# APPLICATION OF MACHINE LEARNING TO THE PREDICTION OF VEGETATION HEALTH

Emily Burchfield[a], John J. Nay[b], Jonathan Gilligan[c]

a. Dept. of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN 37235 USA, emily.k.burchfield@vanderbilt.edu
b. School of Engineering, Vanderbilt University, Nashville, TN 37235 USA, john.j.nay@vanderbilt.edu
c. Dept. of Earth and Environmental Science, Vanderbilt University, Nashville, TN 37235, USA, jonathan.gilligan@vanderbilt.edu

**KEY WORDS:** Machine learning, predictive modeling, decision support, open source software, vegetation index.

**ABSTRACT:**

This project applies machine learning techniques to remotely sensed imagery to train and validate predictive models of vegetation health in Bangladesh and Sri Lanka. For both locations, we downloaded and processed eleven years of imagery from multiple MODIS datasets which were combined and transformed into two-dimensional matrices. We applied a gradient boosted machines model to the lagged dataset values to forecast future values of the Enhanced Vegetation Index (EVI). The predictive power of raw spectral data MODIS products were compared across time periods and land use categories. Our models have significantly more predictive power on held-out datasets than a baseline. Though the tool was built to increase capacity to monitor vegetation health in data scarce regions like South Asia, users may include ancillary spatiotemporal datasets relevant to their region of interest to increase predictive power and to facilitate interpretation of model results. The tool can automatically update predictions as new MODIS data is made available by NASA. The tool is particularly well-suited for decision makers interested in understanding and predicting vegetation health dynamics in countries in which environmental data is scarce and cloud cover is a significant concern.

## 1. INTRODUCTION

Remotely sensed measures of vegetation health, such as the Normalized Difference Vegetation Index (NDVI) or the Enhanced Vegetation Index (EVI), are widely used to monitor agricultural responses to drought (Peters et al., 2002; Rhee, Im, & Carbone, 2010). Providing managers and farmers with accurate information about vegetation health increases system-wide capacity to prepare for and adapt to water scarcity (Dessai, 2009; Ziervoge et al., 2010). In many tropical countries, however, persistent cloud cover causes significant gaps in data availability. In this paper, we describe an open source tool we have developed that predicts vegetation health at a relatively high spatial resolution worldwide.[1] The tool applies a gradient-boosted machine model (GBM) to 16-day 250 meter resolution Moderate Resolution Imaging Spectroradiometer (MODIS) datasets which are openly available on NASA's LP DAAC server. The model learns potentially complex relationships between past remotely sensed variables (and their interactions) and future vegetation health as measured by the Enhanced Vegetation Index (EVI). The tool forecasts vegetation health 16-days out and can be used to impute missing data in highly cloudy locations. In this paper, we apply the tool in two South Asian countries with extremely high levels of seasonal cloud cover: Sri Lanka and Bangladesh.

### 1.1 Sri Lanka

Sri Lanka is a small island nation located off of the eastern coast of India that covers approximately 66,000 square kilometers and is home to nearly 21 million people (Government of Sri Lanka, 2010). The country receives two-thirds of annual rainfall during the northeast monsoon which lasts from October to December. The southwest monsoon lasts from May to October and brings rain primarily to the southwestern region of the island. This rainfall pattern creates two distinct cultivation seasons, the wet Maha season and the dry Yala season (Senaratne & Scarborough, 2011). During the wet season, most farmers cultivate rice, which is a stable of the Sri Lankan diet. Farmers capture wet season rainfall in reservoirs, known locally as tanks, and cultivate rice during the dry season with this stored water. During water scarce dry seasons, farmers cultivate other field crops such as soy, maize, and grain. Field size is small in Sri Lanka, with over 70 percent of farmers cultivating less than 2.5 acres of land (Withananachchi et al., 2014).

### 1.2 Bangladesh

Bangladesh is, apart from a few small city-states, the most densely populated nation on earth, with approximately 160 million people living on 150,000 square kilometers (Lewis, 2011, p. 13). Bangladesh has a monsoon climate, with an average of approximately 2100 mm of rainfall in May through September, compared to 90 mm in November through March. Rice cultivation is the dominant agricultural activity, accounting for roughly 40% percent of total land use (60% of cultivated land) and 10% of GDP (World Bank, 2009). There are three distinct seasons: *aman* rice, traditionally the dominant crop, is planted during the monsoon rains in July/August, and is harvested in November/December; *aus* rice is planted toward the end of the dry season, in March/April, so as to catch the early rains in late April through May, and is harvested in early summer (June/July); *boro* rice is planted in December/January, after the *aman* harvest, and is irrigated with ground water and harvested in the spring, April/May (USDA, 2013). *Boro* rice, which generally consists of high-yielding seed varieties, has significantly greater yields than either *aman* or *aus*. Boro rice

[1] For more information about tool development and testing, please refer to the paper by Nay et al. (2016).

has grown from almost nothing at the time of independence, in 1971 to constitute more than half of national rice production (Lewis, 2011, p. 137-8). Expansion of boro rice, combined with increased productivity of aman rice has tripled annual production since the 1970s (Baffes & Gautam, 2001). Boro production is limited by access to seed and fertilizer, but even more so by access to suitable groundwater and electricity to run irrigation pumps. The number of acres under irrigation has roughly tripled since 1980 (Census of Agriculture, 2008). Throughout the nation, irrigation withdrawals have significantly lowered the water table, and salinity in the groundwater, due both to naturally saline aquifers and to salinity intrusion in the coastal areas, is constraining production (Dasgupta et al., 2014). Field size is also small in Bangladesh with 84 percent of farmers cultivating less than 2.5 acres (Census of Agriculture, 2008).

## 2. METHODS

In this study, we measure variations in vegetation health using the Enhanced Vegetation Index (EVI) which is a proxy for the health of agricultural crops (Cai & Sharma, 2010; Galford et al., 2008; Gumma, 2011; Sakamoto et al., 2005; Xiao et al., 2006), highly correlated with the leaf area index (Huete et al., 2002), and positively linearly related to vegetation fraction estimates (Small & Milesi, 2013). The EVI is measured as:

$$EVI = G\frac{\rho_{NIR}-\rho_{RED}}{\rho_{NIR}+C_1\,x\,\rho_{RED}-C_2\,x\,\rho_{BLUE}+L} \quad (1)$$

where ρ is atmospherically corrected surface reflectance, L is the canopy background adjustment, and C1 and C2 are the coefficients of the aerosol resistance term, which uses the blue band to correct for aerosols in the red band (Huete et al., 2002). EVI values approaching one indicate high levels of photosynthetic activity.

We created a set of Python scripts to automate downloading and processing MODIS data from the MOD09A1 and MOD13Q1 datasets. The software downloads, mosaics, clips, and projects HDF files downloaded from the LP DAAC server and masks all pixels not flagged as "good quality" by each dataset's quality mask. 8-day datasets are transformed to a 16-day time step by computing the average of two quality-masked 8-day pixels. All datasets are resampled to match the spatial resolution of the EVI dataset (250 meters). These scripts, along with the modeling and validation scripts, are open source and can be used to download any MODIS tile for any MODIS dataset found on NASA's LP DAAC server (http://johnjnay.com/forecastVeg). The user has the option of including ancillary geospatial datasets such as land use information, socioeconomic data, or climate data to increase the predictive power of the model. For our analysis, we included gridded world population (CIESIN, 2005), land use (Survey Department, 2011) and an El Niño sea surface temperature index (Rayner et al., 2003). The Niño 3.4 SST Index was used in Sri Lanka and Bangladesh.

We computed the spatial autocorrelation functions of the MOD13Q1 imagery to divide the final matrix into a grid of independent areas. In both regions, autocorrelation functions approached zero at a lag of 150 pixels (approximately 35 kilometers). We divided each image into a grid of 150-pixel by 150-pixel cells and randomly assigned a subset of these cells to training and testing data.

We selected a model that performs well in supervised learning tasks where complex functions link the predictor and outcome variables and there is missing predictor variable data: the gradient boosted machine (GBM) model. To contextualize quantitative performance measures of our model, we compared its performance to a simple model that uses approximate nearest neighbor search to find *k* spatial-temporally close pixel-time observations in the hold-out data and averages their EVI values.

We used a GBM implementation in h2o, an open-source library that allowed us to hold hundreds of millions of rows of data in memory (H2O.ai Team, 2015). The GBM iteratively adjusts model parameters in the direction of lower prediction errors by using gradient computations and improves an ensemble of base models by adjusting the training data. The base models are binary split trees that divide predictor variable values into distinct regions (Hastie, Tibshirani, & Friedman, 2009). The GBM is ideal for our regions of interest because it can handle missing predictor variables by incorporating missing values in the overall tree structure. The model can also automatically learn interactions between predictor variables. Manually specifying all potential interactions would be time-intensive and the interactions that lead to the best predictive performance can vary by location. The algorithm learns which interactions are useful.

The GBM has hyper-parameters that need to be tuned. We used a Tree of Parzen Estimators search algorithm to model the effect of the hyper-parameters on the mean-squared error of the model's predictions of a held-out subset of the training data (Bergstra, Yamins, & Davis, 2013). We selected the hyper-parameters with the highest performance on the training data. Then we trained the model with those hyper-parameter settings on the full training data and used this model to forecast EVI in the hold-out validation data to test our best model on unseen observations.

## 3. RESULTS

### 3.1 Performance across Space

We measured the performance of the model by calculating the correlation between the vector of 16-day ahead predictions of EVI and vector of actual values of EVI in the held-out data. We computed the correlation for each land use category and found that model performance relative to the baseline is high in all categories of land use (Figure 1). In both regions, the correlation in agricultural areas is above 0.75 (0.86 in Bangladesh and 0.76 in Sri Lanka). Predictive power more than doubles in agricultural areas compared to the baseline model.

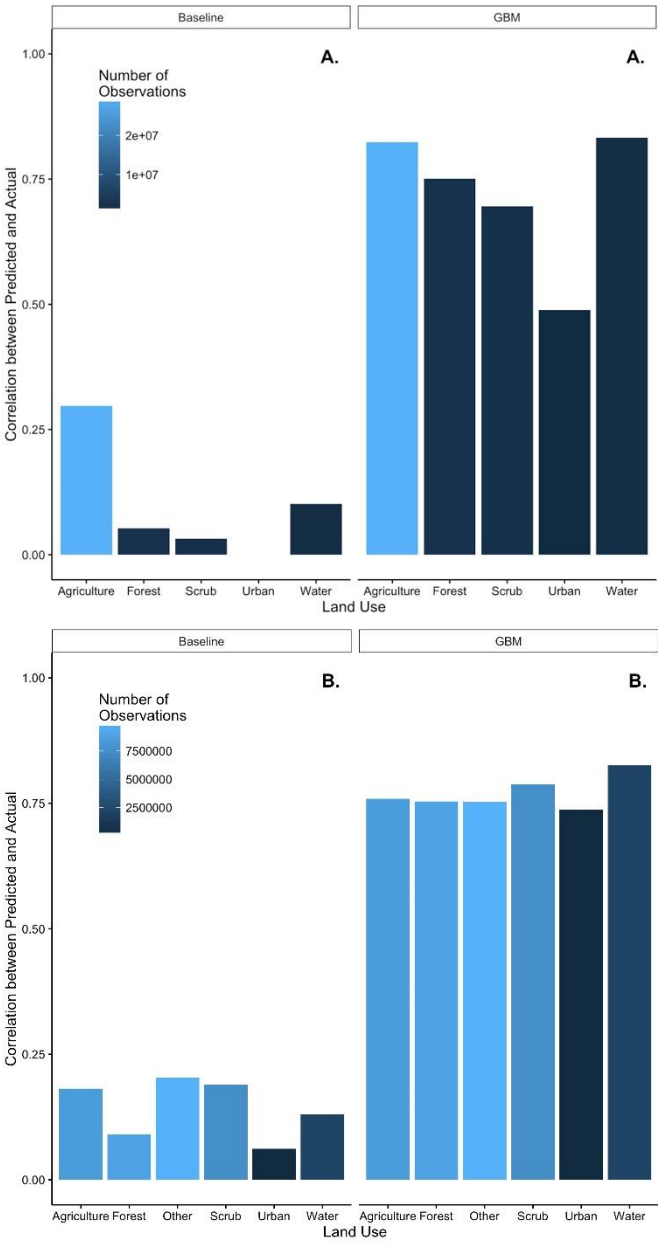



Figure 1: Correlation between predicted and actual EVI in (A) Bangladesh and (B) Sri Lanka across land use categories

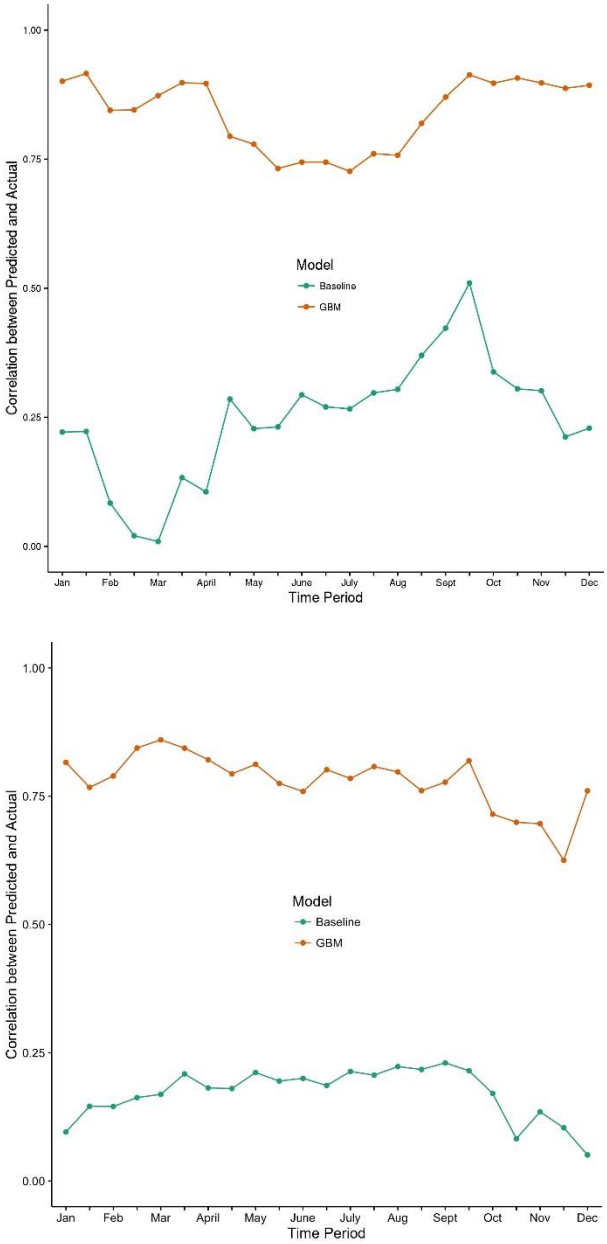



Figure 2: Correlation between predicted and actual EVI in (A) Bangladesh and (B) Sri Lanka across time periods

### 3.2 Performance across time

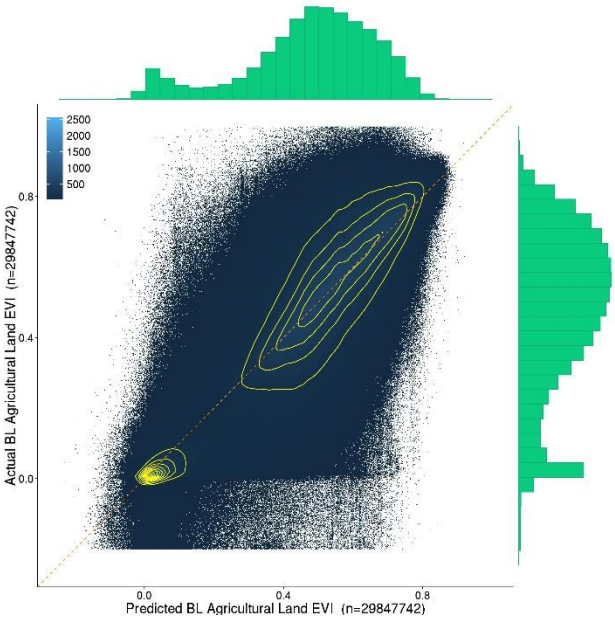
In Sri Lanka, there was variation in the performance of our model across periods of the year (Figure 2). We computed the average percent of missing data at each time period of the year and found that the drops in correlation occurred after increases in the percent of missing data. Many of the lowest drops in correlation occurred during the wet season (October – February), during which the majority of the island is covered in clouds. Similarly, in Bangladesh the largest drop in correlation between actual and predicted EVI values through time occurred during the wet season, which lasts from May to September. Even during these highly cloudy periods, correlations were generally at or above 0.75.

### 3.3 Performance across values of EVI

In Figure 3 we plot the performance for held-out agricultural pixels. The x-axis histogram displays the distribution of hold-out *predicted* agricultural EVI values, and the y-axis displays the distribution of *actual* agricultural EVI values. If our model made perfect predictions, all points in the scatter plot would line up on the dotted line. In Sri Lanka, the strongest predictions of EVI are at values indicative of healthy vegetation, between 0.5 and 0.8. Predictive performance decreases for low EVI values, which are suggestive of stressed vegetation or atmospheric noise. The low predictive performance for extreme EVI values in Sri Lanka may be due

to high levels of atmospheric noise. In Bangladesh, we see far more actual EVI values at or below zero. This is likely due to the fact that a large portion of the county located in the Ganges-Brahmaputra Delta floods seasonally, causing pixel values to drop. This seasonal flooding is often sudden and difficult to predict. In some cases, flooding is actively managed by humans. In both regions, the highest density of points (indicated by the contour lines) falls along the dotted line, suggesting that our predictive power is high.

**A.**




**B.**

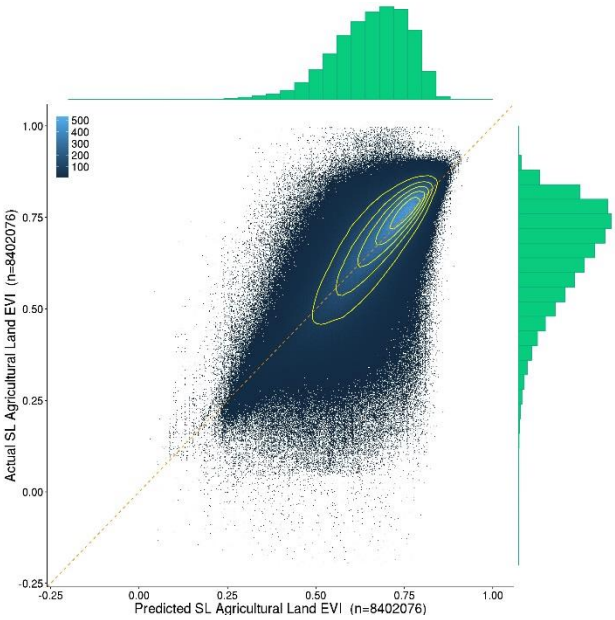



Figure 3: Correlation between predicted and actual EVI in (A) Bangladesh and (B) Sri Lanka across measured values of EVI

## CONCLUSION

In this paper, we have tested a user-friendly and open source set of scripts that download, process and predict high resolution values of vegetation health for any MODIS tile. All scripts and data are open source (http://johnjnay.com/forecastVeg/) and well-documented. The final tool can be used to capture field-level variations in vegetation health and support local and regional decision-making. We have tested the tool in two locations with known data availability issues, where cloud cover is a serious concern that prevents decision makers from using remotely sensed data to support regular decisions. We propose that this tool can be used to "nowcast" remotely sensed data in regions in which large sections of data are regularly missing from remotely sensed images. The tool can be used to monitor and predict vegetation health at a high resolution in regions in which no local data is available, where it could support agricultural decision-making.

Future research could combine our scripts with additional ancillary data to model the effects of particular social and institutional factors on vegetation health. In addition, the integration of machine learning techniques and remote sensing data could be used to predict other environmental phenomena.

## ACKNOWLEDGEMENTS

United States National Science Foundation grant EAR-1204685 funded this research.

## REFERENCES

Baffes, J., & Gautam, M. (2001). Assessing the sustainability of rice production in Bangladesh. *Food Policy*, *26*, 515–542.

Bergstra, J. S., Yamins, D., & Davis, C. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proceedings of the 30th International Conference on Machine Learning*, 115–123.

Brown, M. E., Pinzón, J. E., Didan, K., Morisette, J. T., & Tucker, C. J. (2006). Evaluation of the Consistency of Long-Term NDVI Time Series Derived From AVHRR , and Landsat ETM + Sensors. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(7), 1787–1793.

Cai, X. L., & Sharma, B. R. (2010). Integrating remote sensing, census and weather data for an assessment of rice yield, water consumption and water productivity in the Indo-Gangetic river basin. *Agricultural Water Management*, *97*(2), 309–316. doi:10.1016/j.agwat.2009.09.021

Center for International Earth Science Information Network - CIESIN - Columbia University, United Nations Food and Agriculture Programme - FAO, and Centro Internacional de Agricultura Tropical - CIAT. 2005. Gridded Population of the World, Version 3 (GPWv3): Population Count Grid. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).

http://dx.doi.org/10.7927/H4639MPP. Accessed 01 February 2015.

Dasgupta, S., Hossain, M., Huq, M., & Wheeler, D. (2014). *Climate change, soil salinity, and the economics of high yield rice production in coastal Bangladesh* (No. WPS 7140). Washington, D.C.

Dessai, S. (2009). Climate prediction: A limit to adaptation. In *Adapting to climate change: Thresholds, values, and governance* (pp. 64–78).

Galford, G. L., Mustard, J. F., Melillo, J., Gendrin, A., Cerri, C. C., & Cerri, C. E. P. (2008). Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil. *Remote Sensing of Environment*, *112*(2), 576–587. doi:10.1016/j.rse.2007.05.017

Government of Sri Lanka. (2010). *National climate change adaptation strategy for Sri Lanka - 2011 to 2016*. Colombo, Sri Lanka.

Gumma, M. K. (2011). Mapping rice areas of South Asia using MODIS multitemporal data. *Journal of Applied Remote Sensing*, *5*(1), 053547. doi:10.1117/1.3619838

H2O.ai Team. (2015). H2O Documentation. URL: http://docs.h2o.ai.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction*. Springer.

Huete, A., Didan, K., Miura, T., Rodriguez, E. ., Gao, X., & Ferreira, L. . (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1-2), 195–213. doi:10.1016/S0034-4257(02)00096-2

Ji, L., Peters, A. J. (2004). Forecasting vegetation greenness with satellite and climate data. *IEEE Geoscience and Remote Sensing Letters*, *1*(1), 3–8.

Lewis, D. (2011). *Bangladesh: Politics, economy and civil society*. Cambridge University Press.

Nay, J., Burchfield, E., & Gilligan, J. (2016). Forecasting vegetation health at high spatial resolution. eprint arXiv:1602.06335.

Peters, A. J., Walters hea, E. A., Ji, L., Vliia, A., Hayes, M., Svoboda, M. D., & Nir, R. E. D. (2002). Drought Monitoring with NDVI-Based Standardized Vegetation Index. *Photogrammatic Engineering & Remote Sensing*, *68*(1), 71–75.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., … Kaplan, A. (2003). Global analysis of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, *108*(D14), 4407.

Rhee, J., Im, J., & Carbone, G. J. (2010). Monitoring agricultural drought for arid and humid regions using multi-sensor remote sensing data. *Remote Sensing of Environment*, *114*(12), 2875–2887. doi:10.1016/j.rse.2010.07.005

Sakamoto, T., Yokozawa, M., Toritani, H., Shibayama, M., Ishitsuka, N., & Ohno, H. (2005). A crop phenology detection method using time-series MODIS data. *Remote Sensing of Environment*, *96*(3-4), 366–374. doi:10.1016/j.rse.2005.03.008

Senaratne, A., & Scarborough, H. (2011). Coping with climate variability by rain-fed farmers in Dry Zone, Sri Lanka: Towards understanding adaptation to climate change. In *AARES: Australian Agricultural & Resource Economics Society 55th Annual Conference Handbook* (pp. 1–22). Melbourne, Australia.

Small, C., & Milesi, C. (2013). Multi-scale standardized spectral mixture models. *Remote Sensing of Environment*, *136*, 442–454. doi:10.1016/j.rse.2013.05.024

Survey Department of Sri Lanka. (2011). Land use map of Sri Lanka.

Thenkabail, P., Gamage, M., & Smakhtin, V. (2004). *The use of remote sensing data for drought assessment and monitoring in Southwest Asia*. Colombo, Sri Lanka. Retrieved from http://books.google.com/books?hl=en&lr=&id=BiG6G4am-WEC&oi=fnd&pg=PR5&dq=The+Use+of+Remote+Sensing+Data+for+Drought+Assessment+and+Monitoring+in+Southwest+Asia&ots=FJMokpUD2N&sig=Sax4TznnZClRJdZ_N9lCu6SaHQw

Withananachchi, S. S., Kopke, S., Withanachchi, C. R., Pathiranage, R., & Ploeger, A. (2014). Water resource management in dry zonal paddy cultivation in Mahaweli River Basin, Sri Lanka: An analysis of spatial and temporal climate change impacts and traditional knowledge. *Climate*, *2*(4), 329–354.

Xiao, X., Boles, S., Frolking, S., Li, C., Babu, J. Y., Salas, W., & Moore, B. (2006). Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images. *Remote Sensing of Environment*, *100*(1), 95–113. doi:10.1016/j.rse.2005.10.004

Ziervogel, G., Johnston, P., Matthew, M., & Mukheibir, P. (2010). Using climate information for supporting climate change adaptation in water resource management in South Africa. *Climatic Change*, *103*(3-4), 537–554.