# Targeted Training Data to Improve Site Potential Ecological Models for the Upper Colorado River Basin

*The Upper Colorado River Basin*

**Stephen P. Boyte[1], Trenton D. Benedict[2], Dinesh Shrestha[2], Jesslyn F. Brown[1]**

[1] U.S. Geological Survey (USGS) Earth Resources Observation & Science (EROS) Center, Sioux Falls, SD 57198 USA.
[2] KBR, contractor to the USGS EROS Center, Sioux Falls, SD 57198 USA. Work performed under USGS contract 140G0121D0001.
*** The shrub and grass site potential datasets can be downloaded at https://doi.org/10.5066/P9IKI4XV**

## Main Statement

We improve site potential models and maps by judiciously selecting training data locations filtered by land cover and disturbance datasets

## Objectives and Purpose

- Test the efficacy of using land cover and disturbance datasets to help select training data locations for site potential mapping models.
- Develop more accurate grass and shrub site potential mapping models that estimate expected long-term vegetation cover in an undisturbed state.
- Improve ecological models that estimate expected vegetation performance and identify locations that perform worse than or better than expected based on site potential and weather.

## Definitions and Data

- Site potential represents the long-term vegetation cover that averages out weather variation while accounting for spatial variation in long-term vegetation cover. Site potential is associated with site conditions represented by the independent variables listed below.
- Land Change Monitoring Assessment & Projection (LCMAP) is an annual time series (1985-2017) of 10 land cover or related datasets developed by the USGS. We used two of the LCMAP datasets – spectral change magnitude and spectral model quality. For more details please see: (https://www.usgs.gov/media/files/lcmap-science-product-guide). LCMAP data can be viewed and downloaded at https://eros.usgs.gov/lcmap/viewer/index.html.
- National Land Cover Database (NLCD) datasets used from 2001, 2006, 2011, 2016 (https://www.mrlc.gov/).
- Monitoring Trends in Burn Severity (MTBS) is a time series of data that has mapped the severity and perimeters of fires in the U.S. since 1984 (https://www.mtbs.gov/).
- Independent variables: Elevation, soils – available water capacity, organic matter, sand, clay, silt – 30-yr climate, slope and aspect, land resource unit, solar radiation.
- Dependent variable: The 75th percentile of 250-m enhanced Moderate Resolution Imaging Spectroradiometer (eMODIS) Normalized Difference Vegetation Index (NDVI) from 2000 – 2018. The NDVI data was scaled and converted to an integer using this equation: (NDVI*100) +100.

## Methods

- The Upper Colorado River Basin (UCRB) is primarily an arid and semi-arid environment with elevations ranging from 990 m to 4123 m. The 30-year precipitation average equals 398 mm with an average temperature of about 7° C. The dominant vegetation type is classified as shrub and grass.
- We integrated biophysical and satellite data into regression-tree software to develop machine-learning algorithms that we used in mapping software along with raster data to create spatially explicit datasets (Figs. 1 & 3).
- All native resolution 30-m data were spatially averaged and resampled to 250 m using nearest neighbor. Climate data were resampled from 1000 m to 250 m using bilinear interpolation.
- Training and test data pixels met four criteria: 1) the 2001, 2006, 2011, and 2016 NLCD iterations all classified a pixel as shrub (for shrub model) or grassland/herbaceous (for grass model), 2) MTBS recorded no fires during the period of 1984 – 2017, 3) the change magnitude dataset indicated there were no unpredicted deviations from 1985 – 2017, and 4) model quality data were developed by a full model, indicating high model quality.
- To mitigate spatial autocorrelation issues, we selected training and test pixels that were separate from other training and test pixels by >250 m to 500 m (grass model) and 500 m to 1000 m (shrub model). **For maximum spatial autocorrelation mitigation, the grass model selected for map development separated training and test pixels by 500 m, and the shrub model selected for map development separated training and test pixels by 1000 m (Tables 1 & 2).** Far fewer grass pixels existed in the study area, hence the closer proximity of training and test pixels and the smaller sample sizes.
- We measured model accuracies and map validations using the coefficient of determination ($R^2$) and mean absolute error (MAE). Map test data were withheld independent of all modeling and mapping procedures and then compared to the mapping result for a true independent test.

## Results

- All four shrub models that we developed show strong model accuracy. The associated maps show strong validation metrics. For all models, the $R^2$ values range from 0.81 to 0.90 and MAE values range from 3.92 to 5.54. The selected shrub model has similar training and test accuracy metrics (Train $R^2$ = 0.90; MAE = 4.10/Test $R^2$ = 0.88; MAE = 4.42), suggesting that it is neither overfit nor underfit. Figure 1 shows the selected model's associated map. The independent test (Fig. 2) for the map had the highest $R^2$ (0.89) and the second-best MAE (4.34) of the shrub maps. The shrub model and map selected had a slightly higher MAE values than the model and map from another model, but because the selected model was robust with more training pixels that were spread further apart (1000 m) to mitigate spatial autocorrelation, we consider this model and map a stronger representation of actual shrub site potential. The 1:1 line and the regression line in Figure 2 align closely, indicating little model bias. The majority of the UCRB area is classified as shrub (Fig. 1), and most of the shrub area has relatively low site potential. The shrub areas with higher shrub site potential likely occur in more mesic areas at higher elevations and may represent more complex and diverse ecosystems.
- All three grass models that we developed show strong model accuracy. The associated maps show strong validation metrics. The $R^2$ values from all models range from 0.87 to 0.94 and MAE values range from 3.02 to 5.94. The selected grass model (Table 2) (map shown in Fig. 3) has similar training and test accuracy metrics (Train $R^2$ = 0.94; MAE = 3.72/Test $R^2$ = 0.94; MAE = 4.11), suggesting that the grass model is neither overfit nor underfit. The selected model's associated map has the highest $R^2$ (0.94) and the best MAE (4.13) of the grass maps. The selected grass model has slightly lower accuracy metrics than one of the unselected model, but because the selected model's training data mitigates spatial autocorrelation better and the selected MAE is substantially better, we consider this map a stronger representation of grass site potential. The 1:1 line and the regression line align closely, indicating little model bias. Figure 3 shows that the grass/herbaceous classification, as defined by NLCD, is relatively uncommon in this region. This is likely because of the high percent of grass cover required to be considered the grass/herbaceous class. The grass site potential levels are well distributed across the NDVI spectrum.

## Conclusion

- The accuracy for both selected models was high, and the validation of both selected maps was strong. This demonstrates the added value of filtering training and test data in the UCRB so that only recently undisturbed pixels are used when developing site potential models. The inclusion of the two LCMAP land surface change datasets in the filter rendered the most accurate models and maps.
- LCMAP products should be considered for use in the development of other ecological modeling projects, including as used in this study and as independent variables.

*Photo courtesy: Matthew Miller, USGS*

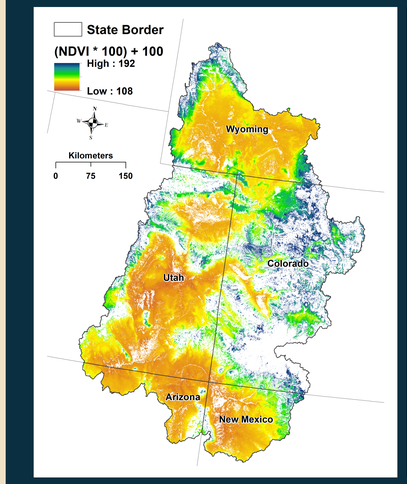


**Figure 1.** The shrub site potential map from the selected model in the Upper Colorado River Basin. Masked pixels (white) indicate where 2016 NLCD classified the pixel as something other than shrub.

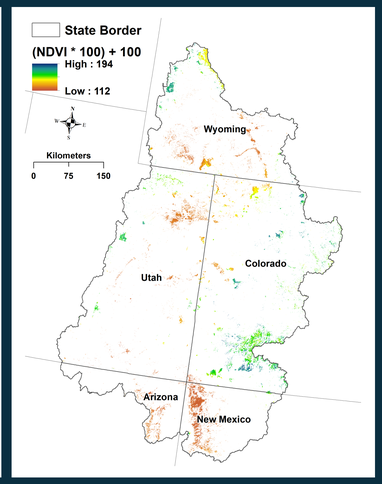


**Figure 3.** The grass site potential map from the selected model in the Upper Colorado River Basin. Masked pixels (white) indicate where 2016 NLCD classified the pixel as something other than grassland/herbaceous.

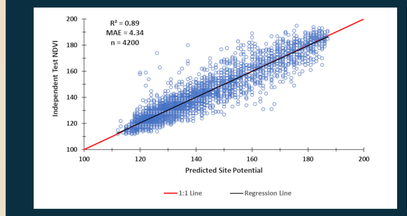


**Figure 2.** The scatterplot from the shrub map's independent test where 4200 data points were withheld from model development and testing. The $R^2$ = 0.89 and the MAE = 4.34. The 1:1 line and regression line are closely aligned, indicating little model bias.

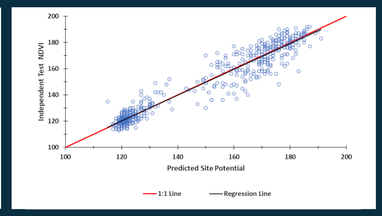


**Figure 4.** The scatterplot from the grass map's independent test where 705 data points were withheld from model development and testing. The $R^2$ = 0.94 and the MAE = 4.13. The 1:1 line and regression line are closely aligned, indicating little model bias.

**Table 1.** The selected scenario developed by different combinations of datasets to locate training and test data emphasizing model accuracy for the shrub model and validation of the shrub map.

| **Shrub Model Accuracy** | | | |
|---|---|---|---|
| Scenario (training data combinations) | $R^2$ Train/Test | Mean Absolute Error (MAE) Train/Test | Sample size Train/Test |
| NLCD, MTBS, LCMAP land surface change magnitude, LCMAP land surface change model quality | 0.90/0.88 | 4.10/4.42 | 66,600/7400 |
| **Shrub Map Validation** | $R^2$ | MAE | Sample size |
| NLCD, MTBS, LCMAP land surface change magnitude, LCMAP land surface change model quality | 0.89 | 4.34 | 4200 |

**Table 2.** The selected scenario developed by different combinations of datasets to locate training and test data emphasizing model accuracy for the grass model and validation of the grass map.

| **Grass Model Accuracy** | | | |
|---|---|---|---|
| Scenario (training data combinations) | $R^2$ Train/Test | MAE Train/Test Train/Test | Sample size |
| 3) NLCD, MTBS, LCMAP land surface change magnitude, LCMAP land surface change model quality | 0.94 / 0.94 | 3.72 / 4.11 | 6220 / 327 |
| **Grass Map Validation** | $R^2$ | MAE | Sample size |
| NLCD, MTBS, LCMAP land surface change magnitude, LCMAP land surface change model quality | 0.94 | 4.13 | 705 |

U.S. Department of the Interior
U.S. Geological Survey

February 2021