Mapping Fractional Land Use and Land Cover in a Monsoon Region: The Effects of Data Processing Options

Bumsuk Seo, Christina Bogner, Thomas Koellner, and Björn Reineking

***Abstract*—Existing global land use/land cover (LULC) raster maps have limited spatial and thematic resolution relative to the strong heterogeneity of agricultural landscapes. One promising approach to derive more informative maps is using fractional cover instead of hard classification. Here, we evaluate the effect of three key data processing options on the performance of Random Forest fractional cover models for MODIS data in a heterogeneous agricultural landscape in a monsoon region: (i) selection of spectral predictor sets [Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), surface reflectance (SR), and all combined (Full)], (ii) time interval (8-day vs. 16-day), and (iii) smoothing (no smoothing vs. Savitzky-Golay filter). Model performance was assessed with spatially stratified RMSE, Spearman's rank correlation, and R^2, per LULC type and averaged over all types. We found adequate performance of the best model (avg. ρ = 0.62) that used all predictors, 8-day interval and no smoothing. Among the different alternatives, the choice of predictors accounted for 36.3% of the variation, smoothing for 19.0% and time interval for 17.9%. The intrinsic dimensionalities of the spectral predictors were investigated to complement the variable importance analyses. Although predicting LULC fractions for minor types remained difficult, our results suggest that existing satellite products can be a useful source of information about LULC at sub-pixel level provided the data-processing options are properly chosen.**

B. Seo is with the Department of Environmental Science, Kangwon National University, Gangwondaehak-ro 1, 200-701 Chuncheon, Republic of Korea, and with the Biogeographical modeling, Bayreuth Center of Ecology and Environmental Research BayCEER, University of Bayreuth, Universitätsstraße 30, D-95440 Bayreuth, Germany, e-mail: bumsukseo@kangwon.ac.kr.

C. Bogner is with the Ecological Modelling, Bayreuth Center of Ecology and Environmental Research BayCEER, University of Bayreuth, Dr.-Hans-Frisch-Straße 1–3, D-95448 Bayreuth, Germany, e-mail: christina.bogner@uni-bayreuth.de

T. Koellner is with the Professorship of Ecological Services, Bayreuth Center of Ecology and Environmental Research BayCEER, University of Bayreuth, Universitätsstraße 30, D-95440 Bayreuth, Germany, e-mail: thomas.koellner@uni-bayreuth.de

B. Reineking is with the Université Grenoble Alpes, Irstea, UR EMGR, 2 rue de la Papeterie-BP 76, F-38402 St-Martin-d'Hères, France and with Biogeographical Modelling, Bayreuth Center of Ecology and Environmental Research BayCEER, University of Bayreuth, Universitïtsstraße 30, D-95440 Bayreuth, Germany, email: bjoern.reineking@irstea.fr

This research was supported by an International Research Training Group between Germany and South Korea (DFG/KOSEF, Complex TERRain and ECOlogical Heterogeneity - TERRECO, GRK 1565/1). The first author has been partially supported by the Agenda Program of the Rural Development Administration (PJ00997802) in Republic of Korea.

This is an author's post-print version of the article published in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (ISSN: 1939-1404).

Manuscript received Aug 16, 2015

***Index Terms*—Land use/land cover; sub-pixel mapping; fractional land cover; Random Forest; agricultural land use; Monsoon**

I. INTRODUCTION

CONVENTIONAL global land cover (GLC) maps are discrete raster maps assigning land cover types to each pixel. Usually, these discrete products are coarse in spatial resolution due to large cell sizes (e.g., 1 km). Recent techniques such as fractional cover allow continuous mapping of land use. Fractional land cover consists of proportions of non-overlapping land cover types in pixels of a given raster grid [1]–[3]. It is often called sub-pixel land cover as it can be conceived as an interpretation of land cover types at the sub-pixel level [4]. It is also called 'continuous fields' [3], [5]. Fractional land cover is increasingly used as a key descriptor of ecosystems and their functions (e.g., [4], [6]–[9]). Yet, currently available GLC databases such as GlobCover 2009 or Moderate Resolution Imaging Spectroradiometer (MODIS) land cover generally lack high resolution maps or fractional land cover data [8], [10]. Therefore, the GLC products are generally limited in representing heterogeneous LULC (e.g., unable to discriminate mixed trees, shrubs, and herbaceous vegetation) [11]. MODIS Vegetation Continuous Fields product (MOD44B) is the only GLC product that provides fractional cover data. However, in the current version (V005) it is limited to tree-related land cover types, namely "tree", "non-tree", and "bare soil".

In cultivated landscapes such as mixed agricultural areas, a large number of LULC types often occur in a relatively small area. Despite the significance of LULC information for studies in cultivated landscapes [12], the existing GLC products are generally limited in cultivated landscapes [11], [13]–[15] due to the small number of crop-related types [16]–[18]. For instance, GlobCover 2009 is provided at 300 m resolution and has four crop-related types, and MODIS Land Cover Type (MCD12Q2) product provides five raster land cover layers at 500 m [16]–[18], each of which with only one or two cropland types. Especially for heterogeneous arable zones like irrigated fields (e.g., [19]), GLC products are underdeveloped [10]. The above mentioned spatial and thematic limitations of the GLC databases are particularly pronounced in heterogeneous agricultural landscapes due to the mosaic of crop/non-crop LULC types [15]. These limitations make it difficult to monitor crop production, land degradation, and other agriculture associated land use based on the GLC products.

To retrieve thematically and spatially rich land cover data, we can attempt to extract additional information from existing multi-spectral medium-resolution sensors. Deriving fractional land cover from existing satellite products can enrich the information contents with little additional cost. Furthermore, it can be applied to the past-time data. Accordingly, there have been continuous efforts to derive fractional land cover information from existing raster data [3], [5]. Among various existing sensors, NASA's MODIS (MODerate Resolution Imaging Spectroradiometer) sensor possesses temporal continuity and global coverage. Despite their limited spectral and spatial resolution, MODIS multi-spectral products provide good temporal resolution and can be useful to map agricultural areas [8], [20]. Indeed, MODIS time series contain the complete seasonal dynamics and therefore potentially useful information to distinguish land cover types (e.g., [21], [22]) and has been used to map agricultural LULC types (e.g., [23]–[25]). Regarding fractional cover, Lu *et al.* [26] showed that MODIS time series are suitable to map fractional woody and herbaceous covers.

Fractional cover estimation of multi-crop LULC would be an important step in LULC studies. However, fractional land cover modeling is still challenging, especially with multiple types [20]. There have been many successful fractional cover mapping but often with a small number of LULC types (e.g. few green vegetation types) [3]–[7], [26]–[29]. Furthermore, some models were not trained on and validated against ground observation and/or with cross-validation (e.g., [30], [31]). For future applications, it is important to develop fraction LULC mapping framework necessarily with 1) multi-type LULC, 2) ground observations, and 3) an appropriate validation scheme.

To develop a fractional land cover model, a number of decisions at the model formulation stage need to be made. First, one needs appropriate predictor data – a difficult choice due to an increasing number of satellite products (e.g., [32]). Second, a suitable algorithm and training parameters should be chosen to avoid sub-optimal performance. Third, pre- and post-processing strategies should be determined (e.g., [7]). We will denote all these decisions 'data-processing options' hereafter.

Improperly selected data-processing options can degrade the model performance by reducing information contained in the data. Optimal data-processing options are case-specific (i.e., dependent on the purpose, cost and processing capacities [33]), thus cannot be universally evaluated. Therefore, in the course of model building, the modeler should select proper data-processing options.

In monsoonal areas, there is a specific problem undermining model performance. In these areas, acquisition of cloud-free data during monsoon is generally difficult due to long-lasting rainfalls [7], [34]. For example, South Korean summer shows typical East Asian monsoon weather with persistent and intensive raining period from June to September. This period is called "Changma" (i.e., long lasting rain) in Korean literature [35].

In a heterogeneous agricultural landscape in South Korea, we aim to derive fractional LULC from multi-spectral satellite data using a data mining algorithm. It is challenging because the study area is a complex heterogeneous agricultural landscape. Spectral datasets are supposedly cloud-contaminated because the study area is situated in a monsoon region. In this context, we set up the main objectives as 1) to develop a fractional LULC modeling framework with globally available data (i.e., multi-spectral data) and 2) to evaluate relevant data-processing options, namely selection of spectral predictor sets, time intervals, and smoothing options.

The study is based on the following hypotheses: 1) the full information of a spectral data product (e.g., all available reflectance bands) perform better than a subset of it (e.g., a single reflectance band) or an index function (e.g., NDVI), 2) multi-day composited data with a narrow (e.g., 8-day) composite window [36] produce a better regression performance due to more details in the data, and 3) smoothing of input data improves the regression performance because it reduces possible cloud contamination. These hypotheses were chosen in accordance with the characteristics of the study area.

In addition to the main analysis, we assess the relative importance of the spectral bands and the data acquisition dates. Based on the results, we discuss the current capacity and potential of the multi-type fractional cover model in heterogeneous agricultural landscapes.

II. Materials and Methods

A. Study area

The study area Haean-myeon is located at the border between North and South Korea (128°1′33.101″E, 38°28′6.231″N) (Fig. 1). It is a small agricultural catchment (64.4 km^2) with elevations ranging between 500 m and 1200 m above see level. The catchment is a heterogeneous agricultural landscape with a variety of natural and artificial LULC types. Seo *et al.* [14] reported 67 LULC types from a three-year field-level LULC census.

The average air temperature of the study area is 8.5° C at the central plateau. The annual average rainfall equals 1599 mm and the maximum daily rainfall was 223 mm between 1999 and 2010 (Korean Meteorological Administration, http://web.kma.go.kr/eng). The study site belongs to the East Asian summer monsoon (EASM) region [34]. More than 60% of annual precipitation is concentrated during the monsoon period from June to August and extreme rainfall events occur frequently.

B. Data

1) Land use/land cover and fractional cover data: For the analysis, we used the LULC polygon data censused in 2010 for the site. The reference LULC polygon data consists of spatial polygons with the observed LULC information and is archived at the public repository *Pangaea* [37]. Additionally to the raw LULC type labels, it provides reclassified type labels based on four classification schemes. We used a reclassified LULC labels in a 10-class scheme, which was designed to describe the edaphic and socio-economic conditions of the area. The scheme includes "Barren", "Dry field", "Forest", "Greenhouse", "Inland water", "Inland wetland", "Orchard field", "Paddy field", "Semi natural", and "Urban". This scheme was selected as it distinguishes paddy field from other

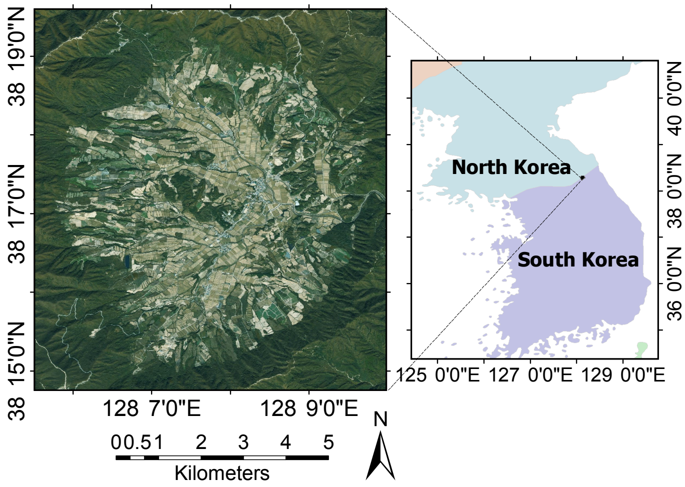

Fig. 1. Map and the location of the study site 'Haean' on the Korean peninsula. The satellite image is a SPOTMaps mosaic product (Astrium Services, http://www.astrium-geo.com) acquired in 2009.

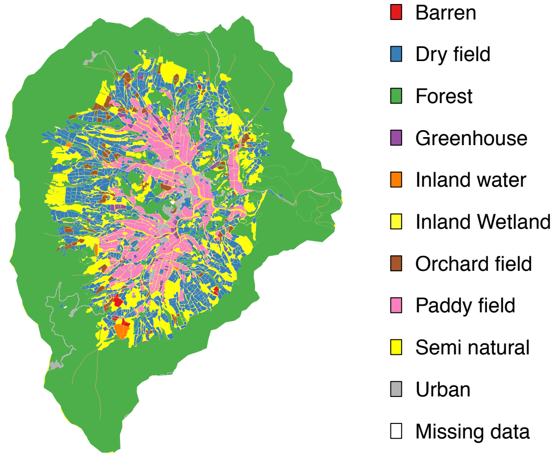

Fig. 2. The reference land use/land cover in the Haean catchment in 2010. The reference LULC in cover fraction is shown in Appendix Figure A1.

agricultural types. More details about the LULC reference data are provided in the meta information of the dataset [37].

Due to the bowl-shaped topography of the catchment, LULC types are unevenly distributed (Fig. 2). The steep slopes and the encompassing mountain ridges are covered by "Forest". The lower area is dominated by the managed land use types. "Paddy field" occurs at the central plateau whereas "Dry field" and "Semi natural" dominate on the surrounding lower slopes. The aforementioned four LULC types are large or moderately large in area proportions ($> 8\%$) and cover 95.0% of the total area (Table I). We will denote these types as 'major types'. The rest of the LULC types are smaller in area proportions ($< 2\%$). We denote the next five types "Urban", "Orchard field", "Inland water", "Greenhouse" and "Barren" as 'minor types'. "Inland wetland" was excluded from the analysis due to its extreme rarity. The selected 9 types make up 99.9% of the study area.

TABLE I
THE LAND USE/LAND COVER TYPES IN THE HAEAN CATCHMENT IN 2010. "INLAND WETLAND" WAS EXCLUDED FROM THE ANALYSIS DUE TO ITS EXTREME RARITY.

Type	Area (km^2)	Area (%)	Category
Forest	37.195	57.805	Major types
Dry field	9.543	14.831	
Semi natural	9.124	14.180	
Paddy field	5.178	8.047	
Urban	1.108	1.723	Minor types
Orchard field	0.952	1.480	
Inland water	0.556	0.864	
Greenhouse	0.544	0.845	
Barren	0.144	0.224	
Inland wetland	0.0004	0.0007	-

Fractional vegetation cover is defined as the sum of the vegetated patch area divided by the total area [1], [29]. In a satellite image, it is calculated per pixel and ranges from 0 (0% cover) to 1 (100% cover) [38]. Similarly, fractional LULC can be defined as the sum of the LULC patch area divided by the total area in each pixel of a given raster grid [4]. The study site is located in the MODIS tile H28V5 and covered 299 pixels of the 500 m sinusoidal grid (SR-ORG:6842). We chose the 500 m grid as the base grid and derived a per pixel fractional cover data from the observed LULC data. To derive per pixel LULC fractions, we first converted the MODIS raster grid to polygons by pixel (i.e., one polygon per pixel). Then we projected the grid polygons into the WGS84/UTM52N space (EPSG:32652) and overlaid the observed LULC polygons. In the projected space, we calculated the area fractions of the LULC types in all grid polygons (Appendix Figure A1).

2) MODIS spectral data: We used multi-spectral data products as predictors of the fractional LULC model. We chose MODIS collection 5 MOD13A1/MYD13A1 products. These MODIS products have a horizontal resolution of 500 m with a good overall geo-location accuracy ($RMSE < 50$ m) [39]. Other satellite products such as Landsat Thematic Mapper (https://lta.cr.usgs.gov/TM) are also often used for land monitoring [40], [41]. However, due to its 16-day repeating interval, Landsat products are often severely cloud contaminated in the monsoon region. In our study area, the Landsat 5 collection at NASA EOSDIS system (http://reverb.echo.nasa.gov) provides only a few cloud free images in 2010. In contrast, the MODIS 16-day products are less cloud contaminated due to its daily acquisition interval and the composition procedure [42].

MOD13A1/MYD13A1 products supply 23 scenes/year at 500 m resolution each. A time series of MOD13A1 starts from the first day of a year but MYD13A1 from the 9th day. Hence, there is an 8-day difference in acquisition date [43]. Each product contains 12 Science Data Sets (SDS) [42]. Among the SDSs, we chose four surface reflectance (SR) bands (B1–3, B7), Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and vegetation index quality assurance (QA). The six biophysical SDSs were used as predictors for regression. The QA SDS was used as data weight only for smoothing. For simplicity, we will denote all the biophysical SDSs as spectral bands in the following.

Each spectral band delivers specific information about land cover [44]. The red band 1 (B1) is sensitive to vegetation chlorophyll and its wavelength is 620–670 nm. The near-infrared (NIR) band 2 (B2) covers 841–876 nm and has been widely used to evaluate ground vegetation viability together with B1. Band 3 (B3) is commonly called the blue band as it is sensitive to water vapor; its wavelength ranges between 459 and 479 nm. The mid-infrared band 7 (B7) with wavelengths between 2105 nm and 2155 nm contains information about land and cloud properties.

NDVI and EVI are vegetation indices designed to capture above ground vegetation properties and biophysical processes [36], [44]. NDVI is a function of the red and the NIR bands. EVI additionally uses the blue band to remove soil and atmospheric contamination [36].

We acquired the MODIS products from NASA Land Processes Distributed Active Archive Center (LP DAAC) at the USGS/Earth Resources Observation and Science (EROS) Center (https://lpdaac.usgs.gov).

C. Scenarios

We considered three key data-processing options: predictor set, time interval, and smoothing. Each option comprises several choices. From all combinations of the three options, we formulated 16 scenarios (Table II) and evaluated them using a 16-fold cross-validation (CV) (Section II-D2). The efficacy of a data-processing option was estimated by the average performance of the associated scenarios. The modeling procedure is illustrated in Fig. 3.

Predictor set We prepared four predictor sets to compare model performance based on different spectral data. The predictor sets 'NDVI' and 'EVI' contained a corresponding vegetation index data. The Surface Reflectance (SR) predictor set 'SR' contained the four surface reflectance bands (B1–B3, and B7). The 'Full' predictor set incorporates all the six available data bands.

Time interval Spectral input data was prepared in 8-day and 16-day intervals. For 16-day input, we simply used MOD13A1 data. For 8-day input, we merged MOD13A1 and MYD13A1 products to produce a quasi 8-day MODIS 13A1 data using the 8-day difference in acquisition date described in Section II-B2. This results in 46 (8-day) or 23 (16-day) data points per band for each MODIS pixel. Note that we used the quasi 8-day data instead of the 8-day MODIS products (MOD/MYD09A1). This is because we want to use the 8-day data most similar to the 16-day data. Additionally, the 09A1 products lack NDVI and EVI data sets.

Smoothing We prepared spectral input data with and without smoothing. By comparing the two input data sets, we evaluated the efficacy of data smoothing in a monsoonal catchment. We chose the 'Savitzky-Golay' (SG) filter [45], which is widely used for smoothing time series data in remote sensing (e.g., [46]). The filter is designed to retrieve the upper envelope of a time series by using a local polynomial regression iteratively to fit the time series [47]. It can filter out negatively biased noise (e.g., NDVI decreases due to cloud contamination), which can be useful in monsoonal regions.

We used the adaptive SG filter provided by the software TIMESAT 3.1 [48], [49]. The seasonal course of the spectral data was smoothed separately for each pixel. The MODIS QA data layer was used to weight the values; data points acquired under non-optimal conditions (e.g., cloudy weather) had only 10% of influence during the smoothing process, compared with the data acquired under optimal conditions. The TIMESAT smoothing parameters were determined according to the software manual [49]. The size of the fitting window was 3 for 16-day data and 5 for 8-day data. The adaptation strength was 1.5 and the number of envelope iterations was 3. Note that the 3-year data (2009–2011) was processed concurrently as the software encourages using a longer time series than the target period.

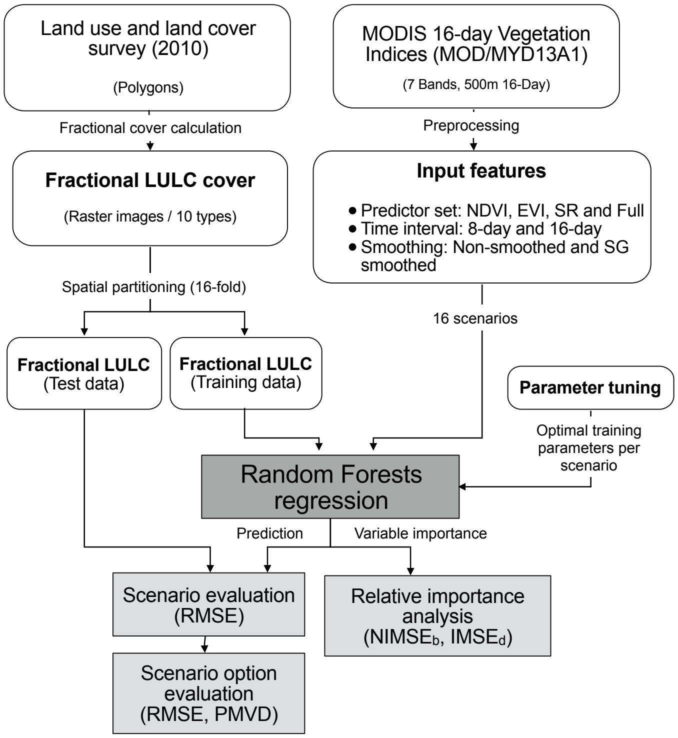
Land use and land cover survey (2010)
(Polygons)
MODIS 16-day Vegetation Indices (MOD/MYD13A1)
(7 Bands, 500m 16-Day)
Fractional cover calculation
Preprocessing
Fractional LULC cover
(Raster images / 10 types)
Input features
• Predictor set: NDVI, EVI, SR and Full
• Time interval: 8-day and 16-day
• Smoothing: Non-smoothed and SG smoothed
Spatial partitioning (16-fold)
16 scenarios
Fractional LULC (Test data)
Fractional LULC (Training data)
Parameter tuning
Optimal training parameters per scenario
Random Forests regression
Prediction
Variable importance
Scenario evaluation (RMSE)
Relative importance analysis (NIMSE$_b$, IMSE$_d$)
Scenario option evaluation (RMSE, PMVD)

Fig. 3. Overview of the fractional cover regression model building and evaluation procedure.

TABLE II
SPECIFICATION OF THE SCENARIOS IN COMBINATIONS OF THE PREDICTOR SET, TIME INTERVAL, AND SMOOTHING OPTIONS.

Smoothing		No smoothing		SG smoothing	
Time interval		8-day	16-day	8-day	16-day
Predictor set	NDVI	S1	S5	S9	S13
	EVI	S2	S6	S10	S14
	SR	S3	S7	S11	S15
	Full	S4	S8	S12	S16

D. Model construction

1) Random Forest regression: We hypothesized that per pixel LULC fractions can be retrieved from spectral data in line with the previous studies (e.g., [6], [7], [10], [26], [38]). Modeling multi-type LULC fractions can be conceived as a multi-output regression task. This task can be accomplished either by simultaneously modeling a multi-output response, or by separately modeling single-output responses and aggregating the outcomes [50], [51]. In this study, we used the latter approach.

Fractional cover regression can be implemented via various techniques. The techniques include the fuzzy classifier [52], the time series model [26], linear models [5], [28], data mining algorithms [4], [6], [53], and spectral unmixing analysis [7], [29]. Here, we used the regression mode of Random Forest (RF). RF is a decision-tree based ensemble algorithm that uses bootstrap aggregation (i.e., bagging) and the random subspace method [54], [55]. It is suitable for modeling non-linear relationships and can handle a large number of covariates as it tends not to overfit the data [50], [54], [55]. Its performance is comparable to the other state-of-the-art learning algorithms such as support vector machine or neural networks [6], [55]–[58]. It is convenient to set up in comparison with other data mining algorithms as it has a small number of hyper parameters [59].

In remote sensing, RF has been used to classify land cover [21], [22], [32], [56], [60], [61], vegetation type [21], [62], [63], and also crop type [64], [65]. In fractional land cover regression, Schwieder *et al.* [6] used RF to estimate shrub cover fractions in which RF showed comparable performance with support vector machine and partial least squares regression.

2) Spatial cross-validation: We used a spatial leave-one out cross-validation scheme in the study. Due to the bagging of RF [54], the bootstrap samples for training (i.e., in-bag data) can be correlated with the test samples (i.e., out-of-bag data), especially for spatial models [66]. To reduce dependencies between training and test data, we externally partitioned training and test data using the spatial cross-validation scheme introduced in [67].

This ‘Checkerboard’ spatial cross-validation was implemented in our study as follows. First, we binary split the whole area six-times recursively, resulting in 64 sub-clusters. Second, we form 16 clusters by randomly sampling four sub-clusters for each; one cluster is composed of four spatially disjointed sub-clusters as distinguished by different colors in Appendix Figure A2.

In each of the 16 CV folds, we reserve one cluster for test and trained a RF regression model on the remaining clusters. The trained RF model is used to predict the hold-out cluster. We obtained cross-validated predictions for the whole area by aggregating the hold-out clusters.

3) Fractional cover estimation: Let T be the number of LULC types such that each type i has a set $F_i = \{f_{i,1}, ..., f_{i,n}\}$ of n observed LULC fractions, where $f_{i,j}$ is the fractional area of the pixel j covered by the LULC type i, and n is the total number of pixels belonging to the study area.

A LULC fraction $f_{i,j} \in [0, 1]$ and all fractions of one pixel sum up to one

$$\sum_{i=1}^{T} f_{i,j} = 1 \tag{1}$$

for all $j = \{1, ..., n\}$.

First we built a RF regression model per type. Given a type i, we used the observed fraction $F_i = \{f_{i,1}, ..., f_{i,n}\}$ as response and a set of feature vectors $P = \{p_1, ..., p_n\}$ as predictor. Each feature vector contained $n_{feature}$ features varied by the spectral data used (Appendix Table B1).

The regression model was trained/tested with a 16-fold cross validation. By accumulating test pixels of all CV folds, we obtained the predicted fractions $\hat{F}_i = \{\hat{f}_{i,1}, ..., \hat{f}_{i,n}\}$ of the type i over the entire study area. RF produces predictions from all regression trees [54], therefore for each pixel n_{tree} fractions were predicted, where n_{tree} is the total number of regression trees. We took the mean value of the n_{tree} predictions and regarded as the predicted LULC fraction for the pixel.

Then we normalized the type-wise predictions by 1. The normalized prediction $\hat{F}_i^*$ was calculated as

$$\hat{F}_i^* = \frac{\hat{F}_i}{\sum_{j=1}^{T} \hat{f}_{i,j}}, \tag{2}$$

where $\hat{F}_{i,j}$ is the type-wise prediction of the type i for the pixel j. Finally, we obtain the predicted LULC fractions $\hat{F}^* = \{\hat{f}_1^*, ..., \hat{f}_T^*\}$.

4) Training parameters: RF has three training parameters: the number of trees in the forest (n_{tree}), the number of randomly selected variables on each split (m_{try}), and the number of minimal samples in terminal nodes ($nodesize$). These parameters need to be tuned to avoid sub-optimal model performance [60], [68].

To find the optimal n_{tree} and $nodesize$ we performed a grid search on the training folds. We used a grid from all combinations of n_{tree}= $\{100, 200, ..., 1000\}$ and $nodesize$= $\{1, 2, 3, 4, 5\}$. Grid searching was implemented using an internal validation. We re-partitioned the training data folds into a new training data and a new test data. The new test data contained two spatial clusters, randomly selected without replacement. We trained the model on the new training data with different parameter values and predicted the hold-out data. This was repeated for all 9 types and we averaged the root mean square error ($RMSE$) over the all types. Overall, the model performance improved with large n_{tree} and small $nodesize$ (Appendix Figure A3).

We optimized n_{tree} and $nodesize$ separately based on its marginal $RMSE$ on the tuning grid. We chose parameters by minimizing the marginal error metric unlike [60] or [69] who used the joint error metric on the grid. Compared with the joint error based selection, the marginal error based selection was less sensitive to the between-partition variations and led to more stable parameter selection between scenarios.

The parameter m_{try} was determined by the square root of $n_{feature}$ without grid searching as in [70]. Since the scenarios have unequal numbers of input features, m_{try} varied between scenarios. The chosen parameter values are summarized in Appendix Table B1.

E. Model evaluation

1) Overall regression performance: We used the cross-validation error metrics instead of the default out-of-bag

(OOB) error of RF. As discussed in Section II-D2, the OOB error can be biased due to a possible correlation between in-bag training samples and out-of-bag test samples, especially for spatial models. Instead, we used cross-validation $RMSE$ to evaluate regression performance. The $RMSE$ of the LULC type i is calculated as

$$RMSE_i = \sqrt{\frac{\sum_{j=1}^{n}(f_{i,j} - \hat{f}^{*}_{i,j})^2}{n}}, \tag{3}$$

where $f_{i,j}$ is the observed and $\hat{f}^{*}_{i,j}$ is the predicted LULC fraction for the type i in pixel j, and n is the total number of pixels.

Furthermore, we used the coefficient of determination (R^2) and Spearman's rank correlation coefficient (ρ) [71]. The R^2 was used to compare our results with the previous studies on fractional cover estimation (e.g., [4]). Spearman's ρ was used to estimate the association between observed and predicted fractions [71].

2) Relative contribution of data-processing options: In addition to cross-validation error, we examined the relationship between the data-processing options and the performance of the fractional cover regression models. For this analysis, we built a linear model explaining the $RMSE$ of the regression model for each LULC type by the different data-processing options:

$$RMSE_i = \beta_0 + \beta_1 O_p + \beta_2 O_t + \beta_3 O_s + \epsilon, \tag{4}$$

where $RMSE_i$ is the $RMSE$ of the type i; O_p is a categorical variable denoting the chosen predictor set option, O_t time interval option, and O_s smoothing option; ϵ is the error term. We did not include interaction terms based on a preliminary model selection using F-statistics (not shown here). Each linear model (per type) was estimated based on the 16 samples from all 16 scenarios and the statistical significance of the type-wise models were tested using F-statistics to validate the model structure.

We assumed that the 'relative contribution' (cf. 'relative importance' in [72]) of a modeling option is that of the corresponding regressor to the linear model. Then we quantified relative contributions of the regressors by decomposing the amount of explained variance of the linear model due to regressors. We used proportional marginal variance decomposition (PMVD) method [72], [73] which decomposes the explained variance of the linear model into non-negative contributions, which sum to the total variance explained. PMVD is able to deal with correlated regressors by averaging over different orderings. Moreover, it has desirable properties such as 'admissibility'.

3) Marginal performance of data-processing options: The efficacy of a data-processing option was estimated by average regression performance of the scenarios using the option. We will call it 'marginal performance' in the following. The marginal performance (M) of a data-processing option k for a performance metric q is calculated as

$$M_{k,q} = \frac{\sum_{x \in s^k} q(x)}{|s^k|}, \tag{5}$$

where s^k is a set of scenarios using the option k and $|s^k|$ is the number of elements of s^k.

4) Relative importance of spectral bands and acquisition dates: We quantified the relative importance of the spectral bands and the acquisition dates on the regression performance. We derived the importance of the features using a RF variable importance metric and grouped them by band and acquisition date. RF provides two importance metrics for quantifying the influence of input features [54], [74]. Among the metrics, we used the 'increased mean square error ($IMSE$)', which is a permutation-based measure. Another metric namely 'increased node purity (INP)' is measured by node purity, in case of regression the residual sum of squares. We avoided using INP because of the possible bias due to the random sub-spacing (i.e., random selection of features). For classification problems, the INP is known to be biased as the impurity measure (i.e., Gini index) favors predictor variables with many categories [75], [76]. $IMSE$ of a feature f is calculated as

$$IMSE_f = \frac{\sum_{k=1}^{n_{tree}}(\overline{MSE_k} - MSE_{f,k})}{n_{tree}} \times \frac{1}{\sqrt{s^2/n_{tree}}}, \tag{6}$$

where n_{tree} is the size of the forest, $\overline{MSE_k}$ is the mean squared OOB error of tree k, $MSE_{f,k}$ is the error after permuting the feature f and s^2 is the standard deviation of the differences between the two errors; if s^2 is zero, the division is omitted. We computed $IMSE_f$ in each cross-validation fold and averaged them. The variable importance metric itself is calculated based on the OOB samples [54].

We defined the importance of a band as the sum of the importance metrics of the features belonging to the band. Let a predictor set $X = \{x_1, ..., x_l\}$ have l features some of which belong to a spectral band b. We calculated importance of the band b as

$$IMSE_b = \frac{\sum_{x \in b} IMSE_x}{l_b}, \tag{7}$$

where l_b is the number of the features belonging to the band.

To facilitate comparisons between different bands, we normalized $IMSE_b$ as

$$NIMSE_b = \frac{IMSE_b}{\sum_{b=1}^{n_{band}} IMSE_b} \tag{8}$$

where n_{band} is the number of the bands in a predictor set. To derive $NIMSE_b$ we used the two groups of the scenarios: scenarios using 'SR' predictor set (S3, S7, S11, and S15) and scenarios using 'Full' predictor set (S4, S8, S12, and S16). As they are different in the number of spectral bands, we calculated two sets of $NIMSE_b$. For each group individually, we calculated the mean importance measures from the included scenarios.

Likewise the importance of an acquisition date is defined as the sum of the importance metrics of the features acquired at a particular date d as

$$IMSE_d = \frac{\sum_{x \in d} IMSE_x}{l_d}, \tag{9}$$

where l_d is the number of the features acquired at the date d. To derive $IMSE_d$ we used the 'Full' predictor set based scenarios (S4, S8, S12 and S16). As 8-day and 16-day data differ in the number of data points, we extracted two seasonal $IMSE_d$ curves individually by interval.

F. Dimensionality of the raw reflectance data

We used time series of the reflectance bands. In our dataset, correlations among the dates as well as the reflectance bands are likely to be very high. The intrinsic dimension of the dataset may be different from the number of the data columns due to the redundant information. In order to assess how much of information each band has, we evaluated the dimensionality of the raw reflectance bands data [77]–[79]. We used the Intrinsic Dimensionality (ID) approach [78], in which the dimensionality is defined as the minimum number of parameters required to account for the observed properties of the data. We evaluated ID of the four raw reflectance bands (B1, B2, B3, and B7) with both 8-day and 16-day intervals. In addition to the ID of the individual reflectance band, the ID of the dataset with the all four bands ('All' bands) was evaluated.

We evaluated ID of the multispectral input data using the algorithm HySime [78], [80]. Since the algorithm requires a hyperspectral input image with m bands (i.e., $1{\times}m$ matrix), we transformed the time series of our input data (i.e., $1{\times}l \times n_{band}$ matrix) into a single pseudo-hyperspectral image with the number of hyperspectral bands equaled to the length of the time series multiplied by the number of the reflectance bands [i.e., 1 pixel $\times(l \cdot n_{band})$ matrix]. The dimensionality identification algorithm was applied to the transformed images.

G. Software

We used `GNU R` version 3.1.2 [81] and the `R` packages `randomForest` version 4.6–7 [82], `raster` version 2.3–40 [83], and `relaimpo` version 2.2–2 [72] for the fractional cover regression. The geometry engine `GEOS` 3.4.2 [84] was used via the `R` package `rgeos` 0.3–8 [85] for the LULC data pre-processing. The software TIMESAT version 3.1 [49] and HyperMix version 2.13 [80] were used for the smoothing and the dimensionality identification of the input spectral data.

III. Results

A. Overall regression performance

The average performance of all scenarios in $RMSE$, ρ, and R^2 were 0.057, 0.624, and 0.414, respectively (Table III). The best scenario S4 used 'Full' predictor set in '8-day' interval with 'No smoothing'. The worst scenario S14 used 'EVI' predictor set in '16-day' interval with 'SG smoothing'. Maps of the modeled LULC fractions are provided in Appendix Figure A4 and A5 for averaged and for the best scenario, respectively.

B. Type-wise regression performance

Spearman's rank correlation between the observed and the predicted LULC fractions was high on average (avg. $\rho = 0.624$; Table III and Appendix Figure A6), not only for the major types but also for some of the minor types (Appendix Table B2). For example, ρ was 0.48 for "Orchard field" and 0.54 for "Inland water", for which predicting absolute fractions were unsuccessful ($R^2 < 0.10$). Similarly, for "Greenhouse" the rank correlation ($\rho = 0.59$) indicates a better model performance than the $R^2 (= 0.25)$.

TABLE III
Fractional LULC regression performance by scenario. All the performance metrics were averaged over LULC types.

Name	Data-processing options			Model performance		
	Predictor set	Time interval	Smoothing	$RMSE$	ρ	R^2
S1	NDVI	8-day	No smoothing	0.056	0.658	0.428
S2	EVI			0.057	0.639	0.438
S3	SR			0.054	0.657	0.441
S4	Full			0.053	0.663	0.455
S5	NDVI	16-day		0.056	0.630	0.410
S6	EVI			0.060	0.601	0.395
S7	SR			0.056	0.638	0.430
S8	Full			0.055	0.634	0.434
S9	NDVI	8-day	SG smoothing	0.058	0.618	0.399
S10	EVI			0.059	0.601	0.389
S11	SR			0.054	0.633	0.411
S12	Full			0.053	0.634	0.434
S13	NDVI	16-day		0.061	0.588	0.364
S14	EVI			0.064	0.572	0.347
S15	SR			0.057	0.611	0.418
S16	Full			0.056	0.609	0.424
		Avg.		0.057	0.624	0.414

To further investigate the performance degradation of the minor type models, we analyzed the relationship between R^2 and the total area proportions of the LULC types (Fig. 4). R^2 increased with increasing area proportion. Since the minor LULC types occurred only sporadically over the area, a large number of pixels have zero fraction for the minor types. Therefore, the distribution of the observed fractions of minor types was right-skewed [Appendix Figure A7 (a)].

C. Relative contribution of data-processing options

Relative contributions of the data-processing options are shown in Fig. 5 and Appendix Table B5. The linear models explaining type-wise $RMSE$ by data-processing options were all significant ($p < 0.05$) except for "Barren".

For the 9 types averaged, 73.2% of the variance of the $RMSE$ was explained by predictor set (O_p; 36.3%), time interval (O_t; 19.0%) and smoothing (O_s; 17.9%), respectively.

Among the three options, O_p was of the highest contribution for "Forest", "Dry field", "Paddy field", "Urban", and "Greenhouse". "Semi natural" and "Inland water" were most attributed by O_t and "Orchard field" by O_s.

D. Marginal performance of data-processing options

Among the four predictor set options, 'Full' predictor set based scenarios achieved the best average $RMSE$ (0.054) followed by 'SR' predictor set based scenarios (0.055). Between the vegetation indices, the marginal $RMSE$ of the predictor

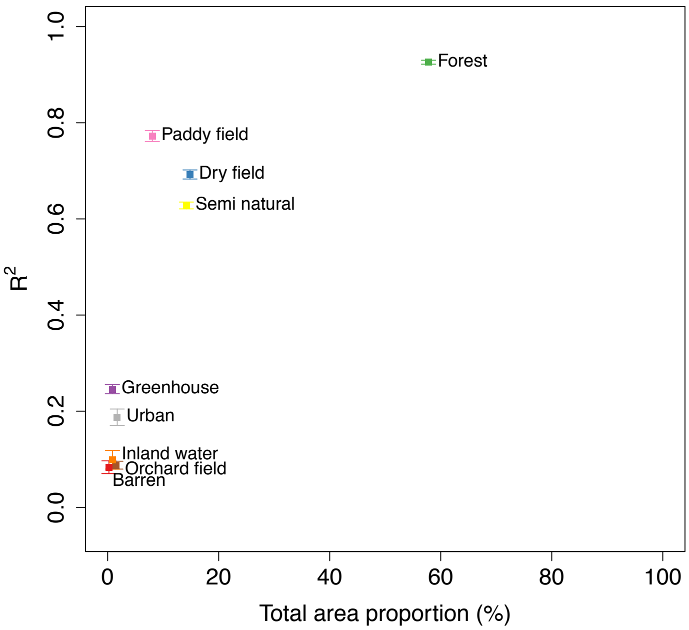

Fig. 4. Observed total area proportions of the LULC types are plotted against the avg. type-wise R^2 over all scenarios. The area proportions were calculated at the catchment level. The error bars indicate the standard errors of the means over the scenarios.

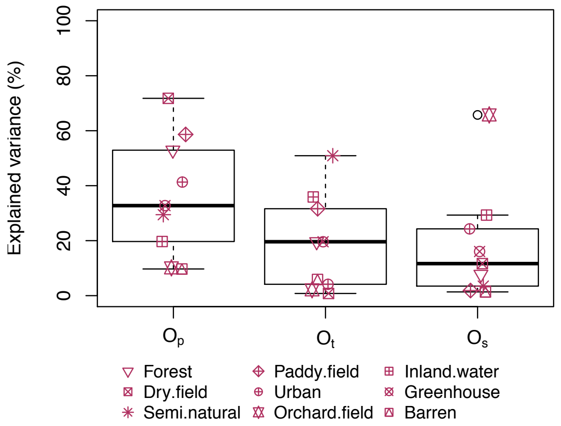

Fig. 5. Relative contribution of the data-processing options in explaining $RMSE$ in a linear regression model per type. O_p is a categorical variable denoting the chosen predictor set option, O_t time interval option, and O_s smoothing option. The relative contributions were calculated by proportional marginal variance decomposition (PMVD) [73]. The 9 points per option represent the 9 LULC types.

set 'NDVI' was smaller (0.058) compared with the 'EVI' set (0.060).

The ranks of the predictor sets varied between the LULC types (Fig. 6a). The 'Full' predictor set was the best set for 6 out of 9 types. Although, "Greenhouse" and "Barren" were best predicted by 'SR' predictor set, the differences between the predictor sets were small. The single vegetation index predictor set 'EVI' was the best predictor set for "Inland Water".

Regarding the time interval, the 8-day scenarios (avg. $RMSE$=0.056) marginally outperformed the 16-day scenarios (avg. $RMSE$=0.058) (Fig. 6c and Table III). This does not hold for the LULC types "Dry field", "Orchard field" and "Greenhouse". These types are minor types except "Dry field".

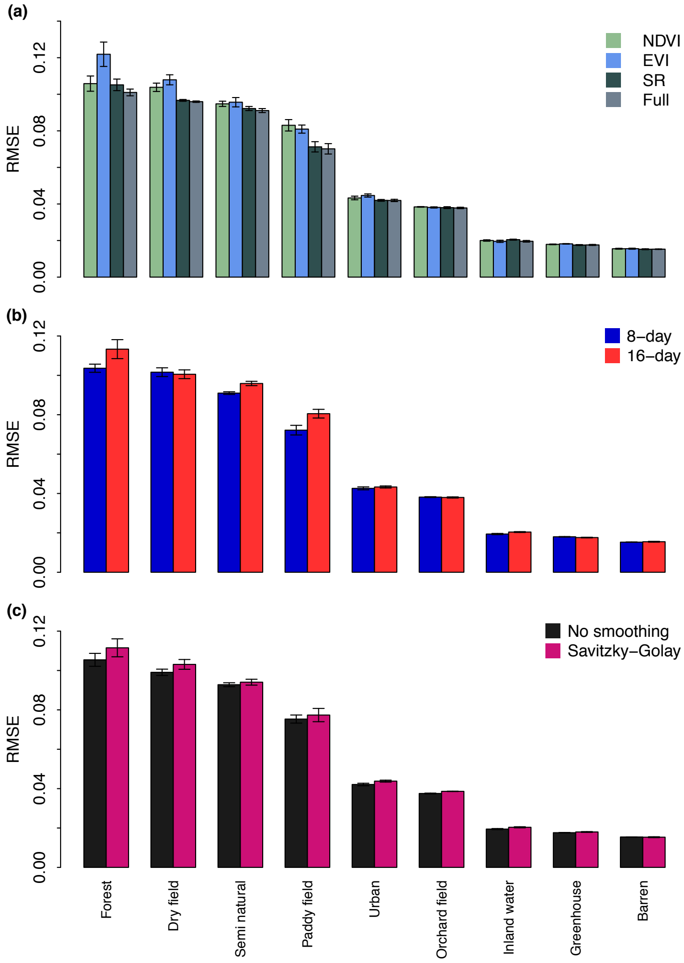

Fig. 6. Performance of the data-processing options measured by marginal $RMSE$: (a) predictor set, (b) time interval, and (c) smoothing. The cross-validated regression metrics were averaged over the other data-processing options to derive marginal performance metrics (5). The error bars indicate the standard errors of the means over the scenarios.

The scenarios with 'No smoothing' performed better (avg. $RMSE$ = 0.056) than the SG smoothed scenarios (avg. $RMSE$ = 0.058) (Fig. 6c). For the individual types, the non-smoothed predictors performed better except for "Barren" (Table III).

E. Relative importance of spectral bands

The average relative importance of the spectral bands were calculated with the 'Full' predictor set based scenarios (Fig. 7a) and 'SR' predictor set based scenarios (Fig. 7b).

Using the variable importance metric from 'SR' predictor set based scenarios, we assessed the relative importance of the four reflectance bands when used with no vegetation index (Fig. 7b). On average, the $NIMSE_b$ of B1 (48.6%) and B2 (46.9%) were substantially higher than that of B3 (2.2%) and B7 (2.3%) and made up 95.5% of the total $IMSE$ (Appendix Table B3). The two bands were almost equally important among all LULC types.

For the most dominant type “Forest”, $NIMSE_b$ of B3 (11.0%) and B7 (12.3%) were larger than that of the rarer types. However, especially for the five rarest types, B3 and B7 were negligible with less than 0.5% of $NIMSE_b$.

In ‘Full’ predictor set based scenarios, NDVI, EVI and B1 bands were similar in $NIMSE_b$ (31–33%) and made up 96.5% of the total $IMSE$ (Fig. 7a and Appendix Table B4). After including NDVI and EVI, B2 became negligible (1.3%), whereas B1 remained important (31.8%). The contribution of B3 and B7 stayed small with a $NIMSE_b$ equaled to 1.0% and 1.1%, respectively.

Only the major types such as “Forest” or “Dry field” benefited from the bands B2, B3 and B7. The $NIMSE_b$ of these three bands were smaller than 0.2% for the minor types.

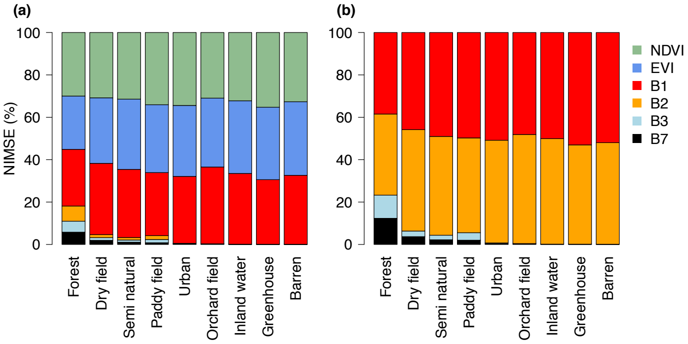

Fig. 7. Normalized increased mean square error ($NIMSE_b$) of spectral bands from (a) ‘Full’ predictor set based scenarios (S4, S8, S12, and S16) and (b) ‘SR’ predictor set based scenarios (S3, S7, S11, and S15).

F. Seasonal variation of relative importance

Fig. 8 shows seasonal variation of $IMSE_d$ by type. Both in 8-day and 16-day intervals, we observed large variable importance in the off-monsoon periods like the start and the end of the growing season. The $IMSE_d$ during the summer monsoon season around day of year (DOY) 200 were rather low for most of the LULC types, suggesting that the features representing this period were less influential on the regression performance.

In a large portion of the types, peaks are found in March (around DOY 90), which is the sowing season in the study area. Other peaks commonly occurred in September, which is the harvest season for most of the local crops (e.g., paddy rice and annual dry field crops) as well as the senescence of natural vegetation types.

The shapes of the seasonal $IMSE_d$ curves differed between the LULC types. For instance, the seasonal $IMSE_d$ of “Paddy field” showed the highest peak in September (around DOY 260) (Fig. 8d), which shows that the model is most sensitive to the harvest season. In contrast, “Forest” exhibits the highest peak in late February (around DOY 80) (Fig. 8a).

The number of major peaks of relative importance was different between types. The $IMSE_d$ of “Dry field” and “Semi natural” can be characterized as bimodal because of the two peaks around the sowing season (around DOY 60) and the harvest season (around DOY 260). However, for rarer types such as “Inland water” or “Greenhouse”, relative importance curves display multiple peaks both in 8-day and 16-day $IMSE_d$ curves.

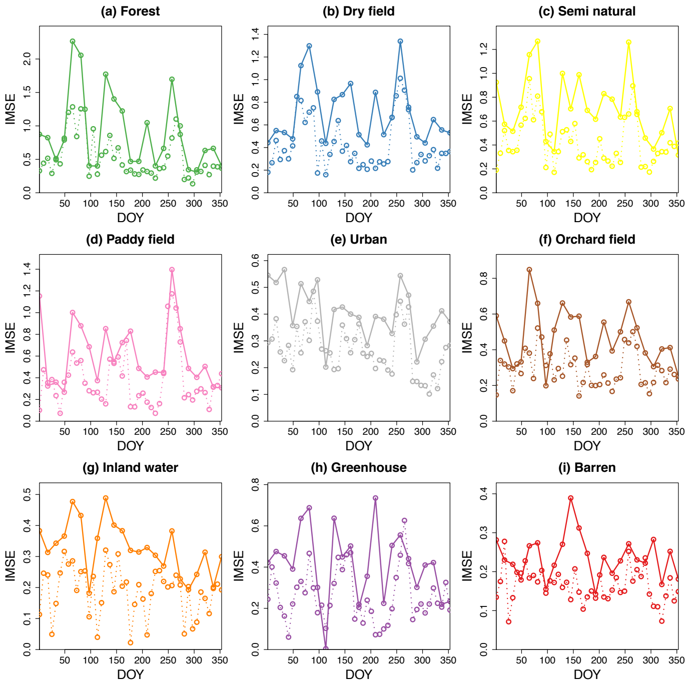

Fig. 8. Seasonal variations of increased mean square error ($IMSE_d$) are displayed to visualize relative importance of the acquisition dates; dotted line indicates the $IMSE_d$ from the 8-day data based scenarios and solid line from the 16-day data based scenarios. Note that we used only ‘Full’ predictor set based scenarios (S4, S8, S12 and S16).

G. Intrinsic dimensionality

The estimated intrinsic dimensionality (k^*) of the input spectral reflectance dataset is shown in Table IV. The ID of the all four bands together were 27 for the 8-day data and 18 for the 16-day data. The average ID of the four bands was higher for the 8-day reflectance data (avg. $k^* = 9.25$) than for the 16-day data (avg. $k^* = 6.75$). The sums of the ID of the single bands were larger than that of the ‘All’ 4-band sets both in the 8- and 16-day data.

TABLE IV
INTRINSIC DIMENSIONALITY (ID) OF THE RAW SURFACE REFLECTANCE BANDS CALCULATED BY HYSIME ALGORITHM [78], [80]. THE BAND ‘ALL’ INCLUDES THE ALL FOUR BANDS: B1, B2, B3, AND B7. THE AVG. VALUES WERE CALCULATED WITHOUT THE ‘ALL’ BANDS.

Band	ID	
	8-day	16-day
B1	10	7
B2	9	5
B3	10	10
B7	8	5
All	27	18
Avg.	9.25	6.75

IV. Discussion

A. Regression performance

The regression performance of the major type models was comparable to previously published studies. [6], for example, reported the avg. $R^2 = 0.60$ for a fractional shrub cover model using three machine learning algorithms including RF. Verbeiren *et al.* [20] confirmed that, at sub-pixel level, land cover estimation with multiple types is challenging; the avg. R^2 of the fractional cover estimation with 8 types were 0.41 using a neural network model and 0.29 using a linear mixture model. These are comparable to the R^2 of the major type models in our study (> 0.6) (Appendix Figure A6 and Appendix Table B2). [30] reported higher $R^2 (> 0.75)$ for fractional green vegetation cover estimations, however their model was not validated against ground observation and/or with cross-validation.

A regression task with multiple responses is inherently more difficult than a single-response regression. Our results are comparable to the work by [10], for example, who used a 14-class land cover system in South Africa (total accuracy = 55.0%) and Germany (total accuracy = 51.6%). Type-wise regression performance was missing in their study. Fernandes *et al.* [4] reported that the regression of fractional covers of minor types was more difficult; the average predictive R^2 was 0.57 for the two dominant types (i.e., “Conifer forest” and “Shrub”) whereas 0.33 for the three minor types (i.e., “Deciduous forest”, “Barren” and“Water”). [86] reported comparable overall accuracy (55.9%) from a fractional cover model with 6 LULC types. Note that their models were evaluated without cross-validation.

We attribute the low performance to the right-skewed distributions of LULC fractions in the training data (Appendix Figure A7a). Since the minor LULC types occurred only sporadically over the area, many pixels have zero fraction for the minor types.

When training data are skewed, a RF regression model has a limitation in prediction due to the way how regression trees are constructed. If the training data is right-skewed or even zero-inflated, the model is insufficiently trained on the high response values (e.g., high LULC fractions). As discussed in Section II-D3, the prediction is the average response of all trees. Thus, RF does not search for the best tree but averages all trees. When trained with the skewed data, it can cause an underestimation bias in prediction. O’Leary *et al.* [87] noted the same issue in RF classification when training data is imbalanced.

Our result confirm that minor types are difficult to estimate in fractional cover studies and thus need more attention. It is even more important to resolve the issues related to minor types in agricultural areas. Due to fragmented land use patterns and heterogeneities embedded in land cover classification systems (e.g., lumped cropland types), minor types are inevitably occurring in this type of landscape. To our knowledge, there were only few studies dealing with multiple LULC types in continuous land cover studies and the case studies generally suffer from poor performance regarding agricultural types (e.g., [20], [86]) and often lack appropriate model validation (e.g., [10], [27], [30]).

The regression model reproduced spatial distributions of the LULC fractions. However, predicting absolute fractions remained difficult especially for the minor types. The high ρ values of the some minor LULC types imply that the presented framework may be useful to detect minor LULC types (e.g., binary classification). As suggested by the high ρ values for the some minor LULC types, with elaborations such as the use of the Hurdle formulation or the use of data-balancing techniques, the regression performance of the minor types may be further improved. In the Hurdle model approach, first the occurrence of a desired response (e.g., LULC type) is modeled and the degree of the response is estimated for the instances passed the first ‘hurdle’. This approach may alleviate the issue of the right-skewed training data. However, the issue of the missing high response values in training data needs to be resolved independently. The Hurdle model can be used in combination with machine learning (e.g., [88], [89]) and fractional LULC regression with the Hurdle formulation would be an interesting future work.

B. Spectral unmixing analysis

Spectral unmixing analysis [29], [38] has been frequently used to derive continuous land cover as well as in many other similar disciplines [90], [91]. In this approach, mixed spectral signals are decomposed into spectral endmembers and thereby sub-pixel fractions of land cover types are estimated [7], [38], [92], [93]. In our study, we did not use the linear spectral unmixing approaches mainly for two reasons. First, unmixing approaches generally necessitate hyperspectral data instead of multi-spectral data (i.e., MODIS reflectance data) which is still deficient at the global scale especially with a short acquisition interval [7], [29], [90]. Hyperspectral data was unavailable for our study area, for example. Instead, we used the time series of the multi-spectral data to secure a large number of data points per pixel. Second, the linear unmixing approach is under the assumption that there are linear relationships between the area fractions of spectral sources (e.g., land cover types) and spectral signals (e.g., surface reflectance) [29], [30], [92], [94]. This assumption is violated when non-linear functions such as NDVI or EVI are used as predictors [92]. While there may be some difficulties regarding the spatio-temporal heterogeneities and non-linearities in the vegetation signal time series, it will be an interesting work to apply linear or non-linear unmixing methods to such multi-spectral time series data.

C. Relative importance

In our case, the information contained in the red channel (B1) was not perfectly encapsulated in the vegetation indices. This implies that we will lose some information if we use only the vegetation indices. The blue (B3) and MIR (B7) channels influenced only subtly the regression performance especially for the minor types. This contradicts our initial assumption that these bands could be useful to distinguish LULC types.

MODIS EVI utilizes an extra band B3 compared with NDVI. However, ‘EVI’ predictor set based scenarios were outperformed by 'NDVI' predictor set based scenarios as if B3 did not supply any incremental information about the vegetation activity or land cover status. It may be due to the way MODIS EVI is parametrized. In principle, the parameters in the EVI formula should be determined on-site. However, fixed parameter values are used for the MODIS EVI product for convenience. EVI may be a better predictor with site-specific calibration.

In agricultural fields, land use can be altered in a short time period leading to abruptly changed spectral signals. Therefore, it appears natural that the 8-day scenarios outperformed the 16-day scenarios. Vegetative LULC types are continuously changing within a single year. Therefore, it is difficult to capture its characteristics using satellite images from a small number of overpasses [21], [22]. Moreover, crops have a relatively short life-cycles as well as frequent human interventions, thus may not be fully characterized by a few images [25], [95]. We therefore recommend using a full time series of satellite data to model multi-type LULC data.

Additional features may further improve regression performance. For example, phenology metrics such as green-on or green-off dates are used to identify vegetation and land cover types (e.g., [8], [96]). However, costs of adding features (i.e., computing time) should be carefully considered. The intrinsic dimensionality approaches can be useful tools for such considerations [77]–[79].

D. *Best strategy*

For complex cultivated landscapes such as an agricultural mosaic catchment, appropriate data processing options should be adopted to boost LULC modeling performance. In this study we demonstrated how to evaluate and choose data-processing options based on a rigorous cross-validation scheme.

As discussed in Section III-A, the best regression performance was obtained by using the entire available predictors without the data smoothing (S4). The relative importance revealed that the most influential periods varied by LULC type. This result supports that it is important to use the whole time series of the predictors for multi-type fractional LULC modeling. Therefore, the use of a full time series and all available predictors is recommended in future applications. RF dealt well with the highly correlated input data with no evidence of overfitting given the number of predictors in our study. The relative importance and the intrinsic dimensionality of the spectral data could provide useful information guiding the selection of predictors in settings with many more predictors.

Smoothing by the Savitzky-Golay filter was disadvantageous. It suggests that the original MODIS maximum value composite algorithm already sufficiently suppressed noise in the 16-day MODIS products. We would recommend to be careful to denoise these MODIS products.

V. Conclusion

Existing global land use/land cover (LULC) raster maps have limited spatial and thematic resolution particularly unfavorable to complex agricultural landscapes. As a contribution to resolving this issue, we developed a fractional cover regression model and a strategy to set up the model with globally available satellite products. When properly chosen and processed, coarse satellite products can yield useful information at the sub-pixel level such as fractional land cover. Among the data processing options, choice of predictor sets was the most important.

In estimating absolute fraction, the model performance differed among LULC types depending on the distributions of the observed fraction data. For the minor types, predicting absolute fractions remained difficult. The monsoon period was not the most important period on the regression performance but the critical periods varied by land cover type.

Estimating fractional land cover is a useful strategy for obtaining continuous representation of LULC. It may also alleviate computational burden related to the use of high-resolution raster images. However, fractional cover estimation especially with multiple land cover types is still underdeveloped. With possible elaborations such as the Hurdle formulation, it may be possible to extract useful land cover information from coarse multi-spectral satellite products.

Our study demonstrated how to build a reliable fractional cover regression model by choosing optimal data-processing options. Our evaluation framework and findings can be a useful guide to make informed decisions in similar studies.

Appendix A
Appendix Figures

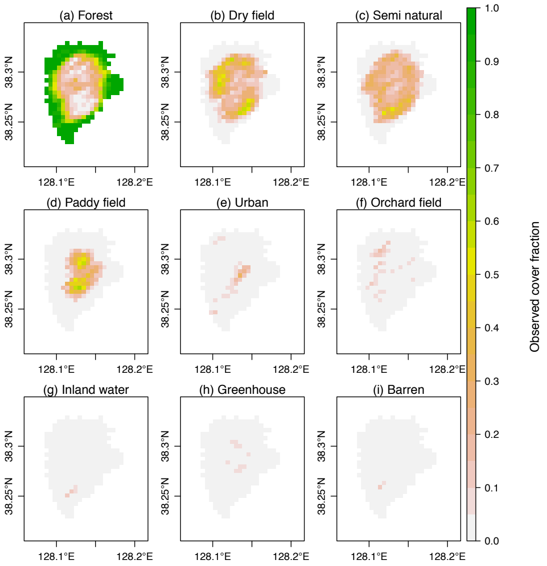

Fig. A1. The reference land use/land cover (LULC) fractions of the study site in 2010. LULC fractions were calculated from the original polygon data [37] to fit the MODIS 500 m sinusoidal grid (SR-ORG:6842) and range from 0 (0% cover) to 1 (100% cover).

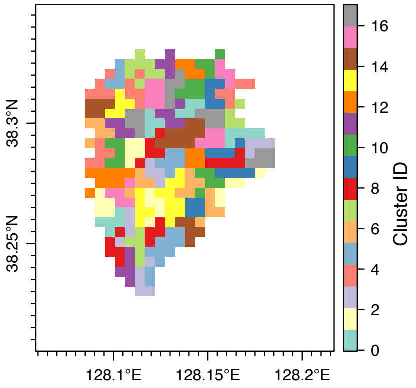

Fig. A2. Location of the 16 clusters and the 64 sub-clusters used for spatial cross-validation. Adjacent pixels in the same color indicate a sub-cluster and four of the sub-clusters comprise a cluster. In each cross-validation fold, one cluster was hold-out as test data and the rest 15 clusters trained a Random Forest regression model. The average area size of the clusters was 4.00 km^2 and the sub-clusters was 1.00 km^2.

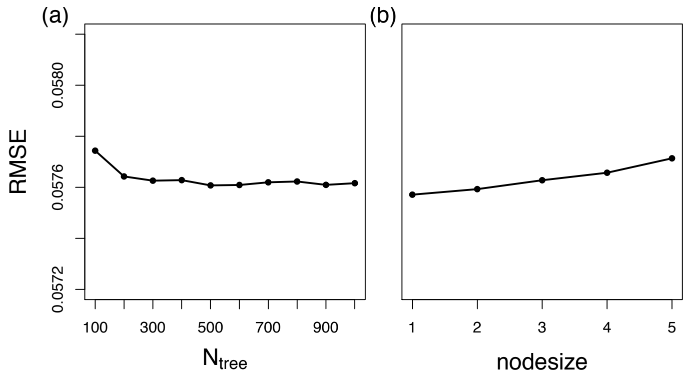

Fig. A3. Variations of $RMSE$ with changing Random Forest parameters (a) N_{tree} and (b) $nodesize$ during the parameter tuning based on the repartitioning of the training data. For illustrating the general response of the model, the avg. $RMSE$ of all scenarios and the LULC types are displayed. Note that the optimal n_{tree} and $RMSE$ were determined individually per scenario.

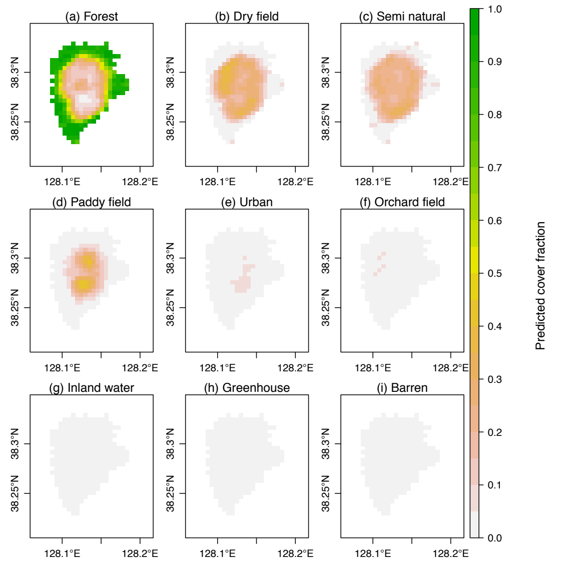

Fig. A4. Mean predicted LULC fractions of the study area. Maps from the averaged fractions over the all 16 scenarios.

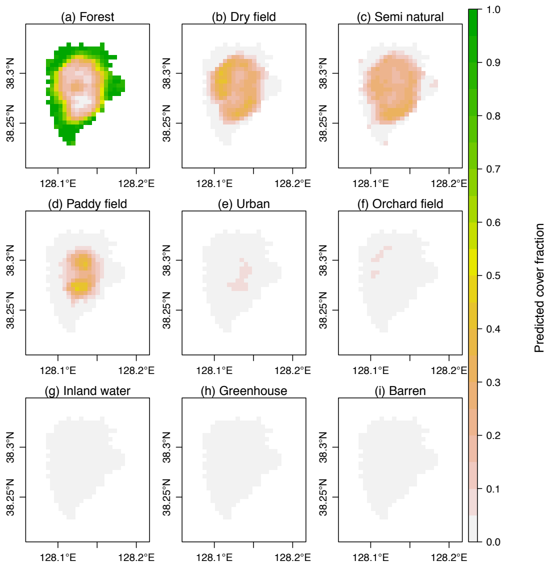

Fig. A5. Predicted LULC fractions from the best performed scenario (S4). This scenario used the non-smoothed full features in 8-day interval as predictor.

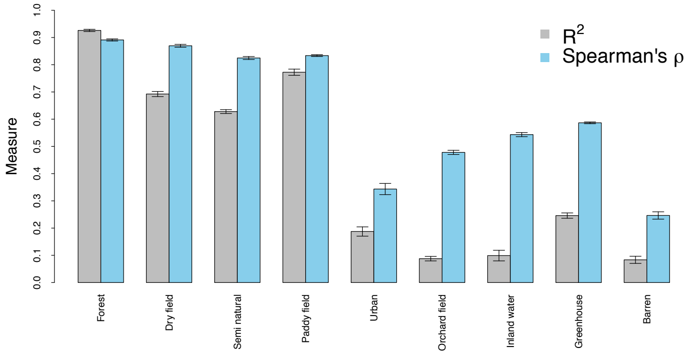

Fig. A6. R^2 and Spearman's rank correlation coefficients between observed and predicted fractions. The error bars indicate the standard errors of the means over the scenarios.

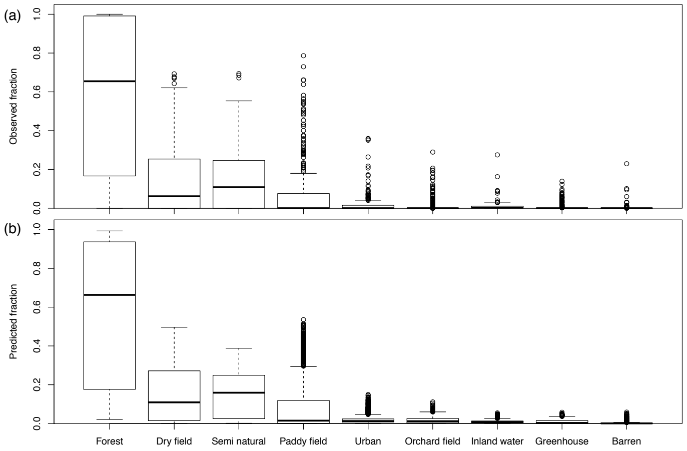

Fig. A7. Distributions of cover fractions of (a) the ground LULC observations and (b) the averaged predictions from scenarios S1 through S16.

APPENDIX B
APPENDIX TABLES

TABLE B1
SPECIFICATION OF THE SCENARIOS AND THE RANDOM FOREST TRAINING PARAMETERS. THE PARAMETERS n_{tree} AND $nodesize$ WERE TUNED AND m_{try} WAS DETERMINED BY THE SQUARE ROOT OF $n_{feature}$ [70], [97].

Name	Data-processing options					Parameters		
	Predictor set	Time interval	Smoothing	n_{band}	$n_{feature}$	n_{tree}	m_{try}	$nodesize$
S1	NDVI	8-day	None	1	46	600	6	1
S2	EVI			1	46	700	6	2
S3	SR			4	184	400	13	3
S4	Full			6	276	700	16	1
S5	NDVI	16-day		1	23	200	4	1
S6	EVI			1	23	500	4	1
S7	SR			4	92	300	9	1
S8	Full			6	138	800	11	2
S9	NDVI	8-day	SG	1	46	600	6	4
S10	EVI			1	46	500	6	1
S11	SR			4	184	500	13	1
S12	Full			6	276	400	16	1
S13	NDVI	16-day		1	23	800	4	1
S14	EVI			1	23	300	4	1
S15	SR			4	92	900	9	1
S16	Full			6	138	600	11	1

TABLE B2
TYPE-WISE PERFORMANCE MEASURES BETWEEN OBSERVED AND PREDICTED FRACTIONS AVERAGED OVER ALL SCENARIOS.

Classes	$RMSE$	ρ	R^2
Forest	0.11	0.89	0.93
Dry field	0.10	0.87	0.69
Semi natural	0.09	0.82	0.63
Paddy field	0.08	0.83	0.77
Urban	0.04	0.34	0.19
Orchard field	0.04	0.48	0.09
Inland water	0.02	0.54	0.10
Greenhouse	0.02	0.59	0.25
Barren	0.02	0.25	0.08
Avg.	0.06	0.62	0.41

TABLE B3
NORMALISED INCREASED MEAN SQUARE ERROR ($NIMSE_b$) OF THE FOUR SPECTRAL BANDS EXTRACTED FROM THE 'SR' PREDICTOR SET BASED SCENARIOS (S3, S7, S11, AND S15).

Classes	$NIMSE_b$ (%)			
	B1	B2	B3	B7
Forest	38.5	38.3	10.9	12.3
Dry field	45.8	47.9	2.6	3.7
Semi natural	49.0	46.6	2.1	2.2
Paddy field	49.7	44.7	3.5	2.0
Urban	50.8	48.4	0.4	0.4
Orchard field	48.1	51.4	0.2	0.2
Inland water	50.1	49.8	0.1	0.1
Greenhouse	53.0	46.8	0.1	0.1
Barren	52.0	47.9	0.1	0.0
Avg.	48.6	46.9	2.2	2.3

TABLE B4
$NIMSE_b$ OF THE SIX BANDS EXTRACTED FROM THE 'FULL' PREDICTOR SET BASED SCENARIOS (S4, S8, S12, AND S16).

Classes	$NIMSE_b$ (%)					
	NDVI	EVI	B1	B2	B3	B7
Forest	30.0	25.1	26.7	7.2	5.2	5.8
Dry field	30.8	30.9	33.6	1.5	1.3	1.9
Semi natural	31.4	33.1	32.1	1.2	1.0	1.1
Paddy field	34.1	32.0	29.7	1.8	1.5	0.8
Urban	34.4	33.5	31.6	0.2	0.2	0.2
Orchard field	31.0	32.5	36.2	0.1	0.1	0.1
Inland water	32.2	34.3	33.4	0.1	0.0	0.0
Greenhouse	35.3	34.2	30.5	0.0	0.0	0.0
Barren	32.7	34.7	32.5	0.0	0.0	0.0
Avg.	32.4	32.3	31.8	1.3	1.0	1.1

TABLE B5
SUMMARY OF THE LINEAR MODELS EXPLAINING THE MODEL'S $RMSE$ BY THE THREE DATA-PROCESSING OPTIONS; O_p IS A CATEGORICAL VARIABLE DENOTING THE CHOSEN PREDICTOR SET OPTION, O_t TIME INTERVAL OPTION, AND O_s SMOOTHING OPTION.

Type	Pr(>F)	Explained variance (%)		
		O_p	O_t	O_s
Forest	0.00	52.92	19.61	7.75
Dry field	0.00	71.78	0.80	11.65
Semi natural	0.00	29.45	50.90	3.47
Paddy field	0.00	58.65	31.62	1.86
Urban	0.02	41.32	4.14	24.30
Orchard field	0.00	10.10	2.11	65.74
Inland water	0.00	19.69	35.90	29.31
Greenhouse	0.02	32.75	19.61	16.06
Barren	0.83	9.70	5.90	1.36
Avg.	-	36.26	18.96	17.94

ACKNOWLEDGMENT

We are grateful to David Harter, Timothy Thrippleton, Andreas Schweiger, Wanmo Kang, Heera Lee, and John Tenhunen for providing advice and technical help.

MODIS information obtained from https://lpdaac.usgs.gov, maintained by the NASA Land Processes Distributed Active Archive Center (LP DAAC) at the USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota. 2003. Data for the image were provided by NASA.

References

[1] M. O. Smith, S. L. Ustin, J. B. Adams, and A. R. Gillespie, “Vegetation in deserts: I. a regional measure of abundance from multispectral images,” *Remote Sensing of Environment*, vol. 31, no. 1, pp. 1 – 26, 1990.

[2] J. Price, “Estimating vegetation amount from visible and near infrared reflectances,” *Remote Sensing of Environment*, vol. 41, no. 1, pp. 29–34, 1992.

[3] R. S. Defries, M. C. Hansen, and J. Townshend, “Global continuous fields of vegetation characteristics: a linear mixture model applied to multi-year 8 km AVHRR data,” *International Journal Of Remote Sensing*, vol. 21, no. 6-7, pp. 1389–1414, 2000.

[4] R. Fernandes, R. Fraser, R. Latifovic, J. Cihlar, J. Beaubien, and Y. Du, “Approaches to fractional land cover and continuous field mapping: A comparative assessment over the BOREAS study region,” *Remote Sensing of Environment*, vol. 89, no. 2, pp. 234–251, Jan. 2004.

[5] M. Schwarz and N. E. Zimmermann, “A new GLM-based method for mapping tree cover continuous fields using regional MODIS reflectance data,” *Remote Sensing of Environment*, vol. 95, no. 4, pp. 428–443, Apr. 2005.

[6] M. Schwieder, P. Leitão, S. Suess, C. Senf, and P. Hostert, “Estimating Fractional Shrub Cover Using Simulated EnMAP Data: A Comparison of Three Machine Learning Regression Techniques,” *Remote Sensing*, vol. 6, no. 4, pp. 3427–3445, Apr. 2014.

[7] J. P. Guerschman, M. J. Hill, L. J. Renzullo, D. J. Barrett, A. S. Marks, and E. J. Botha, “Estimating fractional cover of photosynthetic vegetation, non-photosynthetic vegetation and bare soil in the Australian tropical savanna region upscaling the EO-1 Hyperion and MODIS sensors,” *Remote Sensing of Environment*, vol. 113, no. 5, pp. 928–945, May 2009.

[8] K. Pittman, M. C. Hansen, I. Becker-Reshef, P. V. Potapov, and C. O. Justice, “Estimating Global Cropland Extent with Multi-year MODIS Data,” *Remote Sensing*, vol. 2, no. 7, pp. 1844–1863, Jul. 2010.

[9] M. Bevanda, N. Horning, B. Reineking, M. Heurich, M. Wegmann, and J. Mueller, “Adding structure to land cover - using fractional cover to study animal habitat use,” *Movement Ecology*, vol. 2, no. 1, p. 26, 2014.

[10] R. R. Colditz, M. Schmidt, C. Conrad, M. C. Hansen, and S. Dech, “Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions,” *Remote Sensing of Environment*, vol. 115, no. 12, pp. 3264–3275, 2011.

[11] M. Herold, P. Mayaux, C. E. Woodcock, A. Baccini, and C. Schmullius, “Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets,” *Remote Sensing of Environment*, vol. 112, no. 5, pp. 2538–2556, May 2008.

[12] E. Bartholomé and A. S. Belward, “GLC2000: a new approach to global land cover mapping from Earth observation data,” *International Journal Of Remote Sensing*, vol. 26, no. 9, pp. 1959–1977, May 2005.

[13] S. Fritz, L. See, L. You, C. Justice, I. Becker Reshef, L. Bydekerke, R. Cumani, P. Defourny, K. Erb, J. Foley, S. Gilliams, P. Gong, M. Hansen, T. Hertel, M. Herold, M. Herrero, F. Kayitakire, J. Latham, O. Leo, I. McCallum, M. Obersteiner, N. Ramankutty, J. Rocha, H. Tang, P. Thornton, C. Vancutsem, M. Velde, S. Wood, and C. Woodcock, “The Need for Improved Maps of Global Cropland,” *Eos, Transactions American Geophysical Union*, vol. 94, no. 3, pp. 31–32, Jan. 2013.

[14] B. Seo, C. Bogner, P. Poppenborg, E. Martin, M. Hoffmeister, M. Jun, T. Koellner, B. Reineking, C. L. Shope, and J. Tenhunen, “Deriving a per-field land use and land cover map in an agricultural mosaic catchment,” *Earth System Science Data*, vol. 6, pp. 339–352, 2014.

[15] B. Mora, N.-E. Tsendbazar, M. Herold, and O. Arino, “Global Land Cover Mapping: Current Status and Future Trends,” in *Land Use and Land Cover Mapping in Europe*. Dordrecht: Springer Netherlands, Jan. 2014, pp. 11–30.

[16] T. R. Loveland, B. C. Reed, J. F. Brown, D. O. Ohlen, Z. Zhu, L. Yang, and J. W. Merchant, “Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data,” *International Journal Of Remote Sensing*, vol. 21, no. 6-7, pp. 1303–1330, 2000.

[17] S. Bontemps, P. Defourny, E. Bogaert, O. Arino, V. Kalogirou, and J. Perez, “GLOBCOVER 2009 - Products Description and Validation Report,” European Space Agency, Tech. Rep., Feb. 2011.

[18] U.S. Geological Survey, “Global Land Cover Characteristics Data Base Version 2.0,” U.S. Geological Survey, Tech. Rep., 2012.

[19] C. Conrad, S. Fritsch, J. Zeidler, G. Rücker, and S. Dech, “Per-Field Irrigated Crop Classification in Arid Central Asia Using SPOT and ASTER Data,” *Remote Sensing*, vol. 2, no. 4, pp. 1035–1056, 2010.

[20] S. Verbeiren, H. Eerens, I. Piccard, I. Bauwens, and J. Van Orshoven, “Sub-pixel classification of SPOT-VEGETATION time series for the assessment of regional crop areas in Belgium,” *International Journal Of Applied Earth Observation And Geoinformation*, vol. 10, no. 4, pp. 486–497, Dec. 2008.

[21] C. Hüttich, U. Gessner, M. Herold, B. J. Strohbach, M. Schmidt, M. Keil, and S. Dech, “On the Suitability of MODIS Time Series Metrics to Map Vegetation Types in Dry Savanna Ecosystems: A Case Study in the Kalahari of NE Namibia,” *Remote Sensing*, vol. 1, no. 4, pp. 620–643, Dec. 2009.

[22] P. S. Thenkabail, M. Schull, and H. Turral, “Ganges and Indus river basin land use/land cover (LULC) and irrigated area mapping using continuous streams of MODIS data,” *Remote Sensing of Environment*, vol. 95, no. 3, pp. 317–341, Apr. 2005.

[23] J. C. Brown, J. H. Kastens, A. C. Coutinho, D. d. C. Victoria, and C. R. Bishop, “Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data,” *Remote Sensing of Environment*, vol. 130, pp. 39–50, 2013.

[24] T. W. Biggs, P. S. Thenkabail, M. K. Gumma, C. A. Scott, G. R. Parthasaradhi, and H. N. Turral, “Irrigated area mapping in heterogeneous landscapes with MODIS time series, ground truth and census data, Krishna Basin, India,” *International Journal Of Remote Sensing*, vol. 27, no. 19, pp. 4245–4266, Oct. 2006.

[25] M. K. Gumma, P. S. Thenkabail, and A. Nelson, “Mapping Irrigated Areas Using MODIS 250 Meter Time-Series Data: A Study on Krishna River Basin (India),” *Water*, vol. 3, no. 1, pp. 113–131, Mar. 2011.

[26] H. Lu, M. R. Raupach, T. McVicar, and D. J. Barrett, “Decomposition of vegetation cover into woody and herbaceous components using AVHRR NDVI time series,” *Remote Sensing of Environment*, vol. 86, no. 1, pp. 1–18, 2003.

[27] B. Johnson, R. Tateishi, and T. Kobayashi, “Remote Sensing of Fractional Green Vegetation Cover Using Spatially-Interpolated Endmembers,” *Remote Sensing*, vol. 4, no. 12, pp. 2619–2634, Dec. 2012.

[28] R. S. DeFries, C. B. Field, I. Fung, C. O. Justice, S. Los, P. A. Matson, E. Matthews, H. A. Mooney, C. S. Potter, K. Prentice *et al.*, “Mapping the land surface for global atmosphere-biosphere models: Toward continuous distributions of vegetation’s functional properties,” *Journal of Geophysical Research: Atmospheres (1984–2012)*, vol. 100, no. D10, pp. 20 867–20 882, 1995.

[29] G. Asner and D. B. Lobell, “A biogeophysical approach for automated SWIR unmixing of soils and vegetation,” *Remote Sensing of Environment*, vol. 74, no. 1, pp. 99–112, 2000.

[30] J. Xiao and A. Moody, “A comparison of methods for estimating fractional green vegetation cover within a desert-to-upland transition zone in central New Mexico, USA,” *Remote Sensing of Environment*, vol. 98, no. 2-3, pp. 237–250, Oct. 2005.

[31] X. Zhang, C. Liao, J. Li, and Q. Sun, “Fractional vegetation cover estimation in arid and semi-arid environments using HJ-1 satellite hyperspectral data,” *International Journal Of Applied Earth Observation And Geoinformation*, vol. 21, pp. 506—512, 2013.

[32] M. L. Clark, T. M. Aide, H. R. Grau, and G. Riner, “A scalable approach to mapping annual land cover at 250 m using MODIS time series data: A case study in the Dry Chaco ecoregion of South America,” *Remote Sensing of Environment*, vol. 114, no. 11, pp. 2816–2832, 2010.

[33] R. Thackway, L. Lymburner, and J. P. Guerschman, “Dynamic land cover information: bridging the gap between remote sensing and natural resource management,” *Ecology And Society*, vol. 18, no. 1, 2013.

[34] D. Yihui and J. C. L. Chan, “The East Asian summer monsoon: an overview,” *Meteorology and Atmospheric Physics*, vol. 89, no. 1-4, pp. 117–142, Jun. 2005.

[35] M. Kang, S. Park, H. Kwon, H. T. Choi, Y. J. Choi, and J. Kim, “Evapotranspiration from a deciduous forest in a complex terrain and a heterogeneous farmland under monsoon climate,” *Asia-Pacific Journal of Atmospheric Sciences*, vol. 45, no. 2, pp. 175–191, 2009.

[36] A. Huete, C. Justice, and W. Van Leeuwen, “MODIS Vegetation Index (MOD 13): Algorithm Theoretical Basis Document,” Tech. Rep., 1999.

[37] B. Seo, P. Poppenborg, E. Martin, M. Hoffmeister, C. Bogner, H. Elsayed Ali, B. Reineking, and J. Tenhunen, “Per-field land use and land cover data set of the haean catchment, south korea. doi:10.1594/pangaea.823677,” 2014, supplement to: Seo, B.; Bogner, C.; Poppenborg, P. Martin, E.; Hoffmeister, M.; Jun, M. Koellner, T.; Reineking, B.; Shope, C.L.; Tenhunen, J. (2014): Deriving a per-field land use and land cover map in an agricultural mosaic catchment. Earth System Science Data, 6, 339-352.

[38] K. Obata, T. Miura, and H. Yoshioka, "Analysis of the Scaling Effects in the Area-Averaged Fraction of Vegetation Cover Retrieved Using an NDVI-Isoline-Based Linear Mixture Model," *Remote Sensing*, vol. 4, no. 7, pp. 2156–2180, Jul. 2012.

[39] R. E. Wolfe, M. Nishihama, and J. R. Kuyper, "Improving Satellite Moderate Resolution Instrument Geolocation Accuracy in Rough Terrain," in *Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on*. IEEE, Jul. 2006, pp. 1123–1125.

[40] J. D. Watts, S. L. Powell, R. L. Lawrence, and T. Hilker, "Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery," *Remote Sensing of Environment*, vol. 115, no. 1, pp. 66–75, Sep. 2010.

[41] M. Vittek, A. Brink, F. Donnay, D. Simonetti, and B. Desclée, "Land Cover Change Monitoring Using Landsat MSS/TM Satellite Image Data over West Africa between 1975 and 1990," *Remote Sensing*, vol. 6, no. 1, pp. 658–676, Jan. 2014.

[42] NASA Land Processes Distributed Active Archive Center (LP DAAC), "Mod13a1 vegetation indices 16-day l3 global 500m," 47914 252nd Street, Sioux Falls, South Dakota, Tech. Rep., 12 2013.

[43] K. Didan and A. Huete, "MODIS vegetation index product series collection 5 change summary," *Terrestrial Biophysics and Remote Sensing (TBRS) laboratory, The University of Arizona June*, vol. 29, p. 2006, 2006.

[44] R. Solano, K. Didan, A. Jacobson, and A. Huete, "MODIS vegetation indices (MOD13) C5 user's guide," Tech. Rep., 2010.

[45] A. Savitzky and M. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical chemistry*, vol. 36, no. 8, pp. 1627—1639, 1964.

[46] F. Fontana, C. Rixen, T. Jonas, G. Aberegg, and S. Wunderle, "Alpine grassland phenology as seen in AVHRR, VEGETATION, and MODIS NDVI time series - a comparison with in situ measurements," *Sensors*, vol. 8, no. 4, pp. 2833–2853, 2008.

[47] J. N. Hird and G. J. McDermid, "Noise reduction of NDVI time series: An empirical comparison of selected techniques," *Remote Sensing of Environment*, vol. 113, no. 1, pp. 248–258, Jan. 2009.

[48] P. Jonsson and L. Eklundh, "TIMESAT-a program for analyzing time-series of satellite sensor data," *Computers & Geosciences*, vol. 30, no. 8, pp. 833–845, 2004.

[49] L. Eklundh and P. Jönsson, "TIMESAT 3.1 Software Manual," Lund University and Malmö University, Tech. Rep., 2012.

[50] M. R. Segal, "Machine learning benchmarks and random forest regression," Center for Bioinformatics Molecular Biostatistics, Tech. Rep., 2004.

[51] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 1–21, Sep. 2006.

[52] G. M. Foody and M. K. Arora, "Incorporating mixed pixels in the training, allocation and testing stages of supervised classifications," *Pattern Recognition Letters*, vol. 17, no. 13, pp. 1389–1398, 1996.

[53] D. K. McIver and M. A. Friedl, "Estimating pixel-scale land cover classification confidence using nonparametric machine learning methods," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 9, pp. 1959–1968, Sep. 2001.

[54] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[55] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, Mar. 2006.

[56] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random Forests for land cover classification," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, Mar. 2006.

[57] S. Attarchi and R. Gloaguen, "Classifying complex mountainous forests with l-band sar and landsat data integration: A comparison among different machine learning methods in the hyrcanian forest," *Remote Sensing*, vol. 6, no. 5, pp. 3624–3647, 2014.

[58] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, Jan. 2014.

[59] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[60] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93–104, 2012.

[61] I. Nitze, B. Barrett, and F. Cawkwell, "Temporal optimisation of image acquisition for land cover classification with Random Forest and MODIS time-series," *International Journal Of Applied Earth Observation And Geoinformation*, vol. 34, pp. 136–146, 2015.

[62] M. Immitzer, C. Atzberger, and T. Koukal, "Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data," *Remote Sensing*, vol. 4, no. 12, pp. 2661–2693, Sep. 2012.

[63] C. Senf, D. Pflugmacher, S. van der Linden, and P. Hostert, "Mapping Rubber Plantations and Natural Forests in Xishuangbanna (Southwest China) Using Multi-Spectral Phenological Metrics from MODIS Time Series," *Remote Sensing*, vol. 5, no. 6, pp. 2795–2812, May 2013.

[64] I. Nitze, U. Schulthess, and H. Asche, "Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification," *Proc of the 4th GEOBIA*, 2012.

[65] B. Ghimire, J. Rogan, and J. Miller, "Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic," *Remote Sensing Letters*, vol. 1, no. 1, pp. 45–54, 2010.

[66] A. Brenning, "Spatial prediction models for landslide hazards: review, comparison and evaluation," *Natural Hazards and Earth System Sciences*, vol. 5, no. 6, pp. 853–862, 2005.

[67] B. Reineking, P. Weibel, M. Conedera, and H. Bugmann, "Environmental determinants of lightning- v.human-induced forest fire ignitions differ in a temperate mountain region of Switzerland," *International Journal of Wildland Fire*, vol. 19, no. 5, pp. 541–557, 2010.

[68] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests." *Bmc Bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[69] B. F. Leutner, B. Reineking, J. Müller, M. Bachmann, C. Beierkuhnlein, S. Dech, and M. Wegmann, "Modelling Forest α-Diversity and Floristic Composition — On the Added Value of LiDAR plus Hyperspectral Remote Sensing," *Remote Sensing*, vol. 4, no. 12, pp. 2818–2845, Dec. 2012.

[70] M. L. Clark and D. A. Roberts, "Species-Level Differences in Hyperspectral Metrics among Tropical Rainforest Trees as Determined by a Tree-Based Classifier," *Remote Sensing*, vol. 4, no. 12, pp. 1820–1855, Dec. 2012.

[71] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference, Fourth Edition*, ser. Revised and Expanded. Marcel Dekker, May 2003.

[72] U. Grömping, "Relative importance for linear regression in r: The package relaimpo," *Journal of Statistical Software*, vol. 17, no. 1, pp. 1–27, 2006.

[73] B. E. Feldman, "Relative Importance and Value," *SSRN Electronic Journal*, 2005.

[74] M. Segal and Y. Xiao, "Multivariate random forests," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 80–87, 2011.

[75] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution." *Bmc Bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[76] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[77] C. I. Chang and Q. Du, "Estimation of Number of Spectrally Distinct Signal Sources in Hyperspectral Imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 3, pp. 608–619, Mar. 2004.

[78] J. Bioucas-Dias and J. Nascimento, "Hyperspectral subspace identification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 46, no. 8, pp. 2435–2445, Aug 2008.

[79] K. Cawse-Nicholson, S. B. Damelin, A. Robin, and M. Sears, "Determining the Intrinsic Dimension of a Hyperspectral Image Using Random Matrix Theory," *Image Processing, IEEE Transactions on*, vol. 22, no. 4, pp. 1301–1310, Dec. 2012.

[80] L. I. Jiménez and A. Plaza, "HyperMix: An Open-Source Tool for Fast Spectral Unmixing on Graphics Processing Units," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 9, pp. 1883–1887, 2015.

[81] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.

[82] A. Liaw, *randomForest: Breiman and Cutler's random forests for classification and regression*, 4th ed., Oct. 2012.

[83] R. J. Hijmans, *raster: raster: Geographic data analysis and modeling*, 2014.

[84] GEOS Development Team, *GEOS - Geometry Engine, Open Source*, Open Source Geospatial Foundation, 2014.

[85] R. Bivand and C. Rundel, *rgeos: Interface to Geometry Engine - Open Source (GEOS)*, 2014.

[86] P. Dennison and D. Roberts, "Endmember selection for multiple endmember spectral mixture analysis using endmember average RMSE," *Remote Sensing of Environment*, vol. 87, pp. 123–135, 2003.

[87] R. A. O'Leary, R. W. Francis, K. W. Carter, M. J. Firth, U. R. Kees, and N. H. de Klerk, "A comparison of Bayesian classification trees and random forest to identify classifiers for childhood leukaemia," *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation*, pp. 4276–4282, 2009.

[88] N. A. Povak, P. F. Hessburg, K. M. Reynolds, T. J. Sullivan, T. C. McDonnell, and R. B. Salter, "Machine learning and hurdle models for improving regional predictions of stream water acid neutralizing capacity," *Water Resources Research*, vol. 49, no. 6, pp. 3531–3546, Jun. 2013.

[89] D. J. Lieske, D. A. Fifield, and C. Gjerdrum, "Maps, models, and marine vulnerability: Assessing the community distribution of seabirds at-sea," *Biological Conservation*, vol. 172, no. C, pp. 15–28, Apr. 2014.

[90] I. Dopido, A. Villa, A. Plaza, and P. Gamba, "A Quantitative and Comparative Assessment of Unmixing-Based Feature Extraction Techniques for Hyperspectral Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 421–435, Dec. 2011.

[91] I. Dopido, A. Villa, and A. Plaza, "Unsupervised clustering and spectral unmixing for feature extraction prior to supervised classification of hyperspectral images," in *SPIE Optical Engineering + Applications*. SPIE, Jun. 2012, pp. 81 570M–8.

[92] D. B. Lobell and G. Asner, "Cropland distributions from temporal unmixing of MODIS data," *Remote Sensing of Environment*, vol. 93, no. 3, pp. 412–422, 2004.

[93] J. C. Jiménez-Muñoz, J. A. Sobrino, A. Plaza, L. Guanter, J. Moreno, and P. Martinez, "Comparison Between Fractional Vegetation Cover Retrievals from Vegetation Indices and Spectral Mixture Analysis: Case Study of PROBA/CHRIS Data Over an Agricultural Area," *Sensors*, vol. 9, no. 2, pp. 768–793, Feb. 2009.

[94] L. I. Jiménez, G. Martin, and A. Plaza, "A new tool for evaluating spectral unmixing applications for remotely sensed hyperspectral image analysis," *Proc Int Conf GEOBIA*, 2012.

[95] L. Li, M. Friedl, Q. Xin, J. Gray, Y. Pan, and S. Frolking, "Mapping Crop Cycles in China Using MODIS-EVI Time Series," *Remote Sensing*, vol. 6, no. 3, pp. 2473–2493, Mar. 2014.

[96] L. Lu, C. Kuenzer, H. Guo, Q. Li, T. Long, and X. Li, "A Novel Land Cover Classification Map Based on a MODIS Time-Series in Xinjiang, China," *Remote Sensing*, vol. 6, no. 4, pp. 3387–3408, Apr. 2014.

[97] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, p. 51, Jul. 2011.

Bumsuk Seo Bumsuk Seo received the B.E. and the M.C.P. degrees from the Seoul National University, Republic of Korea and the PhD degree in Biogeographical modelling from the University of Bayreuth, Germany, in 2015. Since 2015 he is senior researcher at the Department of Environmental Science, Kangwon National Univeristy, Republic of Korea. His research interests include land use and land cover (LULC) dynamics, especially in complex heterogeneous ecosystems. He frequently uses remote sensing and machine learning techniques to address problems in LULC modeling.

Christina Bogner Christina Bogner received her "DEA" (M.S.) in Hydrology from the University of Avignon and her "magisère" in Geosciences from the Ecole Normale Supérieure in Paris (France) in 2004. She earned her bilateral PhD (*summa cum laude*) in Soil Physics and Hydrogeology from the University of Bayreuth (Germany) and the University of Avignon (France) in 2009.

She is currently a Research Associate at the Department of Ecological Modelling at the University of Bayreuth (Germany). Her main research interests are the analysis of land use changes and their impacts on soils. In particular she combines time series analysis from remote sensing with proximal sensing of soils and ground measurements of soil moisture and erosion.

Thomas Koellner Thomas Koellner received the PhD in Economics in 2001 from the Institute for Economy and the Environment, University St. Gall, Switzerland and a Diploma in biology from the University Göttingen. He has since 2009 the W2-Professorship of Ecological Services (PES) at the University of Bayreuth. Since 2010 he holds the Venia legendi in Human-Environment Systems at the Department of Environmental Sciences, ETH Zurich. Between 2001 and 2009 he was senior researcher and lecturer at the Department for Environmental Sciences, ETH Zurich. He has published many scientific papers in the fields of industrial ecology, ecological economics and environmental finance. To address the global problem of ecosystem degradation and biodiversity loss the three research lines of his research group are A) Regional models of ecosystem services in human-environment systems; B) Global climate change, biodiversity and ecological services; C) Global markets and ecosystem services.

Björn Reineking Björn Reineking received the PhD degree in environmental sciences from ETH Zurich, Switzerland, in 2005. Since 2013 he is research director at Irstea, National Research Institute of Science and Technology for Environment and Agriculture, in Grenoble, France. He is a quantitative ecologist interested in anthropogenic impacts on the structure and dynamics of ecological systems. In his research, he frequently uses remote sensing techniques.