Prediction of Drought Index Based on Random Forest

GGS 656 Hydrosphere

Final Project

Xiaoxuan Li

March 1st ,2021

Motivation

Drought is considered as one of the major natural hazards that affect the environment and economy at local and global scales. Considering local weather data is not sufficient to monitor and evaluate drought conditions, it is therefore important to include satellite images, climate data, and hydrologic information to predict the location and severity of droughts.

Hypothesis

Random forest model can accurately predict drought conditions across the US in a decade (2010~2020).

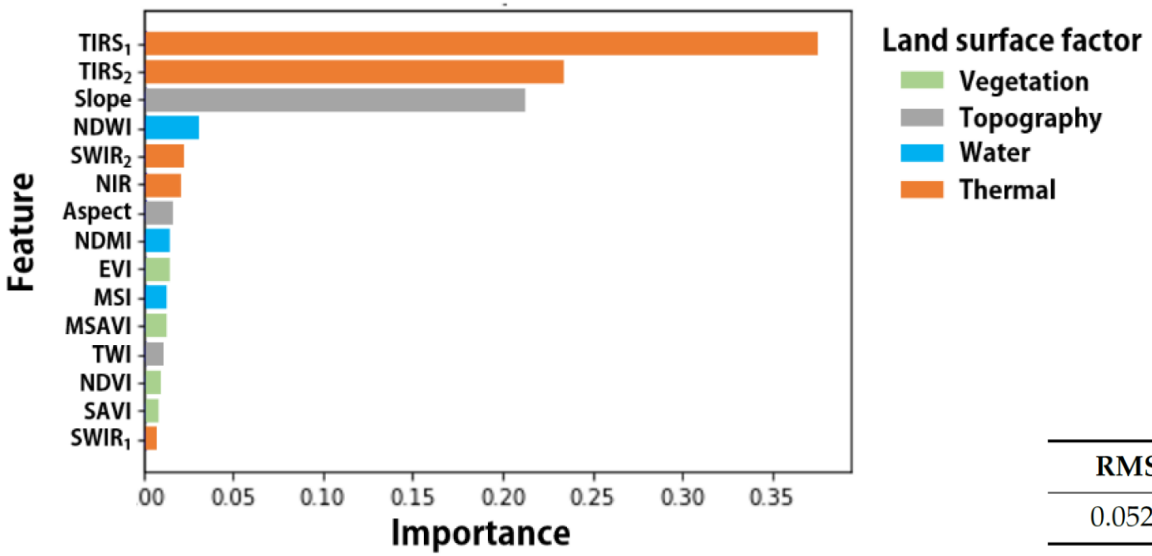
Objective

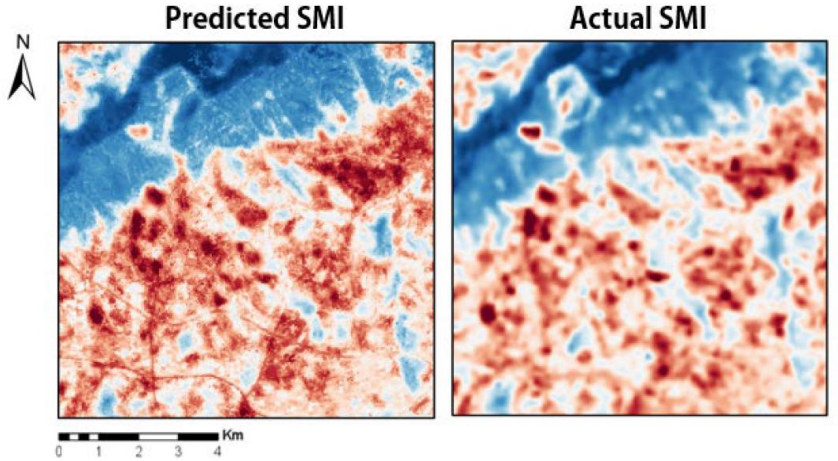
- Implement RF algorithm to predict drought conditions in the US and identify important factors that affect droughts the most.
- Compare RF with other machine learning algorithms (RF, SVM, ANN, GLM, RPART).

Literature review

Prediction of Severe Drought Area Based on Random Forest: Using Satellite Image and Topography Data [1]

- Generated drought function with the RF algorithm using input variables and three months afterward SMI (drought index) to train agricultural drought.
- Identified the order of feature importance that affects drought training (regression) among the input variables (features).





RMSE	NRMSE	MAE	R^2	Max. SMI [1]	Min. SMI [2]	Max. Error
0.05294	5.29%	0.03980	0.91	0.97940	0.05684	0.30352

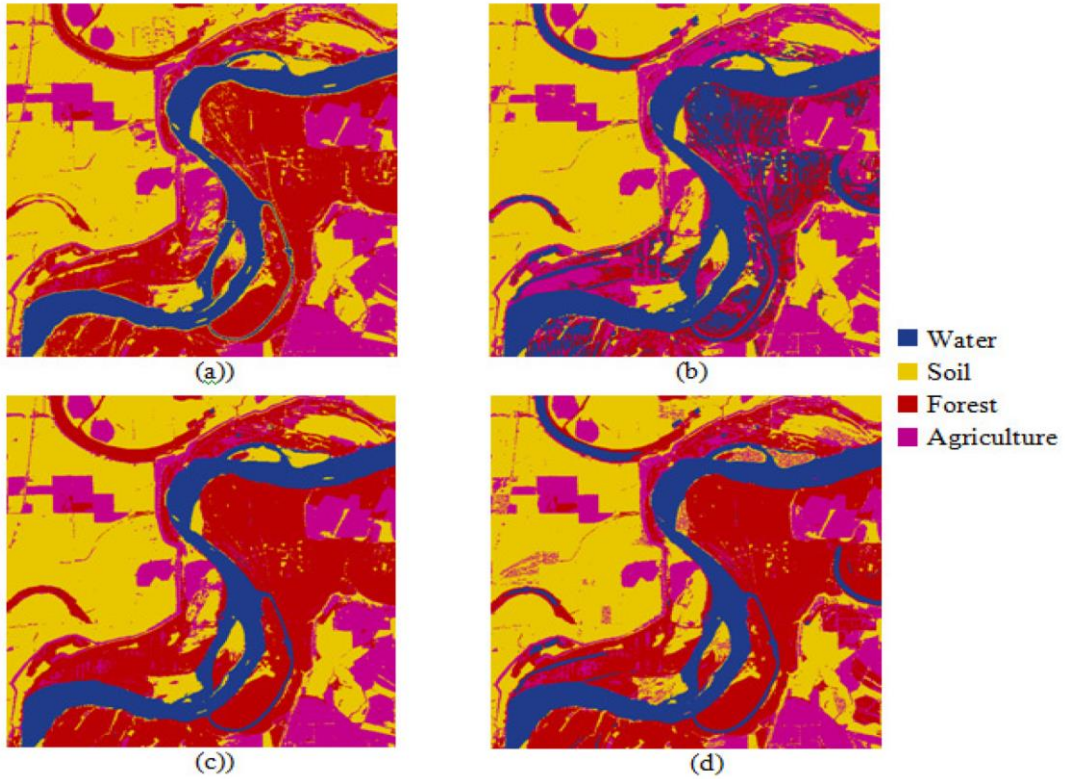
[1] The actual maximum SMI value is 1; [2] The actual minimum SMI value is 0.

Literature review

MULTISPECTRAL IMAGE ANALYSIS USING RANDOM FOREST [2]

- Introduced Random Forest Algorithms and provided implementation of Random Forest and examples of classification of pixels in multispectral images.
- Compare performance of the Random Forest algorithm with other classification algorithms

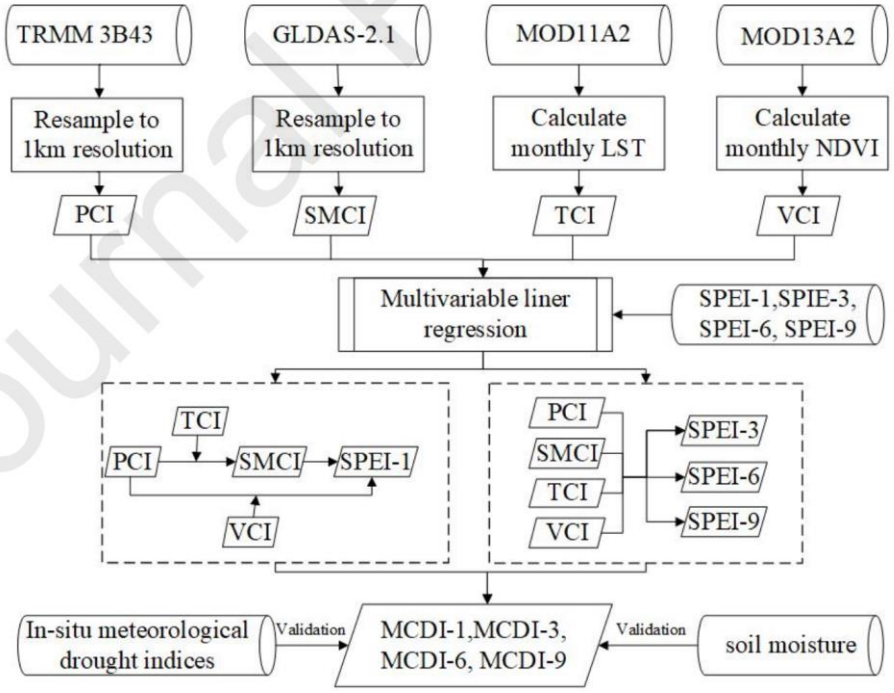
	Random Forest	Neural Network	Support Vector Machine	Maximum Likelihood
Overall Accuracy	96.25%	76.87%	86.88%	83.13%
Kappa	0.95	0.6917	0.825	0.775
Forest User's Accuracy	88.64%	100%	66.1%	64.91%
Water User's Accuracy	100%	64.51%	100%	100%
Soil User's Accuracy	100%	72.72%	100%	90.91%
Vegetation User's Accuracy	97.22%	100%	97.56%	93.02%
Forest Producer's Accuracy	97.5%	7.5%	97.5%	92.5%
Water Producer's Accuracy	100%	100%	50%	40%
Soil Producer's Accuracy	100%	100%	100%	100%
Vegetation Producer's Accuracy	87.5%	100%	100%	100%

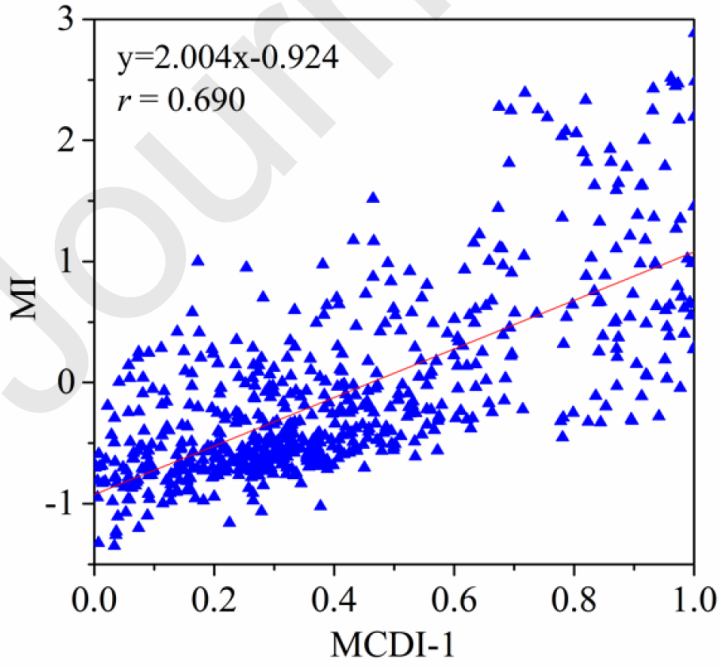


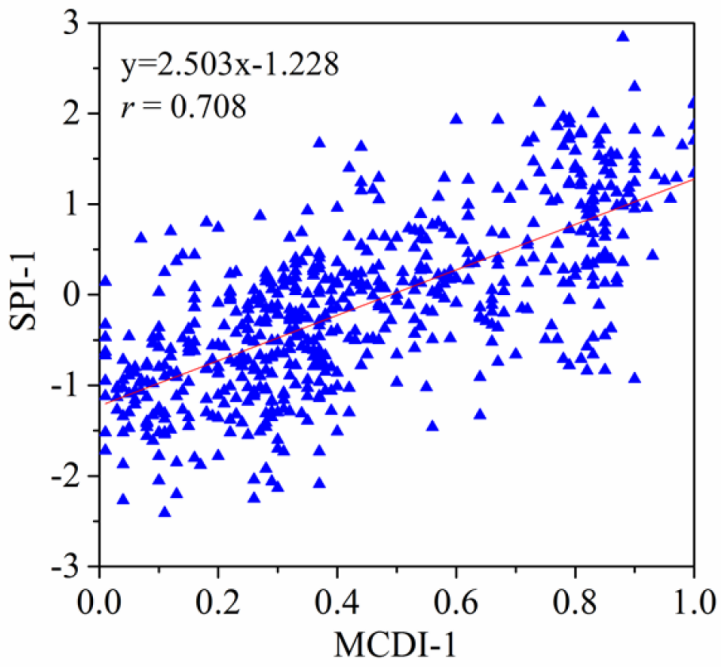
Literature review

Monitoring drought using composite drought indices based on remote sensing [3]

- Established composited drought index compare the correlation coefficients between composite indices and SPEIs with that between single indices and SPEIs
- Verified the ability of the composite index in monitoring meteorological drought and agricultural drought and investigated drought condition by in China over space and time.







Method

Time range: 2010~2020
Study area: the Contiguous USA
Tools: ArcGIS Pro, R

Table 1. Details of input variables

Variable (abbreviation)	Category	Dataset	Resolution (spatial/temporal)	Unit
Drought [4]	Drought	gridMET	3980 m/daily	Index
NDVI [5, 6]	Vegetation	MODIS Terra 16 Days	3980 m/16 days	Index
NDWI [7]	Vegetation	MODIS Terra/Aqua 16 Days	3980 m/16 days	Index
EVI [8]	Vegetation	MODIS Terra 16 Days	3980 m/16 days	Index
Temperature	Climate	gridMET	4660 m/daily	Degree
Precipitation	Climate	gridMET	4660 m/daily	Millimeters
Windspeed	Climate	gridMET	4660 m/daily	Meters/second
Humidity	Climate	gridMET	4660 m/daily	Grams/m^3
Burn	Fire	gridMET	4660 m/daily	Index
Soil Moisture	Hydrology	TerraClimate	4660 m/monthly	Millimeter
Runoff	Hydrology	TerraClimate	4660 m/monthly	Millimeter
Evapotranspiration	Hydrology	TerraClimate	4660 m/monthly	Millimeter

Method

- The rasters were clipped, resampled, georeferenced, and composited in ArcGIS Pro to ensure consistent spatial/temporal resolution for further processing.
- All preprocessed rasters were then converted to R dataframe and normalized to (0,1) range for Random Forest (RF) prediction.
- The whole R data frame was split to two parts: prediction datasets (70%) for training the random forest model, and testing datasets (30%) for validating the random forest model.
- The Random Forest tree count was set to 100 to compromise between model performance and quality.
- Several error measures were calculated to evaluate the performance of the RF model.
- Compare performance of the RF algorithm with other machine learning algorithms

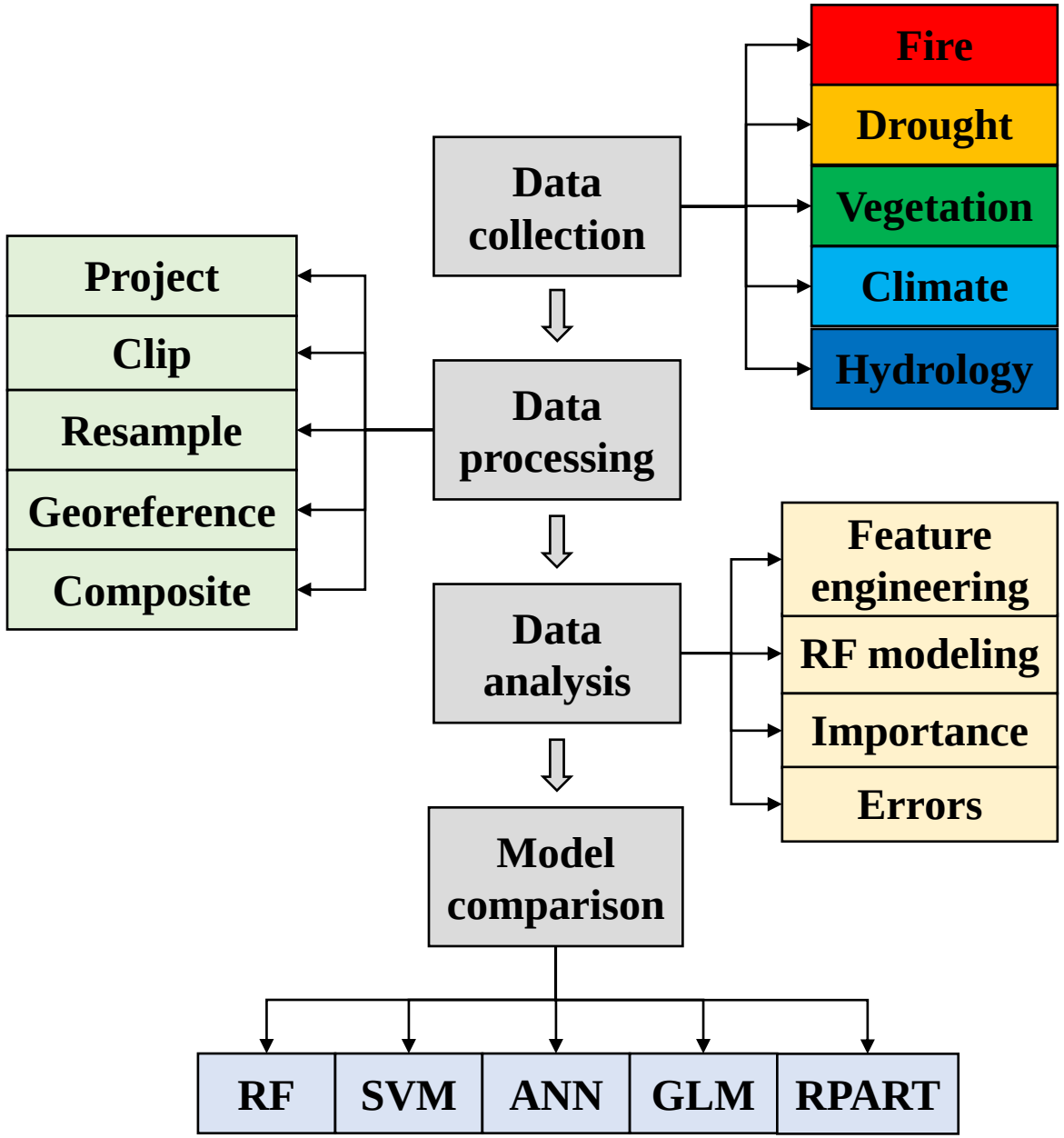


Figure 1. Flow chart

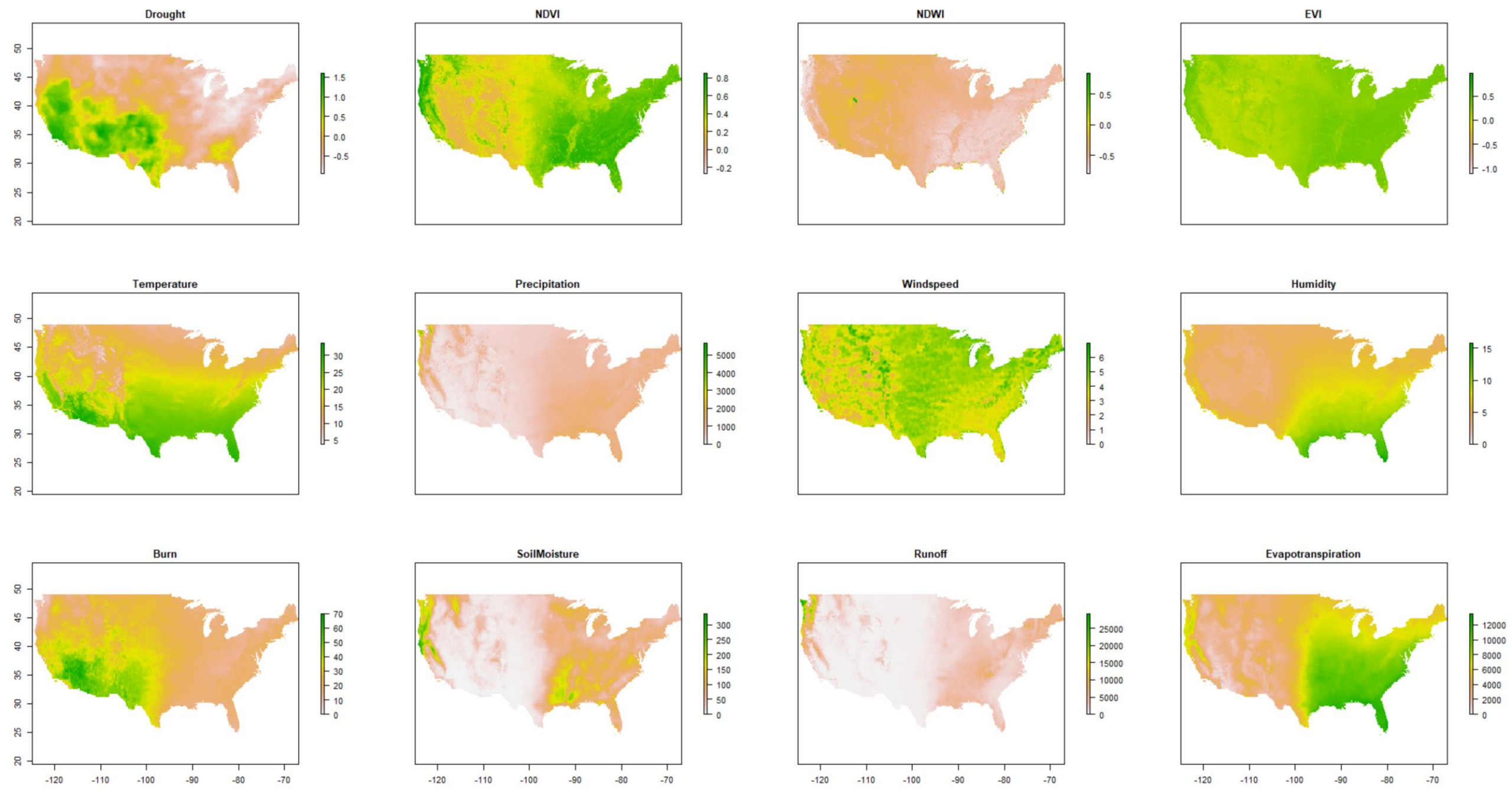


Figure 2. Thematic maps of drought related factors from 2010 to 2020(before normalization)

Result

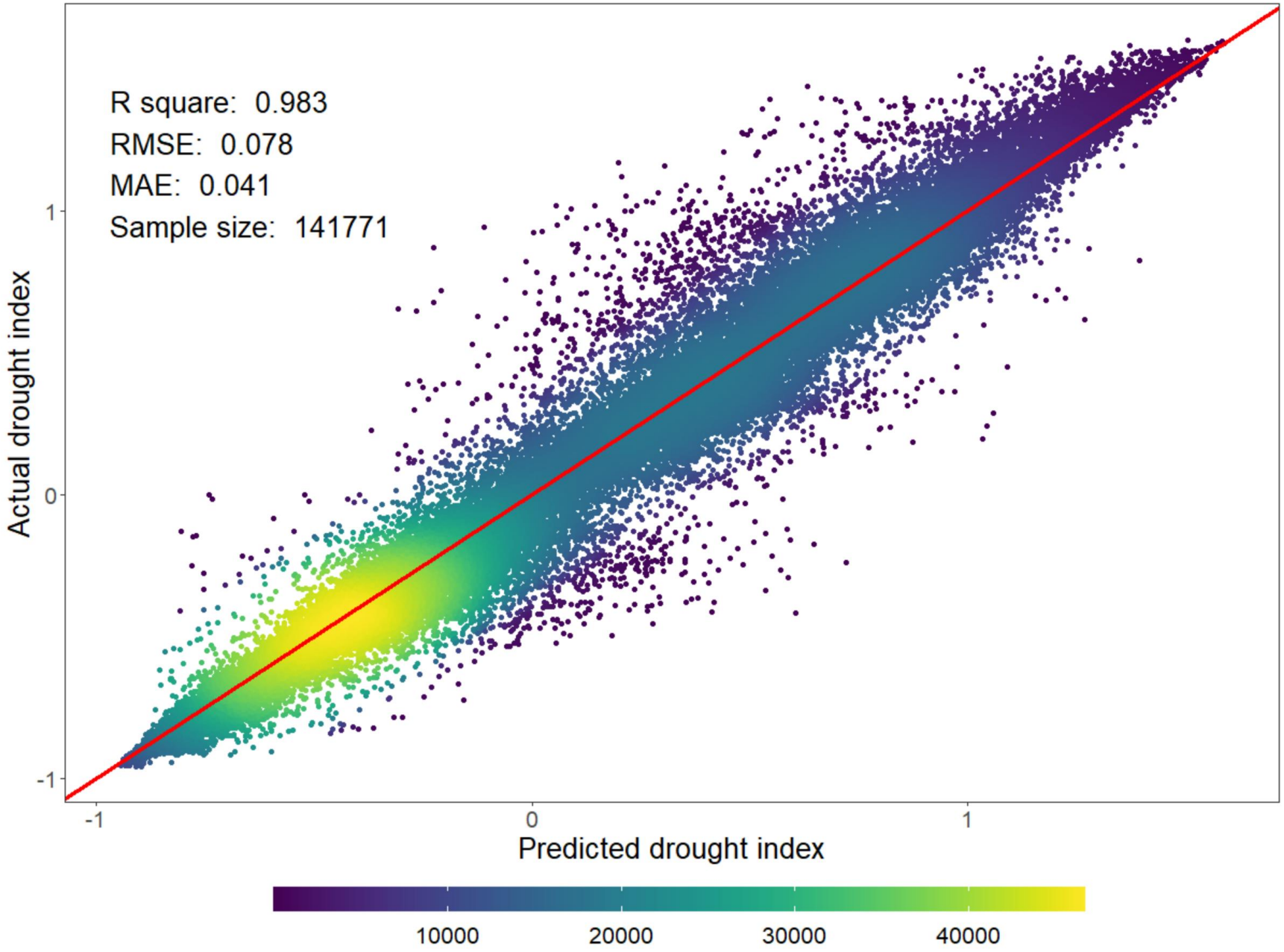


Figure 3. Density plot of actual vs. predicted drought index

Result

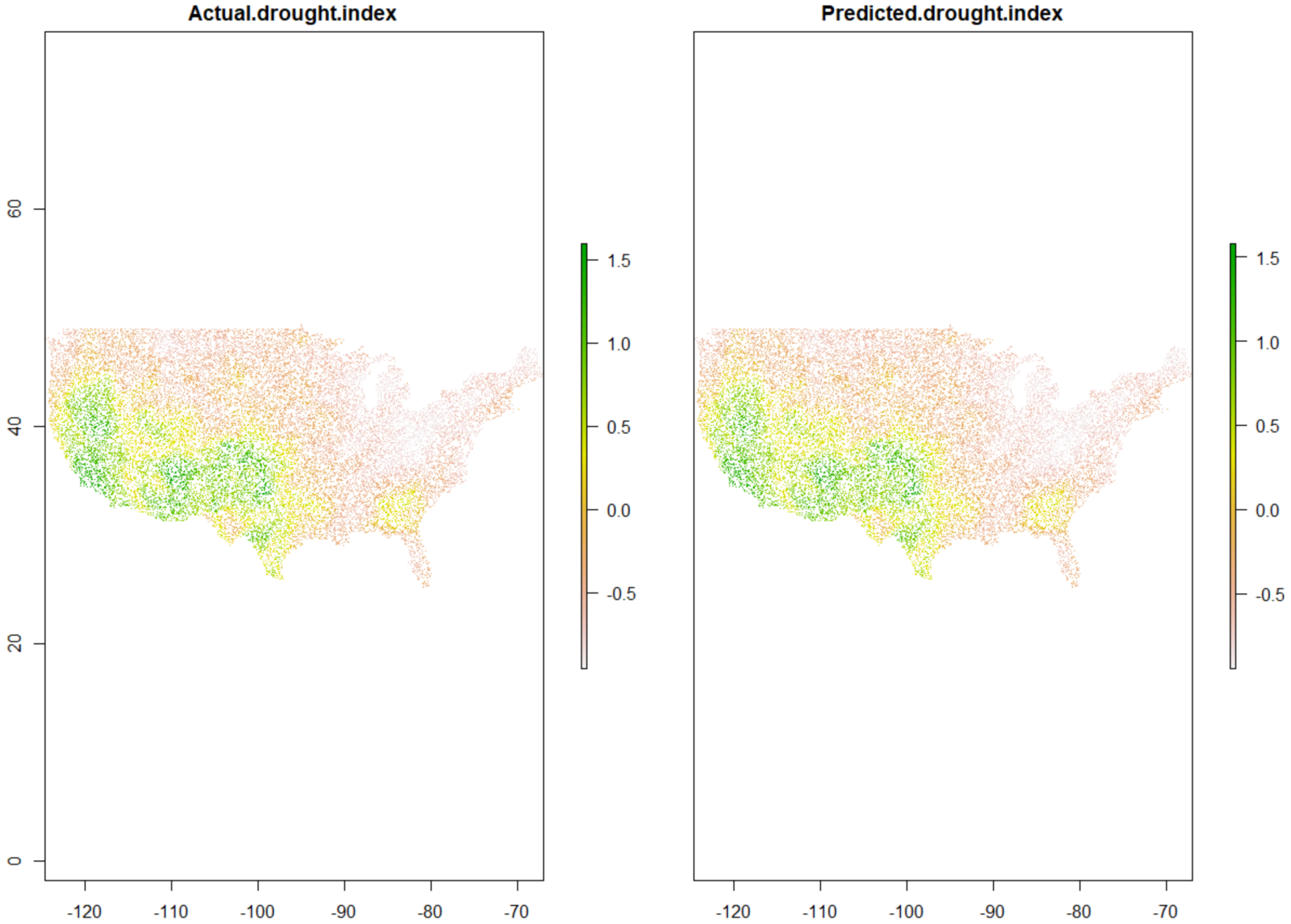
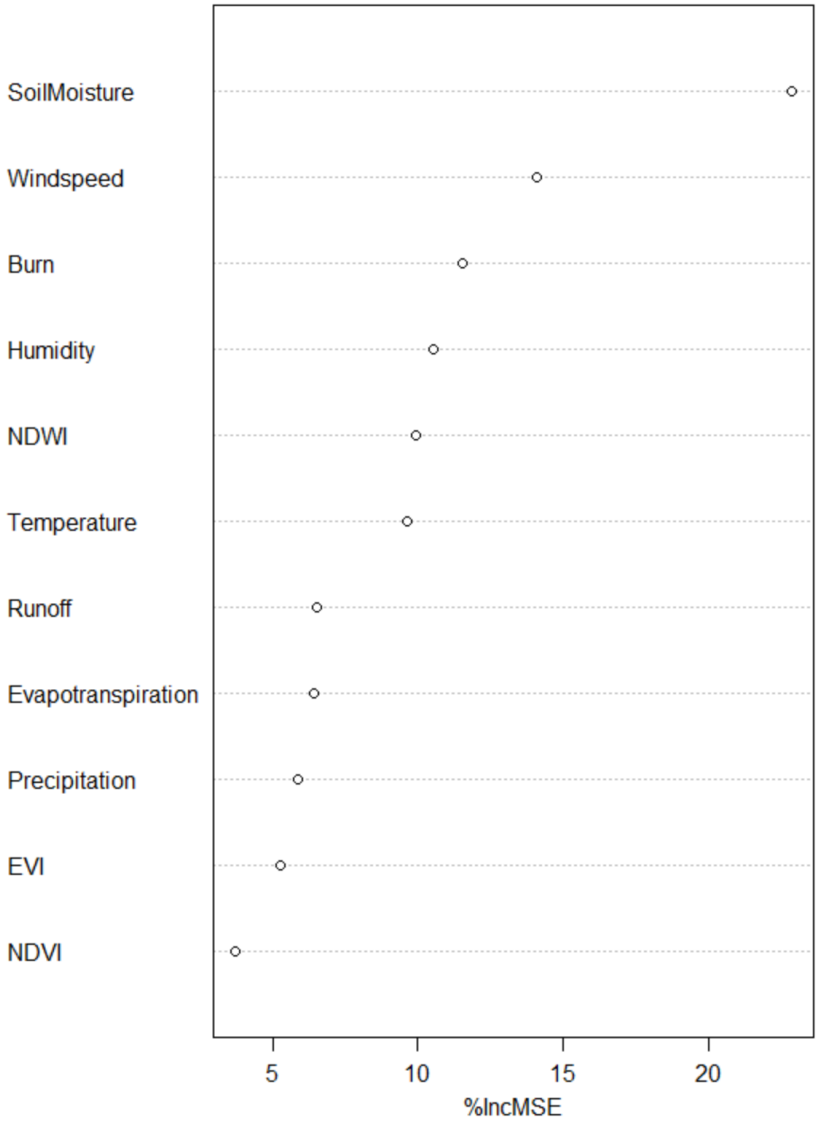
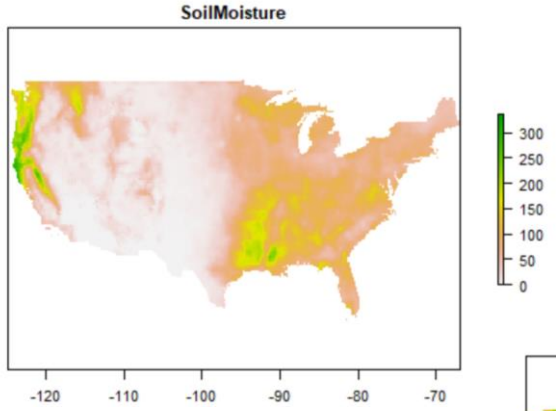


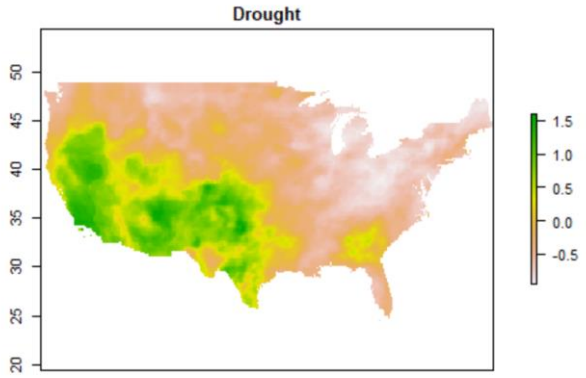
Figure 4. Verification of the RF performance based on test datasets (left: actual drought map; right: predicted drought map based on RF)

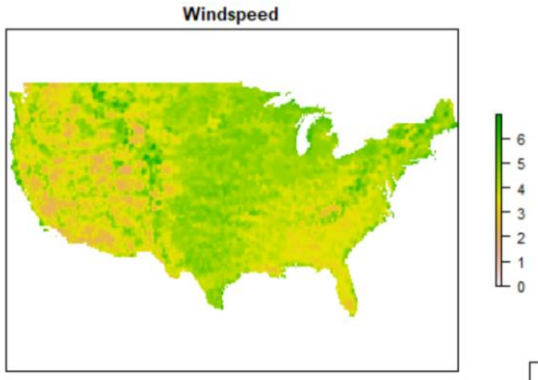
Result

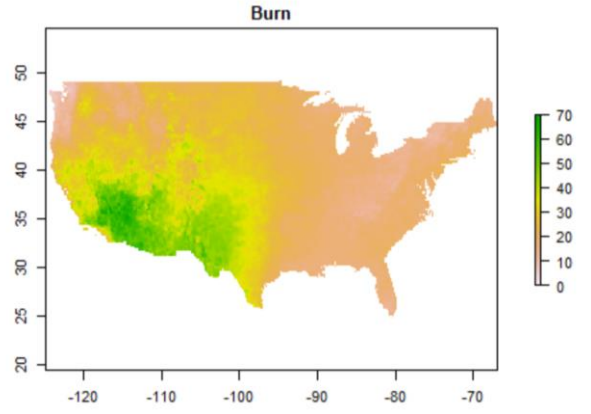


%IncMSE is the most robust and informative measure, which shows how much the model accuracy decreases if we leave out that variable.









Top 3 important factor: soil moisture, windspeed, and burn index.

Figure 5. Importance plot of drought related variables

Result

$$\Delta D = D_{real} - D_{pred} \quad \text{MAE} = \frac{\sum \text{abs}(\sum_{\text{i=1}}^{\text{n}} \Delta D_{\text{i}})}{\text{n}} \qquad \text{RMSE} = \sqrt{\frac{\left(\sum_{\text{i=1}}^{\text{n}} \Delta D_{\text{i}}\right)^2}{\text{n}}}$$

Bias (ΔD) refers to the difference between recorded (D_{real}) and predicted (D_{pred}) drought index. MAE is the absolute mean error, RMSE is the root mean square error, and n is the total number of drought index observations.

Table 2. Comparison of machine learning algorithms

Method	R^2	MAE	RMSE
RF [9]	0.983	0.041	0.078
RPART [10]	0.737	0.229	0.303
SVM [11]	0.848	0.162	0.228
ANN [12]	0.755	0.219	0.289
GLM [13]	0.666	0.254	0.338

Note: machine learning methods listed above refer to Random Forest (RF, 100 trees), R Recursive Partitioning and Regression Trees (RPART), Support Vector Machine (SVM), Artificial Neural Network (ANN), Generalized Linear Models (GLM). To compromise between performance and quality, resample process was applied to some time-consuming algorithms like RF and SVM.

Discussion & Conclusion

- RF and other machine learning algorithms were successfully implemented to predict drought conditions in the US.
- The importance of variables to predict drought index was also examined. Top 3 important factors were: soil moisture, wind speed, and burn index.
- RF results were compared with those from other machine learning algorithms (RF, SVM, ANN, GLM, RPART). The results showed that RF model performed the best, with an R^2 greater than 0.98 and very small errors (MAE, RMSE).
- Future work will focus on predicting other drought index (SPI, SPEI, etc.) using socio-economic, climate, hydrologic and vegetation factors.
- Future work will also consider raster-based machine learning models instead of vector-based models.

Reference

[1] Park, H., & Kim, K. (2019). Prediction of severe drought area based on random forest: Using satellite image and topography data. *Water*, *11*(4), 705.

[2] Lowe, B., & Kulkarni, A. (2015). Multispectral image analysis using random forest.

[3] Liu, Q., Zhang, S., Zhang, H., Bai, Y., & Zhang, J. (2020). Monitoring drought using composite drought indices based on remote sensing. *Science of The Total Environment*, *711*, 134585.

[4] Current map. (n.d.). Retrieved April 26, 2021, from https://droughtmonitor.unl.edu/

[5] Huete, A. R., Jackson, R. D., & Post, D. F. (1985). Spectral response of a plant canopy with different soil backgrounds. *Remote sensing of environment*, *17*(1), 37-53.

[6] Jackson, R. D., & Huete, A. R. (1991). Interpreting vegetation indices. *Preventive veterinary medicine*, *11*(3-4), 185-200.

[7] Gao, B. C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, *58*(3), 257-266.

[8] Liu, H. Q., & Huete, A. (1995). A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. *IEEE transactions on geoscience and remote sensing*, *33*(2), 457-465.

[9] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

[10] Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

[11] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-6. https://CRAN.R-project.org/package=e1071

[12] Stefan Fritsch, Frauke Guenther and Marvin N. Wright (2019). neuralnet: Training of Neural Networks. R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet

[13] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.