



A machine-learning approach to forecasting remotely sensed vegetation health

John Nay, Emily Burchfield & Jonathan Gilligan

To cite this article: John Nay, Emily Burchfield & Jonathan Gilligan (2018) A machine-learning approach to forecasting remotely sensed vegetation health, International Journal of Remote Sensing, 39:6, 1800-1816, DOI: [10.1080/01431161.2017.1410296](https://doi.org/10.1080/01431161.2017.1410296)

To link to this article: <https://doi.org/10.1080/01431161.2017.1410296>



© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 18 Dec 2017.



Submit your article to this journal [↗](#)



Article views: 1704



View Crossmark data [↗](#)



A machine-learning approach to forecasting remotely sensed vegetation health

John Nay^{a,b,c}, Emily Burchfield^d and Jonathan Gilligan^e

^aInformation Law Institute, New York University, New York, NY, USA; ^bBerkman Klein Center, Harvard University, Cambridge, MA, USA; ^cSkopos Labs, Inc., New York, NY, USA; ^dDepartment of Environment and Society, Quinney College of Natural Resources, Logan, UT, USA; ^eDepartment of Earth & Environmental Sciences, Vanderbilt University, Nashville, TN, USA

ABSTRACT

Drought threatens food and water security around the world, and this threat is likely to become more severe under climate change. High-resolution predictive information can help farmers, water managers, and others to manage the effects of drought. We have created an open-source tool to produce short-term forecasts of vegetation health at high spatial resolution, using data that are global in coverage. The tool automates downloading and processing Moderate Resolution Imaging Spectroradiometer (MODIS) data sets and training gradient-boosted machine models on hundreds of millions of observations to predict future values of the enhanced vegetation index. We compared the predictive power of different sets of variables (MODIS surface reflectance data and Level-3 MODIS products) in two regions with distinct agro-ecological systems, climates, and cloud coverage: Sri Lanka and California. Performance in California is higher because of more cloud-free days and less missing data. In both regions, the correlation between the actual and model predicted vegetation health values in agricultural areas is above 0.75. Predictive power more than doubles in agricultural areas compared to a baseline model.



ARTICLE HISTORY


Received 20 June 2017

Accepted 16 November 2017

1. Introduction

Drought significantly reduces agricultural production, destabilizing food systems and threatening food security (Lesk, Rowhani, and Ramankutty 2016). 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 spatiotemporal variations in agricultural responses to drought (Peters et al. 2002; Rhee, Im, and 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 et al. 2009; Ziervogel et al. 2010). These indices can be used to identify vulnerable agricultural systems, to understand past agricultural responses to drought, and to guide efforts to increase resilience to future drought.

CONTACT John Nay  jn1886@nyu.edu  New York University, New York, NY, USA; Berkman Klein Center, Harvard University, Cambridge, MA, USA; Skopos Labs, Inc., New York, NY, USA

 The code for full replication of data download and processing and modelling is available online at <https://github.com/JohnNay/forecastVeg>.

© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Agricultural systems often exhibit non-linear responses to sudden changes in water availability or human activity. However, many agricultural prediction tools rely on linear models to predict future vegetation health (Asoka and Vimal 2015; Bolton and Friedl 2013; Doraiswamy et al. 2014; Peters et al. 2002). Although more complex, non-linear models have been used to predict rainfall in agricultural systems (Chattopadhyay and Chattopadhyay 2008; Singh and Borah 2013), metrics of agricultural drought such as vegetation health better capture changes in farmer livelihoods than the coarse resolution meteorological metrics of drought used in these studies. Coarse resolution models are not able to examine fine-grained intra-system dynamics and justify resource transfers. Higher resolution models tend to rely on data sets only available in data-rich regions of the world (Bolton and Friedl 2013; Kogan, Salazar, and Roytman 2012). Furthermore, data scarce regions tend to lack the economic resources required to buffer against the effects of drought.

Our objective was to create a user-friendly predictive software tool that will increase the capacity of data-scarce agricultural systems to prepare for and respond to drought in the future. We have created a tool that predicts future vegetation health values at a high spatial resolution using open source tools and data that are global in coverage. All scripts and documentation can be downloaded from <https://github.com/JohnNay/forecastVeg>. With simple user inputs, our software downloads, processes, models, and forecasts vegetation health at 16 days intervals at a 250 m resolution anywhere in the world. The tool applies a gradient-boosted machine model to Moderate Resolution Imaging Spectroradiometer (MODIS) data sets openly available on National Aeronautic and Space Administration's (NASA) Land Processes Distributed Active Archive Center server. The model learns potentially complex relationships between past remotely sensed variables (and their interactions) and future vegetation health as measured by the EVI.

We used state-of-the-art machine-learning techniques for training models and measuring performance of those models. We trained gradient boosted machine (GBM) models, which combine gradient-based optimization and boosting, which improves an ensemble of weaker base models by adjusting the training data. The base models are trees that divide predictor variable values into distinct regions by choosing variables to make binary splits on and the threshold values of those variables where the split should be made. An important criterion for our modelling algorithms was automatic handling of missing predictor variable values because remotely sensed data sets often have many missing values.

In this article, we apply the tool in two locations: Sri Lanka and California. We selected these regions based on their distinct agro-ecological systems, climates, and levels of cloud cover. We compared the predictive performance of the model using past values of MODIS surface reflectance data (MOD09A1) and Level 3 MODIS products (MOD11A2, MOD13Q1, MOD15A2, MOD17A2) as predictor variables.

2. Materials and methods

We designed an experiment across location and data dimensions to assess how well our process performs under different conditions. Table 1 illustrates the experimental design of our analyses and the hypothesized relative performance of each model. In terms of location, we hypothesized that within each data category, the model would perform

Table 1. Experimental design and anticipated performance.

Location/data	Lagged EVI	Land use and time	Level 3	Spectral
CA	4A	3A	2A	1A
SL	4B	3B	2B	1B

EVI: enhanced vegetation index.
See [Section 2.1](#) for location description and [Section 2.2](#) for data details.

Table 2. Description of the data sets used in the predictor sets.

	MODIS product	Layer	Description
Spectral model	MOD09A1.005	B1_lag	Lag of MOD09 band 1, 620–670 nm
		B2_lag	Lag of MOD09 band 2, 841–876 nm
		B3_lag	Lag of MOD09 band 3, 459–479 nm
		B4_lag	Lag of MOD09 band 4, 545–565 nm
		B5_lag	Lag of MOD09 band 5, 1230–1250 nm
		B6_lag	Lag of MOD09 band 6, 1628–1652 nm
		B7_lag	Lag of MOD09 band 7, 2105–2155 nm
Level 3 model	MOD11A2.005	LST_Day_1km_lag	Lag of daytime land surface temperature
		QC_Day_lag	Lag of quality control for daytime LST
	MOD13Q1.005	EVI_lag	Lag of enhanced vegetation index
		NDVI_lag	Lag of normalized difference vegetation index
		VI_Quality_lag	Lag of quality control for vegetation indices
	MOD15A2.005	Fpar_1km_lag	Lag of fraction of photosynthetically active radiation
		Lai_1km_lag	Lag of leaf area index
		Fpar_Lai_QC_lag	Lag of quality control for FPAR and LAI
	MOD17A2.005	GPP_lag	Lag of gross primary productivity
		PSN_lag	Lag of net photosynthesis
Level 3 and spectral	Ancillary data	Land_use	SL Survey Department, National Land Cover Database
		nino_lag	Lag of El Niño sea surface temperature index
		GWP_lag	Lag of population

better in California, where there are fewer clouds than in tropical Sri Lanka. [Table 2](#) describes the predictor variables used in each model. We anticipated that the MODIS surface reflectance data (MOD09A1) would predict vegetation health better than Level 3 MODIS data products (land surface temperature, leaf area index, etc.), which are derived from the spectral data, because the flexible models will, in effect, learn intermediate representations of the underlying data that are more suited to predicting future EVI values than the NASA-derived representations of that same underlying spectral data. From a machine-learning perspective, the Level 3 products are part of a feature engineering process orthogonal to the learning task of mapping spectral data to future EVI. We hypothesized that models with only lagged EVI as a predictor will have the lowest performance because all the other predictor sets are multivariate supersets, containing the underlying data from which lagged EVI is computed and more. If the additional variables add little predictive power, we anticipated that the model would learn to ignore them. We included this univariate lagged EVI model to measure relative prediction error reductions associated with land use and time, Level 3, and spectral data. Similarly, because land use and time are included in both the Level 3 and spectral models, we tested models including lagged EVI, land-use classification, and time of the year.

Specifically, we hypothesized that within locations, the performance of the predictor data sets from highest predictive power to lowest predictive power would be Spectral, Level 3, Land Use and Time, lagged EVI (the numbers in the cells of [Table 1](#)). We also

hypothesized that across locations, on average, performance would be higher in California, i.e. the mean of 1A–4A would be greater than the mean of 1B–4B.

2.1. Experimental variable: location

We selected two regions with distinct agro-ecologies, climates, and data availability: Sri Lanka and the San Joaquin Valley in California. Sri Lanka is a small island nation located off of the eastern coast of India that covers approximately 66,000 km² and is home to nearly 21 million people (Government of Sri Lanka 2010). The country receives rainfall during two monsoon periods. The northeast monsoon lasts from October to December and brings two-thirds of annual rainfall to Sri Lanka. The southwest monsoon lasts from May to October and brings rain primarily to the southwestern region of the island. This rainfall pattern divides the island into wet and dry zones and creates two distinct cultivation seasons, the wet Maha season and the dry Yala season (Samad 2005; Senaratne and Scarborough 2011). During the wet season, most farmers cultivate rice. Rice is a staple of the Sri Lankan diet and an estimated 30% of the total labour force is involved in rice production (Mahaweli Authority of Sri Lanka 2012). Farmers capture wet season rainfall in reservoirs and cultivate rice during the dry season with stored water. During water scarce dry seasons, farmers cultivate other field crops such as soy, maize, and grain. Increasing numbers of dry zone farmers pump groundwater to irrigate other field crops (Weligamage et al. 2003). Field size is small in Sri Lanka, with over 70% of farmers cultivating less than 2.5 ac of land (Withananachchi et al. 2014). Persistent cloud cover year-round significantly reduces remotely sensed data availability.

The San Joaquin Valley in California covers approximately 40,000 km² and is home to over 1.6 million people (California Department of Water Resources 2013). This valley is one of the most productive agricultural systems in the world, with an annual gross production of more than 25 thousand million dollars (Environmental Protection Agency 2011–14). The average farm size is 162 ac, significantly larger than the small plots held by Sri Lankan farmers (California Department of Water Resources 2013). The primary crops cultivated in the area are grapes, walnuts, almonds, and cherries (California Department of Water Resources 2013). As in Sri Lanka, many of the agricultural fields in the valley receive water from surface water irrigation systems. Heavy groundwater pumping also provides a significant amount of agricultural water in the region (California Department of Water Resources 2013). The climate in the valley is Mediterranean, with moderate temperatures throughout the year. Cloud cover is significantly lower than in tropical Sri Lanka.

These two regions were selected for the following reasons. First, in both regions, irrigation infrastructures allow decision makers to move large amounts of water over considerable distances. Decision makers may have the capacity to respond to our predictions by moving water to areas we predict to have relatively low vegetation health. Second, the differences in agricultural field size and crops cultivated test the performance of our models in regions with markedly different agro-ecological systems. Finally, by comparing model performance in the cloudy tropics and relatively cloud-free California, we can analyse the effect of data availability over a fixed-time interval (11 years) on predictive performance.

2.2. Experimental variable: data type

Remotely sensed measures of vegetation conditions have been used in many studies to monitor the agricultural effects of drought (Brown et al. 2002; Ji and Peters 2004; Thenkabail, Gamage, and Smakhtin 2004). We measured these effects using the EVI which is a proxy for the health of agricultural crops (Galford et al. 2008; Gumma 2011; Sakamoto et al. 2005), highly correlated with the leaf area index (Sakamoto et al. 2005; Huete et al. 2002) and positively linearly related to vegetation fraction estimates (Small and Milesi 2013). The EVI is measured as

$$EVI = G \frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + C_1 \times \rho_{red} - C_2 \times \rho_{blue} + L} \quad (1)$$

where ρ_{NIR} , ρ_{red} , and ρ_{blue} are the atmospherically corrected surface reflectances or near-infrared, red, and blue light, respectively; L is the canopy background adjustment; and C_1 and C_2 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.

For predicting EVI, our analysis compares the performance of four sets of predictor variables

- (1) land use, time period, the value of EVI from the last time period, and spectral data from the previous time period,
- (2) land use, time period, the value of EVI from the last time period, and Level-3 MODIS products (land surface temperature, NDVI, leaf area index, the fraction of photosynthetically active radiation, net photosynthesis, and gross primary productivity) from the previous time period,
- (3) land use, time period, and the value of EVI from the previous time period, and
- (4) the value of EVI from the previous time period.

We included the third and fourth options because simple univariate models leveraging past values of a variable are often effective in forecasting future values of the same variable, especially if those values are adjusted for the time period (seasonal effects). For options 1 and 2, we also included the lagged population and El Niño sea surface temperature index.

2.3. Data

We downloaded and processed 11 years of remotely sensed imagery (2004–2014). We combined these data with ancillary data sets and reshaped it into a single matrix where each row corresponds to a pixel at one time and each column is a measured variable. We divided the observations into Training Data 1 (80% of the pixels) and Testing Data 1 (20% of the pixels) by sampling from large spatial grid indices without replacement (Figure 1). An 80/20 split of the data into training and testing is a common proportion split of the data used in machine-learning studies to ensure there are enough training data to train a model effectively and still enough testing data to estimate out-of-sample performance. Each cell has equal probability of being

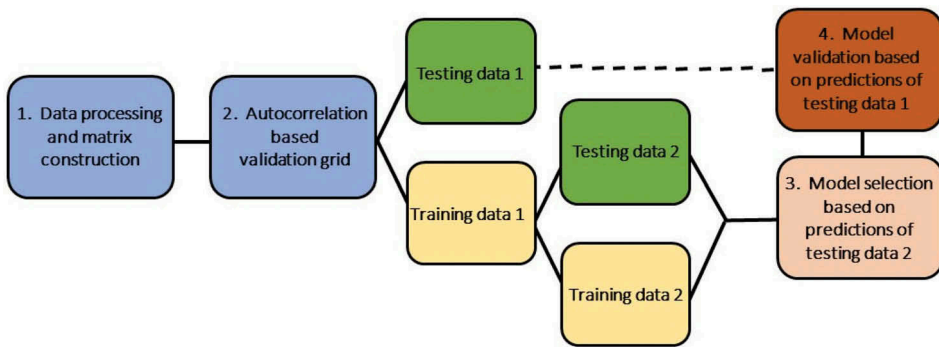


Figure 1. Methods overview.

selected. This was done to increase the chance that our testing of the approach was on a representative data set. We then divided Training Data 1 into Training Data 2 (80% of the pixels) and Testing Data 2 (20% of the pixels) with the same spatial sampling process and trained multiple models on Training Data 2, varying the hyper-parameters for each model estimation. We used Testing Data 2 to assess the performance of each model's predictions. We repeated this loop of learning on Training Data 2 and testing on Testing Data 2 for each of the four different data types and chose the combination of data type and hyper-parameter setting that achieved the highest performance in predicting Testing Data 2. Finally, we validated the best-performing model from the previous step by testing its performance on the held-out data in Testing Data 1. We repeated this entire process separately for Sri Lanka and California. This process is summarized in Figure 1 and detailed in the next subsections.

2.4. Data processing and matrix construction

We automated downloading and processing MODIS data from the MOD09A1, MOD11A2, MOD13Q1, MOD15A2, and MOD17A2 data sets. Our software is open source and can be used on any MODIS data set found on NASA's LP DAAC server, for any region of the world in which MODIS data are collected. The required user inputs include the MODIS tiles for the region of interest, the first and last download dates, and the path to a reference image. The reference image stores information about the desired projection and extent final data set. Users unfamiliar with Python can create the reference image using the MODIS Reprojection Tool or user-friendly software such as ArcGIS. The user has the option of including ancillary geospatial data sets such as land-use information, socioeconomic data, or climate data. The user only needs to add spatial features in the same geographical format as the MODIS data and then our software automatically creates the matrix format needed for the machine-learning models. All data sets are resampled to the resolution of the outcome variable (i.e. 250 m EVI) to allow the machine-learning models to train on a standard matrix format. For our analysis, we included gridded world population (Center for International Earth Science Information Network 2015), land use (Survey Department of Sri Lanka 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 the Niño 4 SST Index in California.

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 data set's quality mask. In both locations, particularly in Sri Lanka, this created a large amount of missing data. Eight-day data sets are transformed to a 16 days time step by computing the average of two quality-masked 8 days pixels. All data sets are resampled to match the spatial resolution of the EVI data set (250 m). The software reshapes the stack of images for each data set into a single column and stacks columns to create a two-dimensional matrix with dimensions' pixel-time by number of variables. The software also creates columns describing the time period of each observation (dividing the year into 16 days periods), the latitude and longitude of each pixel, and the pixel's location in the autocorrelation-based validation grid, which is described in the next subsection.

2.5. Autocorrelation-based validation grid

We computed the spatial autocorrelation functions of the MOD13Q1 imagery to divide the final matrix into a grid of independent areas. In the case of both the San Joaquin Valley and Sri Lanka, the autocorrelation functions approached zero at a lag of 150 pixels (approximately 35 km). We constructed a grid of 150 pixel-by-150 pixel cells, each with a unique identifier. A random subset of these cells was selected as training data and the remainder were used as testing data. This reduces spatial autocorrelation between our testing and training data sets to allow performance of the model on the testing data to estimate how well the model will predict new data that is collected after a model is trained.

2.6. Model training and selection

We selected a model type that has consistently performed well in supervised learning tasks with large amounts of training data where potentially complex functions link the predictor and outcome variables: the GBM. To contextualize quantitative performance measures of our model (correlation and mean-squared error between vectors of predicted and actual EVI), we compared them to a baseline model that serves as a proxy for potentially currently available forecasting undertaken by local residents. Ideally, our baseline model would be a univariate time series model fit to the training data which uses past values of EVI in the hold-out data to forecast future EVI, a standard model for time series forecasting in the environmental sciences, but due to very large amounts of missing EVI in Sri Lanka, this was not feasible. For the same reason, we also cannot use the 'climatological mean' for a given pixel. There are often gaps between observed values of EVI for many consecutive time periods due to cloud contamination. To approximate the desired baseline model, we created a simple model that uses approximate nearest neighbour search to search for k pixel-time observations approximately closest in space and time in the hold-out data (with the condition that the time is in the past) and averages their values of vegetation health to predict the hold-out data EVI. If the search does not return any neighbours because no neighbours without missing EVI

data can be found within the k results, the algorithm uses the average of all EVI values up to that point in time as the prediction.

We used a GBM implementation in H2O, an open-source library of parallelized machine-learning algorithms that use compression techniques that allowed us to hold hundreds of millions of rows of data in memory (H2O.ai 2015). We were able to fit large models on large data (our data matrices are often larger than 60 GB) much more efficiently than most widely used machine-learning libraries such as the scikit-learn Python module. However, there are still significant requirements for RAM (at least 100 GB) because the data are preprocessed with the entire data matrix in memory. This memory requirement could potentially be relaxed through the use of memory-mapped files, but at the cost of a reduction in speed.

The GBM combines gradient-based optimization, which iteratively adjusts model parameters in the direction of lower training data prediction errors by using gradient computations, and boosting, which improves an ensemble of weaker base models by adjusting the training data. The base models are trees that divide predictor variable values into distinct regions by choosing variables to make binary splits on and the threshold values of those variables where the split should be made (Hastie, Tibshirani, and Friedman 2009). An important desideratum for our modelling algorithms was automatic handling of missing predictor variable values. Remotely sensed data sets used to detect vegetation health often have many missing values due to cloud cover. The GBM can handle missing predictor variables by incorporating them in the overall tree structure by always moving missing values to the left at splits in the trees. Furthermore, the model does not rely on one-hot-encoding of categorical variables, so our time and land-use factor variables, which have many levels, are handled efficiently.

Using trees that have multiple splits on different predictor variables allowed the model to automatically learn higher order interactions between predictor variables. Although our largest models only had slightly more than 10 predictor variables, interactions between variables, e.g. lagged EVI and lagged Band 7, may improve predictive power. The level of interaction to which the model may search depends partly on a hyper-parameter that we tuned on the training data (see next paragraph). If we were using a linear regression model, interactions between variables would need to be specified manually; however, manually specifying all such potential interactions would be prohibitively time intensive. Furthermore, the exact interactions that lead to the best predictive performance likely vary by location and thus would need to be specified by local experts each time the model was applied to a new location. The GBM algorithm implicitly automatically tries many interactions and learns which are useful from the data.

There are three important hyper-parameters for the GBM that need to be set for the model to be estimated. They can affect overall model complexity and thus whether the model overfits training data or generalizes well to new data, but their best values depend on the nature of the data and the prediction task and can rarely be effectively determined *a priori*. It is common to conduct an exhaustive grid-search over the entire (suitably discretized) hyper-parameter space. In fact, this is the only automated option available in most statistical software. However, this may be too slow for data this large unless the user has access to a large cluster of powerful computers. It can be more efficient to randomly sample hyper-parameters. Given

some prior distributions that cover all reasonable values of the hyper-parameters, we use a Tree of Parzen Estimators search algorithm (Bergstra, Yamins, and Davis 2013) to search through the hyper-parameter space and record the mean-squared error of the model's predictions of Testing Data 2 for each sampled set of hyper-parameters. This algorithm works by specifying the number of samples desired, e.g. $s = 50$, and the distributions for each of the hyper-parameters. Then, it generates a set of s samples which attempts to explore as much of the hyper-parameter space as possible for the given number of samples and we use these generated hyper-parameter sets to estimate s models and test them on the tuning set. This automates the entire model building process. The user is not required to specify anything other than the location in the world, after which the scripts download the data and train a model specific to that location.

Our model selection process involved selecting (1) the hyper-parameter values for the model and (2) the set of predictor variables (spectral data, Level 3 Products, land use and time period, or lagged EVI). By 'model selection,' we mean a specification of these two components.

2.7. Model validation

We trained models on Training Data 2 and selected the model that performed the best on Testing Data 2. Then, we trained the model with those hyper-parameter settings and data type on the full training data, Training Data 1. Finally, we used this model to forecast all the 16 day-ahead values of EVI in Testing Data 1, the hold-out data. We used a flexible model that can learn complex relationships, such as the interactions discussed above. However, if the model is not tested on data separate from the data it was trained on, there is a risk that the model may have learned structure that is unique to the training data and not generalizable to the ultimate task of predicting EVI for new observations. Although we only used Testing Data 2 for tuning the three hyper-parameters of the GBM and selecting which set of predictor variables is most effective, there was still a risk that we may have overfit Training Data 1 (Training Data 2 and Testing Data 2) and learned characteristics of the noise in this data in addition to the characteristics of the signal. Therefore, to test our best model on fresh, unseen examples, the model predicted the observations in Testing Data 1, which was only used for this purpose.

3. Results

3.1. Model performance on testing data 2: model selection

Figure 2 plots the percentage reduction in mean squared error (MSE) below the MSE of the GBM model using only lagged EVI. We plot the results for the best hyper-parameter setting found for each experiment after using the same model construction and selection scripts for the eight possibilities (four data types and two locations). In both locations, when we included additional data sets, the model learned useful relationships between these data sets and future EVI and error dropped compared to the simple lagged EVI model. All three data types in the plot also used lagged EVI as a predictor

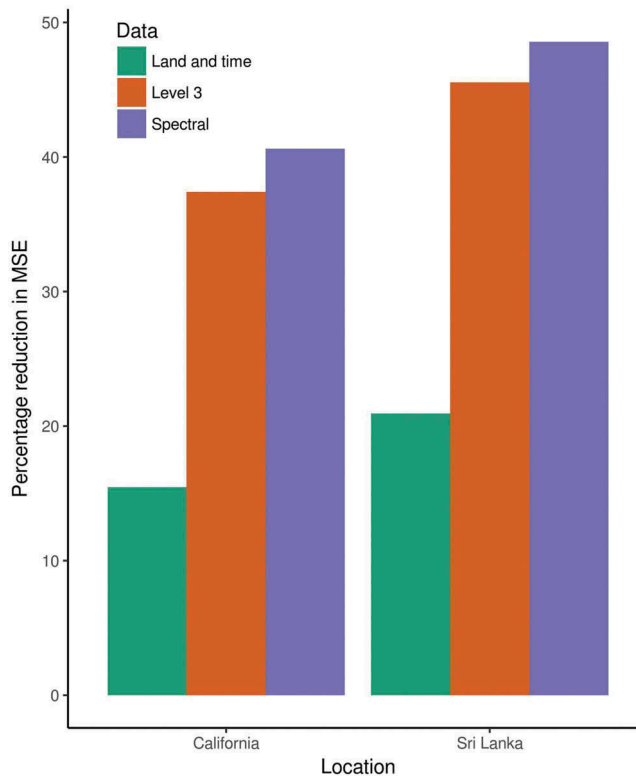


Figure 2. Model performance for each data type in California and Sri Lanka as measured by the percentage reduction in mean squared error below the lagged EVI model for each location.

variable, and the Level 3 and spectral data types used land use and time as predictors, allowing us to determine the relative importance of adding additional data.

Although the absolute performance of models varies across locations with different levels of data availability and agro-ecology (which we explore in the next section), in both locations, the magnitude of error reduction between the predictor variable sets is similar. For instance, error is reduced by between 40% and 50% when moving from only lagged EVI to lagged EVI and spectral data. Overall, these results accord with what we anticipated (see Table 1): the predictor data ordered by performance is spectral, Level 3, land use and time, lagged EVI, and performance is higher in California.

We used these results to select the spectral data for both locations and estimated the model with the chosen best performing set of hyper-parameter values on the full Training Data 1. Finally, we used the estimated models to make predictions on the held-out data in both locations to validate our model and compare to a baseline.

3.2. Model performance on testing data 1: hold-out validation

3.2.1. 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 3). Performance in California is higher because of more cloud-free days and less missing data. In both regions, the correlation in agricultural areas is above 0.75 (0.86 in California and 0.76 in Sri Lanka). Predictive power more than doubles in agricultural areas compared to the baseline model.

3.2.2. Performance across values of true measured EVI

In Figure 4, 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.

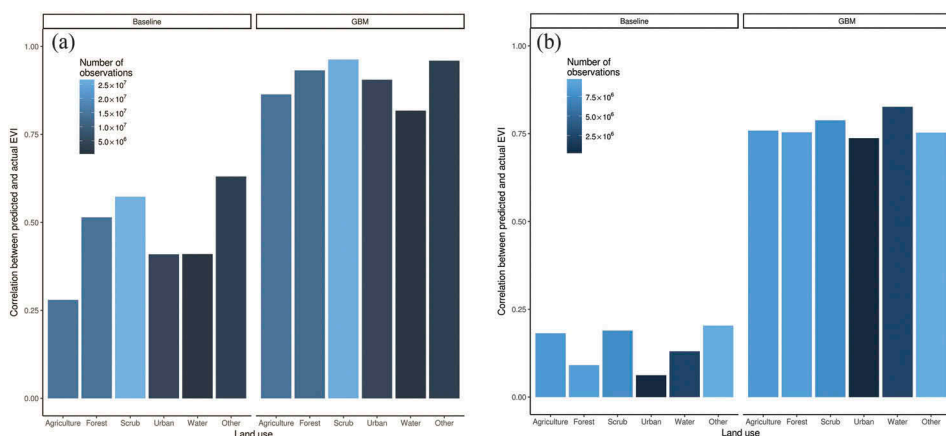


Figure 3. Correlation between predicted and actual EVI in California ((a) $n = 61,681,296$) and Sri Lanka ((b) $n = 36,831,863$).

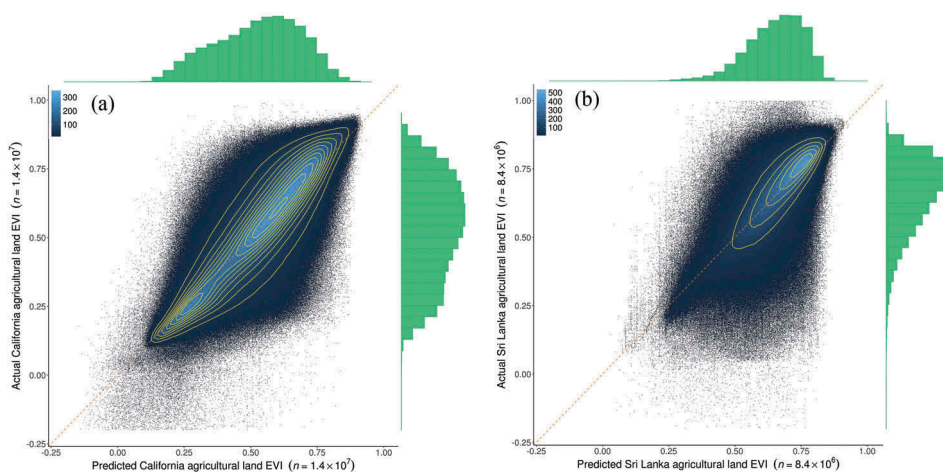


Figure 4. Performance across values of true measured EVI in California agricultural land ((a) $n = 14,414,402$) and in Sri Lanka agricultural land ((b) $n = 8402,076$).

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 California, the drop in performance for low EVI values is very slight.

3.2.3. *Performance across time*

In Sri Lanka, there was variation in the performance of our model across periods of the year (Figure 5). We plotted the average % of missing data at each time period of the year (Figure 6) and found that the drops in correlation occurred after increases in the percentage of missing data. Many of the lowest drops in correlation occurred during the Maha wet season (October–February), during which the majority of the island was covered in clouds. In California, the performance of the model is consistently high across land-use categories and time periods. Periods of lower correlation occur during the winter, when there is also the highest extent of masked data.

4. Discussion

Agricultural communities around the world are experiencing increased climate unpredictability. Scientists have built models to monitor, predict, and explore potential changes in natural and social systems and inform decision makers (Nay et al. 2014). Previous research has explored the application of machine learning to predicting and monitoring drought (Park et al. 2016; Shukla, Funk, and Hoell 2014; Funk et al. 2014; Meroni et al. 2014; Rogan et al. 2008), but many of these models fall short in one of four ways. First, many models rely on proprietary software or data and fail to publish fully reproducible results and software. Second, high-resolution analyses are often only undertaken in specific regions due to data constraints. Third, few analyses are global in coverage. Finally, many existing analyses focus on describing and explaining processes rather than forecasting. Models that do forecast are not often rigorously tested out of sample on held-out data. We have addressed these shortcomings in this article by designing and testing a user-friendly set of open-source scripts that download, process, and predict high-resolution values of vegetation health for any MODIS tile.

Although the scripts were designed for the prediction of EVI, they can be used in a number of ways. The data download and processing scripts that generate the input for the GBM model allow users to create large spatiotemporal data cubes of any MODIS data set with a simple one-line command. These data sets can be used to explore past trends in vegetation, investigate the effects of environmental stressors such as droughts and floods on vegetation health, and monitor inequalities in water access across space and time. The option to include high-resolution local ancillary data sets could significantly increase the predictive power of the models. 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 supervised machine-learning techniques and remote sensing could be used to model human–environmental interactions and predict other environmental phenomena.

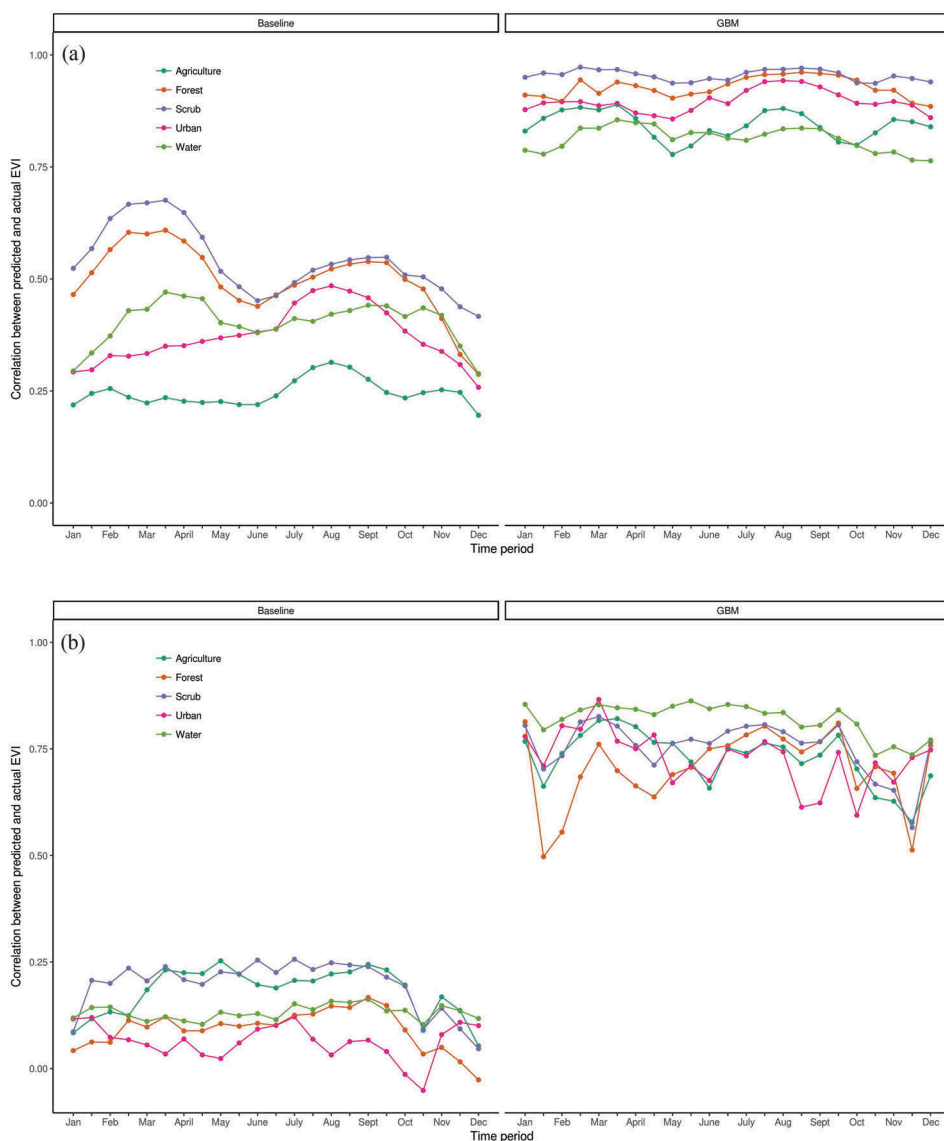


Figure 5. Correlation between predicted and actual EVI over time periods in California (a) and Sri Lanka (b). The periods of lower correlation follow periods with high levels of masked data (see Figure 6).

5. Conclusions

The machine-learning model used, the GBM, effectively handled the missing data and was able to obtain sufficient out-of-sample predictive power in both locations. The mean-squared difference between the actual and model predicted EVI is reduced by between 40% and 50% when moving from using only lagged EVI to using lagged EVI plus spectral data; therefore, in both locations, adding the spectral bands increases prediction power. To contextualize quantitative performance measures of our model

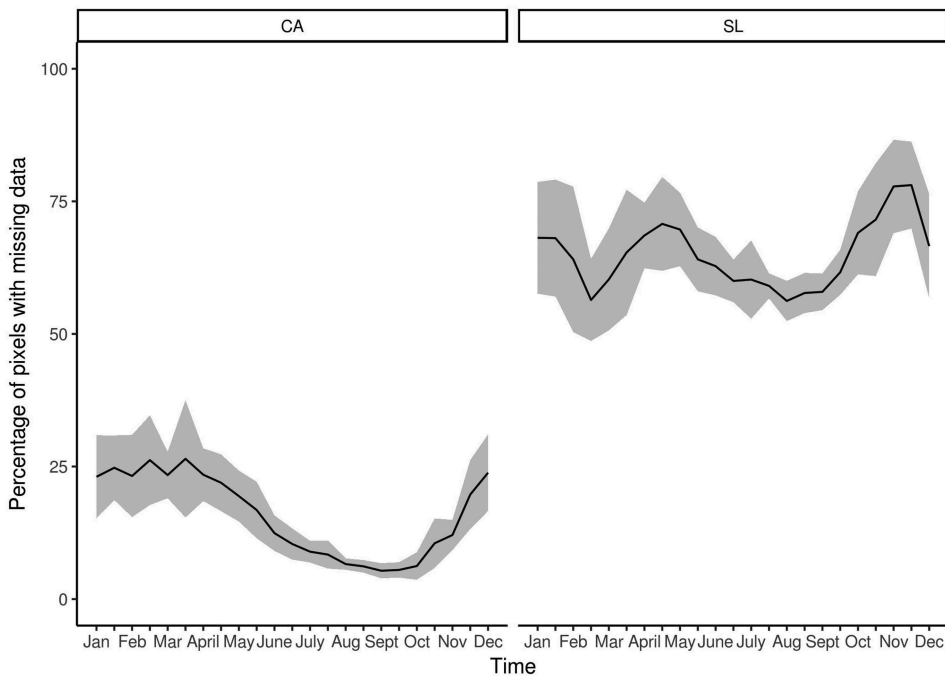


Figure 6. Percentage of pixels with missing data over twenty-three 16 days periods of the year.

(correlation and mean-squared error between vectors of predicted and actual EVI), we compared them to a baseline model that serves as a proxy for potentially currently available forecasting undertaken by local residents. The baseline was a simple model that used approximate nearest neighbour search to search for k pixel-time observations approximately closest in space and time in the hold-out data (with the condition that the time is in the past) and averaged their values of vegetation health to predict the hold-out data EVI. We measured the performance of the model primarily by calculating the correlation between the vector of 16 days ahead predictions of EVI and vector of actual values of EVI in the held-out data. Computing the correlation for each land-use category, we found that model performance relative to the baseline is high in all categories of land use and that performance in California is higher because of more cloud-free days and less missing data. In both regions, the correlation in agricultural areas is above 0.75. Predictive power more than doubles in agricultural areas compared to the baseline model.

Our tool makes predictions at a 250 m resolution, which captures field-level variations in vegetation health and may support local and regional decision-making. All scripts and data are freely available (hosted at <https://github.com/JohnNay/forecastVeg>), well-documented (see the webpage for step-by-step instructions for downloading the free data and modelling it). The tools we have constructed can be applied to any region in which MODIS data are collected. While this tool is best suited for regions with low cloud cover, it performed well in one of the cloudiest regions of the world (Sri Lanka). Finally, our model is tested on held-out data which increases the likelihood that it will perform well in practice and has high predictive power across land-use categories and throughout

time periods. The tool can be used to monitor and predict vegetation health at a high resolution in regions in which no local data are available, where it could inform agricultural decision-making.

Acknowledgement

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

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

United States National Science Foundation [grant number: EAR-1204685] funded this research.

References

- Asoka, A., and M. Vimal. 2015. "Prediction of Vegetation Anomalies to Improve Food Security and Water Management in India." *Geophysical Research Letters*. 42 (13): 5290–5298. doi:[10.1002/2015GL063991](https://doi.org/10.1002/2015GL063991).
- Bergstra, J. S., D. Yamins, and C. Davis. 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, Atlanta, Georgia, USA, 115–123.
- Bolton, D. K., and M. A. Friedl. 2013. "Forecasting Crop Yield Using Remotely Sensed Vegetation Indices and Crop Phenology Metrics." *Agricultural and Forest Meteorology* 173: 74–84. doi:[10.1016/j.agrformet.2013.01.007](https://doi.org/10.1016/j.agrformet.2013.01.007).
- Brown, J. F., B. C. Reed, M. J. Hayes, D. A. Wilhite, and K. Hubbard. 2002. "A Prototype Drought Monitoring System Integrating Climate and Satellite Data". PECORA 15/Land Satellite Information IV/ISPRS Commission I/FIEOS 2002 Conference Proceedings.
- California Department of Water Resources. 2013. *California's Groundwater Update 2013*. Sacramento, CA: Natural Resources Agency, Department of Water Resources.
- 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). doi:[10.7927/H4639MPP](https://doi.org/10.7927/H4639MPP). Accessed 01 February 2015.
- Chattopadhyay, S., and G. Chattopadhyay. 2008. "Comparative Study among Different Neural Net Learning Algorithms Applied to Rainfall Time Series." *Meteorological Applications* 15 (2): 273–280. doi:[10.1002/\(ISSN\)1469-8080](https://doi.org/10.1002/(ISSN)1469-8080).
- Dessai, S., M. Hulme, R. Lempert, and R. Jr. Pielke. 2009. Climate prediction: a limit to adaptation. Adapting to climate change: thresholds, values, governance, 64–78. In *Adapting to climate change: Thresholds, values, governance*, edited by W. N. Adger, I. Lorenzoni, and K. L. O'Brien. Cambridge, UK: Cambridge University Press.
- Doraiswamy, P. C., T. R. Sinclair, S. Hollinger, B. Akhmedov, A. Stern, and J. Prueger. 2014. "Application of MODIS Derived Parameters for Regional Crop Yield Assessment." *Remote Sensing of Environment* 2005 (97): 192–202.
- Environmental Protection Agency. 2011–14. Region 9 Strategic Plan. Accessed 1 May 2016. <http://www3.epa.gov/region09/strategicplan/sanjoaquin.html>

- Funk, C., A. Hoell, S. Shukla, I. Bladé, B. Liebmann, J. B. Roberts, ... G. Husak. 2014. "Predicting East African Spring Droughts Using Pacific and Indian Ocean Sea Surface Temperature Indices." *Hydrol Earth Systems Sciences Discussions* 11 (3): 3111–3136. doi:10.5194/hessd-11-3111-2014.
- Galford, G. L., J. F. Mustard, J. Melillo, A. Gendrin, C. C. Cerri, and C. E. P. Cerri. 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: Climate Change Secretariat.
- 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". Accessed 1 May 2016. <http://docs.h2o.ai>.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY: Springer-Verlag.
- Huete, A., K. Didan, T. Miura, E. Rodriguez, X. Gao, and L. Ferreira. 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., and A. J. Peters. 2004. "Forecasting Vegetation Greenness with Satellite and Climate Data." *IEEE Geoscience and Remote Sensing Letters* 1 (1): 3–8. doi:10.1109/LGRS.2003.821264.
- Kogan, F., L. Salazar, and L. Roytman. 2012. "Forecasting Crop Production Using Satellite-Based Vegetation Health Indices in Kansas, USA." *International Journal of Remote Sensing* 33 (9): 2798–2814. doi:10.1080/01431161.2011.621464.
- Lesk, C., P. Rowhani, and N. Ramankutty. 2016. "Influence of Extreme Weather Disasters on Global Crop Production." *Nature* 529: 84–87. doi:10.1038/nature16467.
- Mahaweli Authority of Sri Lanka. 2012. *Statistical Handbook*. Colombo, Sri Lanka: Mahaweli Authority of Sri Lanka, Government of Sri Lanka.
- Meroni, M., D. Fasbender, F. Kayitakire, G. Pini, F. Rembold, F. Urbano, and M. M. Verstraete. 2014. "Early Detection of Biomass Production Deficit Hot-Spots in Semi-Arid Environment Using FAPAR Time Series and a Probabilistic Approach." *Remote Sensing of Environment* 142: 57–68. doi:10.1016/j.rse.2013.11.012.
- Nay, J. J., M. Abkowitz, E. Chu, D. Gallagher, and H. Wright. 2014. "A Review of Decision-Support Models for Adaptation to Climate Change in the Context of Development." *Climate and Development* 6 (4): 357–367. doi:10.1080/17565529.2014.912196.
- Park, S., J. Im, E. Jang, and J. Rhee. 2016. "Drought Assessment and Monitoring through Blending of Multi-Sensor Indices Using Machine Learning Approaches for Different Climate Regions." *Agricultural and Forest Meteorology* 216: 157–169. doi:10.1016/j.agrformet.2015.10.011.
- Peters, A. J., E. A. Waltershea, L. Ji, A. Vliia, M. Hayes, M. D. Svoboda, and R. E. D. Nir. 2002. "Drought Monitoring with NDVI-Based Standardized Vegetation Index." *Photogrammetric Engineering & Remote Sensing* 68 (1): 71–75.
- Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, ... A. Kaplan. 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. doi:10.1029/2002JD002670.
- Rhee, J., J. Im, and G. J. Carbone. 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.
- Rogan, J., J. Franklin, D. Stow, J. Miller, C. Woodcock, and D. Roberts. 2008. "Mapping Land-Cover Modifications over Large Areas: A Comparison of Machine Learning Algorithms." *Remote Sensing of Environment* 112 (5): 2272–2283. doi:10.1016/j.rse.2007.10.004.
- Sakamoto, T., M. Yokozawa, H. Toritani, M. Shibayama, N. Ishitsuka, and H. Ohno. 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.
- Samad, M. 2005. "Water Institutional Reforms in Sri Lanka." *Water Policy* 7: 125–140.

- Senaratne, A., and H. Scarborough. 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*, 1–22. Melbourne, Australia: AARES.
- Shukla, S., C. Funk, and A. Hoell. 2014. "Using Constructed Analogs to Improve the Skill of National Multi-Model Ensemble March–April–May Precipitation Forecasts in Equatorial East Africa." *Environmental Research Letters* 9 (9): 094009.
- Singh, P., and B. Borah. 2013. "Indian Summer Monsoon Rainfall Prediction Using Artificial Neural Network." 1585–1599. doi:[10.1007/s00477-013-0695-0](https://doi.org/10.1007/s00477-013-0695-0).
- Small, C., and C. Milesi. 2013. "Multi-Scale Standardized Spectral Mixture Models." *Remote Sensing of Environment*. 136: 442–454. doi:[10.1016/j.rse.2013.05.024](https://doi.org/10.1016/j.rse.2013.05.024).
- Survey Department of Sri Lanka. 2011. *Land Use Map of Sri Lanka*. Colombo, SL: Survey Department of Sri Lanka.
- Thenkabail, P., M. Gamage, and V. Smakhtin. 2004. The use of remote sensing data for drought assessment and monitoring in Southwest Asia. In *International Water Management Institute*, 85.
- Weligamage, P., R. Barker, M. Samad, H. Kono, and H. M. Somaratne. 2003. "Agro-well and Pump Diffusion in the Dry Zone of Sri Lanka: Past Trends, Present Status and Future Prospects." In *International Water Management Institute*, 66.
- Withananachchi, S. S., S. Kopke, C. R. Withanachchi, R. Pathiranage, and A. Ploeger. 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. doi:[10.3390/cli2040329](https://doi.org/10.3390/cli2040329).
- Ziervogel, G., P. Johnston, M. Matthew, and P. Mukheibir. 2010. "Using Climate Information for Supporting Climate Change Adaptation in Water Resource Management in South Africa." *Climatic Change* 103 (3–4): 537–554. doi:[10.1007/s10584-009-9771-3](https://doi.org/10.1007/s10584-009-9771-3).