Abstract: Elevated concentrations of dissolved-solids (salinity) including calcium, sodium, sulfate, and chloride, among others, in the Colorado River cause substantial problems for its water users. Previous efforts to reduce dissolved solids in upper Colorado River basin (UCRB) streams often focused on reducing suspended-sediment transport to streams, but few studies have investigated the relationship between suspended sediment and salinity, or evaluated which watershed characteristics might be associated with this relationship. Are there catchment properties that may help in identifying areas where control of suspended sediment will also reduce salinity transport to streams? A random forests classification analysis was performed on topographic, climate, land cover, geology, rock chemistry, soil, and hydrologic information in 163 UCRB catchments. Two random forests models were developed in this study: one for exploring stream and catchment characteristics associated with stream sites where dissolved solids increase with increasing suspended-sediment concentration, and the other for predicting where these sites are located in unmonitored reaches. Results of variable importance from the exploratory random forests models indicate that no simple source, geochemical process, or transport mechanism can easily explain the relationship between dissolved solids and suspended sediment concentrations at UCRB monitoring sites. Among the most important watershed characteristics in both models were measures of soil hydraulic conductivity, soil erodibility, minimum catchment elevation, catchment area, and the silt component of soil in the catchment. Predictions at key locations in the basin were combined with observations from selected monitoring sites, and presented in map-form to give a complete understanding of where catchment sediment control practices would also benefit control of dissolved solids in streams.

Keywords: salinity; Colorado River; suspended sediment; random forests

1. Introduction

The Colorado River and its tributaries supply water to more than 38 million people in the United States and Mexico, provide irrigation to more than 18,200 km² of farmland, and generate about 12 billion kilowatt hours of hydroelectric power annually (Figure 1) [1–3]. The upper Colorado
River basin (UCRB) is the source of much of the more than $8 \times 10^6$ metric tons of dissolved solids (salinity) that flow annually past the Hoover Dam [4], including major cations such as calcium, magnesium, potassium, and sodium, and major anions such as bicarbonate, chloride, and sulfate. High dissolved-solids concentrations in the Colorado River cause substantial economic damage to water users, primarily through corrosion and reduced crop yields, with damages estimated to exceed $300 million dollars annually [2]. The Colorado River Basin Salinity Control Program was created as part of the 1974 Colorado River Basin Salinity Control Act, and charged with investigating and implementing a range of salinity control measures in the basin. Salinity control measures often are based on the assumption that preventing or reducing sediment loading to surface waters in the basin will also reduce dissolved-solids loads. This assumption is most likely based on observations that salinity concentrations remain high during peak flow associated with snowmelt runoff events in some areas [1] and studies that associate sediment and dissolved-solids yield [1,5,6]. Published studies that investigate both suspended sediment and dissolved solids concentrations and loadings are limited, and are often based on data obtained from estuary or coastal systems, where the interest is on sources and the transport of suspended particulate matter, and salinity is used to infer ocean contributions [7–12]. Suspended sediment and dissolved solids data are reported in several studies that estimate the total solids loads and yields of river systems. These studies, however, relate suspended sediment and dissolved-solids concentrations and loads separately to potential contributing factors such as land use, topographic relief, season, and/or river discharge, but they do not relate dissolved solids to suspended-sediment concentrations [13–21].

To test if reducing sediment loading to surface waters may also reduce dissolved-solids loading, Tillman and Anning [22] investigated the statistical relationship between suspended-sediment concentrations and dissolved-solids concentrations at 164 water-quality and streamflow gaging sites in the UCRB (Figure 2). On a site-by-site basis, log-transformed instantaneous specific-conductance (electrical conductivity) measurements representing dissolved-solids concentrations were related to varying combinations of log-transformed mean daily streamflow, suspended-sediment concentrations, and time, using a log-linear regression model. Explanatory variables of sine and cosine of decimal time were included to account for seasonal patterns in dissolved-solids concentrations with either first or second-order harmonics. The long-term trend in dissolved-solids concentration was represented by model parameters of time and time-squared. Results from several statistical tests were used to group the monitoring sites into categories of strong, moderate, weak, and no-evidence of a relation between suspended-sediment and dissolved-solids concentrations, as described in detail in Tillman and Anning [22]. Results indicated that 44 UCRB sites had strong or moderate evidence of a correlation and a positive value for the suspended-sediment term, implying that control measures to reduce suspended sediment would have a beneficial impact on dissolved-solids concentrations at the sites. These 44 monitoring sites were located throughout the basin, and had estimated average dissolved solids loads from 110 kg/day up to >14,000,000 kg/day.

Using a random forests classification analysis, the current study investigates watershed characteristics associated with the long-term relationship between suspended-sediment and dissolved-solids concentrations at water-quality monitoring sites in the UCRB. Random forests is a decision-tree based method that produces multiple decision trees that are then combined to create a single consensus prediction [23]. Decision trees are a type of supervised learning algorithm often used for classification studies. A tree is “learned” by splitting the training dataset into subsets in a recursive manner. Combining large numbers of decision trees greatly increases prediction accuracy, but also increases difficulty in interpreting results. Decision tree methods, including random forests, more closely match human decision-making processes than other regression and classification approaches, and can handle a wide range of qualitative and quantitative data [23]. Random forests is a non-linear, multi-variate classification and regression process that uses a collection of independent decision trees to produce robust (low variance) and low bias predictions [24]. Random forests methods are popular classification tools in ecological studies (for example [25–27]), and are
being used more frequently in hydrologic investigations (for example [28–33]). Random forests methods have advantages over other modeling methods, such as regression, in that they do not require data to be transformed, can use categorical data, can autonomously fit non-linear relations, and can automatically incorporate interactions between explanatory variables [24,34]. During random forests analyses, each decision tree is constructed using a different subsample of the original dataset. This use of a subset of the original data for tree construction allows random forests to test the classification tree on the remaining data, and produce an estimate of the classification error, known as the out-of-bag (OOB) error estimate [24,35]. A random forests approach was selected for this study because it excels at using complex interactions of multiple variables, each with limited but important information, for classifications. Random forests models were used to explore the potential importance of watershed characteristics, and to predict additional areas in the UCRB where suspended-sediment control measures may reduce dissolved solids in streams and rivers. This approach and its results may be useful for water managers in the region, and potentially in other basins, as they seek to reduce impacts of elevated dissolved solids in the Colorado River basin. The selection of catchments for suspended-sediment reduction efforts using this approach may help direct limited financial resources to areas where there is a greater chance of achieving reductions in dissolved solids concentrations.

Figure 1. Study area information for the upper Colorado River basin (UCRB): (a) Location of the UCRB within the southwestern United States; (b) Major tributaries to the Colorado River in the UCRB; (c) Average annual precipitation [36]; (d) Major land-cover classifications [37].
1.1. Study Area

The Colorado River basin comprises catchments in parts of Mexico, Arizona, California, Nevada, Utah, New Mexico, Wyoming, and Colorado. The upper and lower Colorado River basins were divided by the Colorado River Compact of 1922 at the compact point of Lee Ferry, Arizona (Figure 1a,b) [38]. The UCRB is defined for this investigation as the 279,964 km² drainage area upstream of U.S. Geological Survey (USGS) streamflow-gaging station 09380000, Colorado River at Lees Ferry, Arizona (Figure 1a,b). Major upper basin tributaries to the Colorado River include the Yampa, White, San Juan, Gunnison, Green and Dolores Rivers (Figure 1b). Land surface elevation in the UCRB ranges from about 945 m near the Lees Ferry, Arizona streamgage, to more than 4260 m in the Southern Rocky Mountains [39], and average annual precipitation varies with elevation, from less than 250 mm in low elevation areas to more than 1000 mm in the Southern Rocky Mountains (Figure 1c) [36]. Land cover in the UCRB is
predominately classified as shrub/scrub and evergreen forest [37], with few high-population areas (Figure 1d).

1.2. Dissolved Solids and Suspended Sediment in the UCRB

Dissolved solids in UCRB streams and rivers mostly consist of the major cations calcium, magnesium, potassium, and sodium, and the major anions bicarbonate, chloride, and sulfate, as well as neutral silica [39]. There are both natural and anthropogenic sources of dissolved solids in the UCRB. Sedimentary rocks are the largest natural source of dissolved solids to streams in the UCRB [40], including the Upper Cretaceous Mancos Shale, the Paradox Member of the Pennsylvanian Hermosa Formation, and the Eocene Green River Formation [39]. Dissolution of carbonate rocks, including calcite and dolomite, release Ca, Mg, and HCO$_3$; dissolution of gypsum and anhydrite releases Ca and SO$_4$; dissolution of halite releases Na and Cl; and dissolution of silicate minerals releases Na, Ca, Mg, K, and HCO$_3$ [41]. Groundwater that comes into contact with these rocks will dissolve salts from these geologic units, which may then contribute to streamflow either through baseflow or as spring point sources [42]. Additionally, runoff from precipitation or snowmelt may contact these rocks at land surface and contribute dissolved solids loading to streams and rivers. The major anthropogenic activity that increases dissolved solids in UCRB streams is the irrigation of agricultural lands, particularly those derived from the sedimentary rocks described above [40]. Irrigation can contribute additional dissolved solids to groundwater through percolation of oxygenated water from unlined irrigation canals and excess water applied to fields, and subsequent dissolution of mineral salts. Irrigation can also contribute salts to groundwater and surface runoff through the development of efflorescent salt crusts that precipitate onto soil surfaces after the evaporation of excess irrigation water [43]. Runoff contacting these salt crusts may dissolve them or entrain the sediments, with ultimate transport to receiving streams [42]. Dissolution of mineral salts by both surface runoff (including irrigation) and groundwater flow produces dissolved solids that may be transported to streams and rivers. The transport of salinity from sources to UCRB streams and rivers is affected by the amount of precipitation, and by soil type and thickness [40].

Suspended sediment can be generated as precipitation and subsequent runoff entrain soil and erodible geologic material. High energy runoff events may generate and entrain more suspended sediment than low energy events. The amount of sediment that is generated and transported to streams and rivers is affected by the intensity and duration of precipitation and runoff, and by the erodibility of material over which runoff flows. While the source and transport to streams of suspended sediment and dissolved solids maybe be similar in some cases, in others they will differ. Bedrock material may provide dissolved solids to runoff, but not suspended sediment. Alternately, soil material not enriched in easily dissolvable salts may be a source of suspended sediment and not of dissolved solids. As previously mentioned, runoff in contact with efflorescent salt crusts may contribute both dissolved solids and sediments to receiving waters. Dissolved solids loading to streams from groundwater discharge (i.e., baseflow and saline springs) would contribute salts without adding suspended sediment, unless the spring discharge erodes surficial material before entering the stream. Once in the stream, dissolved solids and suspended-sediment concentrations may be affected by similar (e.g., concentration through evaporation) or different (e.g., settling out of suspended sediments in reservoirs) processes.

2. Materials and Methods

2.1. Watershed Characteristics Data

Contributing areas to 163 water-quality and streamflow gaging sites in the UCRB, where the relationship between suspended-sediment and dissolved-solids concentrations was investigated by Tillman and Anning [22], were obtained from the National Hydrography Dataset (NHD Plus V2.1; http://www.horizon-systems.com/NHDPlus/NHDPlusV2_home.php). A single site from
the Tillman and Anning [22] study, site 09216527 “Separation Creek near Riner, WY, USA”, was not used in this investigation because it is located within the Great Divide internally-drained portion of UCRB HUC14, and therefore, does not contribute to the Colorado River or its tributaries. The contributing areas for this study range in size from 3 km$^2$ (site 09306042 Piceance Creek tributary near Rio Blanco, CO, USA) to almost 280,000 km$^2$ (site 09380000 Colorado River at Lees Ferry, AZ, USA), and are distributed throughout the UCRB (Figure 2, Table S1). Watershed characteristics covering a wide range of topographic, climate, geologic, land use, soil, hydrologic, and water-quality information were investigated as potential variables that could explain the relationship between the suspended-sediment and dissolved-solids concentrations in UCRB streams reported in Tillman and Anning [22]. Watershed characteristics that were investigated are summarized here briefly, with detailed descriptions and information on the source and processing of the datasets provided in the Supplementary Materials (File S1), and catchment results for each characteristic provided in Table S1. Catchment topographic information considered includes minimum elevation, maximum elevation, median elevation, elevation range, mean percent slope, and catchment area. Land cover and land use information included the fraction of catchments irrigated by flood or sprinkler methods, catchment fraction that is rangeland, and catchment fraction of each of the 16 land cover designations in the National Land Cover Dataset (NLCD). Climate and related characteristics included actual evapotranspiration, climatic water deficit, excess water, snowmelt, snowpack, potential evapotranspiration, precipitation, sublimation, and snowfall. Geologic information considered included several source variables used in previous SPARROW models [40], such as the fraction of catchment area classified as crystalline and volcanic rocks, high-yield (of dissolved solids) sedimentary Cenozoic rocks, low-yield sedimentary Cenozoic rocks, high-yield sedimentary Mesozoic rocks, low-yield sedimentary Mesozoic rocks, high-yield sedimentary Paleozoic and Precambrian rocks, and low-yield sedimentary Paleozoic and Precambrian rocks. Rock chemistry information included area-weighted means of calcium oxide, iron oxide, potassium oxide, magnesium oxide, phosphorus, sulfur, silicon dioxide, hydraulic conductivity, and uniaxial compressive strength. The fraction of catchment areas underlain by subsurface evaporite deposits (gypsum/anhydrite or halite) also was investigated. Several soil parameters from the National Resources Conservation Service (NRCS) State Soil Geographic (STATSGO) database were investigated, including fractions of the area for each hydrologic soil group, and the weighted area means for erodibility factor, horizon thickness, total clay, total silt, total sand, total organic matter, saturated hydraulic conductivity, and available water capacity. All soil parameters except hydrologic soil group were evaluated for the upper soil horizon only, and as a weighted average for all soil horizons. Hydrology and water-quality information considered included mean annual groundwater recharge, mean annual runoff, base-flow index, fraction of catchment area underlain by saline groundwater less than 500 feet (152 m) below land surface, catchment mean rainfall-runoff erosivity factor, 90th percentile and median specific conductance values at gage sites, and mean daily flow at gage sites.

2.2. Model Development

Random forests models were used to (1) explore the potential importance of watershed characteristics in the classification of UCRB stream monitoring sites, and (2) to predict additional areas in the UCRB where suspended-sediment control measures may reduce dissolved solids in streams and rivers. Random forests analyses were used to model the classification of the relationship between dissolved solids and suspended sediment at 163 water-quality and streamflow gaging sites in the UCRB described in Tillman and Anning [22]. UCRB site classifications are referred to as SM+ (strong or moderate evidence of a relation and positive coefficient of the suspended sediment term in Tillman and Anning [22]), and N (weak or no evidence of a relation or negative coefficient of suspended-sediment term in Tillman and Anning [22]) in this article. Decision trees, on which random forests are based, predict the class of a variable (in our case, the UCRB site classification of SM+ or N) by first training on a source dataset (the 84 watershed characteristics summarized in Table S1).
The randomForest package [44] of the R statistical program [45] was used in this study for classification model development and error evaluation. The randomForest package has several arguments that may be adjusted to “tune” a random forests model to improve classification results [44]. For this study, the number of variables selected at each split (mtry), the minimum size of terminal nodes (nodesize), and two arguments that affect weighting of the different classes during tree construction (cutoff and classwt) were adjusted to improve classification performance. Weighting optimization was required because of the imbalance between the number of catchments in the N (119 sites) and SM+ (44 sites) classes. Because the results from this investigation may be used to deploy suspended-sediment control measures in areas identified as being likely to reduce dissolved-solids concentrations in streams, the misclassification of class N catchments as class SM+ catchments (false positive with respect to SM+) was minimized, while also maintaining a low out-of-bag (OOB) error rate for both classes (SM+ and N). The importance of individual variables in random forests classifications is evaluated in this study as the mean decrease in model accuracy that results from randomly permuting values of the variable [44]. That is, the most important variables contribute the most to the accuracy of model classifications. Because random forests analyses are based on thousands of classification tree results from different combinations of explanatory variables, the interpretation of how individual parameters and parameter levels (i.e., higher or lower) are related to watershed classification is difficult. For this study, the distribution of watershed characteristics between SM+ and N classified catchments are compared using Wilcoxon rank-sum analyses.

Two versions of random forests models were developed for this study. The first goal was to investigate watershed characteristics that might be important to the classification of UCRB stream monitoring sites as either having the potential for suspended-sediment control measures to reduce dissolved-solids concentrations (class SM+), or not (class N). To reduce the N-class classification error, while also maintaining a low OOB error rate, exploratory random forests models were optimized for a range of mtry, nodesize, classwt, and cutoff argument values (Table 1; see randomForest package documentation [44] for a complete description). The randomForest package was repeatedly run using every combination of these four tuning parameters, and results plotted to evaluate the tradeoffs between misclassification errors for each of the models. All potential variables described in the “Watershed Characteristics Data” section were evaluated for the exploratory model. Initial random forests modeling was performed for 2000 decision trees (ntree = 2000) using default randomForest package argument values, including an mtry value of the square root of the number of variables (mtry = 9 for this study), a single terminal node for each tree (nodesize = 1), and equal weights for both SM+ and N classes (classwt = NULL and cutoff = (0.5, 0.5)). Results indicated that classification error rates stabilized around 500–1000 trees (Figure S1 in File S1), so subsequent models were developed for 1000 trees (ntree = 1000).

Table 1. Optimized arguments and values in randomForest package during exploratory and predictive random forests model development.

<table>
<thead>
<tr>
<th>Model Argument</th>
<th>Argument Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>mtry</td>
<td>integers from 1 to 18</td>
</tr>
<tr>
<td>nodesize</td>
<td>integers from 1 to 20</td>
</tr>
<tr>
<td>classwt</td>
<td>integers from 1 to 50 for N class</td>
</tr>
<tr>
<td></td>
<td>integers from 100 to 50 for SM+ class</td>
</tr>
<tr>
<td>cutoff</td>
<td>0.99 to 0.5 for N class</td>
</tr>
<tr>
<td></td>
<td>0.01 to 0.5 for SM+ class</td>
</tr>
</tbody>
</table>

The second goal of this investigation was to predict areas within the UCRB where suspended-sediment control measures may have a beneficial impact on dissolved-solids concentrations in streams. The importance of in-stream characteristics in the optimized exploratory random forests model (discussed below), and the lack of basin-wide data for these characteristics, precludes use of
the exploratory model as a predictive model for this purpose. A second random forests analysis was conducted using none of the three in-stream characteristics to develop a predictive model for the UCRB. Optimization of the model arguments was performed for the predictive model over the same range of argument values as for the exploratory model. Because the predictive model is intended to assist in defining areas where suspended-sediment management practices may be employed, an even lower N-class error rate was chosen for the predictive model compared with the exploratory model, in order to minimize falsely classifying N-class catchments as SM+ catchments.

2.3. Use of Random Forests Class-Prediction Model

To give a complete understanding of where sediment control practices would also benefit control of dissolved solids in the UCRB, predictions at key locations were combined with observations from selected monitoring sites and presented in a map. While the 163 monitoring sites investigated in Tillman and Anning [22] provided good information regarding the locations at which sediment control practices would also benefit the control of dissolved solids, there are several spatial gaps in important areas where limited monitoring data prohibited the multiple-linear regression (MLR) modeling and evaluation process described in Tillman and Anning [22]. To fill in such gaps and illustrate an application of the predictive model, class predictions were made at several key locations throughout the UCRB, defined by the eight-digit hydrologic unit code (HUC8). The UCRB is divided into 58 HUC8 hydrologic unit subregions that define a reach of the Colorado River and its tributaries in that reach [46]. The locations for prediction were selected at the sub-basin outlet (pour point) of HUC8s, for cases where there were no monitoring data nearby on the main river draining the HUC8. In addition, the following constraint was emplaced: that the drainage area upstream of the pour point be three or fewer HUC8 areas; otherwise, the location would be excluded from analysis. This constraint purposefully omits reaches with large contributing areas where it would be less clear which part of the basin produces a dissolved-solids benefit from sediment control measures. Analysis of the proximity of water-quality monitoring sites from Tillman and Anning [22] to HUC8 pour points resulted in the selection of 23 stream locations and their associated drainage areas for class prediction. For each of the sites requiring class prediction, the prediction model generated 1000 total votes for SM+ or N classification. The probability of the stream location being SM+ was determined as the count of SM+ votes divided by 1000; locations with more than a 50 percent probability (>500 votes) were classified as SM+ locations.

3. Results and Discussion

3.1. Exploratory Random Forests Model

Exploratory random forests models were developed to investigate watershed characteristics that might be important to the classification of UCRB stream monitoring sites as either having the potential for suspended-sediment control measures to reduce dissolved-solids concentrations (class SM+), or not (class N). For the exploratory model, an OOB estimate of error [24,35] of 20.3% with an N-class error rate of 19.3% and 77.3% accuracy in classifying SM+ sites was selected from the optimized results as a reasonable balance between misclassification errors for this exploratory investigation (Figure 3a, Table 2). Optimization results (Table S2) indicate eight combinations of optimized parameters produce exploratory random forests models with the selected OOB and N-class error combination.

<table>
<thead>
<tr>
<th>Classification from Tillman and Anning [22]</th>
<th>Random Forests Predicted Class</th>
<th>Class Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>N class</td>
<td>96  23</td>
<td>19.3%</td>
</tr>
<tr>
<td>SM+ class</td>
<td>10  34</td>
<td>22.7%</td>
</tr>
</tbody>
</table>
Figure 3. Results of ~900,000 optimization runs of (a) exploratory and (b) predictive random forests models for select argument values to minimize N class error rate while maintaining a low overall OOB error rate. Note most results overlap visible points in charts.

Overall, variable importance scores for the exploratory models were low, with no indication of a few variables that were clearly much more important in site classification than the bulk of other variables (Table S3). The eight exploratory random forests models that produced the selected N and OOB error rate described above have a similar order of variable importance (Table S3), at least for the most important variables (Figure 4). Among the most important watershed characteristics in the optimized exploratory models were: measures of soil hydraulic conductivity, soil erodibility, minimum catchment elevation, catchment area, and the silt component of soil in the catchment (Figure 4). Comparisons of the distribution of these characteristics between SM+ and N catchments by Wilcoxon rank-sum analyses indicate all but soil erodibility differ significantly ($p$-value < 0.05) between the two classes (Figure 5, Table S4). Higher soil hydraulic conductivity, larger catchment areas, lower minimum catchment elevations, and smaller silt components are indicative of SM+ catchments (Figure 5). Three of the five most important variables (mean daily streamflow, 90th percentile of specific conductance values, and median specific conductance value) are more accurately described as in-stream characteristics instead of watershed characteristics (Figure 4). These parameters are measured at water-quality and streamflow gaging stations and, although certainly a function of attributes of the contributing area to the site, are not values distributed throughout the catchment. Of these most important in-stream characteristics, only mean daily flow has a significantly different distribution for the two classes (Wilcoxon rank-sum test, $p$-value < 0.05), with higher flows for SM+ sites (Figure 5).
Figure 4. Average rank of decrease in exploratory model accuracy for the eight most important variables for eight random forests models that produced the optimized combined out-of-bag and N class error. Whiskers represent minimum and maximum ranks among the eight models.

Figure 5. Distribution of standardized watershed characteristic values for eight most important variables in optimized exploratory random forests models. * Denotes $p$-value < 0.05 for Wilcoxon rank-sum test on SM+ and N classes for unscaled watershed characteristic values. See Table S4 for all standardized values and Wilcoxon rank-sum results.
Results of variable importance from the exploratory random forests models indicate that no simple source, geochemical process, or transport mechanism can easily explain the relation between dissolved solids and suspended-sediment concentrations at UCRB monitoring sites [40]. That is, important variables identified in the random forests classification process do not point to one or more of the simple conceptual models discussed in the introduction [39]. For example, although median soil hydraulic conductivity was higher and median silt component was lower for SM+ sites, individually, these variables might be expected to decrease suspended sediment contributions, while perhaps not affecting dissolved solids. However, both variables were among the most important in classifying the SM+ sites in the explanatory models. Other relatively important variables, like catchment elevation and catchment area, are probably general distinguishers of where SM+ sites occur, and not necessarily explanatory of source or transport mechanisms, and thus, not necessarily useful for a mechanistic understanding of SM+ site locations. Additionally, variables like soil hydraulic conductivity and silt component are probably simple surrogates for more complex soil conditions that lead to the classification of SM+ sites. Using the complex interaction of multiple variables, each with limited but important information for classifying sites, is a strength of the random forests modeling approach, even if the interpretation of results can be challenging. If a few, easy to conceptualize variables with high information value were determined, then another modeling approach, such as additive logistic regression, could be used [23].

3.2. Predictive Random Forests Model

A predictive random forests model was developed to help define areas within the UCRB where suspended-sediment control measures may have a beneficial impact on dissolved-solids concentrations in streams. A separate model was developed for prediction purposes because of the importance of in-stream characteristics in the exploratory model, and the lack of coverage for these in-stream characteristics throughout the basin. An OOB estimate of error [24,35] of 18.4% with an N-class error rate of 10.1% was selected from the optimized predictive random forests results as an effective balance between misclassification errors for the predictive model (Figure 3b, Table 3). Although the selected optimized parameters classify SM+ sites accurately only about 60% of the time, they misclassify N sites at SM+ sites only about 10% of the time, which is of greater importance in the predictive model. A single predictive model with the selected OOB and N error rates was produced from model argument optimization (Table S5).

Table 3. Confusion matrix for the predictive random forests models.

<table>
<thead>
<tr>
<th>Classification from Tillman and Anning [22]</th>
<th>Random Forests Predicted Class</th>
<th>Class Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>N class</td>
<td>107</td>
<td>12</td>
</tr>
<tr>
<td>SM+ class</td>
<td>18</td>
<td>26</td>
</tr>
</tbody>
</table>

The most important variables in the predictive model (Table S6) were similar to those in the exploratory models, with four of the five top eight non-in-stream characteristics in the exploratory models (Figure 4) present in the top eight for the predictive model (Figure 6). The seventh exploratory model variable, “minimum catchment elevation”, is partially represented in the “range in catchment elevation” variable of the predictive model, which is also one of the top eight important variables for the predictive model. Additional predictive model variables among the top eight most important variables are: the fraction of catchment classified as evergreen forest, the area weighted mean of calcium oxide rock, and the fraction of the catchment classified as low-yield sedimentary Cenozoic rocks. Of the eight most important variables in the predictive model, four had significantly different distributions between SM+ and N catchments (Wilcoxon rank-sum analyses, p-value < 0.05), with higher soil hydraulic conductivity, smaller silt components, larger range in catchment elevations, and larger catchment areas indicative of SM+ catchments (Figure S2 in File S1, Table S4).
3.3. Model Application

HUC8 pour-point classification predictions resulted in the identification of three SM+ locations where suspended-sediment control measures upstream of the site may have a beneficial impact on dissolved-solids concentrations at that location (Table S7, Figure 7). These three are at the pour points for HUCs 14010004, 14040105, and 14040109, in the northern and eastern parts of the UCRB (sites P3, P9, and P10 in Figure 7 and Table S7; [39,40,42]). For the remaining 20 stream locations, predictions indicate that upstream suspended-sediment control measures would not be likely to have a beneficial impact on dissolved solids at that location. Two locations classified as N, however, warrant noting, as they had nearly a tie between SM+ and N votes. These sites were at the pour points for 14010003 (site P2; [42]) and 14050001 (site P11), and each had more than 490 class SM+ votes.

Monitoring data and random forests model predictions together provide an indication of where reductions in suspended sediment may also help reduce dissolved-solids concentrations in the UCRB (Figure 7). Monitoring data suggest sediment-control measures would also benefit dissolved-solids concentrations along much of the reach of the Colorado River between monitoring sites 10 (near Dotsero, CO, USA) and 37 (near Cisco, UT, USA) [39,42]. With the exceptions of the lower Roaring Fork River (P3) and the lower Gunnison River (monitoring site 29), reductions of suspended sediment may not benefit dissolved-solids concentration in tributaries to the Colorado in this reach, nor in the Colorado River above this reach (site 1) [42]. These tributaries include the Blue (P1), the Eagle (P2), and the Dolores Rivers (P6, P7, monitoring site 36). On the main stem of the Green River, some reaches may benefit from suspended-sediment reductions, while others likely would not, as indicated by the three SM+ monitoring sites (124, 56, 39) and four N monitoring sites (117, 64, 45, 43; Figure 7) [39,40]. Sediment-control measures generally would not benefit dissolved-solids concentrations in tributaries on the western side of the Green River, with the exception of the Duchesne River (monitoring site 89, Figure 7) and parts of the San Rafael River basin (sites 126 and 128; Figure 2).
More potential for dissolved-solids reductions from sediment-control measures occurs in tributaries draining eastern portions of the upper Green River, including Bitter Creek (P9), Vermillion Creek (P10), the Yampa River between Hayden and Maybell, CO, USA (monitoring sites 73 and 81) and most of the White River (Figure 7) [39,40]. Along the main stem of the San Juan River between monitoring sites 140 (at Hammond Bridge near Bloomfield, NM, USA) and 161 (near Mexican Hat, UT, USA), sediment-control measures may help reduce dissolved solids (Figure 7) [39]. Results indicate that reductions of suspended sediment may not benefit dissolved-solids concentrations in the San Juan River above this reach, and in most tributaries to this reach. A notable exception, however, is the Chaco River in NM, where several monitoring sites (Figure 2) within that basin indicate a benefit in dissolved-solids concentrations from reductions of suspended sediment.

Figure 7. Map of sites with class based on monitoring data (triangles) or random forests model predictions (circles). Only monitoring sites along main stems of major rivers or at HUC8 pour points are presented here for visual clarity of map. See Figure 2, Tables S1 and S7 for all monitoring and prediction sites and results.
4. Summary and Conclusions

A random forests classification analysis was performed on topographic, climate, land cover, geology, rock chemistry, soil, and hydrologic information in 163 UCRB catchments to investigate watershed characteristics that may influence the relationship between suspended-sediment and dissolved-solids concentrations in streams in the region. Random forests models were developed for both exploratory and predictive uses. Model arguments in the randomForest package of the R statistical program were optimized to minimize the misclassification of class N sites as class SM+ sites (false positive with respect to SM+), while also maintaining a low out-of-bag (OOB) error rate for both classes (SM+ and N). The exploratory model was able to correctly predict SM+ sites with 77.3% accuracy. Results of variable importance from the exploratory random forests models indicate that no simple source, geochemical process, or transport mechanism can easily explain the relation between dissolved solids and suspended sediment concentrations at UCRB monitoring sites. Additional watershed characteristics that more precisely describe these processes and mechanisms may be developed for further testing in an exploratory classification model. Also, future exploratory models may be developed using watersheds of similar size to the HUC8 areas used for prioritizing sediment-control efforts. A second random forests model was developed using catchment-wide data for predictive purposes. Calibration parameters for the prediction model were adjusted to maximize the prediction accuracy of sites where dissolved solids were not increased by suspended sediment, so as to reduce the risk of spending unnecessary management resources in those areas. The prediction random forests model was able to correctly classify N sites with 89.9% accuracy, while also correctly classifying SM+ sites with 59.1% accuracy, and resulted in similar important variables as the exploratory model. The predictive model was used to identify UCRB areas that may benefit from sediment control measures, particularly where there were insufficient monitoring data in Tillman and Anning [22] for classification. Predictive model results identified three locations at HUC8 catchment pour points where upstream suspended-sediment control measures may have a beneficial impact on dissolved-solids concentrations in streams, plus two additional locations that were very close to being classified as SM+ sites. These areas, identified through random forests classification analyses, in addition to the catchments identified by multiple linear regression on streamflow data in Tillman and Anning [22], provide water managers in the area with potentially valuable information on where to locate future suspended-sediment control measures in order to reduce dissolved-solids concentrations in the Colorado River.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4441/10/6/676/s1.

File S1: A description of watershed characteristics that were investigated for potential contribution to the relation between suspended-sediment and dissolved-solids concentrations in streams of the upper Colorado River basin.

Figure S1: Out-of-box and SM+ class error rates as a function of the number of trees in initial random forests classification simulations using default argument values. Figure S2: Boxplots showing median and interquartile range (box) and maximum and minimum values not including outliers (whiskers) for standardized watershed characteristic data. Table S1: Watershed characteristics information summarized by catchment area for 163 upper Colorado River basin water-quality monitoring sites. Table S2: Random forests simulation results for optimization of mtry, nodesize, classwt, and cutoff model arguments for exploratory random forests model. Table S3: Mean decrease in accuracy for 84 watershed characteristics for eight exploratory random forests models that produced the optimized combined out-of-bag and N class error. Table S4: Summary watershed characteristics information standardized for catchment areas for 163 upper Colorado River basin water-quality monitoring sites and results from Wilcoxon rank-sum test on the distribution of the original data by class. Table S5: Random forests simulation results for optimization of mtry, nodesize, classwt, and cutoff model arguments for predictive random forests model. Table S6: Mean decrease in accuracy for 81 watershed characteristics for the predictive random forests model. Table S7: Watershed characteristics information summarized for select upper Colorado River basin HUC8 areas (or accumulated HUC8 area as noted) used for prediction of catchment CLASS.

Author Contributions: F.D.T. and D.W.A. conceived and designed the experiments; F.D.T. and D.W.A. performed the experiments; F.D.T. and D.W.A. analyzed the data; J.A.H., S.G.B. and M.P.M. contributed reagents/materials/analysis tools; F.D.T., D.W.A., J.A.H., S.G.B. and M.P.M. wrote the paper.

Acknowledgments: This study was supported by the Colorado River Basin Salinity Control Forum.

Conflicts of Interest: The authors declare no conflict of interest.
References


16. Lewis, W.M.; Saunders, J.F. Concentration and transport of dissolved and suspended substances in the Orinoco River. Biogeochemistry 1989, 7, 203–240. [CrossRef]


21. Subramanian, V. Chemical and suspended-sediment characteristics of rivers of India. *J. Hydrol.* 1979, 44, 37–55. [CrossRef]


29. Reynolds, L.V.; Shafroth, P.B.; Poff, N.L. Modeled intermittency risk for small streams in a North American river basin under climate change. *J. Hydrol.* 2015, 523, 768–780. [CrossRef]


33. Feng, Q.; Liu, J.; Gong, J. Urban flood mapping based on unmanned aerial vehicle remote sensing and random forest classifier—A Case of Yuyao, China. *Water* 2015, 7, 1437–1455. [CrossRef]


© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).