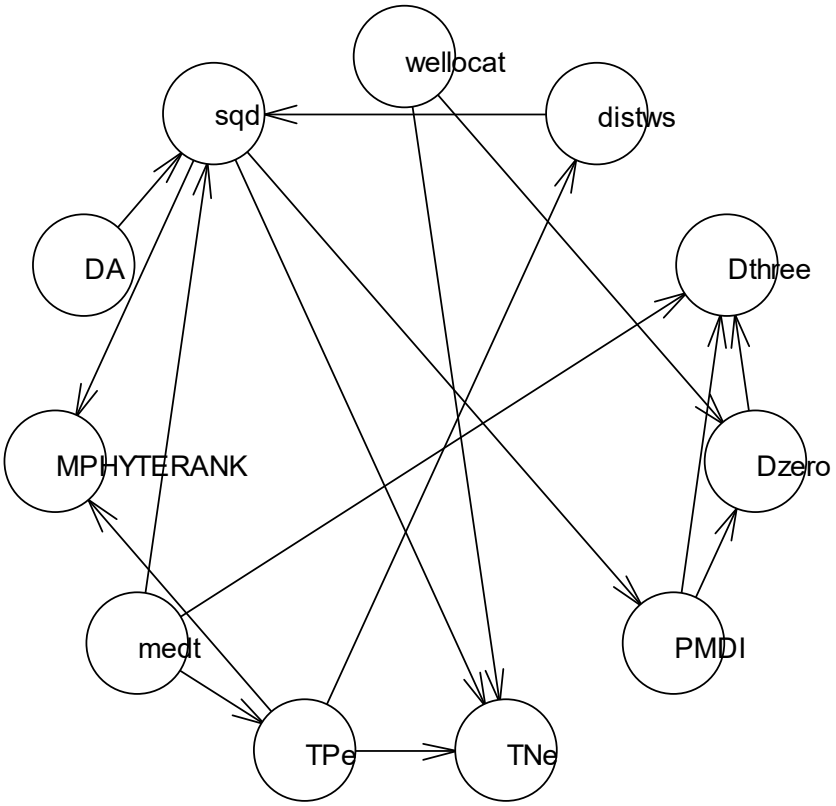


Figure A-2. A directed acyclic graph (DAG) suggested by the call to the hill climbing algorithm in the bnlearn package. Arrows (arcs) show dependencies.

The DAG in Figure A.2 was refined by testing the arcs with 1000, leave-one-out, bootstrapped replicates using the boot.strength function from the bnlearn package. Bootstrapping tests both the frequency of times an arc is drawn in a network and the direction of the arc. Strength values approaching 0.9 are typically retained, but other arcs can be retained based on local expert knowledge or as informed by other analyses. Table A.2 shows the first several rows of computer output to given one a sense of this. The DAG in Figure A.3 shows the final configuration as informed by the steps previously described, and in light of results from regression trees.

Table A-2. Output from a call to `boot.strength`. The yellow highlighted rows are simply examples of nodes and arcs that one would obviously want to retain. The first 12 of 110 rows are shown.

	from	to	strength	direction
1	sqd	DA	0.755	0.282781457
2	sqd	MPHYTERANK	0.964	0.718360996
3	sqd	medt	0.751	0.331557923
4	sqd	TPe	0.089	0.404494382
5	sqd	TNe	0.983	0.995930824
6	sqd	PMDI	0.894	0.491051454
7	sqd	Dzero	0.049	0.244897959
8	sqd	Dthree	0.240	0.316666667
9	sqd	distws	0.994	0.441146881
10	sqd	wellocat	0.006	0.000000000
11	DA	sqd	0.755	0.717218543
12	DA	MPHYTERANK	0.398	0.351758794

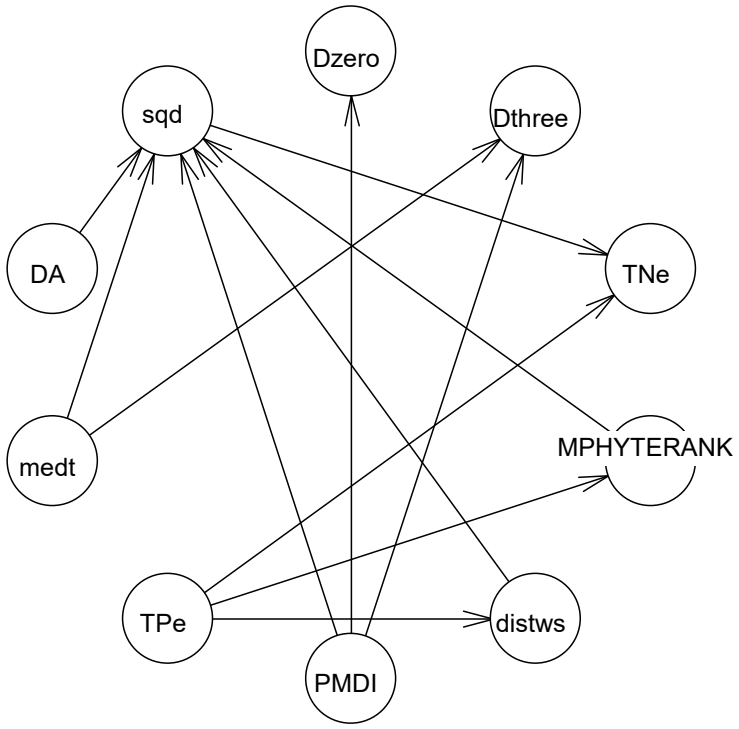


Figure A-3. Directed acyclic diagram (DAG) showing factors related to and predicting weekly average daily dissolved oxygen range. Acronyms are listed in Table A.1. This is the DAG to which the network model was fit. The strengths of the arcs are listed in Table A.3.

Table A-3. Arc strength for the DAG shown in Figure A.3. Strength values less than ~ 4 (absolute value) are marginal.

	from	to	strength
1	sqd	TNe	-11.422960
2	DA	sqd	-4.005337
3	medt	sqd	-5.313140
4	medt	Dthree	-13.110966
5	TPe	distws	-27.363346
6	TPe	MPHYTERANK	-17.728267
7	TPe	TNe	-115.634608
8	PMDI	sqd	-9.034774
9	PMDI	Dthree	-30.343489
10	PMDI	Dzero	-64.800003
11	distws	sqd	-13.729297
12	MPHYTERANK	sqd	-4.580174

The following **output lists the Bayesian network parameters**; nodes with two or more parameters listed are essentially linear regression models. Nodes with only one parameter (i.e., an intercept) show the parameter means.

Parameters of node sqd (Gaussian distribution) - *this essentially a multiple linear regression model*

Conditional density: sqd | DA + medt + PMDI + distws + MPHYTERANK
Coefficients:

(Intercept)	DA	medt	PMDI	distws	MPHYTERANK
1.49459985	-0.29273137	0.03992171	-0.06168268	1.70260688	0.13906104

Standard deviation of the residuals: 0.7010879

Parameters of node DA (Gaussian distribution)

Conditional density: DA - *the mean of the drainage area in log10 units*
Coefficients:
(Intercept)
1.93423
Standard deviation of the residuals: 0.5873918

Parameters of node medt (Gaussian distribution)

Conditional density: medt
Coefficients:
(Intercept)
15.38648
Standard deviation of the residuals: 4.71909

Parameters of node TPe (Gaussian distribution)

Conditional density: TPe
Coefficients:
(Intercept)
1.849869
Standard deviation of the residuals: 0.54906

Parameters of node PMDI (Gaussian distribution)

Conditional density: PMDI
Coefficients:
(Intercept)
1.11637
Standard deviation of the residuals: 3.696618

Parameters of node distws (Gaussian distribution)

Conditional density: distws | TPe
Coefficients:
(Intercept) TPe
0.5276205 -0.1531809
Standard deviation of the residuals: 0.1556268

Parameters of node MPHYTERANK (Gaussian distribution)

Conditional density: MPHYTERANK | TPe
Coefficients:
(Intercept) TPe
3.301853 -1.029487
Standard deviation of the residuals: 1.295871

Parameters of node TNe (Gaussian distribution)

Conditional density: TNe | sqd + TPe
Coefficients:
(Intercept) sqd TPe
1.86376758 0.08789962 0.52588084
Standard deviation of the residuals: 0.2139083

Parameters of node Dthree (Gaussian distribution)

Conditional density: Dthree | medt + PMDI
Coefficients:
(Intercept) medt PMDI
-1.3082387 0.1555311 -0.2980208
Standard deviation of the residuals: 1.935703

Parameters of node Dzero (Gaussian distribution)

Conditional density: Dzero | PMDI
Coefficients:
(Intercept) PMDI
14.474371 -3.253839
Standard deviation of the residuals: 13.63965

Querying the Model

The network model can be queried in two ways. The first yields distributions of a node parameter conditioned on scenarios prescribed for other nodes in the network. The other method yields predicted probabilities of observing a stated set of conditions. Figure A.4 explores the conditional relationship between delta DO, land disturbance and drought conditions. Figure A.5 shows delta DO conditioned on dry climate and macrophyte abundance, whereas Figure A.6 conditions delta DO and wet climate and macrophyte abundance. Figure A.7 explores the probability of observing delta DO greater than several stated thresholds (i.e., 5.3, 7, 8, and 9 mg/l) given low levels of watershed disturbance and normal climate conditions (i.e., normal for that observed in the dataset). An R script and data object is being supplied with comments and instructions for running queries of the network model.

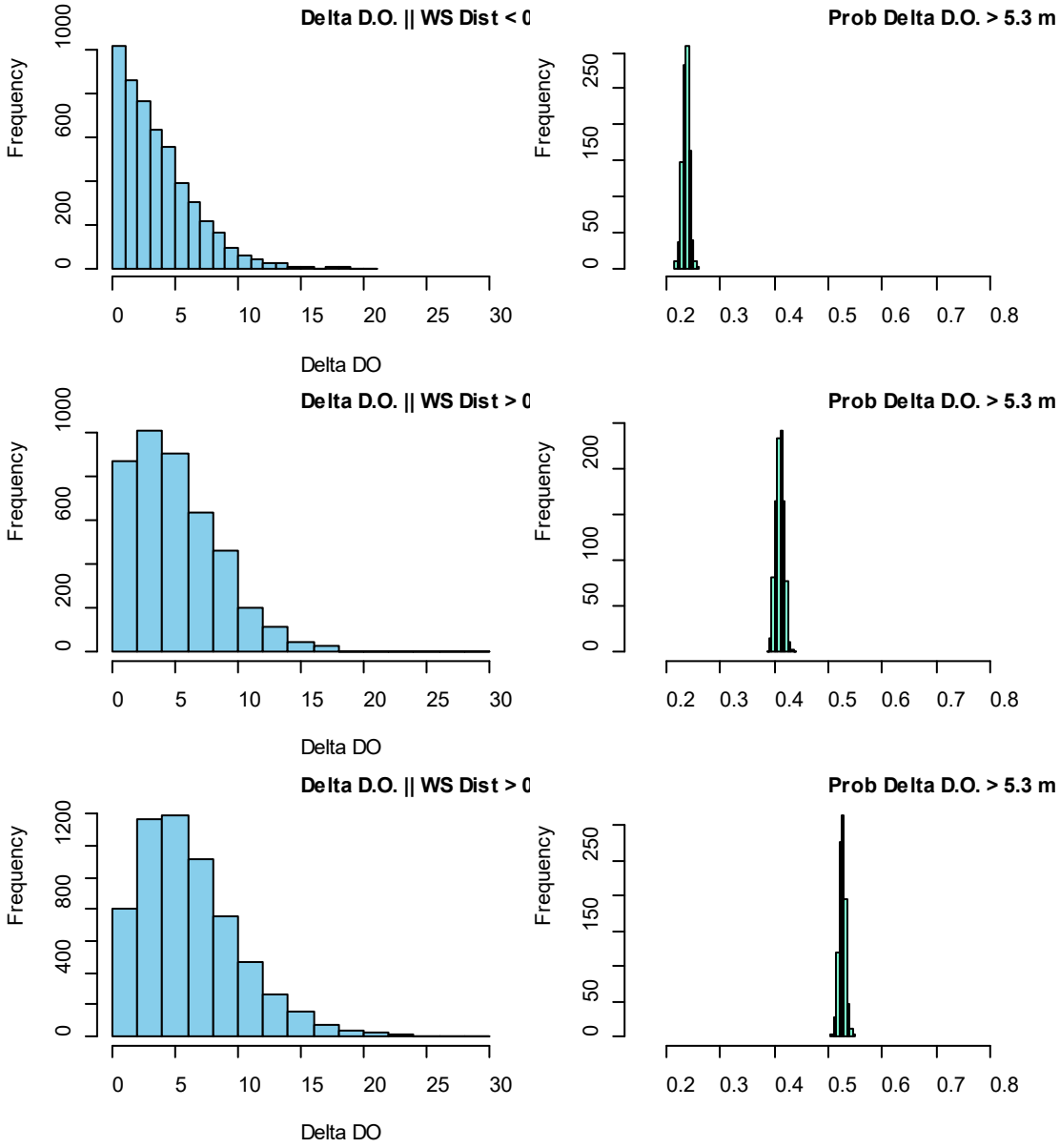


Figure A-4. The left column shows distributions of weekly average delta DO conditioned on land disturbance and drought levels. The level of watershed disturbance was suggested by regression trees. The right-hand column shows the corresponding probabilities of observing the delta DO greater than the threshold value of 5.3 mg/l.

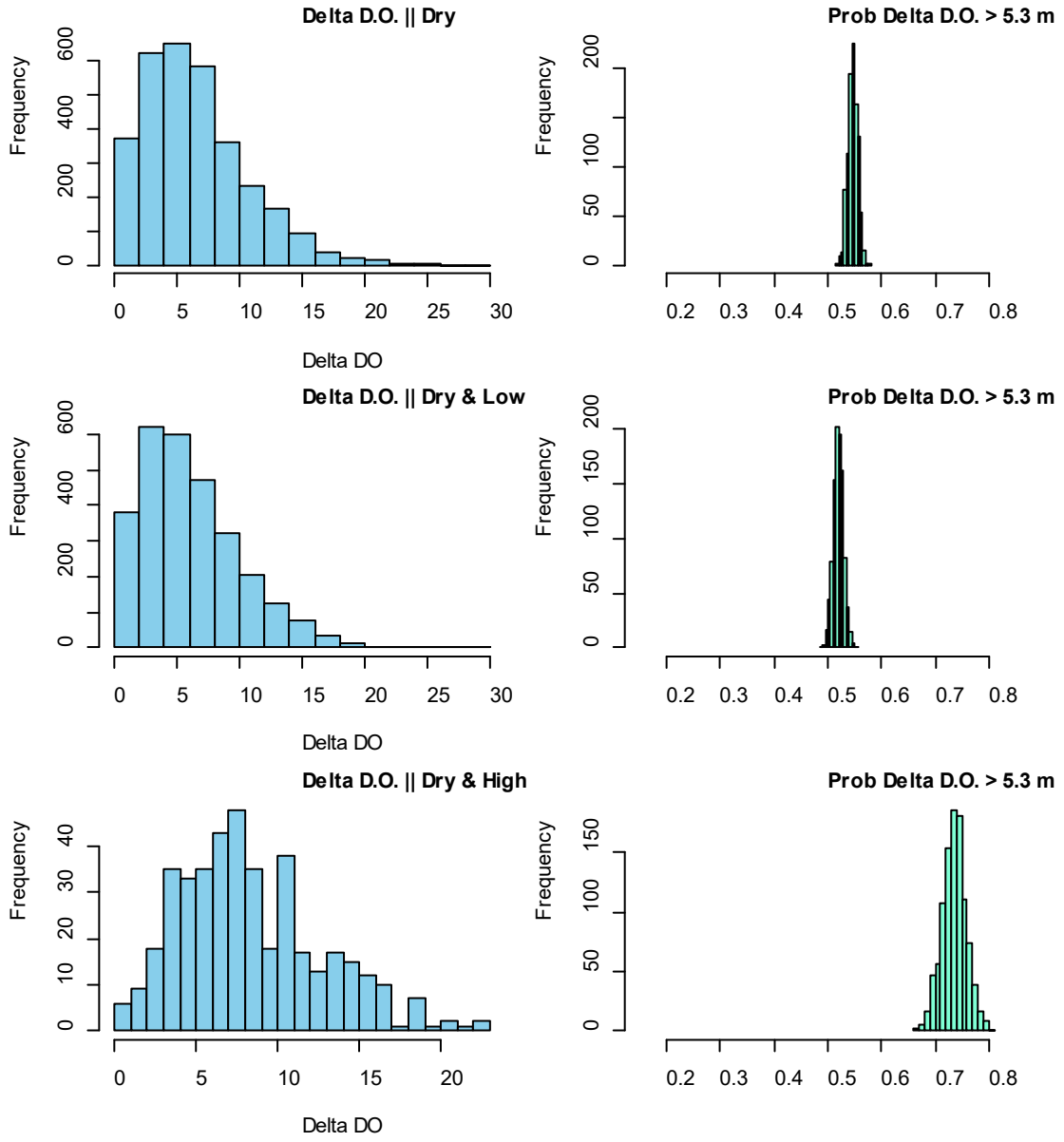


Figure A-5. The left column shows distributions of weekly average delta DO conditioned dry climate, as given by a PMDI of less than -2, and two levels of macrophyte cover. Drought was suggested by observing the distribution of PMDI values (see Figure A.7). The right-hand column shows the corresponding probabilities of observing the delta DO greater than the threshold value of 5.3 mg/l.

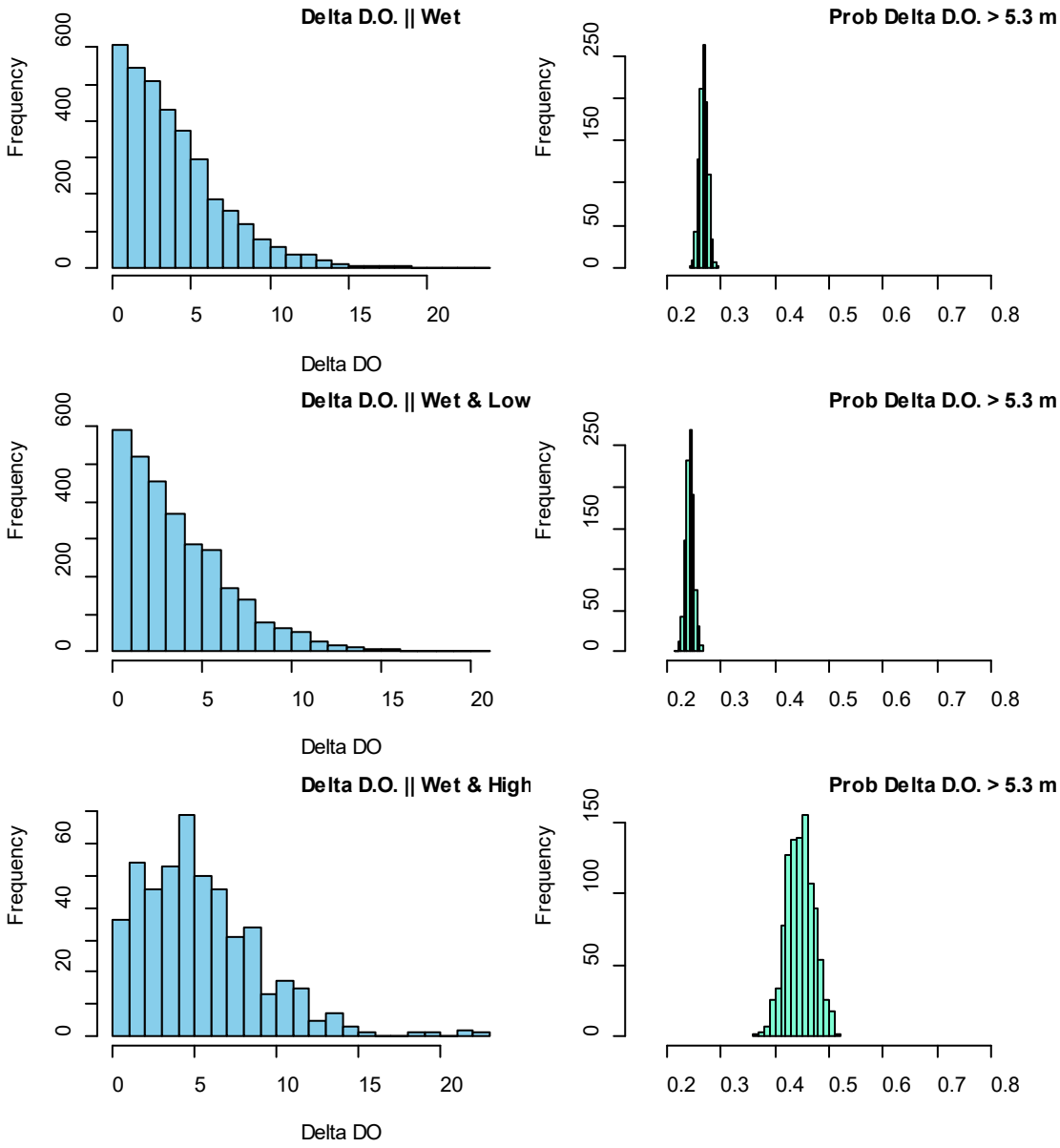


Figure A-6. The left column shows distributions of weekly average delta DO conditioned wet climate, as given by a PMDI greater than 4, and two levels of macrophyte cover. Wet climate was suggested by observing the distribution of PMDI values (see Figure A.7). The right-hand column shows the corresponding probabilities of observing the delta DO greater than the threshold value of 5.3 mg/l.

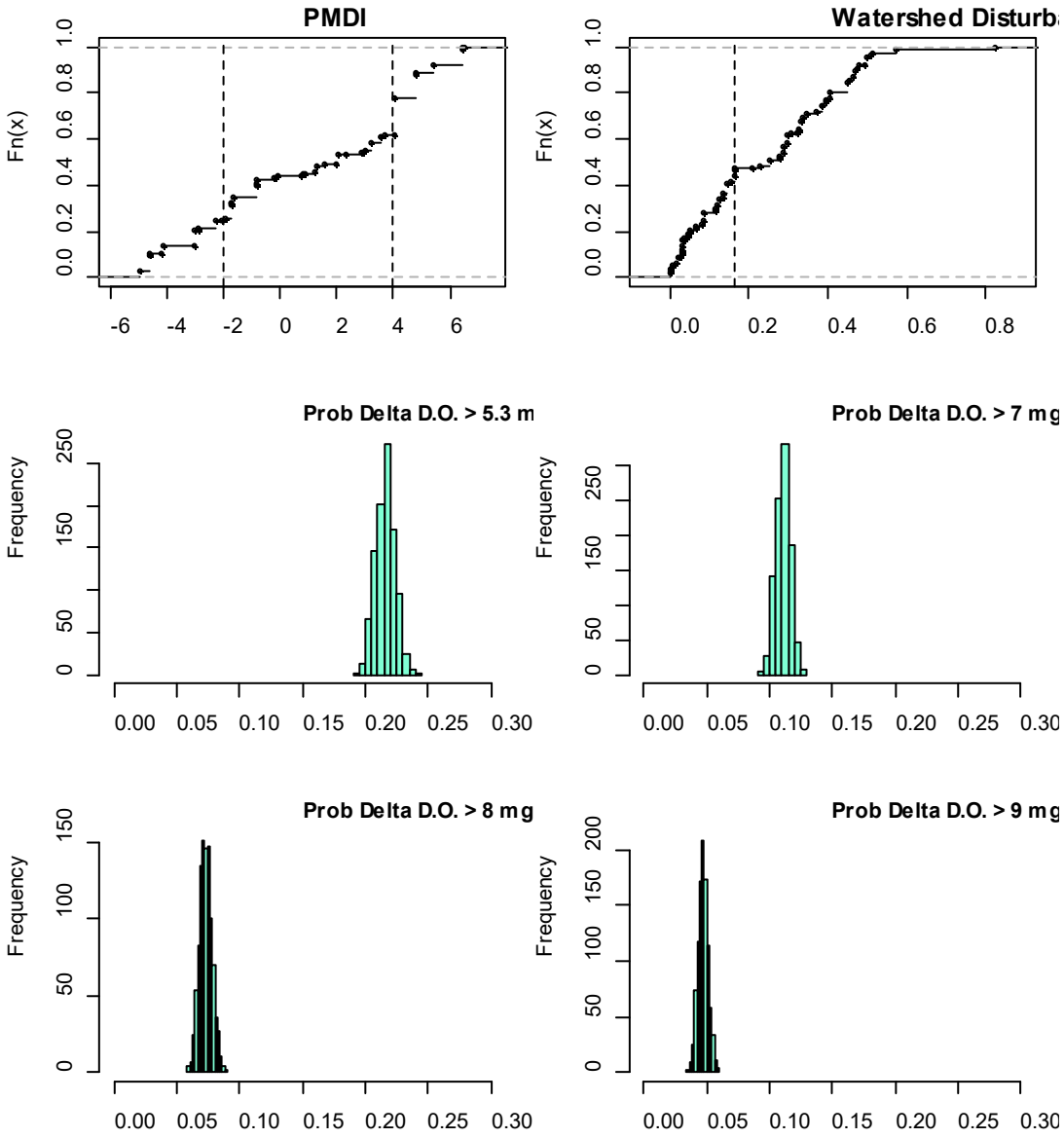


Figure A-7. Probabilities of observing delta DO greater than thresholds shown for each plot for average climate conditions (~the middle quartiles of the PMDI distribution) and low levels (i.e., ≤ 0.163 of land disturbance). The scenarios may help with defining the magnitude and frequency component of a standard appropriate for the eastern Montana study area.

Assessing Model Performance

A separate network model was constructed using only fixed variables in order to compare results with the network model described above. Figure A.8 below shows the correlation of predicted to observed delta DO for 1) the data set containing all variables (n=234), and 2) the data set containing only fixed (i.e., drought indices, land disturbance) variables (n=764). The distributions show the results from 1000 runs of cross fold validations (4 folds of 60 random records for the "all variable"

network model; 5-folds of 152 records for the fixed variable model). The network model containing all variables had better predictive performance but more uncertainty relative to the network containing only fixed variables. The higher degree of uncertainty reflects the lower number of observations included in the folds.

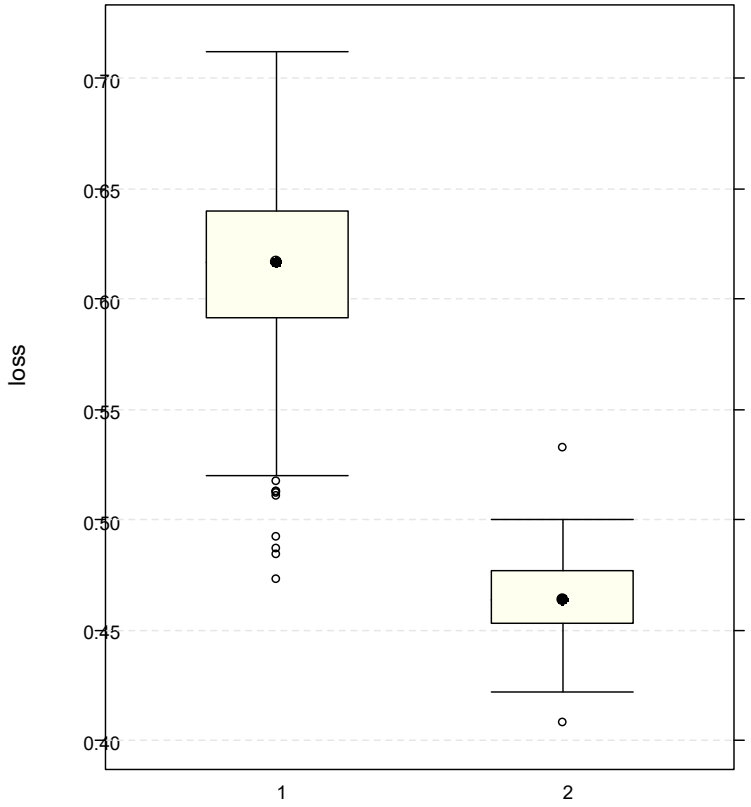


Figure A-8. Correlation of predicted to observed delta DO for (1) all variables (n=234) and (2) only fixed variables (n=764).

A.2. Script Files or Command Files for Running R-based Statistical Analysis

Script / Command File Name	Description
	Regression Trees
`RT_Input.R`	Reads input from MS Excel for the "all variable" dataset and "expanded records" dataset. Sets variable names. Sets variable type necessary for tree modeling. Runs basic exploratory data analysis (Spearman correlation only).
`RT_Model.R`	Performs modeling using classification and regression trees with input from RT_Input.R using either type = anova, poisson, or class. Model formulations established for several response variables.
`RT_Tuning.R`	Performs tree modeling simulation and finds best values for minsplit, cp, and maxdepth parameters.
`RT_Output.R`	Produces output of model runs from RT_Model.R – generates initial cross-validation plot, tree diagram, tree rules, regression surfaces (bivariate plots), summary output, fit plots, and residuals analysis.
	Drought Predictor Development
`WtSum_NDMC_Nweeks.R`	Compute spatially weighted sum for NDMC county-based data (# consecutive weeks) for drought intensities: D0 through D4.
`WtSum_NDMC_Parea.R`	Compute spatially weighted sum for NDMC county-based data (percent area) for drought intensities: D0 through D4.
`WtSum_NOAA.R`	Compute spatially weighted sum for NOAA Climate Division data (Z-Index, PMDI, PHDI).
	Bayesian Network Model
`montana_dag.R`	Imports dodata.rds and bNxx.rds. It draws the network diagram, traces the tuning used to identify the network model, contains examples and instructions for querying the network model, and provides examples for drawing histograms of posterior distributions (i.e., distributions resulting from a query of the model).
`montana_fixed_dag.R`	Imports deltx.rds and dofixN.rds. It traces the step in tuning the initial DAG to a final DAG, and provides instructions and examples for querying the network model, including histograms of posterior distributions resulting from queries to the model.

RT_Input.R

```
## Build File for Statistical Modeling

setwd("C:/Users/…) # work directory

library(readxl)

# read master input file
do_wk <- read_excel("Data_Working.xlsx",sheet="Export_R_All")

# adjust variable types to meet modeling goals
do_wk$ORD <- as.integer(do_wk$ORD)

# set certain variables as factors
do_wk$streamcat <- as.factor(do_wk$streamcat)
do_wk$compsite <- as.factor(do_wk$compsite)

# these are ordered factors
do_wk$exceedMN <- as.ordered(do_wk$exceedMN)
do_wk$exceedMT <- as.ordered(do_wk$exceedMT)
do_wk$exceedOH <- as.ordered(do_wk$exceedOH)
do_wk$wellcat <- as.ordered(do_wk$wellcat)
do_wk$wellocat <- as.ordered(do_wk$wellocat)
do_wk$MARANK <- as.ordered(do_wk$MARANK)
do_wk$FARANK <- as.ordered(do_wk$FARANK)
do_wk$MPHYTERANK <- as.ordered(do_wk$MPHYTERANK)
do_wk$MATHICKMM <- as.ordered(do_wk$MATHICKMM) # thickness assumes 1 of 4 discrete
ordered values

# these variables read-in as character (likely due to NAs present)
do_wk$RELABUND <- as.numeric(do_wk$RELABUND)
do_wk$IMPROB <- as.numeric(do_wk$IMPROB)
do_wk$BP <- as.numeric(do_wk$BP)
do_wk$NH3 <- as.numeric(do_wk$NH3)
do_wk$Nox <- as.numeric(do_wk$Nox)
do_wk$OP <- as.numeric(do_wk$OP)
do_wk$pH <- as.numeric(do_wk$pH)
do_wk$SC <- as.numeric(do_wk$SC)
do_wk$TN <- as.numeric(do_wk$TN)
do_wk$TP <- as.numeric(do_wk$TP)

saveRDS(do_wk, file = "do_wk.rds") # save as an R object for sharing

# read expanded record input files
# expanded weekly
do_exp_wk <- read_excel("Data_Expanded_Working.xlsx",sheet="Export_R_Wk") # weekly -
all DO & matching drought + fixed vars

# set certain variables as factors
do_exp_wk$streamcat <- as.factor(do_exp_wk$streamcat)
do_exp_wk$compsite <- as.factor(do_exp_wk$compsite)

# these are ordered factors
do_exp_wk$exceedMN <- as.ordered(do_exp_wk$exceedMN)
do_exp_wk$exceedMT <- as.ordered(do_exp_wk$exceedMT)
do_exp_wk$exceedOH <- as.ordered(do_exp_wk$exceedOH)
do_exp_wk$wellcat <- as.ordered(do_exp_wk$wellcat)
do_exp_wk$wellocat <- as.ordered(do_exp_wk$wellocat)

# expanded monthly
```

```
do_exp_mon <- read_excel("Data_Expanded_Working.xlsx",sheet="Export_R_Mon") # monthly
- all DO & matching drought + fixed vars

# set certain variables as factors
do_exp_mon$streamcat <- as.factor(do_exp_mon$streamcat)
do_exp_mon$compsite <- as.factor(do_exp_mon$compsite)

# these are ordered factors
do_exp_mon$exceedMN <- as.ordered(do_exp_mon$exceedMN)
do_exp_mon$exceedMT <- as.ordered(do_exp_mon$exceedMT)
do_exp_mon$exceedOH <- as.ordered(do_exp_mon$exceedOH)
do_exp_mon$wellcat <- as.ordered(do_exp_mon$wellcat)
do_exp_mon$wellocat <- as.ordered(do_exp_mon$wellocat)

saveRDS(do_exp_wk, file = "do_exp_wk.rds") # save as an R object for sharing
saveRDS(do_exp_mon, file = "do_exp_mon.rds") # save as an R object for sharing

## exploratory data analysis
par("mar") # should be 5.1 4.1 4.1 2.1
par(mar=c(1,1,1,1))
pairs(do_wk[,-c(1:10)],pch =".",cex = 1.5) # too large - will not plot
pairs(do_wk[,c(11:16)],pch =".",cex = 1.5) # response variables only

summary(do_wk)
summary(do_exp_wk)
summary(do_exp_mon)

library(pastecs)
res <- stat.desc(do_wk[,-c(1:10)])
# round(res,2)
write.csv(res,file="../stat_desc.csv",row.names=TRUE)
```

RT_Model.R

```
## Statistical Modeling - rpart model runs
# See RT_Input.R script for how datasets for model input were built

setwd("C:/Users/…) # work directory

library(rpart) # load the CART library
library(rpart.plot) # load the fancy plotting library
library(plotmo) # for plotting regression surfaces

# tuning parameters...
# cpset=0.015 # set complexity parameter
# mxdep=3 # set maximum depth of any node of tree (depth = 0 at root node)
# sur=2 # how, if at all, to use surrogates -- if =2 mean if all surrogates are
missing, then send the observation in the majority direction
#
rpart.control(minsplit=20,cp=cpset,maxcompete=4,maxsurrogate=5,usesurrogate=sur,xval=1
0,surrogatestyle=0,maxdepth=mxdep)

# initial diagnostic - to find best cp run rpart with low cp (around 0.01 to 0.005)
# then run plotcp() in RT_Output.R

# ------------------------------------------------------------------------------
## Quantitative Response - All Predictors
# set x,y TRUE in rpart to use roundint in rpart.plot()
# set maxdepth
## DO
rpart_out <-
rpart(xdelta~DA+streamcat+compsite+MARANK+MATHICKMM+FARANK+MPHYTERANK+RELABUND+IMPROB+
medt+maxt_avg+
BP+NH3+Nox+OP+pH+SC+TN+TP+TPe+TNe+Zindex+PMDI+PHDI+DSCI+DSCt+Dzero+Done+Dtwo+Dthree+Df
our+
natws+distws+natnf+distnf+wells+wellso+wellcat+wellocat+maxslope+medslope+xslope+devsl
ope,
               data=do_wk,method="anova",x=FALSE,y=FALSE,
               control=rpart.control(cp=0.033,usesurrogate=2))

rpart_out <-
rpart(mxdelta~DA+streamcat+compsite+MARANK+MATHICKMM+FARANK+MPHYTERANK+RELABUND+IMPROB
+medt+maxt_avg+

BP+NH3+Nox+OP+pH+SC+TN+TP+TPe+TNe+Zindex+PMDI+PHDI+DSCI+DSCt+Dzero+Done+Dtwo+Dthree+Df
our+

natws+distws+natnf+distnf+wells+wellso+wellcat+wellocat+maxslope+medslope+xslope+devsl
ope,
                  data=do_wk,method="anova",x=FALSE,y=FALSE,
                  control=rpart.control(cp=0.043,usesurrogate=2))

rpart_out <-
rpart(min_avg~DA+streamcat+compsite+MARANK+MATHICKMM+FARANK+MPHYTERANK+RELABUND+IMPROB
+medt+maxt_avg+

BP+NH3+Nox+OP+pH+SC+TN+TP+TPe+TNe+Zindex+PMDI+PHDI+DSCI+DSCt+Dzero+Done+Dtwo+Dthree+Df
our+

natws+distws+natnf+distnf+wells+wellso+wellcat+wellocat+maxslope+medslope+xslope+devsl
ope,
                  data=do_wk,method="anova",x=FALSE,y=FALSE,
                  control=rpart.control(cp=0.025,usesurrogate=2))
```

```
# ------------------------------------------------------------------------------------
## Ordinal (Factor) Response - All Predictors
## Plant
# removed MARANK from predictor set (obvious strong correlation)
# changed method from anova to class - thickness assumes 1 of 4 discrete ordered
values
rpart_out <-
rpart(MATHICKMM~DA+streamcat+compsite+FARANK+MPHYTERANK+RELABUND+IMPROB+medt+maxt_avg+

BP+NH3+Nox+OP+pH+SC+TN+TP+TPe+TNe+Zindex+PMDI+PHDI+DSCI+DSCt+Dzero+Done+Dtwo+Dthree+Df
our+
natws+distws+natnf+distnf+wells+wellso+wellcat+wellocat+maxslope+medslope+xslope+devsl
ope,
                   data=do_wk,method="class",
                   control=rpart.control(cp=0.069,usesurrogate=2))

rpart_out <-
rpart(MPHYTERANK~DA+streamcat+compsite+MARANK+MATHICKMM+FARANK+RELABUND+IMPROB+medt+ma
xt_avg+

BP+NH3+Nox+OP+pH+SC+TN+TP+TPe+TNe+Zindex+PMDI+PHDI+DSCI+DSCt+Dzero+Done+Dtwo+Dthree+Df
our+

natws+distws+natnf+distnf+wells+wellso+wellcat+wellocat+maxslope+medslope+xslope+devsl
ope,
                   data=do_wk,method="class",
                   control=rpart.control(cp=0.0077,usesurrogate=2))

# ------------------------------------------------------------------------------------
## Count Response - All Predictors
## DO
rpart_out <-
rpart(exceedMT~DA+streamcat+compsite+MARANK+MATHICKMM+FARANK+MPHYTERANK+RELABUND+IMPRO
B+medt+maxt_avg+

BP+NH3+Nox+OP+pH+SC+TN+TP+TPe+TNe+Zindex+PMDI+PHDI+DSCI+DSCt+Dzero+Done+Dtwo+Dthree+Df
our+

natws+distws+natnf+distnf+wells+wellso+wellcat+wellocat+maxslope+medslope+xslope+devsl
ope,
                  data=do_wk,method="poisson",
                  control=rpart.control(cp=0.039,usesurrogate=2))

# ------------------------------------------------------------------------------------
## Quantitative Response - Expanded Rows - Weekly
## DO
rpart_out <-
rpart(xdelta~DA+streamcat+compsite+avgt+maxt_avg+DSCI+DSCt+Dzero+Done+Dtwo+Dthree+Dfou
r+natws+distws+

natnf+distnf+wells+wellso+wellcat+wellocat+maxslope+medslope+xslope+devslope,
                   data=do_exp_wk,method="anova",
                   control=rpart.control(cp=0.011,usesurrogate=2))

## Count Response - Expanded Rows - Weekly
## DO
rpart_out <-
rpart(exceedMT~DA+streamcat+compsite+avgt+maxt_avg+DSCI+DSCt+Dzero+Done+Dtwo+Dthree+Df
our+natws+distws+

natnf+distnf+wells+wellso+wellcat+wellocat+maxslope+medslope+xslope+devslope,
                   data=do_exp_wk,method="poisson",
                   control=rpart.control(cp=0.018,usesurrogate=2))
```

```
# ------------------------------------------------------------------------------------
## Quantitative Response - Expanded Rows - Monthly
## DO
rpart_out <-
rpart(xdelta~DA+streamcat+compsite+avgt+maxt_avg+Zindex+PMDI+PHDI+natws+distws+natnf+d
istnf+wells+wellso+
                      wellcat+wellocat+maxslope+medslope+xslope+devslope,
                    data=do_exp_mon,method="anova")

## Count Response - Expanded Rows - Monthly
## DO
rpart_out <-
rpart(exceedMT~DA+streamcat+compsite+avgt+maxt_avg+Zindex+PMDI+PHDI+natws+distws+natnf
+distnf+wells+wellso+
                      wellcat+wellocat+maxslope+medslope+xslope+devslope,
                    data=do_exp_mon,method="poisson")
```

RT_Tuning.R

```
## Statistical Modeling -- Tuning rpart
# See RT_Input.R script for how datasets for model input were built

setwd("C:/Users/Dale/OneDrive - The Ohio State University/Documents/MT_DEQ/Modeling")
# work directory
setwd("C:/Users/white/OneDrive - The Ohio State University/Documents/MT_DEQ/MOdeling")
# home directory

install.packages("rpart")
install.packages("rpart.plot")

library(rpart) # load the CART library
library(rpart.plot) # load the fancy plotting library
library(plotmo) # for plotting regression surfaces
library(e1071)

## Quantitative Response - All Predictors
## DO
fm <-
formula(xdelta~DA+streamcat+compsite+MARANK+MATHICKMM+FARANK+MPHYTERANK+RELABUND+IMPRO
B+medt+maxt_avg+

BP+NH3+Nox+OP+pH+SC+TN+TP+TPe+TNe+Zindex+PMDI+PHDI+DSCI+DSCt+Dzero+Done+Dtwo+Dthree+Df
our+

natws+distws+natnf+distnf+wells+wellso+wellcat+wellocat+maxslope+medslope+xslope+devsl
ope)

fm <-
formula(MATHICKMM~DA+streamcat+compsite+FARANK+MPHYTERANK+RELABUND+IMPROB+medt+maxt_av
g+BP+NH3+Nox+OP+

pH+SC+TN+TP+TPe+TNe+Zindex+PMDI+PHDI+DSCI+DSCt+Dzero+Done+Dtwo+Dthree+Dfour+natws+dist
ws+

natnf+distnf+wells+wellso+wellcat+wellocat+maxslope+medslope+xslope+devslope)

# ------------------------------------------------------------------------------
# Explore minsplit
audit.rpart <- tune.rpart(fm,data=do_wk,minsplit=seq(from=10,to=100,by=10))
plot(audit.rpart,main="Tune rpart on minsplit")

# Explore cp
audit.rpart <- tune.rpart(fm,data=do_wk,cp=c(0.002,0.005,0.01,0.015,0.02,0.03))
plot(audit.rpart,main="Performance of rpart vs. cp")
readline()

# Explore maxdepth
audit.rpart <- tune.rpart(fm,data=do_wk,maxdepth=1:10)
plot(audit.rpart,main="Performance of rpart vs. maxdepth")
readline()
# ------------------------------------------------------------------------------
```

RT_Output.R

```
## Statistical Modeling - presentation (table, plots) of model results
# See RT_Input.R script for how datasets for model input were built
# See RT_Model.R script for model runs

setwd("C:/Users/Dale/OneDrive - The Ohio State University/Documents/MT_DEQ/Modeling")
# work directory
setwd("C:/Users/white/OneDrive - The Ohio State University/Documents/MT_DEQ/MOdeling")
# home directory

library(rpart) # load the CART library
library(rpart.plot) # load the fancy plotting library
library(plotmo) # for plotting regression surfaces

# ---------------------------------------------------------------------------------

#0 initial diagnostic - to find best cp run rpart with low cp (around 0.01 to 0.005)
# plotcp plots xerror vs. cp & size or nsplits
par(mfrow=c(1,1)) # plot 1 to a page (1row, 1col)
plotcp(rpart_out,minline=TRUE,lty=3,col=1,upper="splits")  # upper "size" or "splits"
or "none"
# minline: whether a horizontal line is drawn 1SE above the minimum of the curve
# good choice of cp for pruning is often leftmost value for which mean lies below
horizontal line

# ---------------------------------------------------------------------------------

#1
# plot a rpart tree
title="Dissolved Oxygen Delta Response"
title="Aquatic Plant Response"
title=NULL

rpart.plot(rpart_out,type=2,extra=101,under=TRUE,faclen=0,varlen=0,digits=3,roundint=T
RUE,yesno=1,main=title)
# roundint is active when access to data used to build model
# extra="auto" lets R decide best

#2
# associated rules - show them
# nn prints node number; cover prints % obs belonging to class
cat("\014") # clear console
rpart.rules(rpart_out,cover=TRUE,nn=FALSE,style="tall")
rules_out <- rpart.rules(rpart_out,cover=TRUE,nn=FALSE,style="wide")

# rules used to make a prediction
rpart.predict(rpart_out,rules=TRUE) # not working...

#3
# plotmo functions
# plotmo(rpart_out,degree2=c("distws","SC","Dzero","maxslope","TPe"),grid.col=TRUE)
# plotmo(rpart_out,grid.col=TRUE,all2=TRUE)
plotmo(rpart_out,grid.col=TRUE) # use this for quantitative (anova) or rate (poisson)
response
plotmo(rpart_out,grid.col=TRUE,type="class") # use this plotmo function for ranked
(class) response

#4
# table output
cat("\014") # clear console
printcp(rpart_out,digits=3) # output of printcp also included in summary command
```

```
summary(rpart_out,digits=3) # can limit output by setting cp in function
# or have summary write to text file
# summary(rpart_out,digits=3,"../Output_Model/rpart_summary.txt") # not sure why not
writing to file

#5
# construct fit plots
par(mfrow=c(1,2)) # plot 2 side-by-side (1row, 2col)
rsq.rpart(rpart_out) # may only be applicable for method=anova

#6
# residuals analysis
plotres(rpart_out,which=2:4,jitter=1) # plot residuals - in plotmo package / jitter=0
no jitter

#7
# terminal nodes - list of sites
# caution:  each record (observation) is a location (SITE) and date (YRMW) so one
record will be
#           one location but may exist in >=1 terminal node (classification result)
rpart_out$frame
do_wk[215,] # yields entire line at this rec# (the case# labeled on the residuals
plot)

# table(rpart_out$where)
# ordinary dataset
map_rpart <-
cbind.data.frame(SITEYRMW=do_wk$SITEYRMW,StationID=do_wk$StationID,DateX=do_wk$DateX,

RecNoFrame=rpart_out$where,NodeName=rpart_out$frame[rpart_out$where,1],
                              NodeNobs=rpart_out$frame[rpart_out$where,2],
                              NodeDeviance=rpart_out$frame[rpart_out$where,4],
                              NodeFittedValue=rpart_out$frame[rpart_out$where,5])
# expanded dataset
map_rpart <-
cbind.data.frame(SITEYRMW=do_exp_wk$SITEYRMW,StationID=do_exp_wk$Station_ID,

RecNoFrame=rpart_out$where,NodeName=rpart_out$frame[rpart_out$where,1],
                              NodeNobs=rpart_out$frame[rpart_out$where,2],
                              NodeDeviance=rpart_out$frame[rpart_out$where,4],
                              NodeFittedValue=rpart_out$frame[rpart_out$where,5])
write.csv(map_rpart,"Output/map_rpart.csv",row.names=FALSE)
write.csv(rules_out,"Output/rules_out.csv",row.names=FALSE)

# ------------------------------------------------------------------------------


# examine distribution of residuals
plot(jitter(residuals(rpart_out))~predict(rpart_out)) # not working...
abline(h=0,col="grey",lwd=1.5,lty=2) # draw straight line through origin

# deviance of the node divided by the number of observations at the node. Also returns
the node number
meanvar(rpart_out, xlab ="average(y)",ylab="average(deviance)")
```

WtSum_NDMC_Nweeks.R

```
## Compute Weighted Sum -- NDMC County Data (# Consecutive Weeks)

setwd("C:/Users/Dale/OneDrive - The Ohio State
University/Documents/MT_DEQ/Precipitation/Drought/Drought Indices/NDMC") # work
directory
setwd("C:/Users/white/OneDrive - The Ohio State
University/Documents/MT_DEQ/Precipitation/Drought/Drought Indices/NDMC") # home
directory

install.packages("xlsx")
library(readxl)

# input station and area percent (of county) data
Stn_WtMC <- read_excel("../Tab_Wshed.xlsx",sheet ="NDMC_Cnty_WshedR")
Stn_WtMCm <- as.matrix(Stn_WtMC[,-1]) # remove Station_ID (for now); produces [73,15]
matrix

# Drought Level = 0
ConWeeks_D0 <- read_excel("ConsecutiveWeeks.xlsx",sheet ="Week_Year_D0_R") # 15
counties by 261 dates [15,261]

# matrix multiplication (inner product)
# remove 1st text col in ConWeeks_Dn dataframe; convert to matrix
Stn_WtD0 <- Stn_WtMCm %*% as.matrix(ConWeeks_D0[,-1]) # produces a [73,261] result
matrix - 73 stations by 261 dates

# Drought Level = 1
ConWeeks_D1 <- read_excel("ConsecutiveWeeks.xlsx",sheet ="Week_Year_D1_R") # 15
counties by 261 dates [15,261]

# matrix multiplication (inner product)
# remove 1st text col in ConWeeks_Dn dataframe; convert to matrix
Stn_WtD1 <- Stn_WtMCm %*% as.matrix(ConWeeks_D1[,-1]) # produces a [73,261] result
matrix - 73 stations by 261 dates

# Drought Level = 2
ConWeeks_D2 <- read_excel("ConsecutiveWeeks.xlsx",sheet ="Week_Year_D2_R") # 15
counties by 261 dates [15,261]

# matrix multiplication (inner product)
# remove 1st text col in ConWeeks_Dn dataframe; convert to matrix
Stn_WtD2 <- Stn_WtMCm %*% as.matrix(ConWeeks_D2[,-1]) # produces a [73,261] result
matrix - 73 stations by 261 dates

# Drought Level = 3
ConWeeks_D3 <- read_excel("ConsecutiveWeeks.xlsx",sheet ="Week_Year_D3_R") # 15
counties by 261 dates [15,261]

# matrix multiplication (inner product)
# remove 1st text col in ConWeeks_Dn dataframe; convert to matrix
Stn_WtD3 <- Stn_WtMCm %*% as.matrix(ConWeeks_D3[,-1]) # produces a [73,261] result
matrix - 73 stations by 261 dates

# Drought Level = 4
ConWeeks_D4 <- read_excel("ConsecutiveWeeks.xlsx",sheet ="Week_Year_D4_R") # 15
counties by 261 dates [15,261]

# matrix multiplication (inner product)
# remove 1st text col in ConWeeks_Dn dataframe; convert to matrix
Stn_WtD4 <- Stn_WtMCm %*% as.matrix(ConWeeks_D4[,-1]) # produces a [73,261] result
matrix - 73 stations by 261 dates
```

```
## prepare datasets for export
# build station ID column
Stn_NamesMCv <- as.vector(t(Stn_WtMC[,1])) # build vector of station names
Stn_NamesMC_WkvR <- sort(rep(Stn_NamesMCv,times=261)) # repeat station names 261 times
(for 261 dates) & sort to match data

# build date column
Week_Year <-
c("1/4/13","1/11/13","1/18/13","1/25/13","2/1/13","2/8/13","2/15/13","2/22/13","3/1/13
","3/8/13","3/15/13","3/22/13","3/29/13",

"4/5/13","4/12/13","4/19/13","4/26/13","5/3/13","5/10/13","5/17/13","5/24/13","5/31/13
","6/7/13","6/14/13","6/21/13","6/28/13","7/5/13",

"7/12/13","7/19/13","7/26/13","8/2/13","8/9/13","8/16/13","8/23/13","8/30/13","9/6/13"
,"9/13/13","9/20/13","9/27/13","10/4/13","10/11/13",

"10/18/13","10/25/13","11/1/13","11/8/13","11/15/13","11/22/13","11/29/13","12/6/13","
12/13/13","12/20/13","12/27/13","1/3/14","1/10/14",

"1/17/14","1/24/14","1/31/14","2/7/14","2/14/14","2/21/14","2/28/14","3/7/14","3/14/14
","3/21/14","3/28/14","4/4/14","4/11/14","4/18/14",

"4/25/14","5/2/14","5/9/14","5/16/14","5/23/14","5/30/14","6/6/14","6/13/14","6/20/14"
,"6/27/14","7/4/14","7/11/14","7/18/14","7/25/14",

"8/1/14","8/8/14","8/15/14","8/22/14","8/29/14","9/5/14","9/12/14","9/19/14","9/26/14"
,"10/3/14","10/10/14","10/17/14","10/24/14","10/31/14",

"11/7/14","11/14/14","11/21/14","11/28/14","12/5/14","12/12/14","12/19/14","12/26/14",
"1/2/15","1/9/15","1/16/15","1/23/15","1/30/15",

"2/6/15","2/13/15","2/20/15","2/27/15","3/6/15","3/13/15","3/20/15","3/27/15","4/3/15"
,"4/10/15","4/17/15","4/24/15","5/1/15","5/8/15",

"5/15/15","5/22/15","5/29/15","6/5/15","6/12/15","6/19/15","6/26/15","7/3/15","7/10/15
","7/17/15","7/24/15","7/31/15","8/7/15","8/14/15",

"8/21/15","8/28/15","9/4/15","9/11/15","9/18/15","9/25/15","10/2/15","10/9/15","10/16/
15","10/23/15","10/30/15","11/6/15","11/13/15",

"11/20/15","11/27/15","12/4/15","12/11/15","12/18/15","12/25/15","1/1/16","1/8/16","1/
15/16","1/22/16","1/29/16","2/5/16","2/12/16",

"2/19/16","2/26/16","3/4/16","3/11/16","3/18/16","3/25/16","4/1/16","4/8/16","4/15/16"
,"4/22/16","4/29/16","5/6/16","5/13/16","5/20/16",

"5/27/16","6/3/16","6/10/16","6/17/16","6/24/16","7/1/16","7/8/16","7/15/16","7/22/16"
,"7/29/16","8/5/16","8/12/16","8/19/16","8/26/16",

"9/2/16","9/9/16","9/16/16","9/23/16","9/30/16","10/7/16","10/14/16","10/21/16","10/28
/16","11/4/16","11/11/16","11/18/16","11/25/16",

"12/2/16","12/9/16","12/16/16","12/23/16","12/30/16","1/6/17","1/13/17","1/20/17","1/2
7/17","2/3/17","2/10/17","2/17/17","2/24/17","3/3/17",

"3/10/17","3/17/17","3/24/17","3/31/17","4/7/17","4/14/17","4/21/17","4/28/17","5/5/17
","5/12/17","5/19/17","5/26/17","6/2/17","6/9/17",

"6/16/17","6/23/17","6/30/17","7/7/17","7/14/17","7/21/17","7/28/17","8/4/17","8/11/17
","8/18/17","8/25/17","9/1/17","9/8/17","9/15/17",
```

```
"9/22/17","9/29/17","10/6/17","10/13/17","10/20/17","10/27/17","11/3/17","11/10/17","1
1/17/17","11/24/17","12/1/17","12/8/17","12/15/17",
               "12/22/17","12/29/17")

# replicate but do not sort (in order to match station names & data)
Week_YearR <- rep(Week_Year,times=73) # repeat date 73 times (73 stations)

# convert drought matrices to column vectors
Stn_WtD0v <- as.vector(t(Stn_WtD0)) # D0
Stn_WtD1v <- as.vector(t(Stn_WtD1)) # D1
Stn_WtD2v <- as.vector(t(Stn_WtD2)) # D2
Stn_WtD3v <- as.vector(t(Stn_WtD3)) # D3
Stn_WtD4v <- as.vector(t(Stn_WtD4)) # D4

# combine column vectors into one dataframe
NDMC_Week_All <-
data.frame(Stn_NamesMC_WkvR,Week_YearR,Stn_WtD0v,Stn_WtD1v,Stn_WtD2v,Stn_WtD3v,Stn_WtD
4v,row.names = NULL)
names(NDMC_Week_All)[1] <- "Station_ID"
names(NDMC_Week_All)[2] <- "Week_Year"
names(NDMC_Week_All)[3] <- "D0_wt"
names(NDMC_Week_All)[4] <- "D1_wt"
names(NDMC_Week_All)[5] <- "D2_wt"
names(NDMC_Week_All)[6] <- "D3_wt"
names(NDMC_Week_All)[7] <- "D4_wt"

# export data
write.csv(NDMC_Week_All,file="../NDMC_Nweeks.csv",row.names=FALSE)

# write.csv(data.frame(Stn_WtD0),file="../Stn_WtD0.csv",row.names=FALSE) # for testing
purposes only
```

WtSum_NDMC_Parea.R

```
## Compute Weighted Sum -- NDMC County Data (% Area County)

setwd("C:/Users/Dale/OneDrive - The Ohio State
University/Documents/MT_DEQ/Precipitation/Drought/Drought Indices/NDMC") # work
directory
setwd("C:/Users/white/OneDrive - The Ohio State
University/Documents/MT_DEQ/Precipitation/Drought/Drought Indices/NDMC") # home
directory

install.packages("xlsx")
library(readxl)

# input station and area percent (of county) data
Stn_WtMC <- read_excel("../Tab_Wshed.xlsx",sheet ="NDMC_Cnty_WshedR")
Stn_WtMCm <- as.matrix(Stn_WtMC[,-1]) # remove Station_ID (for now); produces [73,15]
matrix

# DSCI by County (%area)
# produces a [15,261] result matrix - 15 counties by 261 dates (~52 weeks over 5 yrs)
DSCI_wt <- read_excel("AreaPercent.xlsx",sheet ="DSCI_Rready") # input drought index
data; already conformable

# matrix multiplication (inner product)
# remove 1st col of station ID; convert DSCI to matrix form
Stn_WtDSCI <- Stn_WtMCm %*% as.matrix(DSCI_wt[,-1]) # produces a [73,261] result
matrix - 73 stations by 261 dates (weeks over 5 yrs)


## prepare datasets for export
# build station ID column
Stn_NamesMCv <- as.vector(t(Stn_WtMC[,1])) # build vector of station names
Stn_NamesMCvR <- sort(rep(Stn_NamesMCv,times=261)) # repeat station names 261 times
(for 5 yrs x ~52 of data) & sort to match data

# build date column
Week_Year <-
c("1/4/13","1/11/13","1/18/13","1/25/13","2/1/13","2/8/13","2/15/13","2/22/13","3/1/13
","3/8/13","3/15/13","3/22/13","3/29/13",

"4/5/13","4/12/13","4/19/13","4/26/13","5/3/13","5/10/13","5/17/13","5/24/13","5/31/13
","6/7/13","6/14/13","6/21/13","6/28/13","7/5/13",

"7/12/13","7/19/13","7/26/13","8/2/13","8/9/13","8/16/13","8/23/13","8/30/13","9/6/13"
,"9/13/13","9/20/13","9/27/13","10/4/13","10/11/13",

"10/18/13","10/25/13","11/1/13","11/8/13","11/15/13","11/22/13","11/29/13","12/6/13","
12/13/13","12/20/13","12/27/13","1/3/14","1/10/14",

"1/17/14","1/24/14","1/31/14","2/7/14","2/14/14","2/21/14","2/28/14","3/7/14","3/14/14
","3/21/14","3/28/14","4/4/14","4/11/14","4/18/14",

"4/25/14","5/2/14","5/9/14","5/16/14","5/23/14","5/30/14","6/6/14","6/13/14","6/20/14"
,"6/27/14","7/4/14","7/11/14","7/18/14","7/25/14",

"8/1/14","8/8/14","8/15/14","8/22/14","8/29/14","9/5/14","9/12/14","9/19/14","9/26/14"
,"10/3/14","10/10/14","10/17/14","10/24/14","10/31/14",

"11/7/14","11/14/14","11/21/14","11/28/14","12/5/14","12/12/14","12/19/14","12/26/14",
"1/2/15","1/9/15","1/16/15","1/23/15","1/30/15",
```

```
"2/6/15","2/13/15","2/20/15","2/27/15","3/6/15","3/13/15","3/20/15","3/27/15","4/3/15"
,"4/10/15","4/17/15","4/24/15","5/1/15","5/8/15",

"5/15/15","5/22/15","5/29/15","6/5/15","6/12/15","6/19/15","6/26/15","7/3/15","7/10/15
","7/17/15","7/24/15","7/31/15","8/7/15","8/14/15",

"8/21/15","8/28/15","9/4/15","9/11/15","9/18/15","9/25/15","10/2/15","10/9/15","10/16/
15","10/23/15","10/30/15","11/6/15","11/13/15",

"11/20/15","11/27/15","12/4/15","12/11/15","12/18/15","12/25/15","1/1/16","1/8/16","1/
15/16","1/22/16","1/29/16","2/5/16","2/12/16",

"2/19/16","2/26/16","3/4/16","3/11/16","3/18/16","3/25/16","4/1/16","4/8/16","4/15/16"
,"4/22/16","4/29/16","5/6/16","5/13/16","5/20/16",

"5/27/16","6/3/16","6/10/16","6/17/16","6/24/16","7/1/16","7/8/16","7/15/16","7/22/16"
,"7/29/16","8/5/16","8/12/16","8/19/16","8/26/16",

"9/2/16","9/9/16","9/16/16","9/23/16","9/30/16","10/7/16","10/14/16","10/21/16","10/28
/16","11/4/16","11/11/16","11/18/16","11/25/16",

"12/2/16","12/9/16","12/16/16","12/23/16","12/30/16","1/6/17","1/13/17","1/20/17","1/2
7/17","2/3/17","2/10/17","2/17/17","2/24/17","3/3/17",

"3/10/17","3/17/17","3/24/17","3/31/17","4/7/17","4/14/17","4/21/17","4/28/17","5/5/17
","5/12/17","5/19/17","5/26/17","6/2/17","6/9/17",

"6/16/17","6/23/17","6/30/17","7/7/17","7/14/17","7/21/17","7/28/17","8/4/17","8/11/17
","8/18/17","8/25/17","9/1/17","9/8/17","9/15/17",

"9/22/17","9/29/17","10/6/17","10/13/17","10/20/17","10/27/17","11/3/17","11/10/17","1
1/17/17","11/24/17","12/1/17","12/8/17","12/15/17",
 "12/22/17","12/29/17")

# replicate but do not sort (in order to match station names & data)
Week_YearR <- rep(Week_Year,times=73) # repeat date 73 times (73 stations)

# convert drought matrices to column vectors
Stn_WtDSCIv <- as.vector(t(Stn_WtDSCI)) # DSCI


# combine column vectors into one dataframe
NDMC_Wt_All <- data.frame(Stn_NamesMCvR,Week_YearR,Stn_WtDSCIv,row.names = NULL)
names(NDMC_Wt_All)[1] <- "Station_ID"
names(NDMC_Wt_All)[2] <- "Week_Year"
names(NDMC_Wt_All)[3] <- "DSCI_wt"

# export data
write.csv(NDMC_Wt_All,file="../NDMC_Parea.csv",row.names=FALSE)
```

WtSum_NOAA.R

```
## Compute Weighted Sum -- NOAA Climate Division Data (Z-Index, PMDI, PHDI)

setwd("C:/Users/Dale/OneDrive - The Ohio State
University/Documents/MT_DEQ/Precipitation/Drought/Drought Indices/NOAA") # work
directory
setwd("C:/Users/white/OneDrive - The Ohio State
University/Documents/MT_DEQ/Precipitation/Drought/Drought Indices/NOAA") # home
directory

install.packages("xlsx")
install.packages("readxl")

library(xlsx)
library(xlsxjars)
library(readxl)

# input station and area percent (of climate division) data
Stn_Wt <- read_excel("../Tab_Wshed.xlsx",sheet ="NOAA_CDiv_WshedR")
Stn_Wtm <- as.matrix(Stn_Wt[,-1]) # remove Station_ID (for now); produces [73,6]
matrix

# Z-index
Zindex <- read_excel("data-sub.xlsx",sheet ="Z_Div_TimeSeriesR")

# remove all cols but Z-index, convert to matrix, and transpose
ZindexT <- t(as.matrix(Zindex[,-c(1:4)])) # produces a [6,60] matrix

# matrix multiplication (inner product)
Stn_WtZ <- Stn_Wtm %*% ZindexT # produces a [73,60] result matrix - 73 stations by 60
dates


# PMDI index
Mindex <- read_excel("data-sub.xlsx",sheet ="PMDI_Div_TimeSeriesR")

# remove all cols but Z-index, convert to matrix, and transpose
MindexT <- t(as.matrix(Mindex[,-c(1:4)])) # produces a [6,60] matrix

# matrix multiplication (inner product)
Stn_WtM <- Stn_Wtm %*% MindexT # produces a [73,60] result matrix - 73 stations by 60
dates


# PHDI index
Hindex <- read_excel("data-sub.xlsx",sheet ="PHDI_Div_TimeSeriesR")

# remove all cols but Z-index, convert to matrix, and transpose
HindexT <- t(as.matrix(Hindex[,-c(1:4)])) # produces a [6,60] matrix

# matrix multiplication (inner product)
Stn_WtH <- Stn_Wtm %*% HindexT # produces a [73,60] result matrix - 73 stations by 60
dates


## prepare datasets for export

# convert drought matrices to column vectors
Stn_WtZv <- as.vector(t(Stn_WtZ)) # Z-index
Stn_WtMv <- as.vector(t(Stn_WtM)) # PMDI index
Stn_WtHv <- as.vector(t(Stn_WtH)) # PHDI index
```

```
Stn_Namesv <- as.vector(t(Stn_Wt[,1])) # build vector of station names
Stn_NamesvR <- sort(rep(Stn_Namesv,times=60)) # repeat station names 60 times (for 5
yrs x 12 mos of data) & sort to match data

Month_Year <- c("01-2013","02-2013","03-2013","04-2013","05-2013","06-2013","07-
2013","08-2013","09-2013",
                "10-2013","11-2013","12-2013","01-2014","02-2014","03-2014","04-
2014","05-2014","06-2014","07-2014","08-2014",
                "09-2014","10-2014","11-2014","12-2014","01-2015","02-2015","03-
2015","04-2015","05-2015","06-2015","07-2015",
                "08-2015","09-2015","10-2015","11-2015","12-2015","01-2016","02-
2016","03-2016","04-2016","05-2016","06-2016",
                "07-2016","08-2016","09-2016","10-2016","11-2016","12-2016","01-
2017","02-2017","03-2017","04-2017","05-2017",
                "06-2017","07-2017","08-2017","09-2017","10-2017","11-2017","12-2017")

# replicate but do not sort (in order to match station names & data)
Month_YearR <- rep(Month_Year,times=73) # repeat date 73 times (73 stations)

# combine column vectors into one dataframe
NOAA_Wt_All <- data.frame(Stn_NamesvR,Month_YearR,Stn_WtZv,Stn_WtMv,Stn_WtHv,row.names
= NULL)
names(NOAA_Wt_All)[1] <- "Station_ID"
names(NOAA_Wt_All)[2] <- "Month_Year"
names(NOAA_Wt_All)[3] <- "Zindex_wt"
names(NOAA_Wt_All)[4] <- "PMDI_wt"
names(NOAA_Wt_All)[5] <- "PHDI_wt"

# export data
write.csv(NOAA_Wt_All,file="../NOAA_Wt.csv",row.names=FALSE)
```

montana_dag.R

```
#This script traces the methods used to develop the BN network;
the part where you can query the network is toward the end, obviously;
you don't need to rerun all of this, the first part is just showing
the work and giving examples#

dodata<-readRDS(file = "dodata.rds")
summary(dodata)


#packages to load#
library(bnlearn)
library(boot)

#this reads in the network model arrived at through the work below and;
adds a copy as backup#
bNxx<-readRDS(file="bNxx.rds")
bNx2<-bNxx



#look at the data#
summary(dodata)
dim(dodata)

#example character string matching to identify column in the dataframe#
grep("nat",colnames(dodata))

#scatter plot of two colinear variables#
plot(natws~distws,dodata)

#look at distributions - add transformation if indicated by shape;
in the example below, a log transform might help; note this was alread done#
par(mfrow=c(3,1))
par(mar=c(4,1,1,1))
plot(ecdf(dodata[,59]))
plot(ecdf(log(dodata[,59])))
plot(ecdf(sqrt(dodata[,59]))


dodata$xslopeln<-log(dodata$xslope)
dodata$dslopeln<-log(dodata$devslope)



#create a naive DAG#
dagAll<-hc(dodata[c(63,17:26,36:40,42:46,49,51,54,55,64,65)])
par(mfrow=c(1,1))
par(mar=c(1,1,1,1))
plot(dagAll)
score(dagAll, data =dodata[c(63,17:26,36:40,42:46,49,51,54,55,64,65)],type="bic-cg" )

#handy to find names of columns by position in dataframe#
names(dodata[c(63,17:26,36:40,42:46,49,51,54,55,64,65)])
names(dodata[c(55)])


#test arc strength via bootstrapping#
dN<-boot.strength(dodata[c(63,17:26,36:40,42:46,49,51,54,55,64,65)], algorithm =
"hc",R=1000,m=233)
dN
```

```
#note there are 17 obs in wellocat level 2; caution needed in interpreting that
variable#
table(dodata$wellocat)


dNa<-dN[(dN$strength > 0.85) & (dN$direction > 0.5), ]
dNaa<-averaged.network(dNa)
plot(dNaa)

#variables that don't contribute much info to delta DO; however the DAG shows the
relationships and
depedendencies, so worth a look#
grep("MARANK",colnames(dodata))
grep("FARANK",colnames(dodata))
grep("IMPROB",colnames(dodata))
grep("MATHI",colnames(dodata))
grep("Zin",colnames(dodata))
grep("RELA",colnames(dodata))
grep("Done",colnames(dodata))
grep("Dtwo",colnames(dodata))
grep("DSCt",colnames(dodata))
grep("streamcat",colnames(dodata))
grep("PHDI",colnames(dodata))



dagSub<-hc(dodata[c(63,17,19,23,26,36,37,39,43,46,49,51,54,55,64,65)])
par(mar=c(1,1,1,1))
plot(dagSub)

#the call to boot.strength needs to have the same data as that for dagSub#
dNb<-boot.strength(dodata[c(63,17,19,23,26,36,37,39,43,46,49,51,54,55,64,65)],
algorithm = "hc",R=1000,m=233)

dNbb<-dNb[(dNb$strength > 0.85) & (dNb$direction > 0.5), ]
dNbb<-averaged.network(dNbb)
plot(dNbb)

#thinning the DAG out; identifying column numbers of variables being dropped#
grep("xslopeln",colnames(dodata))
grep("wellcat",colnames(dodata))
grep("distnf",colnames(dodata))
grep("comp",colnames(dodata))
grep("dslope",colnames(dodata))


dagSub2<-hc(dodata[c(63,17,23,26,36,37,39,43,46,49,55)])
par(mar=c(1,1,1,1))
plot(dagSub2)

#have a look at arc strength#
arc.strength(dagSub2, data =
dodata[c(63,17,23,26,36,37,39,43,46,49,55)],criterion="bic-cg")

score(dagSub2, data =dodata[c(63,17,23,26,36,37,39,43,46,49,55)] )

dNc<-boot.strength(dodata[c(63,17,23,26,36,37,39,43,46,49,55)], algorithm =
"hc",R=1000,m=233)

dNcc<-dNc[(dNc$strength > 0.85) & (dNc$direction > 0.5), ]
dNcc<-averaged.network(dNcc)
plot(dNcc)
```

```
#drainage area was dropped during the averaging step#
grep("DA",colnames(dodata))



dagSub3<-hc(dodata[c(63,23,26,36,37,39,43,46,49,55)])
par(mar=c(1,1,1,1))
plot(dagSub3)

dNd<-boot.strength(dodata[c(63,23,26,36,37,39,43,46,49,55)], algorithm =
"hc",R=1000,m=233)
score(dagSub3, data =dodata[c(63,23,26,36,37,39,43,46,49,55)] )



dNd<-dNd[(dNd$strength > 0.85) & (dNd$direction > 0.5), ]
dNdd<-averaged.network(dNd)
plot(dNdd)

grep("PMDI",colnames(dodata))


#substitute drainage area for wellocat#
dagSub4<-hc(dodata[c(63,17,23,26,36,37,39,43,46,49)])
par(mar=c(1,1,1,1))
plot(dagSub4)

dNe<-boot.strength(dodata[c(63,17,23,26,36,37,39,43,46,49)], algorithm =
"hc",R=1000,m=233)

par(mar=c(4,4,1,1))
plot(ecdf(dNe$strength))
abline(v=0.65)

score(dagSub4,dodata[c(63,17,23,26,36,37,39,43,46,49)])


dNx<-dNe[(dNe$strength >= 0.65) & (dNe$direction > 0.5), ]
dNxx<-averaged.network(dNx)
par(mar=c(1,1,1,1))
plot(dNxx)

score(dNxx, data =dodata[c(63,17,23,26,36,37,39,43,46,49)] )
arc.strength(dNxx, data = dodata[c(63,17,23,26,36,37,39,43,46,49)],criterion="bic-cg")

dNxx<-reverse.arc(dNxx,"sqd","MPHYTERANK")
dNxx<-reverse.arc(dNxx,"sqd","PMDI")
score(dNxx, data =dodata[c(63,17,23,26,36,37,39,43,46,49)] )
plot(dNxx)



#fit the network model with the DAG as presently configured#
bNxx<-bn.fit(dNxx,data=dodata[c(63,17,23,26,36,37,39,43,46,49)],method="mle")



docvz<-bn.cv(data=dodata[c(63,17,23,26,36,37,39,43,46,49)],dNxx,method = "hold-out",
k = 4, m = 60, runs = 1000,loss="cor-lw",loss.args = list(target="sqd"))
docvz
```

```
#compare performance of network models with n=234 to network with n=764#
plot(docvz,docv)


#query the network#
#if not already load, load boot#
library(boot)

bNxx


#this function will return means from the distributions generated in the boot sample
and
plot them as a histogram#

R = 1000
boot.x = numeric(R)

boot.x = function(data, i) {
    d = data[i,]
t.test(cpdist(bNxx, nodes = c("sqd"), evidence = distws<=0.163))$estimate

    }

boot.out = boot(data=dodata[c(63,17,23,26,36,37,39,43,46,49)], statistic=boot.x,
R=1000)

hist(boot.out$t^2,main="Mean of Delta D.O. || WS Dist < 0.163",cex.main=0.8)


#compares prior distribution (i.e., the orginal delta DO) to posterior based on the
squareroot transformation#
test2dst<-cpdist(bNxx, nodes = c("sqd"), evidence = distws>0)

par(mar=c(5,4,2,1))
boxplot(dodata$xdelta,test2dst$sqd^2,xlab="",ylab="Delta DO mg/l")
axis(1,c(1:2),c("Prior","Posterior"))

ks.test(dodata$xdelta,test2dst$sqd^2)
#that's not too bad#


#the following was how generated figures 4 through 7#
par(mfrow=c(3,2))
par(mar=c(4,4,2,1))



#run the first line, find the corresponding call to cpquery add that line to the boot
function; plot the first histogram,
then the histogram below the boot function, repeat#
testldst<-cpdist(bNxx, nodes = c("sqd"), evidence = distws<=0.163)
testldst<-cpdist(bNxx, nodes = c("sqd"), evidence = distws>0.163 & Dzero<6)
testldst<-cpdist(bNxx, nodes = c("sqd"), evidence = distws>0.163 & Dzero>=6)



h<-hist(testldst[[1]]^2, breaks=20, col="skyblue", xlab="Delta DO",main="Delta D.O. ||
WS Dist < 0.163",cex.main=1,xlim=c(0,30))
h<-hist(testldst[[1]]^2, breaks=20, col="skyblue", xlab="Delta DO",main="Delta D.O. ||
WS Dist > 0.163 & Dzero<6",cex.main=1,xlim=c(0,30))
h<-hist(testldst[[1]]^2, breaks=20, col="skyblue", xlab="Delta DO",main="Delta D.O. ||
WS Dist > 0.163 & Dzero>=6",cex.main=1,xlim=c(0,30))
```

```
thold<-sqrt(5.3)


doproba<-cpquery(bNxx, sqd>thold, (distws<=0.163))
doprobb<-cpquery(bNxx, sqd>thold, (distws>0.163 & Dzero<6))
doprobc<-cpquery(bNxx, sqd>thold, (distws>0.163 & Dzero>=6))


library(boot)

R = 1000
boot.x = numeric(R)

boot.x = function(data, i) {
    d = data[i,]
doprobc<-cpquery(bNxx, sqd>thold, (distws>0.163 & Dzero>=6))
    }

boot.out = boot(data=dodata[c(63,17,23,26,36,37,39,43,46,49)], statistic=boot.x,
R=1000)

xsd<-function(x) c(mean=mean(x,na.rm=TRUE),sd=sd(x,na.rm=TRUE))

xsd(boot.out$t)

hist(boot.out$t,main="Prob Delta D.O. > 5.3
mg/l",cex.main=1,col="aquamarine",xlab="",xlim=c(0.15,0.85))

t.test(cpdist(bNxx, nodes = c("sqd"), evidence = distws<=0.163))$estimate



#conditioned on drought and macrophyte abundance#

testldst<-cpdist(bNxx, nodes = c("sqd"), evidence = PMDI<(-2))
testldst<-cpdist(bNxx, nodes = c("sqd"), evidence = PMDI<(-2) & MPHYTERANK<3)
testldst<-cpdist(bNxx, nodes = c("sqd"), evidence = PMDI<(-2) & MPHYTERANK>=3)

testldst<-cpdist(bNxx, nodes = c("sqd"), evidence = PMDI>=4)
testldst<-cpdist(bNxx, nodes = c("sqd"), evidence = PMDI>=4 & MPHYTERANK<3)
testldst<-cpdist(bNxx, nodes = c("sqd"), evidence = PMDI>=4 & MPHYTERANK>=3)


h<-hist(testldst[[1]]^2, breaks=20, col="skyblue", xlab="Delta DO",main="Delta D.O. ||
Dry",cex.main=1)
h<-hist(testldst[[1]]^2, breaks=20, col="skyblue", xlab="Delta DO",main="Delta D.O. ||
Dry & Low Macrophyte Cover",cex.main=1)
h<-hist(testldst[[1]]^2, breaks=20, col="skyblue", xlab="Delta DO",main="Delta D.O. ||
Dry & High Macrophyte Cover",cex.main=1)


h<-hist(testldst[[1]]^2, breaks=20, col="skyblue", xlab="Delta DO",main="Delta D.O. ||
Wet",cex.main=1)
h<-hist(testldst[[1]]^2, breaks=20, col="skyblue", xlab="Delta DO",main="Delta D.O. ||
Wet & Low Macrophyte Cover",cex.main=1)
h<-hist(testldst[[1]]^2, breaks=20, col="skyblue", xlab="Delta DO",main="Delta D.O. ||
Wet & High Macrophyte Cover",cex.main=1)
```

```
doproba<-cpquery(bNxx, sqd>thold, (PMDI<(-2)))
doprobb<-cpquery(bNxx, sqd>thold, (PMDI<(-2) & MPHYTERANK<3))
doprobc<-cpquery(bNxx, sqd>thold, (PMDI<(-2) & MPHYTERANK>=3))

doproba<-cpquery(bNxx, sqd>thold, (PMDI>=4))
doprobb<-cpquery(bNxx, sqd>thold, (PMDI>=4 & MPHYTERANK<3))
doprobc<-cpquery(bNxx, sqd>thold, (PMDI>=4 & MPHYTERANK>=3))



boot.x = function(data, i) {
    d = data[i,]
doprobc<-cpquery(bNxx, sqd>thold, (PMDI>=4 & MPHYTERANK>=3))
    }

boot.out = boot(data=dodata[c(63,17,23,26,36,37,39,43,46,49)], statistic=boot.x,
R=1000)
hist(boot.out$t,main="Prob Delta D.O. > 5.3
mg/l",cex.main=1,col="aquamarine",xlab="",xlim=c(0.15,0.85))



#figure 7#
 par(mfrow=c(3,2))
par(mar=c(4,4,2,1))

plot(ecdf(dodata$PMDI),main="PMDI",xlab="")
abline(v=c(-2,4),lty=2)
plot(ecdf(dodata$distws),main="Watershed Disturbance",xlab="")
abline(v=0.163,lty=2)


testldst<-cpdist(bNxx, nodes = c("sqd"), evidence = PMDI>(-2) & PMDI<4 & distws<0.165
& MPHYTERANK<3)
h<-hist(testldst[[1]]^2, breaks=20, col="skyblue", xlab="Delta DO",main="Delta
D.O.",cex.main=1)


#substitute in the cpquery statement after sqd>#
thold
s7<-sqrt(7)
s8<-sqrt(8)
s9<-sqrt(9)


boot.x = function(data, i) {
    d = data[i,]
doprob<-cpquery(bNxx, sqd>s9, (PMDI>(-2) & PMDI<4 & distws<0.165 & MPHYTERANK<3))
    }

boot.out = boot(data=dodata[c(63,17,23,26,36,37,39,43,46,49)], statistic=boot.x,
R=1000)
hist(boot.out$t,main="Prob Delta D.O. > 9
mg/l",cex.main=1,col="aquamarine",xlab="",xlim=c(0.0,0.3))


saveRDS(bNxx, file = "bNxx.rds")
```

montana_fixed_dag.R

```
library(bnlearn)

deltx<-readRDS(file = "deltx.rds")
dofixN<-readRDS(file="dofixN")

par(mar=c(4,4,1,1))
plot(xdelta~PHDI_wt,deltx)

grep("xdelta",colnames(deltx))
deltx$sqd<-sqrt(deltx$xdelta)
grep("sqd",colnames(deltx))


dodag<-hc(deltx[c(24,9,12,14:16,19:23)])
par(mar=c(1,1,1,1))
plot(dodag)

arc.strength(dodag,data=deltx[c(24,9,12,14:16,19:23)],criterion="bic-cg")


dodagst<-boot.strength(deltx[c(24,9,12,14:16,19:23)],algorithm="hc",R=1000,m=500)
dodagst


dobest<-dodagst[(dodagst$strength>0.899) & (dodagst$direction>0.5),]
dobestx<-averaged.network(dobest)
par(mfrow=c(1,1))
par(mar=c(1,1,1,1))
plot(dobestx)



dofixN<-bn.fit(dobestx,data=deltx[c(24,9,12,15,19,22,23,16,20)],method="mle")


#log likelihood loss#
docv<-bn.cv(data=deltx[c(24,9,12,15,19,22,23,16,20)],dobestx,method = "hold-out",k =
5, m = 152, runs = 100)
docv

#correlation loss#
docv<-bn.cv(data=deltx[c(24,9,12,15,19,22,23,16,20)],dobestx,method = "hold-out",
k = 5, m = 152, runs = 100,loss="cor-lw-cg",loss.args = list(target="sqd"))
docv

plot(docv)



#query the network#

cpquery(dofixN, xdelta>5.3, (PMDI_wt>4))



library(boot)

R = 1000
boot.x = numeric(R)

boot.x = function(data, i) {
```

```
    d = data[i,]
cpquery(dofixN, xdelta>5.3, (PMDI_wt<(-2)))
    }

boot.out = boot(data=deltx[c(24,9,12,15,19,22,23,16,20)], statistic=boot.x, R=1000)

xsd<-function(x) c(mean=mean(x,na.rm=TRUE),sd=sd(x,na.rm=TRUE))

xsd(boot.out$t)

hist(boot.out$t,main="Probability of Delta D.O. > 5.3 mg/l || PMDI > 4 & D3 =
0",cex.main=0.8)



saveRDS(dofixN, file = "dofixN.rds")
```

A.3. Datasets Containing Predictor and Response Variables and Model Objects[32]

Dataset Name	Description
Regression Trees	
map_rpart_all.xlsx	Excel dataset with multiple tabs and each tab provides the station-date inventory for leaf nodes of each of the regression tree model runs. Only contains model runs that have DO response (5 total) and includes the "rule" that defines the node. Column headers include: SITEYRMW, StationID, DateX, RecNoFrame, NodeName, NodeNobs, NodeDeviance, NodeFittedValue, Rule. Other value-added material: tabs organized, color added, data filters, and column headers defined.
Data_Working.xlsx	Finalized predictor and response variable dataset (234 rows by 60 columns). Includes data dictionary for all fields and tool to build station-date record ID. Other value-added material: tabs organized, color added, data filters, and column headers defined.
Data_Expanded_Working.xlsx	Finalized fixed-effect predictor and response variable dataset for weekly response (762 rows by 33 columns) and monthly response (318 rows by 29 columns). Other value-added material: tabs organized, color added, data filters, and column headers defined.
do_wk.rds	R-object (imported from Data_Working.xlsx) that contains the data set of weekly average dissolved oxygen, land use/cover, drought indices, stream categories, water chemistry and biological (e.g., macrophyte cover) observations. N = 234
do_exp_wk.rds	R-object (imported from Data_Expanded_Working.xlsx) that contains the data set of fixed variables (i.e., land use/cover, NMDC drought indices, and stream categories among others). N = 762 station-weeks
do_exp_mon.rds	R-object that contains the data set of fixed variables (i.e., land use/cover, NOAA drought indices, and stream categories among others). N = 762 station-months
Predictor or Response Development	
LU_byWshed.xlsx	Excel dataset containing GIS-derived land use/cover area percentages (by watershed or Station ID) for whole watershed (all land use/cover classes, reduced classes, natural vs. disturbed land use/cover), 5000 m radius (all land use/cover classes, reduced classes), 1000 m radius (all land use/cover classes, reduced classes, natural vs. disturbed land use/cover). Also includes: 1) attribute table of land use/cover at various classification scales; 2) definition of "comparison sites"; and 3) comparison of natural vs. disturbed land use/cover. Other value-added material: tabs organized, color added, data filters, and column headers defined.

[32] Shown here as an inventory only; actual digital dataset available from GLEC upload (2/2021) to **ePass Montana File Transfer Service**.

Dataset Name	Description
`Slope_byWshed.xlsx`	Excel dataset containing slope univariate-statistics derived for each watershed from GIS-based digital elevation model. Other value-added material: tabs organized, color added, data filters, and column headers defined.
`Wells_byWshed.xlsx`	Excel dataset with compiled watershed totals of all wells and old wells (i.e., developed prior to year 1990). Worksheet compiles GIS-defined link of well ID to watershed ID and summarizes by watershed. Other value-added material: tabs organized, color added, data filters, and column headers defined.
`Variables_FixedEffect.xlsx`	Compilation of land use/cover, wells, and slope information by station ID, plus factor variables: "comparison site" and stream category. Other value-added material: tabs organized, color added, data filters, and column headers defined.
`DO_delta(combined).xlsx`	Initial dataset of DO surveys measured and compiled by Montana DEQ. Combined DO datasets from two instrument sources: YSI and miniDOT. Includes DO saturation analysis using two independent equations. Includes several pivot tables for producing weekly summaries (e.g., mean delta, max delta, counts above threshold). Other value-added material: tabs organized, color added, data filters, and column headers defined.
`AquaticVisualAssessment.xlsx`	Initial dataset of assessment measured and compiled by Montana DEQ. Value-added material: entries for equivalent numeric values of ranked entries, tabs organized, color added, data filters, and column headers defined.
`Chemistry.xlsx`	Initial dataset of stream water chemistry measured and compiled by Montana DEQ. Value-added material: "staging area" for handling censored data using R-based Regression on Order Statistics (ROS), pivot tables for counts of samples by year by parameter (for selected parameters), tabs organized, color added, data filters, and column headers defined.
`SiteList.xlsx`	Includes drainage area calculation from GIS digital elevation model, "comparison site" designation, and linked stream category. Other value-added material: tabs organized, color added, data filters, and column headers defined.
Bayesian Network Model	
`dodata.rds`	R-object that contains the data set of weekly average dissolved oxygen, land use/cover, drought indices, stream categories, water chemistry and biological (e.g., macrophyte cover) observations. N = 234
`bNxx.rds`	R-object containing the network model of relationships that describe posterior distributions of the weekly average of daily dissolved oxygen ranges. It depends on dodata.rds
`deltx.rds`	R-object that contains the data set of fixed variables (i.e., land use/cover, drought indices, and stream categories). N = 762
`dofixn.rds`	R-object containing the network model for delta DO using variables representing land use/cover, drought indices, and stream categories. It depends on the object deltx.rds.

B.1. Data Dictionary for All Variables Used in Modeling Efforts

Variable Name	Definition	Data Type (in R)
Station and Date Information		
SITEYRMW	StationID + YYYY + M + week (1-4)	character
StationID	Station ID assigned by Montana DEQ to point sampling event (73 total)	character
ORD	vector to order rows by station ID + year + mon + week (SITEYRMW)	integer
xdelta	average of weekly suite of daily averages for DO delta; continuous monitor measurement (with DO & temp)	numeric
sdelta	standard deviation (of sample, N-1) of weekly suite of daily averages for DO delta; continuous monitor measurement (with DO & temp)	numeric
mxdelta	maximum of weekly suite of daily averages for DO delta; continuous monitor measurement (with DO & temp)	numeric
min_avg	average of weekly suite of daily averages for DO minimum (mg/L); continuous monitor measurement (with DO & temp)	numeric
exceedMN	# days where daily delta exceeds specified threshold; continuous monitor measurement (with DO & temp); threshold: > 3.5 mg/L Minnesota; Heiskary & Bouchard (2015) - Table 1 for Central River Nutrient Region	integer
exceedMT	# days where daily delta exceeds specified threshold; continuous monitor measurement (with DO & temp); threshold: > 5.3 mg/L Montana; reference: Suplee & Sada (2016) Table C2-2.	integer
exceedOH	# days where daily delta exceeds specified threshold; continuous monitor measurement (with DO & temp); threshold: > 6.5 mg/L Ohio; reference: Miltner (2010) Table 5	integer
Factor Variables (fixed effect)		
DA	drainage area (sq.mi)	numeric
streamcat	stream category: P: perennial, I: intermittent, E: ephemeral, W: wetland	factor
compsite	reference site type: C: comparison site, R: MT DEQ official reference site, O: ordinary site	factor
Aquatic Plant Predictor-Response Variables (random effect)		
MARANK	% cover of micro-algae (as rank)	factor
MATHICKMM	micro-algae thickness (mm)	numeric
FARANK	% cover of filamentous algae (as rank)	factor
MPHYTERANK	% cover of macrophyte (as rank)	factor
RELABUND	% relative abundance of periphyton (nutrient enricher taxa)	numeric
IMPROB	impairment probability (%); indicates nutrient or sediment problem when > 51%	numeric
Water Chemistry Predictor Variables (random effect)		
medt	median of weekly suite of daily averages for water temperature; continuous monitor measurement (with DO & temp)	numeric
maxt_avg	maximum of weekly suite of daily averages for water temperature; continuous monitor measurement (with DO & temp)	numeric

Variable Name	Definition	Data Type (in R)
BP	barometric pressure (mmHg)	numeric
NH3	total ammonia including NH3 and NH4+ (mg/L)	numeric
Nox	nitrite + nitrate (NO23) (mg/L)	numeric
OP	orthophosphate (mg/L)	numeric
pH	pH (international units)	numeric
SC	specific conductance (µS/cm)	numeric
TN	total nitrogen (mg/L)	numeric
TP	total phosphorus (mg/L)	numeric
TPe	e is for 4 cases estimated by substituting the median TP value (for all stationdates) to deal with missingness	numeric
TNe	e is for 4 cases estimated by substituting the median TN value (for all stationdates) to deal with missingness	numeric
Drought Indices Predictor Variables (random effect)		
Zindex	NOAA Z-Index (drought)	numeric
PMDI	NOAA Palmer Meteorological Drought Index (drought)	numeric
PHDI	NOAA Palmer Hydrological Drought Index (drought)	numeric
DSCI	NDMC Drought Severity and Cover Index -- weighted sum of D0-D4	numeric
DSCt	transformed (square-root) DSCI	numeric
Dzero	# consecutive weeks at drought severity level D0 (source: NDMC)	integer
Done	# consecutive weeks at drought severity level D1 (source: NDMC)	integer
Dtwo	# consecutive weeks at drought severity level D2 (source: NDMC)	integer
Dthree	# consecutive weeks at drought severity level D3 (source: NDMC)	integer
Dfour	# consecutive weeks at drought severity level D4 (source: NDMC)	integer
Land Use/Cover Predictor Variables (fixed effect)		
natws	% natural land cover (watershed scale)	numeric
distws	% disturbed land cover (watershed scale)	numeric
natnf	% natural land cover (near-field scale; <1k m)	numeric
distnf	% disturbed land cover (near-field scale; <1k m)	numeric
Well - Oil and Gas - Predictor Variables (fixed effect)		
wells	total well count within watershed	integer
wellso	old (pre-1990) well count within watershed (previously wellsn)	integer
wellcat	total count all wells (as rank)	factor
wellocat	total count of old wells (as rank) (previously wellnew)	factor
Watershed Slope Predictor Variables (fixed effect)		
maxslope	maximum slope in watershed (% slope = tangent x 100)	numeric

Variable Name	Definition	Data Type (in R)
medslope	median slope in watershed (% slope = tangent x 100)	numeric
xslope	mean slope in watershed (% slope = tangent x 100)	numeric
devslope	standard deviation (sample, N-1 ?) of slope in watershed (% slope = tangent x 100)	numeric