Continuous and consistent land use/cover change estimates using socio-ecological data

Michael Marshall[a†], Michael Norton-Griffiths[a], Harvey Herr[a], Richard Lamprey[b], Justin Sheffield[c], Tor Vagen[a], and Joseph Okotto-Okotto[d]

[a] Climate Research Unit, World Agroforestry Centre, United Nations Ave, Gigiri, P.O. Box 30677-00100, Nairobi, Kenya, Email: m.marshall@cgiar.org

[b] Fauna & Flora International, The David Attenborough Building, Pembroke St, Cambridge, CB2 3QZ, UK

[c] Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, 08544, USA

[d] Lake Basin Development Authority, P.O Box 1516-40100, Kisumu, Kenya

[†] Corresponding Author

Abstract

A growing body of research shows the importance of land use/cover change (LULCC) on modifying the earth system. Land surface models are used to stimulate land-atmosphere dynamics at the macro- (regional to global) scale, but model bias and uncertainty remain that needs to be addressed, before the importance of LULCC is fully realized. In this study, we propose a method of improving LULCC estimates for land surface modelling exercises. The method is driven by projectable socio-ecological geospatial predictors available seamlessly across sub-Saharan Africa and yielded continuous (annual) estimates of LULCC at 5x5 km^2 spatial resolution. The method was developed with 2,252 5x5 km^2 sample area frames of the proportion of several land cover types in Kenya over multiple years. Forty-three socio-ecological predictors were evaluated for model development. Machine learning was used for data reduction and simple (functional) relationships defined by generalized additive models were constructed on a subset of the highest ranked predictors ($p \leq 10$) to estimate LULCC. The predictors explained 62% and 65% of the variance in the proportion of agriculture and natural vegetation, respectively, but were less successful at estimating more descriptive land cover types.

In each case, population density on an annual basis was the highest ranked predictor. The approach was compared to a commonly used remote sensing classification procedure, given the wide use of such techniques for macro-scale LULCC detection, and out-performed it for each land cover type. The approach was used to demonstrate significant trends in expanding (declining) agricultural (natural vegetation) land cover in Kenya from 1983-2012, with the largest increases (declines) occurring in densely populated high agricultural production zones. Future work should address the improvement (development) of existing (new) geospatial predictors and issues of model scalability and transferability.

Key words: earth system; land use/cover change; population-environment; remote sensing; vegetation function

1. Introduction

Land use/cover change (LULCC) is an important concern for global environmental sustainability, because it can adversely affect surface albedo and heating (Davin and de Noblet-Ducoudré, 2010); evapotranspiration and other components of the hydrologic cycle (Sterling et al., 2013); local to regional climate with the coupling or indirect recycling of surface moisture (Makarieva et al., 2013); global climate via carbon and other greenhouse gas emissions (Anderson-Teixeira and DeLucia, 2011; Ward et al., 2014); and ecosystem services worsened by these impacts (Turner et al., 2013). Land surface models, which can be coupled to a regional or global climate model, are used to simulate land-atmosphere interactions retrospectively or prospectively (Pitman, 2003) to identify intervention "hotspots" or develop realistic land management scenarios at the macro- (regional to global) scale (Turner et al., 2007). Traditionally, spatially-explicit LULCC was not an input to land surface models, but was instead represented by structural (e.g. leaf area index) or physiological (e.g. stomatal resistance) changes in vegetation. LULCC was then mapped in parallel to characterize these changes. These early attempts have been replaced by fully coupled LULCC and land surface models (e.g. Shevliakova et al., 2009; Lawrence et al., 2012). Although the impact of LULCC on the earth system is well established and quantifiable, studies remain few, due in part to the inadequacy of LULCC estimates (Pielke et al., 2011). In order to further land-atmosphere interaction research, LULCC models must be developed that provide consistent estimates over long historical time frames, regular (annual) intervals, and large spatial domains at 5x5 km^2 spatial resolution; are projectable 50-100 years into the future; and use a consistent classification approach (Meiyappan et al., 2014; Rounsevell et al., 2014; Verburg et al., 2011).

Heistermann et al. (2006) reviews the two primary categories of macro-scale LULCC models (geographic and economic) while Schaldach and Priess (2008) and Rounsevell et al. (2014) include reviews of blended or integrated approaches. The Conversion of Land Use and its Effects (CLUE) model (Veldkamp and Fresco, 1996; Verburg et al., 2002) is an example of a geographic technique. It identifies important socio- (population, economy, society, politics and planning, culture, and technology) ecological (climate, vegetation, soil, topography, and hydrology) predictors from observed LULCC data, which are related to each other statistically, and then cellular automata are used to simulate competition between the predicted LULC types and neighboring grid cells based on these relationships. Decision rules are typically used iteratively to guarantee realistic LULC transitions occur. LandSHIFT (Alcamo et al., 2011) is an example of an economic approach, because supply (LULC) is distributed on a grid cell basis by demand. Supply is determined from national estimates of crop yield and the net primary productivity of grasslands. Multi-criteria analysis, which involves applying cost functions and LULC constraints based on socio-ecological inputs, is used to define demand hierarchically and disaggregate supply over baseline or projected periods. Integrated approaches (e.g. CLUMondo: van Asselen and Verburg, 2013) are becoming more common, because they more adequately account for LULCC processes and the interaction of demand and trade with supply than economic or geographic models, respectively. Like most geographic and economic models, however, integrated models have a sound theoretical basis, but can be difficult to employ on a grid-cell basis at high spatial resolution at the macro-scale, because of data inconsistencies and incongruities and model complexity that can propagate error, as well as, the time and other resources needed to operate them. Earth observation (remote sensing) models are an important sub-category of the geographic approach, because they overcome many of these challenges,

making their operational use on grid-cell basis at high spatial resolution at the macro-scale more feasible.

Hansen and Loveland (2012) and Ban et al., (2015) present recent reviews of macro-scale remote sensing-based LULCC modeling. Remote sensing approaches use multivariate statistical techniques to classify land cover types based on the spectral or textural characteristics of gridded satellite data (DeFries et al., 1995). These approaches are simpler than integrated approaches, because they tend to capture change at a single resolution directly with no interaction between adjacent pixels. Remote sensing approaches, therefore, tend to be more parsimonious than integrated approaches and require less time for processing. Early remote sensing approaches involved daily coarse spatial resolution (8km) Advanced Very High Resolution Radiometer (AVHRR) data available from 1981. Large disagreement and uncertainties in the models, due to mixed pixel effects from small LULC patch size, as well diverse classification systems and methods, limited their use at the macro-scale in the past (Lepers et al., 2005). Recently, improved computational storage and processing and consensus on classification has facilitated the creation of consistent global LULCC maps at Landsat (30 m) resolution (Giri et al., 2013). GlobeLand30 (Chen et al., 2015), for example, uses a pixel-object-knowledge-based approach to classify Landsat images from spectrally-derived vegetation indices globally in 2000 and 2010. The use of Landsat data alone poses serious challenges to modeling LULCC on an annual basis: persistent cloud cover and a 16-day revisit cycle, makes retrieval of cloud-free pixels difficult; the Landsat platforms have been retired (Landsat 5), have failed (Landsat 6), suffer from technical problems (Landsat 7), or are only recently active (Landsat 8). To improve the temporal resolution and continuity of classification, other remote sensing products, such as the Global Forest Change product (Hansen et al., 2010), fuse Moderate-resolution Imaging

Spectroradiometer (MODIS) data available every 1-2 days at 250-500m spatial resolution with Landsat data. But these products are only available over the MODIS era (2000-present), making long-term classification difficult. In short, the major drawback of remote sensing approaches is that the temporal range and continuity necessary for long-term annual global change detection are often sacrificed for high (≤ 500m) spatial resolution. Finally, remote sensing data is not projectable like other socio-ecological data, such as population density, precipitation, or temperature, limiting their use to retrospective analyses.

The purpose of this study was to propose a simple (functional) way to map LULCC at the macro-scale at 5x5 km^2 spatial resolution on an annual basis using socio-ecological predictors that are available on an annual basis and projectable 50-100 years into the future to facilitate land-atmosphere modeling and research. The method was developed using sample area frames consisting of continuous land cover proportions developed from multi-year aerial and ground surveys in Kenya over a 30-year period. The approach was compared with remote sensing predictors that have been used to classify land cover types based on their phenology. Kenya is an ideal location to develop such a method, because like many countries in sub-Saharan Africa (SSA) data is scarce compared to the Global North, and the impact of land modification on people and the environment is high (Lambin et al., 2003). In addition: 1) population density is highest in the most agriculturally productive areas due to unequitable land distribution and poor infrastructure (Jayne and Muyanga, 2012) making ecological determinants that are generally used to map LULCC potentially less relevant (Pricope et al., 2013); 2) agriculture is the primary source of livelihood and crops are mostly rainfed (Ngetich et al., 2014); and 3) inter-annual rainfall variability is high and frequently causes devastating droughts and floods (Held and Soden, 2006).

2. Data and Methods

2.1 Study area

Aerial surveys were conducted in 1983, 1985, 2012, and 2013, to assess changes in land cover over parts of the Lake Victoria basin and central region of Kenya (Machakos and Makueni areas). The surveys yielded 2,252 5x5 km^2 sample area frames covering 28,150 km^2 or approximately 47% of Kenya's arable lands (**Figure 1**). Olofsson et al. (2012) has suggested that 5x5 km^2 sample area frames are appropriate for evaluating macro-scale LULCC models. The lakeshore and lowlands of Lake Victoria basin are primarily tropical with one long rain season that extends from February to September (UNEP, 2008). The neighboring highlands follow a bimodal pattern and annual totals are higher than near the lakeshore, due to warm moist westerlies during the West African monsoon and orographic uplift. Central Kenya is drier and has two distinct rain seasons: long rains (March-June) and short monsoon rains (October-December). The Machakos area, which includes Muranga', Kiambu, and the northern part of Machakos, is humid subtropical and therefore wetter than Makueni to the south-east, which is semi-arid.

The probability (proportion) of various land cover types within each frame were available at two levels of specificity: level one (agriculture, natural vegetation, urban, and miscellaneous) and level two (crops, fallow, shrubs, savanna, wetlands, forest, and agroforestry). These two levels of specificity were analyzed to determine the level of detail that can be captured using coarse resolution geospatial data. Continuous data were used, because at 5x5 km^2 resolution, spatial heterogeneity makes discrete classification impractical. Agriculture included agroforestry, defined here as trees on a farm; crops (banana, coffee, maize, sugar cane, tea, wheat, and others); and pasture/fallow. Natural vegetation included savanna, shrubs (open and

closed), wetlands (perennial and permanent), and forest (evergreen and deciduous). Urban included built up structures, such as roads, homes, and towns. Miscellaneous included fish ponds and other water bodies, exposed rock, and charcoal pits. The frames were developed using an aerial point sampling approach (Norton-Griffiths, 1988): several thousand geotagged aerial photos were taken over parallel transects spaced 1 km apart at approximately 488 m (height-above-ground) in 1983/1985 and then again in 2012/2013, resulting in approximately 7 aerial natural color analogue photos per frame with a ground-sampling-distance of < 1 cm in 1983/1985 and 5 aerial natural color digital photos per frame with a ground-sampling-distance of 6.5 cm in 2012/2013. The retrieval dates are shown in **Table 1**. A team of six technicians interpreted the photos on a rolling basis to minimize potential bias and errors that can occur from manual classification by different interpreters and for different years. The proportion of each land cover type (0-100%) was determined by manually classifying a grid of 320 randomly distributed points superimposed over each photo. For each year, all land cover types were represented and classified, but not all frames were interpreted and classified (**Figure 1**). The interpretations were validated via site visits and meetings with community stakeholders. The estimates were then averaged over the photos across interpreters to get the proportions for each frame. Further details on the 1983/1985 and 2012/2013 campaigns can be found in EcoSystems Ltd (1983), EcoSystems Ltd (1987), and Lamprey (2013).

2.2 Macro-scale data handling and processing

The development of the functional relationships from the sample area frames involved four major steps illustrated in **Figure 2**. Non-remote sensing and remote sensing predictors were selected after an exhaustive online search that are freely and seamlessly available across SSA, so that the relationships can be used in future studies across the continent for retrospective or

prospective analyses. Given the large number of predictors collected, machine learning was used to identify a subset of the most powerful predictors before constructing the functional relationships. The functional relationships were then evaluated against remote sensing predictors with hold-out samples and finally, used to demonstrate how the relationships can be used to reconstruct LULCC estimates continuously through time.

Forty-three non-remote sensing (climatic, hydrologic, socioeconomic, and topographic) and sixteen remote sensing (phenological) predictors of land cover change were compared and subset for model-building with the sample area frames. Either slowly-changing (long-term average/one-time value) or dynamic predictors were considered. The slowly-changing predictors and their sources are shown in **Table 2**. Using these predictors alone could streamline the modeling process. However, in reality, phenology, climate, and population change frequently, so these predictors were derived on an annual basis as well. The handling and processing of annually-changing or dynamic predictors are discussed in Sections 2.2.1-2.2.3. For the remainder of the paper, dynamic predictors include a ".d" extension. All of the geospatial data was projected to Africa Equidistant Conic (m) to facilitate distance calculations. The predictors were resampled to the finest resolution data (90x90 m^2) and aggregated to 5x5 km^2 resolution for model-building.

2.2.1 Climate

Bioclimatic (BIOCLIM: Hijmans et al., 2005) variables were used to capture climatic differences in land cover types because they 1) provide biologically meaningful information and 2) have been projected mid-21st century at high spatial resolution for SSA (AFRICLIM: Platts et al., 2014). Two additional climate parameters were included in the analysis, because they are potentially relevant and part of the Platts et al. (2014) dataset: atmospheric demand for moisture

(Potential Evapotranspiration- PET) and the Moisture Index. The BIOCLIM variables were computed on an annual basis from 1983-2012 using monthly temperature, shortwave incoming radiation, and precipitation. The variables were estimated using the "biovars" function in the "dismo" package in R (Hijmans et al., 2015). As with the Platts et al. (2014) dataset, PET was estimated using Hargreaves and Samani (1985).

The temperature/radiation and precipitation predictors were taken from the Princeton University high resolution meteorological forcing (PHF) (Chaney et al., 2014) and the Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS) (Funk et al., 2014) datasets, respectively. PHF originally spanned 1979-2008, but was extended to 2012 for this study. It is a downscaled version of the Princeton University global meteorological forcing (PGF) dataset (Sheffield et al., 2006) for SSA. It assimilates new observation data, specifically station data from the U.S. National Climatic Data Center (NCDC) Integrated Surface Database (ISD) and has undergone more rigorous correction than the global dataset. PHF is a blend of the most up-to-date observation-based, remote sensing, and reanalysis data sources: the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP-NCAR) reanalysis, Global Precipitation Climatology Project, Tropical Rainfall Measuring Mission (TRMM), the Climatic Research Unit (CRU), and the Surface Radiation Budget. The data is downscaled using elevation. The dataset includes precipitation, minimum/maximum temperature, pressure, shortwave and longwave radiation, specific humidity, and wind speed at a daily time step and 0.1° (~10 km at the equator) resolution. CHIRPS is available at pentad (5-day) intervals and 0.05° (~ 5km at the equator) spatial resolution from 1981-2012. Like PHF, CHIRPS is a blend of several observation-based, remote sensing, and reanalysis sources: geostationary thermal infrared satellite observations from the Climate Prediction Center and National Climatic Data

Center; TRMM; and NOAA-NCAR. CHIRPS was selected as the precipitation data source over PHF, because it incorporates the largest collection of ground-based precipitation data in East Africa and bias-correction is performed using the Climate Hazards Precipitation Climatology (Funk et al., 2015).

2.2.2 Population density

Population density was derived from the UNEP/GRID-Sioux Falls African Population Distribution Database (APDD) on an annual basis from 1983-2012. APDD consists of population density at a spatial resolution of 2.5 arc-minutes ° (~ 5x5 km^2 at the equator) for base years 1960, 1970, 1980, 1990, and 2000. The grids are derived from population statistics at various administrative (district, province, etc.) levels and temporal scales, depending on the availability of national population statistics. A detailed description of the derivation of gridded population can be found in Deichmann (1996). Each grid cell represents "population potential", based on its proximity to the transportation network (roads, railroads, and navigable rivers, and major towns/cities). Population at a given administrative level is then disaggregated according to the population potential. Grid cells that are closer to the network have higher coefficients and therefore receive a larger proportion of the population than grid cells further away. The base years are then extrapolated with an exponential growth/decay function (Davis, 1995). For consistency, the same function was used to distribute population between base years on an annual basis for each grid cell:

$$\mathbf{P_{i,j,t}} = \mathbf{P_{i,j,T}} \mathbf{e}^{\mathbf{\Delta t k_{i,j}}} \quad (2)$$

$$\mathbf{k_{i,j}} = \mathbf{ln}\left(\mathbf{P_{T+10n}} / \mathbf{P_{T+10(n-1)}}\right) / \mathbf{10} \quad (3)$$

$\mathbf{P_{i,j,t}}$ is the interpolated population/population density for a given year ($\mathbf{t}$) and at grid cell $\mathbf{i}$, $\mathbf{j}$, $\mathbf{P_{i,j,T}}$ is the population/population density for a given base year (period = 10 years), $\Delta\mathbf{t}$ is the change in

time from the base year to the year being interpolated, and $\mathbf{k_{i,j}}$ (**Equation 2**) is the growth/decay coefficient. The growth/decay coefficient is defined by $\mathbf{P_{T+10(n-1)}}$ (initial base year for iteration **n**) and $\mathbf{P_{T+10n}}$ (last base year for iteration **n**). The denominator was set to ten, because $\mathbf{k_{i,j}}$ accounted for decadal trends. After 2000, population statistics were extrapolated to 2012 using the 1990-2000 growth/decay coefficients.

2.2.3 Remote Sensing Predictors

The National Aeronautics and Space Administration's Global Inventory Modeling and Mapping Studies (GIMMS) Normalized Difference Vegetation Index (NDVI) Version 3 (NDVI3g) (Pinzon and Tucker, 2014) was used to estimate the remote sensing predictors. NDVI is a ratio-based vegetation index derived from Earth observation (AVHRR) surface reflectance in the visible red and near infrared (NIR). NDVI approaching one (zero) is indicative of dense vegetation (bare soil). NDVI3g is available at 0.08° (~8x8 km^2 at the equator) spatial resolution and at a 15-day timestep from 1983-2013. NDVI3g has been compared to other long-term global vegetation records and is considered the most appropriate for trend analyses (Tian et al., 2015).

The predictors were derived from NDVI using harmonic regression (Eastman et al., 2009) on an annual basis from 1983-2012. Linear harmonic regression estimates the amplitude (maximum) and phase (timing) of a fitted time series, but unless higher order harmonics are introduced, linear harmonic regression is too rigid to account for outliers and multimodal regimes commonly found in the tropics. To overcome these obstacles, non-linear harmonic regression (Carrão et al., 2010) was used to estimate five phenological predictors:

$$\mathbf{NDVI_{i,j,T}} = \mathbf{M_{i,j}} + \mathbf{A_{i,j}}\,\mathbf{cos}(\boldsymbol{\omega_0}\mathbf{t} + \emptyset + \boldsymbol{\alpha}\,\mathbf{cos}(\boldsymbol{\omega_0}\mathbf{t} + \boldsymbol{\varphi})) \quad (1)$$

Where $\mathbf{NDVI_{i,j,T}}$ is NDVI3g at grid cell **i**, **j** and over period **T**, which in this case was 24, because non-linear harmonic regression was computed on an annual basis from the 15-day data;

M is the intercept (annual mean NDVI); **A** is the amplitude; **ϕ** is the annual phase; and **α** and **φ** are non-linear terms defining the strength of non-linearity (asymmetry) and non-linear phase (deceleration/acceleration of asymmetry), respectively. The frequency (ω_0) equals $2\pi/T$. The approach can be reduced to a linear harmonic oscillator by setting $\alpha\cos(\omega_0 t+\varphi)$ to zero. The non-linear predictors were derived at each grid cell using the "nlsLM" function in the "minipack.lm" package in R (Elzhov et al., 2015). nlsLM uses the Levenberg-Marquardt optimization method (Moré, 1978) to find the non-linear least-squares fit. The function was constrained by the seed and boundary conditions described in Carrão et al., (2010). One thousand iterations at each grid cell were performed to avoid fitting local optima. Linear terms (**A** and **ϕ**) were computed for the analysis as well, using the "lm" function in the "stats" package in R (https://cran.r-project.org/), because they are more efficient and are easier to interpret.

2.3 Land-cover model development using remote sensing and non-remote sensing predictors

Land cover models were developed for each level of specificity. Seventy percent of the samples (N=1,576) were used for model calibration and 30% of the samples (N=676) were used for model validation.

Machine learning was used to omit redundant predictors and determine the feasibility of using the remaining predictors to predict each land cover type, given the large number of predictors and possible inter-correlations. Machine learning techniques lead to stable results when the number of predictors is large and are less affected by non-linearity and multicollinearity than other automated fitting routines (Binder and Tutz, 2008). The Breiman's random forest algorithm (Breiman, 2001) available in the "randomForest" package in R was selected in particular, because it is less susceptible to over-fitting and yields higher prediction accuracy than other machine learning algorithms (Fernández-Delgado et al., 2014). The Random

Forest (RF) algorithm yields an ensemble model, bagged from multiple and independent decision trees consisting of various combinations of predictors and sample subsets. The performance of the ensemble was measured with a pseudo coefficient of determination (pseudo-R^2), which is one minus the ratio of the cross-validated mean squared error (MSE) of the prediction to the variance of the observed data. As MSE or the average error between predicted and observed estimates approaches zero, R^2 approaches one (perfect correlation). The importance of each predictor in the ensemble is also quantified and is defined by the percent increase in cross-validated MSE when a predictor is removed from the ensemble. Once the predictors were ranked, the "rfcv" function was used to determine the number of predictors to use to develop functional relationships for each land cover class. Rfcv computes the cross-validated MSE versus the number of predictors included in the ensemble in descending order of importance.

The drawback of RF is that it results in complex relationships that are difficult to interpret. Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1990) were used to build functional relationships on the subsets of important predictors identified with RF, because a number of studies have successfully estimated the proportion of crop area with socio-ecological predictors and GAMs (Grace et al., 2014; Husak et al., 2008; Marshall et al., 2011); like RF, GAMs are not severely impacted by non-linear data; and unlike RF, GAMs are relatively simple and easy to interpret. Since the response variable (proportion of land cover type) was continuous and bounded from 0-100%, the data was fitted using a quasi-binomial distribution (link=logistic). The logistic GAM predicts the log-likelihood of an event (probability of success/probability of failure) using, in our case, a series of cubic spline functions:

$$\mathbf{log}\left(\frac{\mathbf{p_j}}{\mathbf{1-p_j}}\right) = \boldsymbol{\beta_0} + \sum \mathbf{f_{i,j}}(\mathbf{x_{i,j}}) \quad (4)$$

Where **p** is the probability of a LULC type for sample area frame **j**, β_0 is the intercept, and $f_{i,j}(x_{i,j})$ is the cubic spline function for predictor x_i at sample area frame **j**. The GAMs were developed with the "gam" function in the "mgcv" package in R. Model calibration was evaluated with explained part and overall deviance. Deviance is the log-likelihood (probability space) alternative to variance. Part deviance is the deviance explained when the target predictor is removed from a GAM minus the overall deviance. Another pseudo-R^2 statistic (1-model deviance/null deviance) was also computed to compare calibration with validation.

In order to demonstrate how the models can be used for macro-scale application, the final GAMs developed were employed to reconstruct the annual change in agriculture and natural vegetation and to perform a trend analysis from 1983-2012 at each sample area frame. Trends were estimated using the Theil-Sen technique, which computes the median of all possible pairwise slopes in a time series. The approach has been used, for example, to measure long-term trends in NDVI (de Beurs and Henebry, 2005), because it is not significantly impacted by outliers or non-linearity. The significance of each trend was assessed using the Mann-Kendall statistic. Trends were masked at the 99.9% confidence band to mask statistically insignificant trends.

3. Results

3.1 Land cover sample area frame summary

The distribution of land cover over the sample are frames is illustrated with a boxplot in **Figure 3**. Agriculture and natural vegetation land cover (level one) were normally distributed, with agriculture having a higher median (54.04%) and lower spread (29.32% and 76.33% at the first and third quartiles) than natural vegetation (median=39.72%, first quartile=16.21%, and third quartile=65.67%). The proportion of urban and miscellaneous land cover was considerably

lower (median=4.00% and 0%, respectively) and non-linear, each having several high proportion outliers. The disaggregated land cover (level two) distributions, with the exception of crops, were non-linear with long right tails. Crops represented the largest proportion of land cover (median=37.52%) and had the largest spread (19.43% and 58.46% at the first and third quartiles), followed by savanna (median=15.79%, first quartile=3.80%, and third quartile=31.91%). Wetlands represented the smallest proportion of land cover (median=0%), with sample area frames not exceeding 75%, while forest represented the second smallest proportion of land cover (median=2.22%), but had a large number of outliers.

3.2 Data reduction

The top remote sensing and non-remote sensing predictors considered are ranked in descending order of importance for agriculture and natural vegetation using bar graphs in **Figure 4**. The RF ensemble models using non-remote sensing predictors performed moderately well for agriculture (pseudo-R^2=0.69) and natural vegetation (pseudo-R^2=0.69), but poorly for the more non-linear distributions (urban pseudo-R^2=0.37 and miscellaneous pseudo-R^2=0.50). The RF ensemble models using remote sensing predictors all performed poorly: agriculture (pseudo-R^2=0.49), natural vegetation (pseudo-R^2=0.50), urban (pseudo-R^2=0.22), and miscellaneous (pseudo-R^2=0.33). It should be noted in each case however, the highest ranked remote sensing predictors resulted in lower model error than the highest ranked non-remote sensing predictors. The non-remote sensing predictors were more numerous and generated larger incremental improvements that contributed to overall greater predictive power. For the non-remote sensing ensembles, dynamic predictors were more important than slowly-changing predictors, and population density and climate predictors consistently outranked topographic or hydrologic predictors. Popd.d, popd, bio7.d, bio14.d, and bio3.d were consistently ranked the most

important predictors of agriculture and natural vegetation proportions. Omitting popd.d, the most important predictor for agriculture, for example, led to a more than 65% increase in ensemble MSE. Given that popd.d and popd were both important, model results were compared with popd.d and popd individually and combined as anomalies (popd.d /popd). Ensemble performance was better when the two predictors were considered separately. The most important remote sensing predictors were less influential than popd.d. strn, ampn.d, and ampl.d were more equally important for agriculture and natural vegetation, followed by phsl and phsn.

The importance of predictors of level two (crops, savanna, and forest) proportions are ranked in **Figure 5**. The ranking was more variable for level two classifications, but popd.d remained the most important predictor in each case. The level two RF ensemble models predicted less variability than the level one RF ensemble models and the non-remote sensing predictors outperformed the remote sensing predictors when more than the highest ranked predictors were introduced. The non-remote sensing models performed moderately well for crops (pseudo-R^2=0.63), savanna (pseudo-R^2=0.62), and forest (pseudo-R^2=0.61), but poorly for fallow (pseudo-R^2=0.42), shrubs (pseudo-R^2=0.54), wetlands (pseudo-R^2=0.10), and agroforestry (pseudo-R^2=0.55). Precipitation-based climatic predictors (bio12.d, bio13.d, bio14.d, and bio16.d) were more important in the savanna ensemble than temperature-based climatic determinants driving the crop ensemble. For the forest simulation, topographic predictors (slp and topind) were more important than most of the climatic predictors. The remote sensing ensembles performed poorly for all of the level two land cover classes: crops (pseudo-R^2=0.46), fallow (pseudo-R^2=0.33), shrubs (pseudo-R^2=0.44), savanna (pseudo-R^2=0.44), wetlands (pseudo-R^2<1%), forest (pseudo-R^2=0.46), and agroforestry (pseudo-R^2=0.41). For crops, strn and ampl.d remained the most important predictors. Maximum annual NDVI, as captured by

ampl.d and ampn.d, was much more important for predicting the proportion of savanna. Unlike other ensembles, which were driven by dynamic predictors, the most important remote sensing predictors for forest cover were long-term averages.

3.3 Building functional relationships

The GAMs were developed for moderately performing land cover classes and used considerably fewer predictors than the RF ensembles, because most of the predictors in the ensembles explained very little, if any, variability. This is illustrated in **Figure 6**, which shows MSE versus the number of predictors used in the non-remote sensing and remote sensing ensembles for forest. For the non-remote sensing ensemble, MSE increased from 119.76 to 120.49 after the 10^{th} predictor and leveled off after the 13^{th} predictor were introduced. For the remote sensing ensemble, MSE increased from 120.49 to 163.34 and levelled off after the 7^{th} predictor was introduced. For this reason, the GAMs were built with 10-13 of the highest ranked non-remote sensing predictors and additional predictors, namely popd, were removed after redundancies were identified in the GAM component functional plots and with significance tests (not shown). GAMs were not constructed using the remote sensing predictors, because of the poor results of the ensembles and the inability of additional predictors to substantially improve the accuracy of the GAMs. Similarly, non-remote sensing GAMs were not developed for urban, miscellaneous, fallow, shurbs, or wetlands.

Figures 7 and **8** show the functional relationships of the predictors used for estimating the proportion of agriculture and natural vegetation. Each model explained 61.5% (pseudo-R^2=0.66) and 61.4% (pseudo-R^2=0.66) of model deviance with nine and seven predictors, respectively. The error bars tended to be wider at proportion extremes, because fewer data points were available to train the models. The relative importance of each predictor, as defined

by part deviance and other calibration statistics are shown in **Table 3** for the land cover types that were considered feasible for model-building. Popd.d remained the most important predictor and uniquely explained 7.0-26.2% of model deviance. The log-likelihood of agriculture (natural vegetation) increased (decreased) rapidly as population density increased from 0 to 550 people•km^{-2}, more gradually between 550 and 1200 people•km^{-2}, and reversed beyond 1200 people•km^{-2}. The predictive power of the topographic and climatic variables dropped off sharply after popd.d. For agriculture, bio14.d and topind were the second and third most important predictors, but explained only 1.9% and 1.6% unique deviance. As seen in the partial functional plots, the proportion of agriculture was highest in high production zones (medium population density) on ridges and crests where topind was low and for very wet tropical areas where bio14.d was high and semi-arid areas where bio14.d was low. For natural vegetation, temperature predictors, bio4.d and bio7.d, explained the second and third highest unique deviance after popd.d (2.0% and 1.3%). As seen in the functional plots, low populated areas with more temperature seasonality, or inter-annual variation, and lower bio3.d (isothermality) tended to have higher proportions of natural vegetation (savanna and shrubs). Isothermality is the ratio of mean diurnal temperature range (bio2.d) to the temperature annual range (bio7.d), which is the difference between the annual maximum and minimum temperatures. Areas that are less isothermal essentially have more pronounced seasons and are climatically less tropical. For the level two classifications, calibration was more difficult and yielded poorer relationships. Popd.d was the most important predictor and explained 7.0 – 16.4% unique deviance. The predictive power of the topographic and climatic variables was more equally distributed than for the level one classification.

In all cases, the R^2 for the validation subset was lower than the pseudo-R^2 from the calibration subset: agriculture (ΔR^2=-0.04), natural vegetation (ΔR^2=-0.01), crops (ΔR^2=-0.03), savanna (ΔR^2=-0.01), and forest (ΔR^2=-0.06) (**Figure 9**). With the exception of the crops GAM, level two GAMs tended to under-predict high proportions of land cover (savanna and forest) and contained numerous outliers.

3.4 Trend analysis

The GAMs for agriculture and natural vegetation were used to simulate trends in the annual proportions for the sample area frames from1983-2012 as part of the evaluation to demonstrate how the approach could be used for a retrospective analysis. The proportion of agriculture for 1983 and 2012 are shown in **Figure 10a and 10b**, while trends over the 30-year period are shown in **Figure 10c**. The high potential agricultural zone (wet highlands) in Western Kenya experienced the largest increase in simulated agricultural cover (> 1% per year or 30% over the 30-year period). A time series of the strongest trend (1.68% per year) is shown in **Figure 10d**. Simulated population density was at 145 people•km^{-2} in 1983 for this sample area frame, which steadily increased to 478 people•km^{-2} in 2012. Closer to the lake, which consists of drier marginal mixed farming, trends were insignificant at the 99.9% confidence band or relatively weak (< 1 % per year). Similar patterns were seen for the marginal mixed farming and high potential agricultural zones of central Kenya. The only decrease in agricultural lands was in Kitale town (-1.40% per year). The time series is also shown in **Figure 10d**. Population growth in Kitale was 1,110 people•km^{-2} in 1983, which is near the threshold of declining agriculture cover versus population density at 1,200 people•km^{-2}. By 2009, when the largest decrease in agriculture cover occurred, from 51.0 to 29.5%, population density had steadily increased and surpassed another apparent threshold above 3,000 people•km^{-2}. The direction and relative

magnitude of trends in natural vegetation (not shown) generally corresponded inversely to trends in agriculture, but were negatively-weak (maximum=-0.4% per year or -12% over the 30-year period).

5. Discussion

The results make three important contributions that the land surface modeling community should consider to improve LULCC detection, particularly for SSA: 1) a socioeconomic variable (population density) was the highest ranked predictor of LULCC and had considerably more predictive power than biophysical predictors; 2) non-remote sensing predictors outperformed remote sensing predictors due to their number and the incremental improvement in the predictive power of each; and 3) coarse resolution data was able to capture general classification descriptors, but was unable to capture more detailed descriptors.

The global increase in agricultural land cover has been attributed to the demand for food and other agricultural commodities by a growing population (Pongratz et al., 2008). In SSA, smallholder farms, which support the majority of the labor force, are small (half are < 1.5 ha) and concentrated in densely populated areas, while large portions of arable farmland in sparsely populated areas remain underutilized (Jayne et al., 2003). This underutilization is due primarily to a lack of investment in infrastructure and unequitable tenure systems, which forces farmers to grow more on less land. This relationship is confirmed by rural population survey data in Kenya, which showed that fertilizer input use and net farm income per hectare increase until approximately 550 persons•km^{-2} and then sharply decline, because farm sizes shrink, surplus production decreases, and farmers must adopt costlier strategies (e.g. zero-grazing) to maximize revenue (Jayne and Muyanga, 2012). The functional relationship for population density and steady increase in area under cultivation in high production zones demonstrated by the trend

analysis in this study, corresponds to this finding, as area under cultivation increased rapidly to approximately 550 persons•km^{-2} and then increased more gradually with higher population density until 1200 persons•km^{-2}. Few sample area frames had population densities greater than 1,200 persons•km^{-2}, as in Kitale town, so it is difficult to know if this functional relationship holds for very high population densities. At least to 2008, Kitale experienced a growth rate of 12%, well above the national average (7%), due to persistent drought and out-migration from neighboring high production zones (Majale, 2008). Although the functional relationship for population density corroborates household surveys in Kenya and other agrarian countries in SSA, it should be further scrutinized, because land tenure in SSA is complex (Place, 2009), the dependency of LULCC predictors on location and spatial scale can be high (Rindfuss et al., 2004), and the transition from agrarian to industrialized nations may make 50-100 year projections for SSA obsolete.

The proposed methodology when applied to other regions of the world will undoubtedly result in a different combination of socio-ecological predictors and functional relationships, because access to land varies across agrarian and non-agrarian societies, so further study is required with observed data to develop region-specific models and validate the results for countries in SSA. Kumar et al., 2013, for example, showed that in the United States pre-1900 when the country was largely agrarian and transportation networks were weak, population density and crop area were highly correlated, because crops needed to be grown close to markets. However, as the country became more industrialized and transportation networks improved, farmers moved to more biophysically suitable areas away from city centers, making biophysical determinants of crop area more important than population density in the latter half of the 20th century. Whether the analyses are performed in agrarian or non-agrarian regions, extensive

preparation of observation data will be required, because the data used in this study, namely consistent sample area frames at a spatial resolution appropriate for land surface modeling and spanning multiple climatic zones through time, is quite unique.

Population density estimates vary widely (Wilson, 2014) and given its fundamental importance to the proposed model framework, future work should aim to integrate a more dynamic product that better accounts for inter-annual variability and realistic representation of current and projected population density. To the authors' knowledge, this was the first attempt to make a population product dynamic. However, the approach is essentially tracking decadal trends that explain a significant portion of inter-annual variability. In reality, population density can show high inter-annual variability due to migration and other factors. Regarding the product itself, changes in population density do not necessarily "grow" from transportation networks and are influenced by important feedbacks now and in the future. In addition, the extrapolation method used is efficient and can be projected indefinitely, but does not capture complex demographics that other methods do and can lead to "runaway" growth/decline and unrealistic mid- to late- 21st century projections for scenario-building (Baker et al., 2008). Finally, there is no consensus on which population product to use however, in the future, other products (e.g. Afripop) should be compared against the product used here, used to adjust growth/decay coefficient for population density estimates beyond 2000, or combined to make a model ensemble.

This paper highlights the importance of gridded socioeconomic data in mapping LULCC, but gridded macro-scale datasets are almost exclusively biophysical in nature. The biggest gains in LULCC prediction could be made, therefore, by developing gridded macro-scale socioeconomic data from existing country-level products, such as the Human Development

Index. More minor gains could be made by integrating biophysical predictors not used in this study, such as soil type and properties. Gridded soils data exists globally from the International Soil Reference and Information Center, but was not considered in this study, because it is a one-time value and does not capture the dynamic nature of soils or its complex relationship with LULCC. A dynamic soils product was recently developed for the MODIS era (see Vågen et al., 2016) and could be a powerful tool for LULCC detection, especially if it is back-casted over the full temporal range of other predictors. Many biophysical predictors are available mid- and late-21st century and are therefore widely used for prospective analyses, so methods should be explored to project soils and socio-economic data into the future to improve LULCC estimates.

Grace et al. (2014) developed GAMs to predict cropped area in Kenya using biophysical predictors (rainfall, elevation, NDVI, slope, and the topographic wetness index) and explained much of the deviance in cropped area (41.9-81.4%). Although the models used different predictors for different years and production zones, and the definition of cropped area and the degree of functional smoothing were not explicit, the study highlights that the inter-correlation among predictors may be obscuring the importance of biophysical determinants. Specifically, population density tends to be highly correlated with and could be suppressing the explanatory power of biophysical predictors, though the partial deviance statistics did not reflect this. In addition, the random forest algorithm accounts for inter-correlation to some degree, but other techniques could be introduced to further reduce these effects. For example, Principal Components Analysis could be used to develop temperature and precipitation indices that integrate all or some of the BIOCLIM predictors.

Phenological patterns extracted from continuous Earth observation based NDVI have been widely used to map LULCC over long time periods, given the lack of higher spatial and

spectral resolution data before the MODIS era (Ali et al., 2014; Bie et al., 2012). These studies show that vegetation periodicity is highly variable for a given land cover type and that long-term averages of phenological predictors are more reliable for mapping LULC. In this study, many of the important remote sensing predictors (particularly for forests) were long-term averages, but they still under-predicted LULCC when compared against non-remote sensing predictors, which were more numerous and resulted in larger incremental improvements to model accuracy. Perhaps the main difficulty in using long-term Earth observation data for LULCC estimation is the coarseness of the data and the rapid change in vegetation that often occurs over small spatial scales. Population density, which was a much stronger predictor, on the other hand, may well be captured using coarse resolution data, because this predictor changes more gradually over space. An analysis of the non-remote sensing and remote sensing predictors together revealed that for agriculture, natural vegetation, savanna, and forest cover, Earth observation data provided an additional 1-2% explained deviance. If the long-term average remote sensing predictors could be downscaled using MODIS or Landsat data and then aggregated to 5x5 km^2 resolution with distribution moments as predictors, for example, the explanatory power of non-remote sensing predictors could be further enhanced for retrospective analyses. Another avenue worth exploring could involve using downscaled long-term average remote sensing predictors to develop 5x5 km^2 probabilities as in the Pengra et al., 2015 dataset to evaluate the non-remote sensing models proposed here.

The evaluation of the models at two levels of specificity revealed that coarse resolution is able to better simulate general descriptors, such as natural vegetation, but is poorer at predicting more detailed descriptors, such as forest. Each of the more detailed random forest ensembles with non-remote sensing predictors had ΔR^2s of -0.06, -0.07, and -0.07 for crop, savanna, and

forest over agriculture and natural vegetation, respectively. Part of this discrepancy can be attributed to the increased interpretation uncertainty, as interpreters find it more challenging to distinguish between more detailed LULC types. In addition, coarse resolution data may not be able to capture the level of heterogeneity in the area sample frames needed to distinguish land use/cover-specific socio-ecological patterns and properties.

6. Conclusion

This study developed and evaluated a simple method to provide consistent estimates of LULCC annually over 30 years at 5x5 km^2 resolution using non-parametric functional relationships with a small subset of socio-ecological predictors ($p \leq 10$). Functional relationships were developed after data mining 43 geospatial datasets that are available seamlessly across SSA, which can be used for retrospective or prospective mid- and late-21st century analyses as well. The relationships are intuitive and tunable, making their use practical for decision-makers to identify intervention hotspots and develop land management scenarios. Model validation, performed with the proportion of major land cover types in Kenya over a 30-year period, revealed that a number of activities should be performed to improve the predictive power of the models for practical use. These activities should include integrating improved existing or newly developed geospatial (particularly socioeconomic) datasets into the proposed model framework. With these improvements, land surface and LULCC modelling could be greatly enhanced and the consequence of the latter on the earth system more fully understood. In an upcoming study, the modeling approach proposed here will be used with a newly acquired sample area frame dataset to estimate baseline LULCC and project land suitability across SSA mid-21st century with AFRICLIM and other geospatial data.

Acknowledgements

The work summarized in this manuscript was primarily funded through support by the CGIAR research program on Climate Change, Agriculture and Food Security for the project, titled "Multi-disciplinary species distribution modelling for "climate smart" agriculture in East Africa." Additional support for the early field and aerial surveys was supported by the Kenyan Lake Basin Development Authority and Ministry of Planning and National Development. We would like to extend our special thanks to Sandra Nakibilango, Dorcas Ninsiima, Charles Ngugi, Patrick Ojorot and Samuel Olowo who interpreted the aerial photos taken for this project; Bernadette Apio who coordinated the interpreter team; and Juliet Kyakobyewo who entered the data into our database. Finally, we would like to thank Dr. Eike Luedeling, who initially led and organized the project.

References

Alcamo, J., Schaldach, R., Koch, J., Kölking, C., Lapola, D., Priess, J., 2011. Evaluation of an integrated land use change model including a scenario analysis of land use change for continental Africa. Environ. Model. Softw. 26, 1017–1027.

Ali, A., de Bie, C.A.J.M., Skidmore, A.K., Scarrott, R.G., Lymberakis, P., 2014. Mapping the heterogeneity of natural and semi-natural landscapes. Int. J. Appl. Earth Obs. Geoinformation 26, 176–183.

Anderson-Teixeira, K.J., DeLUCIA, E.H., 2011. The greenhouse gas value of ecosystems. Glob. Change Biol. 17, 425–438.

Baker, J., Ruan, X., Alcantara, A., Jones, T., Watkins, K., McDaniel, M., Frey, M., Crouse, N., Rajbhandari, R., Morehouse, J., Sanchez, J., Inglis, M., Baros, S., Penman, S., Morrison, S., Budge, T., Stallcup, W., 2008. Density-dependence in urban housing unit growth: An evaluation of the Pearl-Reed model for predicting housing unit stock at the census tract level. J. Econ. Soc. Meas. 33, 155–163.

Ban, Y., Gong, P., Giri, C., 2015. Global land cover mapping using Earth observation satellite data: Recent progresses and challenges. ISPRS J. Photogramm. Remote Sens. 103, 1–6.

Bie, C.A.J.M. de, Nguyen, T.T.H., Ali, A., Scarrott, R., Skidmore, A.K., 2012. LaHMa: a landscape heterogeneity mapping method using hyper-temporal datasets. Int. J. Geogr. Inf. Sci. 26, 2177–2192.

Binder, H., Tutz, G., 2007. A comparison of methods for the fitting of generalized additive models. Stat. Comput. 18, 87–99.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.

Carrão, H., Gonalves, P., Caetano, M., 2010. A Nonlinear Harmonic Model for Fitting Satellite Image Time Series: Analysis and Prediction of Land Cover Dynamics. IEEE Trans. Geosci. Remote Sens. 48, 1919–1930.

Chaney, N.W., Sheffield, J., Villarini, G., Wood, E.F., 2014. Development of a High-Resolution Gridded Daily Meteorological Dataset over Sub-Saharan Africa: Spatial Analysis of Trends in Climate Extremes. J. Clim. 27, 5815–5835.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., Mills, J., 2015. Global land cover mapping at 30 m resolution: A POK-based operational approach. ISPRS J. Photogramm. Remote Sens., Global Land Cover Mapping and Monitoring 103, 7–27.

Davin, E.L., de Noblet-Ducoudré, N., 2010. Climatic Impact of Global-Scale Deforestation: Radiative versus Nonradiative Processes. J. Clim. 23, 97–112.

Davis, H.C., 1995. Demographic Projection Techniques for Regions and Smaller Areas: A Primer. UBC Press.

de Beurs, K.M., Henebry, G.M., 2005. A statistical framework for the analysis of long image time series. Int. J. Remote Sens. 26, 1551–1573.

DeFries, R.S., Field, C.B., Fung, I., Justice, C.O., Los, S., Matson, P.A., Matthews, E., Mooney, H.A., Potter, C.S., Prentice, K., Sellers, P.J., Townshend, J.R.G., Tucker, C.J., Ustin, S.L., Vitousek, P.M., 1995. Mapping the land surface for global atmosphere-biosphere models: Toward continuous distributions of vegetation's functional properties. J. Geophys. Res. Atmospheres 100, 20867–20882.

Deichmann, U., 1996. A Review of Spatial Population Database Design and Modeling (Technical Report No. 96–3). National Center for Geographic Information and Analysis, Santa Barbara, CA.

Eastman, R., Sangermano, F., Ghimire, B., Zhu, H., Chen, H., Neeti, N., Cai, Y., Machado, E.A., Crema, S.C., 2009. Seasonal trend analysis of image time series. Int. J. Remote Sens. 30, 2721–2726. doi:10.1080/01431160902755338

EcoSystems Ltd, 1987. Integrated Land Use Database for Kenya. Minstry of Planning & Natural Development, Nairobi, Kenya.

EcoSystems Ltd, 1983. Integrated Land Use Survey: Final Report. Lake Basin Development Authority, Kisumu, Kenya.

Elzhov, T.V., Mullen, K.M., Spiess, A.N., Bolker, B., 2015. Package "minpack.lm."

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? J Mach Learn Res 15, 3133–3181.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Rowland, J., Romero, B., Husak, G.J., Michaelsen, J., Verdin, A., 2014. A Quasi-Global Precipitation Time Series for Drought Monitoring (No. 832), U.S. Geological Survey Data Series. U.S. Geological Survey, Washington, D.C.

Funk, C., Verdin, A., Michaelsen, J., Peterson, P., Pedreros, D., Husak, G., 2015. A global satellite assisted precipitation climatology. Earth Syst. Sci. Data Discuss. 8, 401–425.

Giri, C., Pengra, B., Long, J., Loveland, T.R., 2013. Next generation of global land cover characterization, mapping, and monitoring. Int. J. Appl. Earth Obs. Geoinformation 25, 30–37.

Grace, K., Husak, G., Bogle, S., 2014. Estimating agricultural production in marginal and food insecure areas in Kenya using very high resolution remotely sensed imagery. Appl. Geogr. 55, 257–265.

Grace, K., Husak, G.J., Harrison, L., Pedreros, D., Michaelsen, J., 2012. Using high resolution satellite imagery to estimate cropped area in Guatemala and Haiti. Appl. Geogr. 32, 433–440.

Hansen, M.C., Stehman, S.V., Potapov, P.V., 2010. Quantification of global gross forest cover loss. Proc. Natl. Acad. Sci. 107, 8650–8655.

Hansen, M.C., Loveland, T.R., 2012. A review of large area monitoring of land cover change using Landsat data. Remote Sens. Environ., Landsat Legacy Special Issue 122, 66–74.

Hargreaves, G.H., Samani, Z.A. 1985. Reference Crop Evapotranspiration from Temperature. Appl. Eng. Agric. 1, 96–99.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. CRC Press.

Held, I.M., Soden, B.J., 2006. Robust Responses of the Hydrological Cycle to Global Warming. J. Clim. 19, 5686–5699.

Heistermann, M., Müller, C., Ronneberger, K., 2006. Land in sight?: Achievements, deficits and potentials of continental to global scale land-use modeling. Agric. Ecosyst. Environ. 114, 141–158.

Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J., 2015. Package “dismo.”

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology. 25, 1965-1978.

Husak, G.J., Marshall, M.T., Michaelsen, J., Pedreros, D., Funk, C., Galu, G., 2008. Crop area estimation using high and medium resolution satellite imagery in areas with complex topography. J. Geophys. Res. Atmospheres 113, D14112.

Jayne, T.S., Yamano, T., Weber, M.T., Tschirley, D., Benfica, R., Chapoto, A., Zulu, B., 2003. Smallholder income and land distribution in Africa: implications for poverty reduction strategies. Food Policy 28, 253–275.

Jayne, T.S., Muyanga, M., 2012. Land constraints in Kenya's densely populated rural areas: implications for food policy and institutional reform. Food Secur. 4, 399–421.

Kumar, S., Merwade, V., Rao, P.S.C., Pijanowski, B.C., 2013. Characterizing Long-Term Land Use/Cover Change in the United States from 1850-2000 Using a Nonlinear Bi-analytical Model. Ambio 42, 285-297.

Lambin, E.F., Geist, H.J., Lepers, E., 2003. Dynamics of Land-Use and Land-Cover Change in Tropical Regions. Annu. Rev. Environ. Resour. 28, 205–241.

Lamprey, R.H., 2013. Aerial Point Sampling (APS) Survey: Lake Basin, Machakos and Makueni, Kenya, 2012-13, Nairobi, Kenya.

Lawrence, P. J., Feddema, J. J., Bonan, G. B., Meehl, G. A., O'Neill, B. C., Oleson, K. W., Levis, S., Lawrence, D. M., Kluzek, E., Lindsay, K., and Thornton, P. E., 2012. Simulating the Biogeochemical and Biogeophysical Impacts of Transient Land Cover Change and Wood Harvest in the Community Climate System Model (CCSM4) from 1850 to 2100. J. Climate 25, 3071–3095.

Lepers, E., Lambin, E.F., Janetos, A.C., DeFries, R., Achard, F., Ramankutty, N., Scholes, R.J., 2005. A Synthesis of Information on Rapid Land-cover Change for the Period 1981–2000. BioScience 55, 115–124.

Majale, M., 2008. Employment creation through participatory urban planning and slum upgrading: The case of Kitale, Kenya. Habitat Int., Labour in Urban Areas 32, 270–282.

Makarieva, A.M., Gorshkov, V.G., Li, B.-L., 2013. Revisiting forest impact on atmospheric water vapor transport and precipitation. Theor. Appl. Climatol. 111, 79–96.

Marshall, M.T., Husak, G.J., Michaelsen, J., Funk, C., Pedreros, D., Adoum, A., 2011. Testing a high-resolution satellite interpretation technique for crop area monitoring in developing countries. Int. J. Remote Sens. 32, 7997–8012. doi:10.1080/01431161.2010.532168

Meiyappan, P., Dalton, M., O'Neill, B.C., Jain, A.K., 2014. Spatial modeling of agricultural land use change at global scale. Ecol. Model. 291, 152–174.

Moré, J.J., 1978. The Levenberg-Marquardt algorithm: Implementation and theory, in: Watson, G.A. (Ed.), Numerical Analysis, Lecture Notes in Mathematics. Springer Berlin Heidelberg, pp. 105–116.

Ngetich, K.F., Mucheru-Muna, M., Mugwe, J.N., Shisanya, C.A., Diels, J., Mugendi, D.N., 2014. Length of growing season, rainfall temporal distribution, onset and cessation dates in the Kenyan highlands. Agric. For. Meteorol. 188, 24–32.

Norton-Griffiths, M., 1988. Aerial Point Sampling for Land Use Surveys. J. Biogeogr. 15, 149–156.

Olofsson, P., Stehman, S.V., Woodcock, C.E., Sulla-Menashe, D., Sibley, A.M., Newell, J.D., Friedl, M.A., Herold, M., 2012. A global land-cover validation data set, part I: fundamental design principles. Int. J. Remote Sens. 33, 5768–5788.

Pengra, B., Long, J., Dahal, D., Stehman, S.V., and Loveland, T.R., 2015. A global reference database from very high resolution commercial satellite data and methodology for

application to Landsat derived 30m continuous field tree cover data Remote Sens. Environ. 165, 234-248.

Pielke, R.A., Pitman, A., Niyogi, D., Mahmood, R., McAlpine, C., Hossain, F., Goldewijk, K.K., Nair, U., Betts, R., Fall, S., Reichstein, M., Kabat, P., de Noblet, N., 2011. Land use/land cover changes and climate: modeling analysis and observational evidence. Wiley Interdiscip. Rev. Clim. Change 2, 828–850.

Pinzon, J.E., Tucker, C.J., 2014. A Non-Stationary 1981–2012 AVHRR NDVI3g Time Series. Remote Sens. 6, 6929–6960.

Pitman, A.J., 2003. The evolution of, and revolution in, land surface schemes designed for climate models. Int. J. Climatol. 23, 479–510.

Place, F., 2009. Land Tenure and Agricultural Productivity in Africa: A Comparative Analysis of the Economics Literature and Recent Policy Strategies and Reforms. World Dev., The Limits of State-Led Land Reform 37, 1326–1336.

Platts, P.J., Omeny, P.A., Marchant, R., 2014. AFRICLIM: high-resolution climate projections for ecological applications in Africa. Afr. J. Ecol. 53, 103–108.

Pongratz, J., Reick, C., Raddatz, T., Claussen, M., 2008. A reconstruction of global agricultural areas and land cover for the last millennium. Glob. Biogeochem. Cycles 22, GB3018.

Pricope, N.G., Husak, G., Lopez-Carr, D., Funk, C., Michaelsen, J., 2013. The climate-population nexus in the East African Horn: Emerging degradation trends in rangeland and pastoral livelihood zones. Glob. Environ. Change 23, 1525–1541.

Rindfuss, R.R., Walsh, S.J., Turner, B.L., Fox, J., Mishra, V., 2004. Developing a science of land change: Challenges and methodological issues. Proc. Natl. Acad. Sci. U. S. A. 101, 13976–13981.

Rounsevell, M.D.A., Arneth, A., Alexander, P., Brown, D.G., de Noblet-Ducoudré, N., Ellis, E., Finnigan, J., Galvin, K., Grigg, N., Harman, I., Lennox, J., Magliocca, N., Parker, D., O'Neill, B.C., Verburg, P.H., Young, O., 2014. Towards decision-based global land use models for improved understanding of the Earth system. Earth Syst Dynam 5, 117–137.

Schaldach, R., Priess, J.A., 2008. Integrated Models of the Land System: A Review of Modelling Approaches on the Regional to Global Scale. Living Rev. Landsc. Res. 2. doi:10.12942/lrlr-2008-1

Sheffield, J., Goteti, G., Wood, E.F., 2006. Development of a 50-Year High-Resolution Global Dataset of Meteorological Forcings for Land Surface Modeling. J. Clim. 19, 3088–3111.

Shevliakova, E., Pacala, S. W., Malyshev, S., Hurtt, G. C., Milly, P. C. D., Caspersen, J. P., Sentman, L. T., Fisk, J. P., Wirth, C., and Crevoisier, C., 2009. Carbon cycling under 300 years of land use change: Importance of the secondary vegetation sink. Biogeochem. Cy. 23, GB2022, doi:10.1029/2007gb003176.

Sterling, S.M., Ducharne, A., Polcher, J., 2013. The impact of global land-cover change on the terrestrial water cycle. Nat. Clim. Change 3, 385–390.

Tian, F., Fensholt, R., Verbesselt, J., Grogan, K., Horion, S., Wang, Y., 2015. Evaluating temporal consistency of long-term global NDVI datasets for trend analysis. Remote Sens. Environ. 163, 326–340.

Turner, B.L., Janetos, A.C., Verbug, P.H., Murray, A.T., 2013. Land System Architecture: Using Land Systems to Adapt and Mitigate Global Environmental Change. Glob. Environ. Change 232395-397.

Turner, B.L., Lambin, E.F., Reenberg, A., 2007. The emergence of land change science for global environmental change and sustainability. Proc. Natl. Acad. Sci. 104, 20666–20671.

UNEP, 2008. Africa: Atlas of Our Changing Environment. UN Environment Programme, Nairobi, Kenya.

Vågen, T.-G., Winowiecki, L.A., Tondoh, J.E., Desta, L.T., Gumbricht, T., 2016. Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. Geoderma 263, 216–225.

van Asselen, S., Verburg, P.H., 2013. Land cover change or land-use intensification: simulating land system change with a global-scale land change model. Glob. Change Biol. 19, 3648–3667.

Veldkamp, A., Fresco, L.O., 1996. CLUE-CR: An integrated multi-scale model to simulate land use change scenarios in Costa Rica. Ecol. Model. 91, 231–248.

Verburg, P.H., Neumann, K., Nol, L., 2011. Challenges in using land use and land cover data for global change studies. Glob. Change Biol. 17, 974–989

Verburg, P.H., Soepboer, W., Veldkamp, A., Limpiada, R., Espaldon, V., Mastura, S.S.A., 2002. Modeling the Spatial Dynamics of Regional Land Use: The CLUE-S Model. Environ. Manage. 30, 391–405.

Ward, D. S., Mahowald, N. M., and Kloster, S., 2014. Potential climate forcing of land use and land cover change. Atmos. Chem. Phys. 14, 12701–12724.

Wilson, T., 2014. New Evaluations of Simple Models for Small Area Population Forecasts: Small Area Population Forecasts. Popul. Space Place 21, 335–353.

Yu, L., Liang, L., Wang, J., Zhao, Y., Cheng, Q., Hu, L., Liu, S., Yu, L., Wang, X., Zhu, P., Li, X., Xu, Y., Li, C., Fu, W., Li, X., Li, W., Liu, C., Cong, N., Zhang, H., Sun, F., Bi, X., Xin, Q., Li, D., Yan, D., Zhu, Z., Goodchild, M.F., Gong, P., 2014. Meta-discoveries from a synthesis of satellite-based land-cover mapping research. Int. J. Remote Sens. 35, 4573–4588.