

The Vegetation Outlook (VegOut): A New Tool for Providing Outlooks of General Vegetation Conditions Using Data Mining Techniques

Tsegaye Tadesse and Brian Wardlow

National Drought Mitigation Center, University of Nebraska-Lincoln

E-mail: ttadesse2@unl.edu and bwardlow2@unl.edu

Abstract

The integration of climate, satellite, ocean, and biophysical data holds considerable potential for enhancing our drought monitoring and prediction capabilities beyond the tools that currently exist. Improvements in meteorological observations and prediction methods, increased accuracy of seasonal forecasts using oceanic indicators, and advancements in satellite-based remote sensing have greatly enhanced our capability to monitor vegetation conditions and develop better drought early warning and knowledge-based decision support systems. In this paper, a new prediction tool called the Vegetation Outlook (VegOut) is presented. The VegOut integrates climate, oceanic, and satellite-based vegetation indicators and utilizes a regression tree data mining technique to identify historical patterns between drought intensity and vegetation conditions and predict future vegetation conditions based on these patterns at multiple time steps (2-, 4-, and 6-week outlooks). Cross-validation (withholding years) revealed that the seasonal VegOut models had relatively high prediction accuracy. Correlation coefficient (R^2) values ranged from 0.94 to 0.98 for 2-week, 0.86 to 0.96 for 4-week, and 0.79 to 0.94 for 6-week predictions. The spatial patterns of predicted vegetation conditions also had relatively strong agreement with the observed patterns from satellite at each of the time steps evaluated.

1. Introduction

Drought is responsible for major crop and other agricultural losses in the U.S. each year. Recent extended periods of drought in the U.S. emphasized this vulnerability of the agricultural sector and the need for a more proactive, risk management approach to drought-induced water shortages that negatively affect vegetation conditions [1]. Preparedness and mitigation reduces drought vulnerability and its devastating impacts. A critical component of planning for drought is the provision of timely and reliable information that

aids decision makers at all levels in making critical management decisions [2].

Early warning is the foundation of effective drought planning and mitigation. As a result, thorough studies are underway to monitor and predict drought to reduce its devastating impacts on vegetation. However, predicting drought and its impact on vegetation is challenging because there is inevitable uncertainty in predicting precipitation. Moreover, even if it is possible to get accurate forecasts, the complex spatial and temporal relationships between climate and vegetation makes the prediction of vegetation condition difficult.

However, recent improvements in meteorological observations and forecasts have greatly enhanced our capability to monitor vegetation conditions [3]. This is because of the spatial and temporal patterns that we have learned from long historical records of climate, improved spatial coverage and density of weather observations, and the use of technologically advanced meteorological observation instruments. In addition, remote sensing observations from satellite-based platforms provide spatially continuous and repeat measurements of general vegetation conditions over large areas where no weather observing stations are present. Satellite observations have proven useful over the past 15 years for monitoring vegetation conditions at regional to global scales [4]. Most remote sensing studies have analyzed normalized difference vegetation index (NDVI) data, which provides a multi-spectral-based measure (combined spectral data from the visible red and near infrared wavelength regions) of plant vigor and general vegetation condition. Even though this information is very valuable for monitoring the vegetation condition, drought-affected areas cannot be distinguished from locations affected by other environmental stressors (e.g., pest infestation) solely from satellite-based observations.

There is an increasing demand for improved drought monitoring tools that provide more accurate and spatially detailed information than the traditional methods that have relied on only climate or satellite-derived vegetation index (VI) information [3]. In order

to monitor and predict vegetation conditions, one needs to understand the complex relationships of ocean-atmosphere-vegetation dynamics. Recent studies that investigated ocean-atmosphere relationships found significant improvements in seasonal climate predictions (e.g., precipitation and temperature) with the inclusion of oceanic indicators [5]. Thus, an integrated approach, which incorporates climate, oceanic, and satellite observations and considers the spatio-temporal relationships between these variables, is needed to improve our abilities to monitor and predict vegetation conditions.

The integration of climate, satellite, and ocean data, with other biophysical information such as available soil water capacity, ecoregion, and land cover type holds considerable potential for enhancing our drought monitoring and prediction capabilities beyond the tools that currently exist. Recent studies have shown that data mining techniques are an effective means to integrate this diverse collection of data sets and identify hidden and complex spatio-temporal patterns within the data that are related to drought [3]. As a result, data mining techniques are increasingly being used to assess the impact of drought on vegetation conditions and predict future conditions based on historical climate-ocean-vegetation relationships.

In this paper, we present a new drought monitoring tool called the Vegetation Outlook (VegOut) that provides outlooks of general vegetation conditions based on prior climate and ocean index measurements, satellite-based observations of current vegetation conditions, and other environmental information. A regression-tree modeling technique was used to analyze the time-lag relationships between vegetation conditions and the oceanic and climatic observations and predict future vegetation conditions at multiple time steps. The goal of this paper is to introduce the VegOut methodology and present initial results from a case study performed over a 15-state region in the central U.S. to demonstrate VegOut's utility to predict vegetation conditions at local to regional scales.

2. Data and Methods

2.1. Data

The specific climate, satellite, and oceanic data sets and the data for several static biophysical variables used in the VegOut regression tree models are briefly described below.

2.1.1. Climate-based data. Two commonly used climate-based drought indices, the Palmer Drought Severity Index (PDSI) and the Standardized Precipitation Index (SPI), were used to represent the

climatic variability that affects the vegetation condition in the VegOut models. The SPI is based on precipitation data and has the flexibility to detect both short- and long-term drought. The PDSI is calculated from a soil water balance model that considers precipitation, temperature, and available soil water capacity observations at the station. Both indices were initially calculated at each weather station location and an Inverse Distance Weighting (IDW) method was then applied to these point-based values to produce a continuous 1-km² gridded surface of SPI and PDSI values across the entire study area.

2.1.2. Satellite data. The Standardized Seasonal Greenness (SSG) metric, which represents the general condition of vegetation, was calculated from 1-km² resolution NDVI data over the study area. The SSG is calculated from the Seasonal Greenness (SG) measure, which represents the accumulated NDVI through time from the start of the growing season (as defined from satellite) [6]. From the SG data, the SSG is calculated at 2-week time steps throughout the growing season using a standardization formula (i.e., the current SG minus the average SG divided by the standard deviation). The result is a series of SSG images (which have values ranging from -4.0 to +4.0) that show the vegetation condition at 1-km² spatial resolution that can be compared spatially to the other geospatial data sets.

2.1.3. The oceanic indices. Eight oceanic indices that show ocean-atmosphere dynamics and teleconnections were used in this study. The indices include the Southern Oscillation Index (SOI), Multivariate El Niño and Southern Oscillation Index (MEI), Pacific Decadal Oscillation (PDO), Atlantic Multi-decadal Oscillation (AMO), Pacific/North American index (PNA), North Atlantic Oscillation index (NAO), Madden-Julian Oscillation (MJO), and Sea Surface Temperature anomalies (SST). The observed values for each index on the three dates the Vegetation Outlooks were calculated from had to be converted to 1-km² images, to which the predictive VegOut models were applied.

2.1.4. Biophysical data. The biophysical parameters used in this study included land cover type, available soil water capacity, percent of irrigated land, and ecosystem type. The dominant (or majority) value within a 9-km² window surrounding each weather station was calculated from the 1-km² images for each biophysical variable and used for VegOut model development.

For the dynamic climate, oceanic (2-week values extrapolated from monthly data), and satellite variables (the bi-weekly historical records from 1989 to 2005); and single, static values for the biophysical variables were extracted for each weather station and organized into a database, which would be used in the regression

tree analyses to generate the rules for the VegOut model.

2.2. Weather station selection and development of the training database

More than 3000 weather stations were available for the 15 states to build the historical database. However, only 1402 stations were selected to be used in the VegOut model. Stations that did not have a long historical climate record (i.e., > 30 years of precipitation data and > 20 years of temperature data) and/or were not currently in operation were excluded. Stations that were predominately surrounded by either an urban area or water (i.e., > 50% of the surrounding 3 km x 3 km area) were also eliminated because they would not be representative of vegetation conditions.

2.3. Rule-based regression-tree modeling

The data mining method used in this study was a rule-based regression tree method available in the commercially available *Cubist* data mining software. The technique is generally referred to as regression-tree modeling. *Cubist* analyzes data and generates rule-based linear models that are a collection of rules, each of which is associated with a linear expression for computing a target value. The user determines the dependent and independent variables [7]. The predictive models (e.g., 2-week outlook) were developed using a time-lag relationship in which the model is based on the historical patterns of a delayed response of future vegetation condition to a weather event (e.g., observed precipitation and temperature condition). The dependent variable in the model is the satellite-observed SSG, and all other climate, oceanic, and biophysical parameters were used as input variables. Thus, the VegOut is predicting the SSG into the future at the 3 time steps (i.e., 2-, 4-, and 6-weeks).

The rules of the VegOut model were generated using the historical data from 1989 to 2005. The training data for the model included all data except a randomly selected hold-out year (i.e., 2000) that was used as test data to cross-validate the accuracy of the model.

The VegOut product is generated bi-weekly from the start of the growing season. However, in this paper, only 3 VegOut models developed for spring (April to June), mid-summer (Jun to August), and fall (August to October) dates were presented to illustrate the seasonal predictive ability of the VegOut approach. The biweekly periods selected to test the VegOut models were Period 10 (the first two weeks of May), Period 15 (the last two weeks of July), and Period 17 (the last 2 weeks of August) representing the spring, mid-summer, and fall season, respectively.

After generating the rule-based seasonal models for each of the three time steps, each model was applied to the geospatial gridded data to produce a 1-km² spatial resolution VegOut map of predicted general vegetation conditions over the next 2, 4, and 6 weeks.

3. Model Evaluation and Implementation

3.1. Evaluating with training and test data

A cross-validation technique [8] was used to compare the predicted VegOut model results to the observed SSG values in the test data from the 2000 holdout year. The Mean Absolute Difference (MAD) expressed in SSG data units and the correlation coefficient (R^2) between the observed and predicted SSG calculated during the *Cubist* regression tree runs were also evaluated.

Period	Outlooks	Evaluation on test data	
		MAD	R^2
Period 10	2-week	0.16	0.94
	4-week	0.23	0.86
	6-week	0.29	0.79
Period 15	2-week	0.09	0.98
	4-week	0.14	0.94
	6-week	0.18	0.92
Period 17	2-week	0.07	0.98
	4-week	0.11	0.96
	6-week	0.15	0.94

Table 1. Evaluation of the VegOut Model. The Mean Absolute Difference (MAD) values and the correlation coefficient (R^2) between the observed and predicted SSG are shown for each period and the corresponding outlooks.

Cross-validation (withholding a random year) revealed the seasonal VegOut models had relatively high prediction accuracy (Table 1). R^2 values ranged from 0.94 to 0.98 for the 2-week outlook, 0.86 to 0.96 for the 4-week outlook, and 0.79 to 0.94 for the 6-week outlook predictions. The R^2 value was slightly lower during the spring phase compared to the mid-summer (peak growing season) and fall (senescence) periods of the growing season. The MAD increased as the length of predictions increased from 2 to 6 weeks. The MAD values ranged from 0.07 to 0.16, 0.11 to 0.22, and 0.14 to 0.28 for the 2-, 4-, and 6-week outlooks, respectively. Higher prediction errors (higher values of MAD) were observed in spring (early growing season) when the seasonal greenness is not as stable because of the different timings of initial vegetation growth, which can vary by land cover type and under different general environmental conditions.

3.2. Evaluating the model with the 2006 drought year

The VegOut model was implemented for the 2006 growing season to evaluate the accuracy of the model. Each model was applied to geospatial gridded data to produce a 1-km² spatial resolution VegOut map of predicted general vegetation conditions at the 3 time steps. Figure 1 shows the series of VegOut maps for the July 11-24 bi-weekly period to illustrate the observed and predicted vegetation conditions. Figure 1a shows the observed SSG for bi-week period 15 (July 11-24), from which the predictions of SSG were made. Using the observed satellite, climatic, oceanic, and biophysical data for period 15 as an input, the VegOut model produced the 2-, 4-, and 6-week outlook maps (Figures 1b, 1c, and 1d). These observed SSG for the periods corresponding to each of these outlooks are presented in Figures 1e, 1f, and 1g.

The VegOut maps presented in Figure 1 depicted generally similar spatial patterns of vegetation conditions (or SSG values) as compared to the satellite-based vegetation observations on those corresponding dates. However, there are a few minor, localized differences in the spatial patterns that were modeled. For example, in the 2- and 4-week outlooks (Figures 1b and 1c), drought patterns that were expected to linger in the southern part of Texas actually improved sooner at those times (Figure 1e and 1f). However, the model predicted similar vegetation conditions in the 6-week outlook over this same area (Figures 1d and 1g). There was also some disagreement between the predicted and observed patterns in eastern Illinois and southern Wisconsin, which became more apparent at the longer 4- and 6-week time steps. Similar levels of spatial agreement with some localized differences were observed for the series of outlooks produced for the other two spring and fall periods.

3.3. Evaluating the model with Land Cover

The VegOut maps were also assessed over the 15-state region's three major land cover types to evaluate VegOut's predictive ability for different vegetation classes. The median difference between the predicted and observed SSG values was calculated from all pixels corresponding to cropland, forest, and rangeland across the study region. The median difference between the observed and predicted SSG values was minimal across the 15-state region for the three time periods that were tested (Table 2). The maximum difference was 0.22 SSG units (period 10, 6-week outlook for all land cover and crops), with most differences less than 0.10 SSG units. The difference

slightly increased as the outlook period became longer for several period/land cover combinations, but this increase was not substantial on the SSG's -4.0 to +4.0 scales. Several interesting trends appeared during the land cover-specific comparisons. First, the 6-week outlooks typically had the largest differences. Second, the predicted values were usually less over cropland, with the largest deviations early in the growing season (period 10) when the crops are beginning to emerge. Lastly, the differences for rangeland were considerably greater for the 4- and 6-week outlooks at the beginning (period 10) and end (period 17) of the growing season than during the midsummer (period 15). The trend was to under-predict the SSG values during these periods, with the exception of the 4-week outlook during period 10. In general, the predicted and observed SSG values had strong agreement across the study area and no prominent differences or trends in these values were observed for a specific time period, outlook, or land cover type.

Period	Outlook	All Land Cover	Crops	Forest	Range
Period 10	2-week	2.67	1.13	2.67	0.79
	4-week	3.06	-18.38	3.73	13.54
	6-week	-20.10	-22.21	-13.03	-14.85
Period 15	2-week	-3.15	-4.20	-0.65	-7.93
	4-week	-2.50	-6.00	-3.34	-2.50
	6-week	-5.20	-15.08	5.33	-0.01
Period 17	2-week	1.22	5.11	4.83	1.22
	4-week	-5.19	-11.31	-1.25	-10.03
	6-week	0.78	-1.84	15.66	-10.00
Sample size (n)		3714767	1358694	589757	1592105

Table 2. Median difference between the predicted and observed SSG values across the 15-state study area by land cover type.

4. Future Works

At present, the VegOut uses rule-based regression tree models that are generated by identifying relationships between satellite-derived vegetation conditions, climatic drought indices, oceanic indices, and other biophysical data. Alternative modeling techniques including association rules and neural networks are being investigated to compare with the current VegOut models. Ensemble techniques that base predictions on the results from multiple data mining techniques are also under consideration. In addition, new inputs into the current VegOut models are also being investigated in an effort to provide more accurate predictions of future vegetation conditions. The current VegOut research is focusing on the development of 2-, 4-, and 6-week vegetation outlooks in the U.S. Great

Plains, but expansion of VegOut to other areas of the U.S. is planned in the near future.

5. Conclusion

The VegOut is a new drought monitoring tool that provides outlooks of general vegetation conditions. VegOut integrates climate information and satellite-based observations of current vegetation conditions with oceanic index data and other biophysical information about the environment to produce 1-km² resolution maps of projected general vegetation conditions. The data mining method builds numerical rule-based models that find temporal and spatial relationships among the input variables. The VegOut utilizes the inherent time-lag relationship between climate and vegetation response and considers teleconnections between the ocean and climate patterns over the continental U.S. Because the models can be applied iteratively with input data from previous time periods, the method enables predictability of vegetation conditions farther into the growing season based on earlier conditions.

The results from this study illustrate the substantial potential VegOut holds for predicting general vegetation conditions at multiple time steps. The predictive accuracy of the seasonal VegOut models was quite high ($R^2 > 0.90$) for the three periods tested. The spatial patterns of future vegetation conditions characterized in the series of VegOut maps also had a high level of agreement with the patterns observed from satellite on those dates. Most of the differences between the predicted and observed patterns were relatively localized and/or restricted to a limited number of outlook periods. In addition, the predictive accuracy of VegOut did not substantially decline as the outlook interval became longer, which suggests that relatively stable vegetation outlook results can be attained for predictions made up to 6 weeks in advance.

The experimental results from this study suggest that there is a strong potential to use data mining in monitoring drought and its impact on vegetation conditions ahead of time over large geographic areas. VegOut represents a potentially valuable tool for the agricultural sector, which can use the VegOut results to assess potential impacts of drought on crop and rangeland production into the growing season. Considerable progress and improvements to VegOut are expected in the future as the research initiatives outlined in section 4 are undertaken.

6. Acknowledgements

This study is supported in part by the U.S.D.A.'s Federal Crop Insurance Corporation (FCIC) through the Risk Management Agency (RMA) under USDA partnership (02-IE-0831-0228) with the National Drought Mitigation Center, University of Nebraska-Lincoln.

7. References

- [1] D. Wilhite. Preparing for drought: a methodology. In: Wilhite, D.A. (Ed.), *Drought: A Global Assessment*. Routledge Hazards and Disaster Series Vol. II, pp. 89-104, 2002.
- [2] D. Wilhite and M. Svoboda. *Drought Early Warning Systems in the Context of Drought Preparedness and Mitigation*. Preparedness and Mitigation Proceedings of an Expert Group, Lisbon, Portugal, 2000.
- [3] T. Tadesse, J.F. Brown, and M.J. Hayes. A new approach for predicting drought-related vegetation stress: Integrating satellite, climate, and biophysical data over the U.S. central plains. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(4):244-253, 2005.
- [4] Y. Bayarjargala, A. Karnieli, M. Bayasgalan, S. Khudulmurb, C. Gandush, and C.J. Tucker. A comparative study of NOAA-AVHRR derived drought indices using change vector analysis. *Remote Sensing of Environment*, 105(1):9-22, November 2006.
- [5] T. Tadesse, D.A. Wilhite, S.K. Harms, M.J. Hayes, S. Goddard. Drought monitoring using data mining techniques: A case study for Nebraska, U.S.A., *Natural Hazards Journal*, 33: 137-159, 2004.
- [6] B.C. Reed, J.F. Brown, D. VanderZee, T.R. Loveland, J.W. Merchant, and D.O. Ohlen. Measuring phenological variability from satellite imagery. *Journal of Vegetation Science*, 5: 703-714, 1994.
- [7] Rulequest Research. An overview of Cubist. <http://rulequest.com/cubist-win.html> (Accessed on July 9, 2007).
- [8] B.K. Wylie, E.A. Fosnight, T.G. Gilmanov, A.B. Frank, J.A. Morgan, M.R. Haferkamp, and T.P. Meyers. Adaptive data-driven models for estimating carbon fluxes in the Northern Great Plains. *Remote Sensing of Environment*, 106(4): 399-413, February 2007.

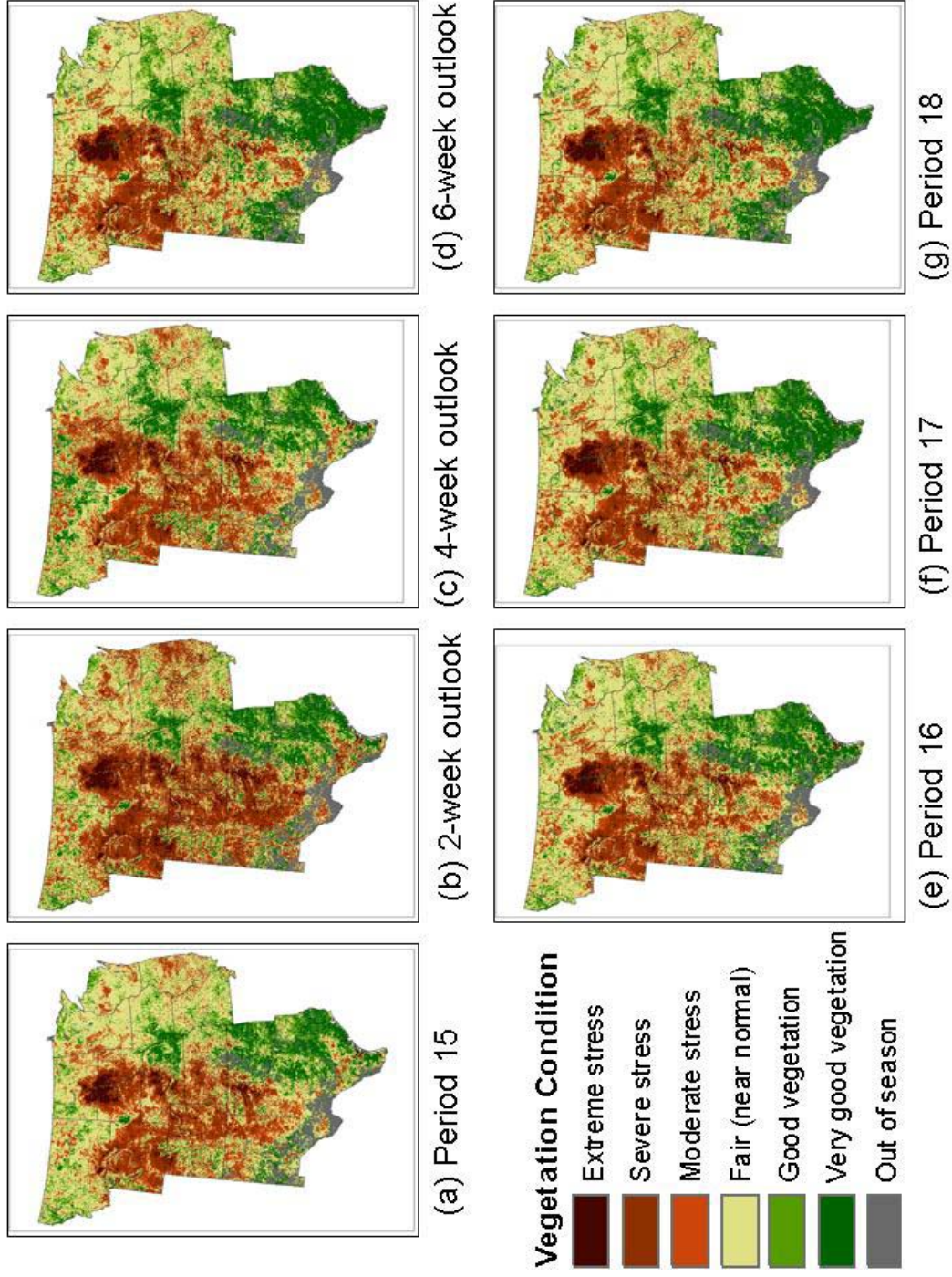


Figure 1. (a) Observed Seasonal Greenness (SSG) for period 15 (July 11 to 24) in 2006; (b) to (d) are 2-, 4-, and 6-week outlooks; (e) to (g) are observed SSG for periods 16 (July 25 to August 6), 17 (August 7 to August 20), and 18 (August 21 to September 4) that correspond to the 2-, 4-, 6-week outlooks, respectively.