Synthesis of a Complete Land Use/Land Cover Data Set for the Conterminous United States

N. A. Best[*] J. W. Elliott[†] and I. T. Foster[‡]

January 12, 2012

Abstract

The PEEL_0 land cover data set for the conterminous USA characterizes each of its five arc-minute ($5'$) cells in terms of sub-pixel area fractions for fifteen land use and natural cover classes, with the fractions for each cell summing to unity. This dataset has three advantages relative to other products. First, its cover classes address distinctions important in studies of the economic dimensions of land use and land cover change, by for example distinguishing between land uses associated with human and natural processes and among different crops while simultaneously providing a *complete* representation of cover in each cell. Second, aggregates for the various cover classes in PEEL_0 compare more favorably with national level statistics from the USDA Major Land Uses (MLU) census data for 2002 than do other sources. Aggregate cultivated land is within 0.8% of the MLU value, as compared to 16.2% for the Modis Land Cover Type (MLCT) primary cover product, 1.8% for the National Land-Cover Database (NLCD), and 1.2% for the Agricultural Lands in the Year 2000 dataset. Aggregate water, natural, and urban cover classes are within 2.2, 1.0, and 6.1% resp. as compared to deviations of 24.0, 0.01, and 69.2% for MLCT primary and 2.2, 1.35, and 6.1% for NLCD. Third, the spatial distribution of cultivated land is also substantially improved; PEEL_0's per-cell sub-pixel fraction root mean square error for cropland relative to the NLCD is 0.149 versus 0.175 for the 2001 MLCT primary classification, a 16% improvement in RMSE. PEEL_0's improved performance is due to the *multi-source guided aggregation-decomposition* method used to construct this dataset. This method combines information from multiple sources, including spatial data sets and agricultural production statistics, to guide both *aggregation* of multiple fine-scale land use/land cover classifications, and *decomposition* of hybrid classes to separate their constitutive land uses and natural covers. We describe how we use this method to construct PEEL_0 by incorporating information from the 2001 MODIS Land Cover Type data set, the 2001 NLCD, and the Agricultural Lands in the Year 2000 data set. In the case of the MODIS data set we demonstrate that considering all of its information components (primary classification, secondary classification, and confidence level) leads to more accurate aggregate measures of cultivated acreage. We also describe our evaluation process. This procedure is adaptable to other regions, data sets, and requirements.

[*] Computation Institute, University of Chicago and Argonne National Laboratory, email: nbest@ci.uchicago.edu
[†] Computation Institute, University of Chicago and Argonne National Laboratory,
[‡] Computation Institute, University of Chicago and Argonne National Laboratory,

Contents

1 Introduction

Understanding the factors that drive land use change and developing better methods for projecting future land use change are vitally important problems in the context of both climate change and agricultural production. The 3rd Assessment Report of the UN Intergovernmental Panel on Climate Change (IPCC) noted that "the emissions scenarios considered in future climate change studies need to integrate high resolution representations of land use change," and that increased coupling among the various relevant components, such as mitigation and adaptation responses to climate change, and climate response to land use, should be included in a consistent framework for integrated assessment [*Jones et al.*, 2001]. For these and other reasons, models that integrate physical and socio-economic factors at resolutions that resolve important variations in the spatial and temporal patterns of land use decisions, environmental conditions, and climate impacts are necessary. Such models can assist policy and decision makers to evaluate the life cycle impacts of programs intended to, for example, subsidize the production of biofuels on industrial scales; encourage the adoption of sustainable farming practices that will increase carbon uptake in the biosphere; or incentivize research and development to both increase crop yields and decrease fertilizer use while anticipating shifts in the distributions of particular crops in response to climate change. Relevant projects include IMAGE [*van Vuuren et al.*, 2006, 2007], AIM [*Matsuoka et al.*, 1995], and, at the University of Chicago, the Partial Equilibrium Economic Land-use (PEEL) model.

Such modeling can benefit greatly from accurate, high-resolution, and global characterizations of current and past land use. Many projects have leveraged a variety of information sources to develop such characterizations for different purposes, regions, and geographical and temporal scales. The emergence of high-resolution multi-spectral satellite-borne instruments has transformed the field, enabling the construction of such products as the MODIS Land Cover Type v005 (MLCT) dataset, which provides a 15-arc-second ($15''$) resolution characterization of the entire globe, with each pixel characterized by a primary class, secondary class, and confidence level for the primary class assignment.

The MLCT product is constructed by applying an automated decision-making algorithm to satellite sensor data. The algorithm is trained by spatially sparse ground truth data sets that provide exemplars for the various classification schemes. This approach enables higher resolution and accuracy than approaches based on, for example, census data. However, it also has limitations when it comes to specific applications. For example, an investigator studying economic factors driving land use change may need to know what area is used in different regions for different crops—information that MLCT, with its single "cropland" class, does not provide. It is feasible, in principle, to differentiate between different crops, and indeed that information is available in other datasets, such as 175Crops2000 or SPAM, even the 2001 National Land Cover Database (NLCD) to a lesser degree. However 175Crops2000 and SPAM are completely silent on non-agricultural use and cover types and the NLCD is the product of a piecewise synthesis that required extensive manipulation to match the edges of its mapping zones and therefore certain classes in certain areas have systematic idiosyncracies and cannot be considered spatially homogeneous in its formulation. Indeed, every dataset, however produced, will inevitably exhibit limitations for many purposes by virtue of the design of the classification scheme and algorithm that was used to produce it.

This situation was understandable when dataset sizes and associated computational requirements meant that the production of a new dataset was extremely expensive. However, rapid increases in storage capacity and computational power mean that it is now feasible for even small teams to produce their own datasets. Thus, we suggest that the production of land use datasets should become a commonplace activity. An investigator requiring data of a certain type should be able to synthesize the required dataset via the application of a specified set of transformations to data obtained from various sources. These transformations may include, for example, the aggregation or disaggregation of data in order to convert between resolutions; the reclassification of pixels, either through a simple reassignment or more complex logic; the decomposition of compound classes into their constitutive classes; and guided corrections in order to satisfy global or regional constraints implied by other datasets, such as an agricultural census. Regardless of the method, what is important is that it is well-defined, reproducible, and suitable for the purpose at hand.

Such considerations led us to produce our own custom dataset, PEEL$_0$, which we produce via a multi-step process that uses four distinct processes to combine information from three different data sets, MLCT, NLCD, and 175Crops2000:

Reclassification: Because we are dealing with LULC classifications based on different sets of class defintions we seek to find a common basis on which to compare their contents. This is done by defining a simplified set of LULC classes and mapping the classes in the original data sets to them. This simplified classification is suitable for our economic land use change forecast models because subtle distinctions between ecological roles of different types of forests, for example.

Aggregation: For purposes of this discussion aggregation simply refers to moving to a lower resolution raster representation while preserving as much information about the composition of the pixels as possible. This includes translating categorized pixels directly to fractional areas for the larger grid cell in which they fall, or calulating sub-pixel fractions for pixels whose classification is less than certain and an alternative is offered, which we will treat as a mixture, and then upscaling to the coarser resolution. Although this results in a loss of locational precision it preserves more information than a more naive aggregation, such as accepting the class that occurs most frequently, i.e. the mode, as the representative class.

Decomposition: The need for decmposition arises when a class defintion is a hybrid of the fundamental types that we have chosen for our simplified classification. This occurs in two cases in our study. First, MLCT features a "cropland / natural vegetation mosaic" class that we wish to unpack in order to assign a portion of its area to agricultural production and the rest to natural covers. Upon deriving a more complete estimate of total cropland area, cropland is further decomposed into crop sub-classes according to local farming practices. Decomposition is distinguished from disaggregation, or downscaling, by holding the spatial resolution constant and making no attempt to impart additional spatial precision to sub-cell distributions of LULC phenomena.

Correction: By selectively incorporating aspects of higher-resolution data sets whose complete representation is unsatisfactory for some reason, but has an advantage in resolving particular classes by virute of resolution and ancillary data used as prior probabilities in its classification algorithm, we are able to make adjustments to the landscape composition indicated by the data set chosen as the foundation for this method. These corrections are primarily motivated by a recognition of biases against finely detailed features that get washed out by the coarser resolutions of the starting data set's characterization.

We restrict our construction of PEEL$_0$ to the USA because it is in that region that we have access to the NLCD, 175Crops2000, and AgLand2000 data sets that we leverage to improve on the global MLCT product. We further restrict our analysis to the conterminous US because of the relative insignificance of agricultural activity in Hawaii and Alaska. However, similar methods can be applied in other regions, given suitable data sources. The technique presented in Section 4 can then be applied globally and also extended in time to convert the proceeding years of the MLCT time series to a form useful to various analyses including our own LULC analysis.

In the sections that follow, we describe in greater detail the data sets that we use to construct and evaluate PEEL$_0$ (Section 2); the algorithm that applies the procedural concepts outlined above (Section 3) to merge information derived from both MLCT and NLCD; and an evaluation process (Section 4). We summarize our computational tools in Section 5 and conclude in Section 6 with a discussion of the merits of this endeavor and of future avenues of research based thereon.

We next describe a procedure for combining information from the data sets described in Section 2 using the same sub-pixel analysis data structure at $5'$ resolution for cUSA to produce a data set that exhibits high accuracy in the distribution of agricultural production according to both NLCD and Agland2000; provides a realistic characterization of other uses and covers as suggested by MLCT; achieves agreement with particular fine-scale aspects of NLCD; and is thus able to resolve crop types using the 175Crops2000 dataset.

In Section 3.3 we describe a method for selectively incorporating cover fractions for particular classes from NLCD due to a perceived underestimation of those classes by MLCT due primarily to its lower resolution. Those classes are water, wetland, and urban, occasionally abbreviated as WWU in what follows. In the PEEL$_0$ classification we have broadened the "urban" class relative to the MLCT/IGBP definition to include rural transportation infrastructure, small towns, farmsteads, etc. thereby including lower-density development. Accepting NLCD's quantification of these classes as truth is intended to counteract a perceived overestimation of cropland area in MLCT caused, at least in part, by the fact that substantial ancillary land that may be associated with cultivated

Data set	Sensor	Resolution	Time Span
UMD Global Land Cover 1998 [a]	AVHRR	1km	1981 – 1994 (composite)
Global Land Cover 2000 [b]	SPOT	1km	Nov 1999 – Dec 2000 (composite)
National Land Cover Database [c]	Landsat	30 m	2001
MODIS Land Cover Type v005 [d]	MODIS (Aqua & Terra)	500m	2001 – 2008 (annual)

[a] *Hansen et al.* [2000]

[b] *Bartholomé and Belward* [2005]

[c] *Homer et al.* [2004, 2007]

[d] *Friedl et al.* [2010], *LP DAAC* [2008]

Table 1: Summary of global LULC data sets.

land but is not directly involved in crop production is included in the MLCT crop class. The result is an adjusted version of MLCT as amended by these NLCD offsets.

Finally Section 3.5 shows how information from the 175Crops2000 data set is used to decompose the cropland layer in order to characterize the distribution of production of major crop commodities, corn (maize), soybean, wheat, rice, sugarcane, other cereals, and other field crops.

Throughout, we use decreasing root mean squared error (RMSE) figures and correlation hexbin plots to show that our complete characterization of the landscape is improving in accuracy with respect to both the NLCD and Agland2000 datasets.

2 Land Use and Land Cover Datasets

Recent years have seen a significant increase in the availability of global land cover data sets including the University of Maryland Global Land Cover Classification, Global Land Cover 2000 (GLC2000), and MODIS Land Cover Type (MLCT). At the regional level the National Land cover Database (NLCD) provides high-resolution LULC data for the United States and Puerto Rico. We summarize several such data sets in Table 1. The proliferation of these data sets reflects the diversification and technological advances among space-borne sensors in recent years, resulting in improved resolution, both spatial and temporal, as well as innovation in post-processing and classification algorithms that transform raw sensor data into the thematic data that is readily applicable to modeling.

There has also been a proliferation of data sets that describe the distribution and intensity of global agricultural activity. Some such as the Global Irrigated Areas Map (GIAM) [*Thenkabail et al.*, 2008] and the Global Map of Rainfed Crop Areas (GMRCA) [*Biradar et al.*, 2009] are the product of applying classification techniques to large collections of remote sensing and GIS data. Others such as Agricultural Land in the Year 2000 (Agland200) [*Ramankutty et al.*, 2008], Harvested Area and Yields of 175 Crops (175Crops2000) [*Monfreda et al.*, 2008], and the Spatial Production Allocation Model (SPAM) [*You et al.*, 2006] are further informed by agricultural production data published at national and sub-national levels and disaggregated to grid cells within those boundaries according to an optimization method described by *You and Wood* [2006]. Such data sets can complement those of the general comprehensive LULC category by offering additional information on how to differentiate areas of cropland according to cultivars, and farming practices such as crop rotation, multiple cropping, and irrigation.

We use four data sets in this work: three (MLCT, NLCd, and 175Crops2000) to construct PEEL_0, and one (AgLand2000) to evaluate the quality of the PEEL_0 product. We describe these four data sets here.

The **2001 MODIS Land Cover Type v005** (MLCT) dataset is a 500m ($\sim 15''$) resolution land cover dataset. We base our method on MLCT because of its global coverage, its annual time series, and its free availability. In the future we plan to extend PEEL_0 in both time and space, so MLCT is the clear choice as a foundation data set. MLCT provides three data values for each pixel: a primary classification, a percentage measure of classification confidence, and a secondary classification. Among other alternatives MLCT classifies its pixels according to the International Geosphere-Biosphere Programme (IGBP) classification scheme [*Friedl*, 2002]. The confidence level is intended as a measure of the likelihood of classification error. As we describe in Section XX, we reinterpret this information as an estimate of the fraction of sub pixel area that is covered by the primary class.

As we will see, MLCT characterizations for land cover classes in the US are somewhat unsatisfying for those classes that frequently occur at scales too small to resolve with the 500m MODIS instrument on which the MLCT is based. We turn to the NLCD to resolve these higher resolution features, but since a similar dataset is not yet available globally, this step of the algorithm is designed as a validation/feedback step, rather than a core piece of the algorithm, which can be performed regionally wherever high resolution data sets are available.

The **National Land-cover Database 2001** (NLCD) provides a higher-resolution (30m, $\sim 1.25''$) snapshot of LULC across the cUSA study area, plus Alaska, Hawaii and Puerto Rico, circa 2001. Because NLCD's classification was informed by ancillary data sets such as population density, buffered roads, and the National Wetland Inventory [*Homer et al.*, 2004], we expect that it will give better estimations of aggregate area for detailed features like rural transportation networks and small stream and wetland features. Although it is unclear from *Homer et al.* [2004] what ancillary data was applied in what constituent mapping zones of NLCD, we accept its representation of these fine details to be the best available. As we describe in Section 3.3, we apply differences between NLCT and MLCT data as a correction, in order to compensate for MLCT's bias against these finely detailed structures due to its resolution.

The **Agricultural Lands in the Year 2000** (Agland2000) *Ramankutty et al.* [2008] is a $5'$-resolution data set that merges satellite-derived LULC classifications with census data of arable land and permanent crops compiled at national or sub-national levels according to availability of such data at or near the turn of the last century. It uses two LULC classification data sets derived from remote sensing data as inputs, an older version of the MLCT (known as BU-MODIS) and the GLC2000 data set mentioned in Section 1. The "pasture" class in Agland2000 likely has much in common with the "open" class from MLCT but we do not employ that data in this analysis. Note that the cropland aspect of Agland2000 is used as an input into the classification algorithm of the MLCT version that we use here. We acknowledge the possibility of circularity when comparing the two, but because of its basis in census data we will use the cropland component of Agland2000 as a type of observational product for the purposes of evaluating our incremental adjustments to the maps we derive from MLCT in Section 4. Figures 1 and 18 show the distribution of both cropland and pasture areas for the detail and full study areas respectively.

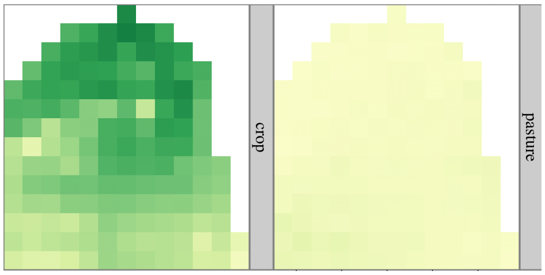


Figure 1: Agland2000 distribution in detail area.

The **Harvested Area and Yields of 175 Crops** (175Crops2000) dataset [*Monfreda et al.*, 2008] provides 5′-resolution information about crops grown globally. As we describe in Section 3.5, we use this data to decompose PEEL_0 cropland area fractions. It is not possible to use this data directly because it reflects only harvested area and so ignores various types of ancillary agricultural land. For our purposes it provides relative proportions for the decomposition of the cropland fractions at the grid cell level. Rather than considering the full array of 175 crops we consider only corn, soy, wheat, rice, and sugarcane individually, combine other cereals into their own class, and combine all remaining crops as a catch-all "other" category. Field crops are distinguished from orchard / plantation crops that would likely fall under areas classified by MLCT as forest or shrub in this step. Table 2 specifies how the 175 crops in the *Monfreda et al.* [2008] data set are collected into crop sub-classes in PEEL_0. Realistically many of these crops are not present in the cUSA study area, but we provide the complete mapping to our crop sub-classes for sake of completeness.

Sub-class	Crops
maize	maize
wheat	wheat
rice	rice
other cereals	barley; buckwheat; canary seed; cereals nes; fonio; millet; mixed grain; oats; pop corn; quinoa; rye; sorghum; triticale
soybean	soybean
sugarcane	sugarcane
forage	alfalfa for forage; beets for fodder; cabbage for fodder; carrots for fodder; clover; forage products nes; grasses nes; green oilseeds for fodder; legumes nes; maize for forage; mixed grasses and legumes; rye grass for forage; sorghum for forage; swedes for fodder; turnips for fodder; vegetables and roots for fodder
other field crops	anise, badian and fennel; artichokes; asparagus; bambara beans; beans, dry; beans, green; broad beans, dry; broad beans, green; cabbages; cantaloupes and other melons; carrots; cassava; castor beans; cauliflower; chickpeas; chicory roots; chillies and peppers, green; coir; cotton; cow peas, dry; cucumbers and gherkins; eggplants; fibre crops nes; flax fibre and tow; garlic; ginger; green corn (maize); groundnuts in shell; hemp fibre and tow; hempseed; jute; jute-like fibres; lentils; lettuce; linseed; lupins; melonseed; mushrooms; mustard seed; oilseeds nes; okra; onions and shallots, green; onions, dry; peas, dry; peas, green; peppermint; pigeon peas; pimento; poppy seed; potatoes; pulses nes; pumpkins, squash, gourds; rapeseed; roots and tubers nes; safflower seed; sesame seed; spinach; string beans; sugar beets; sugar crops nes; sunflower seed; sweet potatoes; taro; tobacco leaves; tomatoes; vegetables fresh nes; vetches; watermelons; yams; yautia
shrub crops	abaca (manila hemp); agave fibres nes; bananas; berries nes; blueberries; cocoa beans; coffee, green; cranberries; gooseberries; grapes; hops; mate; nutmeg and mace and cardamons; pepper; pineapples; plantains; pyrethrum, dried flowers; ramie; raspberries; sisal; strawberries; tea; vanilla
tree crops	almonds; apples; apricots; areca nuts (betel); avocados; brazil nuts; carobs; cashewapple; cashew nuts; cherries; chestnuts; cinnamon (canella); citrus fruit nes; cloves; coconuts; currants; dates; figs; fruit fresh nes; fruit tropical fresh nes; grapefruit and pomelos; hazelnuts (filberts); kapok fibre; kapokseed in shell; karite nuts (sheanuts); kiwi fruit; kolanuts; lemons and limes; mangoes; natural gums; natural rubber; nuts nes; oil palm fruit; olives; oranges; papayas; peaches and nectarines; pears; persimmons; pistachios; plums; quinces; sour cherries; spices nes; stone fruit nes, fresh; tang.mand.clement.satsma; tung nuts; walnuts

nes: "not elsewhere specified"

Adapted from *Monfreda et al.* [2008].

Table 2: Crop sub-classes for simplifying 175Crops2000.

The **USDA Major Land Uses** (MLU) dataset [*Lubowski et al.*, 2006] is a consistent census-based time-series record of US agricultural land uses produced by the USDA Economic Research Service (ERS) going back to 1945. The MLU contains data by state and distinguishes land use among 6 broad categories: cropland; grassland, pasture, and range; forest-use land; special-use land; urban land; and other uses. Additionally, important sub-categories for relating these census-based land-use categories to satellite based land-cover categories include cropland pasture, rural transportation areas, and farmsteads, farm roads, and lanes. For example, we include parks and other protected forest areas with the forest class, we distinguish cropland pasture from both cropland and pasture land for better comparison with NLCD statistics, we combine rural transportation networks and other developed areas, farmsteads, and other developed uses with the urban cover class, and we include miscellaneous lands such as marches, swamps, deserts, and lands designated for defense or other special purposes as other land cover. We note that MLU does not include coastal or inland water bodies, and so has a lower overall count of land area as compared to the satellite sets.

3 Our Algorithm

Our general algorithmic approach can be summarized as follows. First, reclassify the categories of the MLCT and NLCD data sets to a common scheme, calculate per-pixel, per-class areas at the native resolutions, and aggregate the new classification to the $5'$ grid. A challenge is that classification definitions are sometimes subtly different between data sets, making direct comparison across data sets somewhat subjective.

1. Prepare MLCT data
   (a) Reproject to geographic coordinates and mask cUSA study area
   (b) Reclassify to PEEL_0 classification (Table 3)
   (c) Calculate per-pixel, per-class areas at native resolution as a function of parameter A_{min} (see Sec. 3.1.2 for details)
   (d) Aggregate the new classification to the $5'$ grid, combining MLCT primary class, confidence, and secondary class values
2. Prepare NLCD data for use as correction layer
   (a) Reproject to geographic coordinates and mask cUSA study area
   (b) Reclassify to PEEL_0 classification (Table 4)
   (c) Calculate per-pixel, per-class areas at native resolution
   (d) Aggregate the new classification to the $5'$ grid (no secondary layer, so set $A_{min} = 1$)
3. Combine data sets to produce PEEL_0
   (a) Adopt MLCT as PEEL_0^a
   (b) Selectively incorporate into PEEL_0^a cover fractions for water, wetland, and urban classes from NLCD, to produce PEEL_0^b.
   (c) Decompose mosaic fraction into crop and natural cover components
   (d) Use 175Crops2000 data to decompose the cropland class in PEEL_0^b to produce the final PEEL_0.

We begin with the global MLCT product which provides categorical LULC classification at $15''$ ($\sim$500m) resolution plus a secondary cover type for each pixel, and a confidence measure for the primary type. The first step in the algorithm is a systematic attempt to incorporate this full depth of information offered by MLCT. Rather than interpret the secondary classification as the next most likely possibility we accept the triplet as an expression of the sub-pixel composition of that area. Aggregation of MLCT from $15''$ to $5'$ blurs the spatial precision implied by this formula. We treat the local $20 \times 20 \times 3$ array as a probabilistic expression of the local landscape composition. We will show that this approach, given a principled assumption about the relationship between confidence level and the allocation of sub-pixel area among the detected classes, improves the estimates of acreages in aggregate as well as their spatial distributions, particularly for cropland. See Section 3.1.

3.1 Preparing the MODIS Land Cover Type

The 2001 MLCT data set comprises a set of tiles in a global equal-area sinusoidal projection. To prepare this data set for use in this study, we first patch those tiles together and reproject the resulting mosaic to geographic coordinates. We then extract the conterminous USA study area, which we define as those $5'$ grid cells that intersect with the cUSA polygons in version 1 of the Global Administrative Areas (GADM) vector data set [*Hijmans et al.*, 2009]. This area includes the water bodies on the American side of the international border across the Great Lakes, but not oceanic waters beyond the coastal grid cells that intersect with any land mass. To illustrate the process of converting these data sets from their original representation, we shall also often include maps of an area of southeastern Michigan to show greater detail through each step of the process. We chose this region for its diversity of land covers and uses, its relative diversity of agricultural commodities across its significant cropland area, the significant presence of the MLCT/IGBP mosaic class to illustrate our method for its decomposition, and its familiarity to the authors. Where space allows, we also present limited maps of variables over the conterminous USA. A more detailed set of maps will be provided online upon publication.

3.1.1 Reclassification of MLCT to PEEL$_0$ classes

Table 3 shows the mapping of the IGBP classes used in MLCT to our PEEL$_0$ classification. Because our primary interest is in agriculture we collapse the five forest categories into a single class. We assign woody savannas and savannas to the shrub and open classes, respectively; these assignments are supported by the IGBP class definitions due to the overlap in the forest canopy cover for those classes. These assignments makes sense in the context of LULC modeling because the ecological roles, potential uses, and conversion costs of the two savanna types are dissimilar. We combine areas of ‘permanent snow and ice’ with ‘barren or sparsely vegetated’ areas, which includes deserts, to form the PEEL$_0$ ‘barren’ class, based on their shared characteristics of low population density and low intensity of economic activity.

Figure 2 shows the result of reclassifying the MLCT data for our detailed study area. This area is dominated by the crop class in the north and the mosaic class to the south with scattered forests and pockets of development throughout. The urban complex of Port Huron, Michigan and Sarnia, Ontario is visible in the southeast corner. Areas in the northern and central sections of the map, classified as crop with 100% confidence have null values in the secondary class (white cells).

Note that areas in the northern and central sections of the map that were classified as crop in the primary layer have null values in the secondary class, shown as white cells. These cells coincide with values of 100% confidence in the third layer of the data, so their areas are assigned entirely to the primary class. The high prevalence of 100% confidence cells for the crop class leads us to expect that MLCT will over-estimate cropland, since any such large areas of cultivated land are certainly

Adapted from *Friedl* [2002].

	MLCT/IGBP	PEEL$_0$
0	water	water
1	evergreen needleleaf forest	forest
2	deciduous needleleaf forest	
3	evergreen broadleaf forest	
4	deciduous broadleaf forest	
5	mixed forests	
6	closed shrublands	shrub
7	open shrublands	
8	woody savannas	
9	savannas	open
10	grasslands	
11	permanent wetlands	wetland
12	croplands	crop
13	urban	urban
14	cropland / natural vegetation mosaics	mosaic
15	permanent snow and ice	barren
16	barren or sparsely vegetated	

Table 3: Reclassification of MLCT/IGBP to PEEL$_0$.

interspersed with homesteads, fence lines, small wood lots, roads, and other such cultural features. For example, in areas such as this that were made available for settlement in the 19th century according to the Public Land Survey System (PLSS) we expect to find a more-or-less regular grid of rural roads at one-mile intervals, which is readily visible in maps and aerial photography of the region. It is admissible to argue that the IGBP cropland class is intended to indicate an overall function of the landscape rather than to measure areas directly employed in production, so in that sense we can say that the PEEL cropland has a different definition because we seek to reconcile cropland areas with agricultural census data.

It is possible for primary and secondary classes to be assigned to the same category because of the reclassification step. When a pixel indicates the forest class for both its primary and secondary classifications it simply reflects a distinction between sub-types of forest in the original data, for

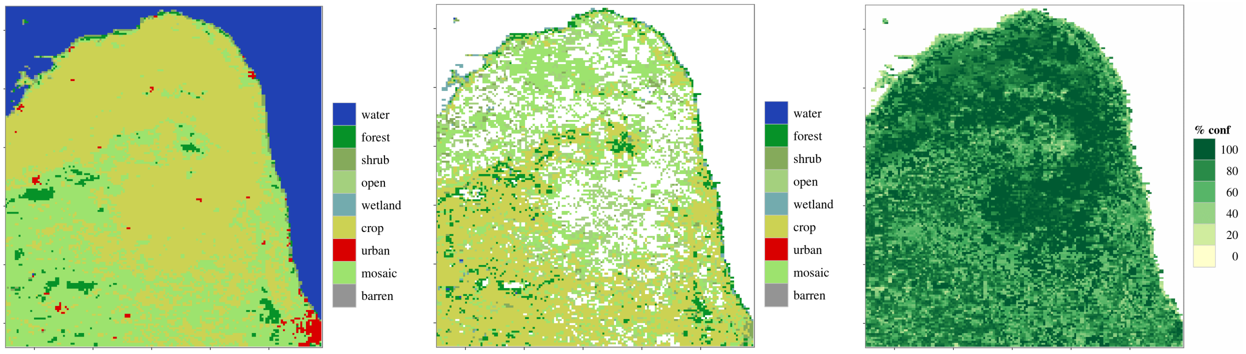


Figure 2: MLCT primary reclassified cover (left), secondary reclassified cover (middle), and primary cover classification confidence(right) for detailed evaluation area.

example evergreen and deciduous. It is worth noting that crop and mosaic classes often appear in pairs (primary crop, secondary mosaic and vice versa). This coupling is not surprising given that mosaic areas are comprised of 40–60% cropland by definition so their exemplars must necessarily be near one another in the classification space. We will explore this dynamic further in 3.4.

Figure 15 shows maps of the primary classification, secondary layer, and confidence level for the entire cUSA study area. For improved visualization of the relative distributions of particular classes, we also provide, in Figure 16, facet maps for the individual classes. Familiar generalities of cUSA geography are apparent in this maps, such as the prevalence of forests in the east and northwest, cropland in the midwest, shrub lands in the southwest, and open lands across the west. The mosaic class is concentrated in the eastern portion of the study area; we attribute this phenomenon to greater population density, topography, and historical patterns of settlement resulting in characteristically smaller parcels and a greater degree of mixing among agricultural uses and natural covers.

3.1.2 Aggregation of MLCT to PEEL$_0$ resolution

MLCT has a nominal resolution of roughly 500m that equates to 15″ at the equator, a conveniently even factor-of-20 division of the 5′ grid to which we wish to aggregate. Recall that for each pixel, MLCT provides three data values: a primary classification, a measure of confidence up to 100%, and a secondary classification. Our aggregation strategy aims to extract as much information as possible from the (20 × 20 = 400 MLCT pixels) × (3 values) = 1200 data values that MLCT provides for each PEEL$_0$ cell. MLCT's secondary cover type was originally intended to express the most likely alternative to the primary type [*Friedl et al.*, 2010], with the confidence level providing an indication of per-pixel classification error. We propose an alternative interpretation where the sub-pixels areas for the primary and secondary cover types in pixel x are given by:

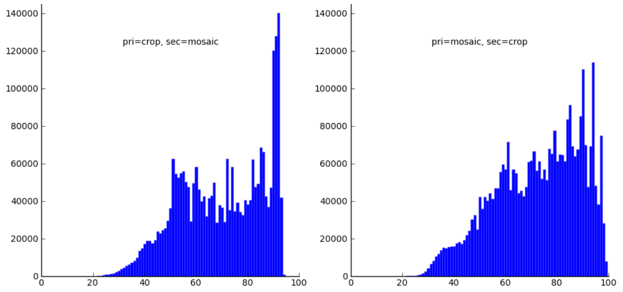
$$\begin{aligned} A_p(x) &= A_{min} + (1 - A_{min})c(x) && (1) \\ A_s(x) &= 1 - A_p(x) && (2) \end{aligned}$$

where $c(x)$ is the confidence level of the primary classification on $(0, 1]$ and A_{min} represents the minimum area fraction to be assigned to the primary class given $c(x)$=0. A_p and A_s are the fractional areas assigned to the primary and secondary classes respectively. A given class must comprise at least 50% of the pixel area in order to be considered primary, therefore setting A_{min}=0.5 affords maximum consideration to the secondary class in this scheme. Simplifying the equations by substituting this value gives:

$$\begin{aligned} A_p(x) &= \frac{1 + c(x)}{2} && (3) \\ A_s(x) &= 1 - A_p(x) = \frac{1 - c(x)}{2} && (4) \end{aligned}$$

Instances of c<0.20 are rare as shown by Figure 3 for a particular subset of MLCT pixels (see Section 3.4), so generally the primary class will be assigned more than 60% of the MLCT pixel area–which is consistent with MLCT's definition of "primary class," which is that it covers no less than 60% of a given pixel x [*Friedl*, 2002]. These definitions assume that the relationship between classification confidence and the sub-pixel fraction of the primary class is a linear, monotonically increasing function. Other monotonic functions could be used, but the differences would be second-order refinements to this formulation.

In the analysis that follows we compare the product of these assumptions with the case of A_{min}=1.0, which gives zero consideration to the secondary class and is therefore indifferent to the



we will examine

Figure 3: Histograms of the confidence measure for all cells in the cUSA classified as primary type crop, secondary type mosaic (left) and as primary type mosaic, secondary type crop (right).

confidence level. In the interest of brevity, we do not consider here intermediate parameterizations in which the secondary class is used, but given less than maximum consideration. Still, an advantage of our algorithm is that inclusion of the secondary class can be varied continuously by A_{min}.

For a clearer intuition of this sub-pixel area allocation procedure, consider the histograms of the confidence measures for MLCT pixels with primary/secondary classes equal to crop/mosaic or mosaic/crop, shown in Figure 3. These pixels are atypical in a sense because the mosaic class is itself defined as a hybrid of crop and natural cover that contains about 40–60% cropland. By pinning this fraction at 50% our algorithm says that the fractional area of crop in crop/mosaic cells as a function of the confidence $c(x)$ is:

$$A_{crop}(x) = \frac{1 + c(x)}{2} + \frac{1 - c(x)}{4} = \frac{3 + c(x)}{4} \tag{5}$$

since half the mosaic class is going to crop. Similarly for mosaic/crop cells:

$$A_{crop}(x) = \frac{1 + c(x)}{4} + \frac{1 - c(x)}{2} = \frac{3 - c(x)}{4} \tag{6}$$

implying that, at minimum, these cells are majority cropland.

Computationally the process of converting the reclassified maps to sub-pixel fractions at the desired $5'$ resolution is a three-step process. The first step is to calculate the fraction of the primary cover type as a function of the classification confidence as described above, independently of the primary and secondary classifications. Next, a sub-pixel fraction for each cover type is calculated at the original $15''$ resolution, which is zero for all classes but the one or two classes indicated. Aggregating to a coarser resolution is a simple matter of calculating the means of these fractions over the intersecting $15''$ pixels within a given $5'$ grid cell. Figure 4 emphasizes the difference between the choice of $A_{min} = 0.5$ and $A_{min} = 1.0$ for the calculation of the sub-pixel fractions and their aggregation to $5'$ with a difference map. Positive values in the map indicate areas where $A_{min} = 0.5$ produced a greater value. Considering the secondary class results in a shift of up to 10% of total cell area from crop to mosaic in the north of the detail area and vice versa for the southern portion.

3.2 Preparing NLCD

Our reclassification for NLCD, shown in Table 4, is more complicated than that used for MLCT (Table 3). Although NLCD has fewer forest classes than MLCT, they are equally unambiguous. We equate four NLCD developed land classes with the PEEL_0 "urban" class, to represent developed areas of all densities. The result of this reclassification is shown in Figure 5 at native NLCD

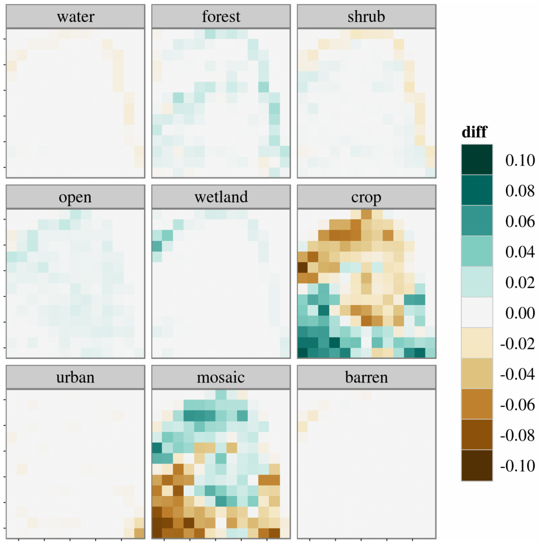


Figure 4: Difference of aggregated sub-pixel fractions for $A_{min} = 1.0$ vs. $A_{min} = 0.5$, positive when $f(A_{min} = 0.5)$ is greater.

resolution. Many detailed features missing from the 500m MLCT product are apparent in this figure, including rural transportation networks and small wooded, water, and wetland features. Figure 17 shows the result of reclassifying NLCD and aggregating it to 5′ sub-pixel fractions for the full cUSA.

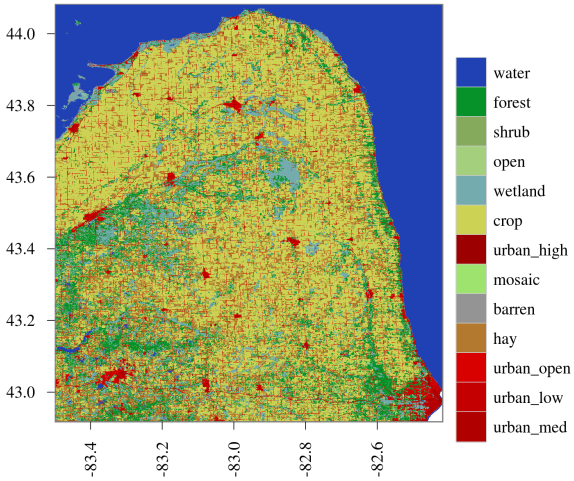


Figure 5: NLCD reclassified at native resolution.

11[a]	water	water
41	deciduous forest	forest
42	evergreen forest	
43	mixed forest	
52	shrub/scrub	shrub
71	grassland / herbaceous	open
90[a]	woody wetlands	wetland
95	emergent herbaceous wetlands	
82	cultivated crops	crop
81	pasture / hay	hay
21	developed, open space	urban_{open}
22	developed, low intensity	urban_{low}
23	developed, medium intensity	urban_{med}
24	developed, high intensity	urban_{high}
12	perennial ice/snow	barren
31[a]	barren land	

Adapted from *Homer et al.* [2004].
[a] Additional coastal classes exist in NLCD but are not present in the lower 48 states.

Table 4: Reclassification of NLCD.

Though repeating the aggregation process for the entire study area is computationally expensive due to the NLCD's high resolution, the algorithm is precisely the same as for refactoring the MLCT when considering only the primary cover type, i.e. setting $A_{min} = 1$, so we will not describe it in further detail here. The effect of the aggregation for the detailed study area is shown in Figure 6.

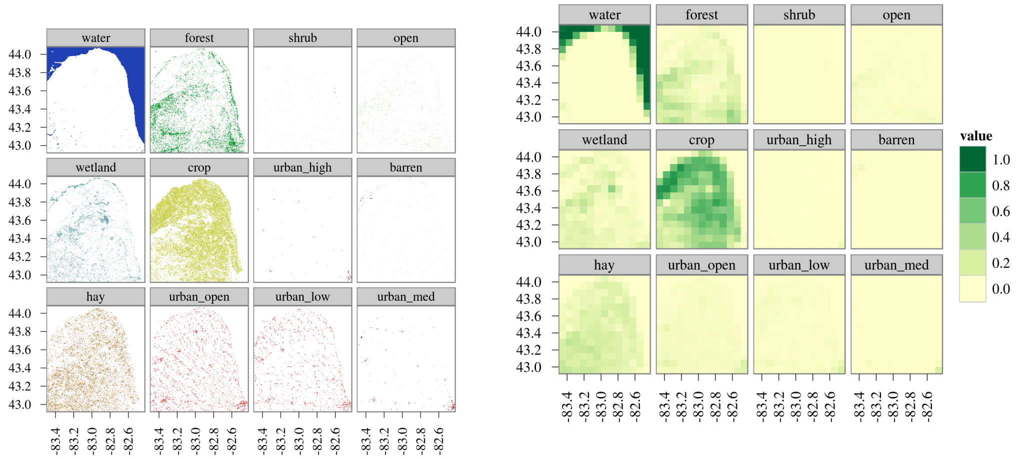


Figure 6: NLCD covers (left) and aggregated cover fractions (right) shown separately for the detailed study area.

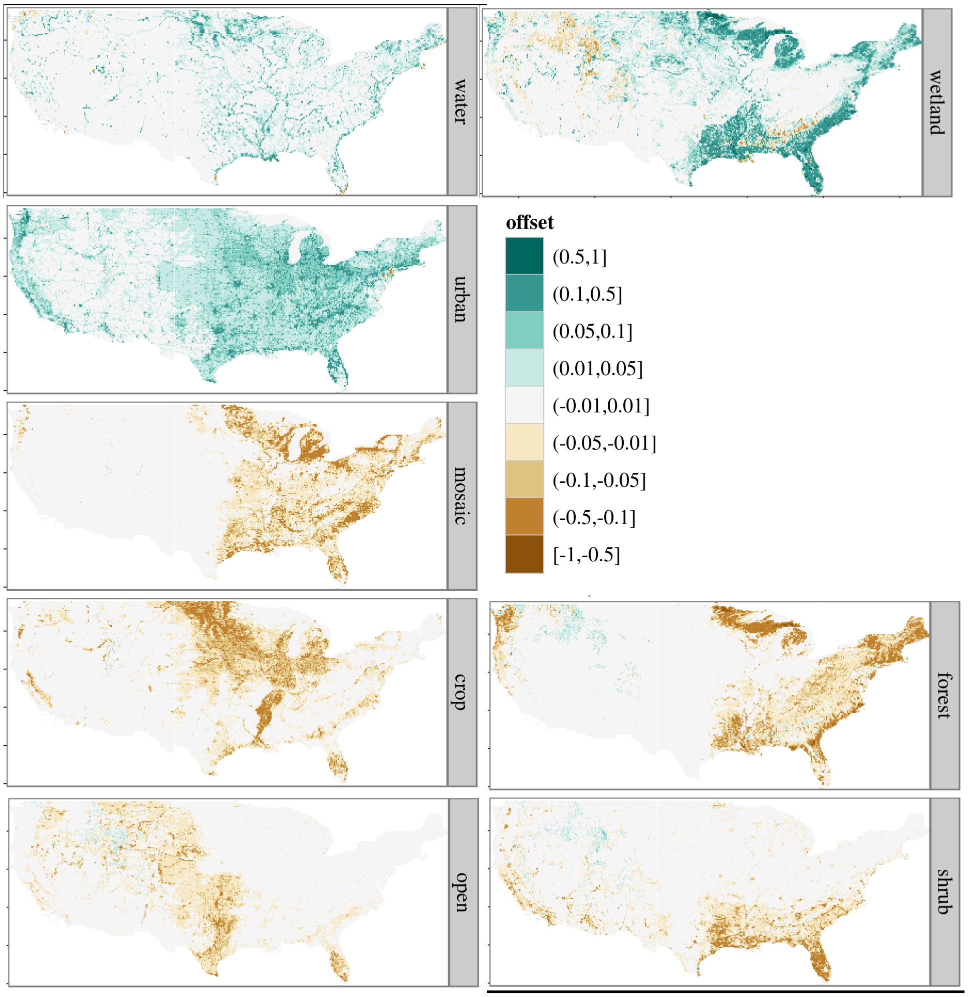


Figure 7: NLCD offsets (log scale). Water, wetland, and urban cover fractions are tuned to match those in NLCD, and the difference is accommodated as described in Section 3.3.

3.3 Applying NLCD Offsets to PEEL$_0$

Table 5 shows that MLCT is negatively biased in the total areas assigned to water, wetland, and urban features relative to NLCD. Visual inspection shows that features of these classes tend to have smaller characteristic dimensions causing them to be overlooked in MLCT due to its resolution. The most obvious example is the rural transportation networks in areas surveyed under the Public Land Survey System (PLSS) where roads have been laid out on a generally regular grid of square miles. PEEL$_0$ includes this infrastructure in the urban class as another form of developed land.

To merge this information from NLCD we first accept the areas for water, wetland, and urban classes in the reclassified, $5'$-aggregated version of NLCD that we have computed as truth, and then calculate offsets for those classes versus our $5'$ MLCT data by direct subtraction. Where NLCD is greater the difference will be positive and so a positive offset will be added to the fraction already present for any one of the "truth" classes from NLCD. The other classes are then adjusted so that they are present in proportion to each other as indicated by MLCT but in the area remaining after accepting the water, wetland, and urban areas from NLCD. Figure 7 shows the spatial distributions of the offsets calculated based on our assumptions about the water, wetland, and urban classes in NLCD. The logarithmic scale used in these maps makes apparent both areas of significant adjustment, greater than 10%, and the extent to which small adjustments on the order of 1–5% occur. We see the detailed structure of drainage networks in the water class and population centers in the urban class which could easily be confused with the vegetative classes in the MLCT classification, perhaps in heavily wooded suburbs where transportation infrastructure is obscured and difficult to resolve, for example. The offsets for the NLCD truth classes (urban, water, and wetland) are generally positive, although not strictly so because the algorithm does not preclude the possibility that MLCT may locally overestimate these classes in particular regions and still suffer an aggregate deficit relative to NLCD.

3.4 Decomposing the MLCT mosaic class

The MLCT "cropland/natural vegetation mosaic" class is problematic for the economic models for which PEEL$_0$ is intended, because it combines developed land use and natural land cover. This class is defined as a hybrid of cropland and some mixture of natural covers (forest, shrub, or open) with no single component exceeding 60% [*Friedl*, 2002]. We wish to differentiate the cropland from the natural vegetation in order to calculate a more meaningful total for cropland area and thereby eliminate the mosaic class from the final tabulation. To this end, we make three simple assumptions about the composition of area identified as mosaic lands:

1. 50% of mosaic area is assigned to the crop class.
2. The other 50% is a blend of forest, open, and shrub in relative proportion to the expression of those classes in the same $5'$ cell.
3. In the absence of any natural classes in the $5'$ cell the natural component of the mosaic is an equal blend of all three.

We make these simplifying assumptions so that we can proceed with the evaluation of this analytical framework. It might be interesting to vary the proportion of mosaic land allocated to crop land. However, we have no principled basis for doing so, despite the definition's implication that this proportion is variable. Our chosen 50% level reflects the assertion that the mosaic is a cultural class grouped with cropland and urban in the IGBP classification scheme without overstating the degree of development. MLCT provides adequate variability in this dimension by commonly pairing cropland and mosaic in the primary/secondary class data. The second assumption imposes that $15''$ mosaic cells' non-crop portion will have the same relative composition of forest, open, and shrub as the non-mosaic portion of the $5'$ grid cell in which it falls. Therefore mosaic pixels in a

$5'$ cell where of the three non-crop mosaic components only forest is found are assigned 50% crop and 50% forest.

3.5 Decomposition of PEEL$_0$ Crop Fractions

The final step is to further dissaggregate the cropland layer into different types or categories of crops using cover fractions calculated from 175Crops2000. Double-cropping is ignored by normalizing the crop fractions by the sum of all crops, which can exceed unity in instances of intense double-cropping. The predominant double-cropping system in the cUSA to our knowledge is soy followed by winter wheat, but there may be others such as multiple cropping of rice in the southern extremes of its range. In cells where soy and wheat are double-cropped alongside other cropping systems, their relative prevalence among the cultivated land in the cell will likely be somewhat underestimated as compared to that given in the 175Crops2000 data set. This issue bears further study. The crop sub-class layers are shown as facets in Fig. 21.

4 Analysis

We have hypothesized that there is information worth capturing in the secondary class and classification confidence level provided by MLCT. We test this hypothesis via comparison with AgLand2000, which we view as representing "truth" for the purposes of this analysis. We also evaluate the accuracy of cropland distribution in MLCT as a function of the A_{min} parameter, contrast accuracy obtained with maximum incorporation of the secondary class in our analytical framework, as modulated by the MLCT classification confidence data ($A_{min} = 0.5$) versus considering only the primary classification ($A_{min} = 1.0$). We evaluate results on the basis of root mean square error (RMSE) metrics calculated relative to the distribution of cropland given in both NLCD and Agland2000.

4.1 Comparison of Aggregate Areas

We start by tabulating the aggregate areas by class for MLCT, NLCD, and Agland2000. After decomposing the mosaic class MLCT indicates 495.4 Ma (200.5 Mha) of cropland for A_{min}=0.5 and 488.1 Ma (197.5 Mha) for A_{min}=1.0 in the cUSA in 2001, the NLCD indicates roughly 448.9 Ma (181.7 Mha) combined of 'cultivated crops' and intensively managed 'pasture/hay', Aglands2000 indicates roughly 446.5 Ma (180.7 Mha) of cropland, and the USDA MLU [*Lubowski et al.*, 2006] dataset indicates roughly 441.3 Ma (178.6 Mha) of combined cultivated crops and 'cropland pasture'. The areas for all the major comparison and intermediate data sets, and most parameter choices considered here, are shown in Figure 8, and key values are collected in Table 5.

We are primarily concerned here with the ∼10% overestimation of cropland in MLCT, especially in light of the relative agreement among the other products considered here. The inability of MLCT to resolve rural transportation networks, minor settlements, and small water or wetland features is a major contribution to its surplus of cropland acreage indicated. Due to its greater resolution, ∼30m vs. ∼500m, NLCD is better suited to discerning developed areas in rural landscapes, ranging from rural roads to farmsteads to small communities that do not show up in MLCT. A total area of roughly 75.4 Ma (30.5 Mha) of land developed to one extent or another that remains after subtracting the MLCT urban class from all developed classes in NLCD. Applying this difference as an offset brings us signficantly closer to the expected acreage under cultivation in 2001; see Section 3.3. This offset also brings the national area of urban and developed cover much more in line with the USDA MLU census data.

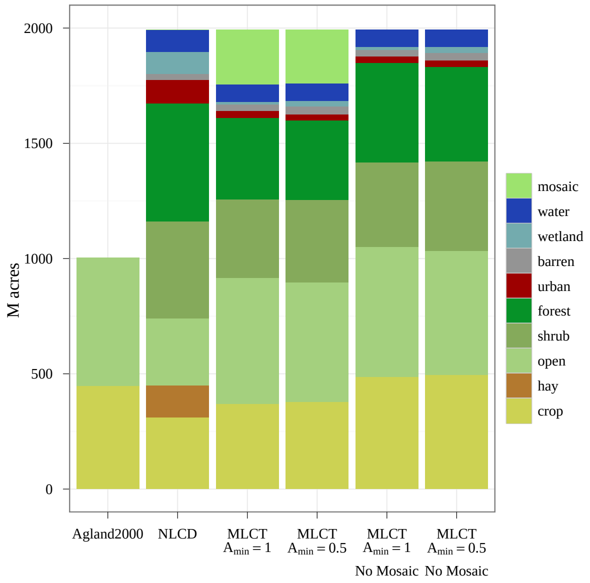


Figure 8: Total Acreages by Map and Cover.

	MLU 2002[a]	Agland 2000	NLCD	MLCT A_{min}=1	MLCT A_{min}=.5	MLCT No Mosaic	WWU Offset	MLCT Adjusted	PEEL_0
water			96.5	75.0	74.3	74.3	22.3	96.5	96.5
forest	657.1[b]		513.2	353.6	344.7	410.8	-44.7	300.1	355.7
shrub			420.1	341.8	358.7	387.2	-23.8	334.9	358.0
open	584.2	557.1	291.2	545.8	516.9	538.7	-21.0	495.9	514.9
wetland			95.0	11.0	26.0	26.0	69.0	95.0	95.0
crop	379.5[c]	446.5	310.8	369.6	378.9	495.4	-39.0	339.9	437.6
pasture	61.8 [c]		138.4						
urban	96.9[d]		102.8[e]	29.8	27.3	27.3	75.4	102.8	102.8
mosaic				237.0	232.9		-37.4	195.5	
barren			24.5	28.9	32.8	32.8	-0.9	31.9	31.9
other	112.0[f]								
(all)	1893.8	1003.7	1992.5	1992.5	1992.5	1992.5	0.0	1992.5	1992.5

[a] Data from the USDA MLU database does not include coastal or large inland water bodies.
[b] We include parks and other protected forest areas with the forest class.
[c] For MLU we distinguish 'cropland pasture' from harvested and idle cropland for better comparison.
[d] We combine rural transportation networks and other development with the urban class.
[e] This includes all NLCD developed classes: high, medium, and low density and 'developed open'.
[f] This includes misc. lands, marshes, swamp, desert, etc., plus 14.8 Ma designated for defense or other special purpose.

Table 5: Effect of NLCD offsets on total acreages, $A_{min} = 0.5$ everywhere, except where explicitly noted.

4.2 Comparison of Root Mean Square Errors

The purpose for processing the MLCT for two values of A_{min} as described in Section 2 is to evaluate whether or not information from the secondary cover type contributes positively to the accuracy of the data set we seek to synthesize. The primary objective of this synthesis is to achieve accuracy in cropland distribution. Although MLCT overstates cropland acreage for both A_{min}=0.5 and A_{min}=1.0, the discrimination among the two is made by the distribution of errors rather than the aggregate error. Figure 9 shows the cell-by-cell differences in cropland fractions between the A_{min}=0.5 MLCT-derived data set that we have calculated after mosaic decomposition and the Agland2000 and NLCD cropland layers (the map for A_{min}=1.0 cannot be distinguished from this visually). To summarize and compare these errors we calculate the root of the mean squared error (RMSE) given by:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}} \tag{7}$$

where $\hat{\theta}_i$ are the predictions derived from the respective MLCT derivations and θ_i are the observations, taken from the Agland2000 or NLCD data sets in turn. These RMSE values are collected later in Table 6, from which its clear that choosing A_{min}=0.5 improves the spatial correlation of the MLCT cropland layer compared to both Agland2000 and NLCD.

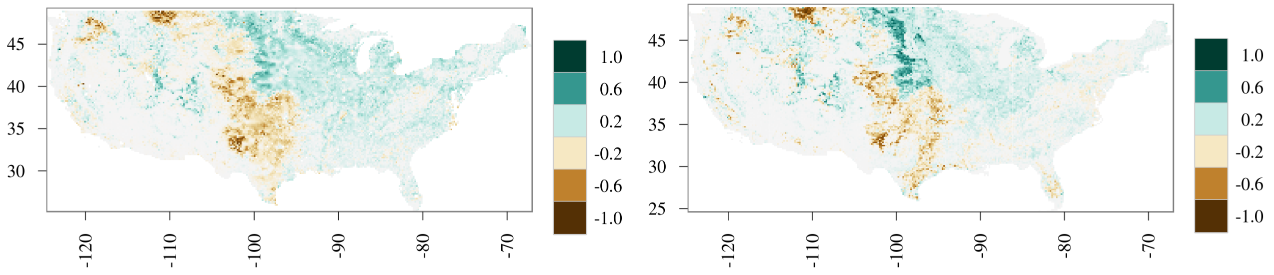


Figure 9: Difference between MLCT (no mosaic, A_{min}=0.5) and Agland2000 (left) and NLCD (right) crop.

WWU[a] offset	A_{min}	RMSE (Agland2000)	RMSE (NLCD)
no	1.0	0.180	0.175
no	0.5	0.165	0.161
yes	0.5	0.151	0.149

[a] "WWU offset" refers to the water, wetland, and urban offsets calculated from NLCD and applied to the data product in Sec. 3.3.

Table 6: RMSE vs. Agland2000 and NLCD for a variety of intermediate and final data products.

4.3 Comparison of Hexbin Plots

To further examine the relationships between the distributions of cropland that we derive from MLCT data and the Agland2000 and NLCD data we create "hexbin" plots. These are essentially two-dimensional histograms that show the number of grid cells that occur within discrete regions of the space defined by coordinates that are cropland fractions for the two data sets. This operates much like a common scatter plot but for data sets with as many observations as we wish to include it gives a cleaner representation of that structure. We employ a logarithmic scale for the bin counts to obtain a more complete picture of the overall dispersion and local concentration of the observations.

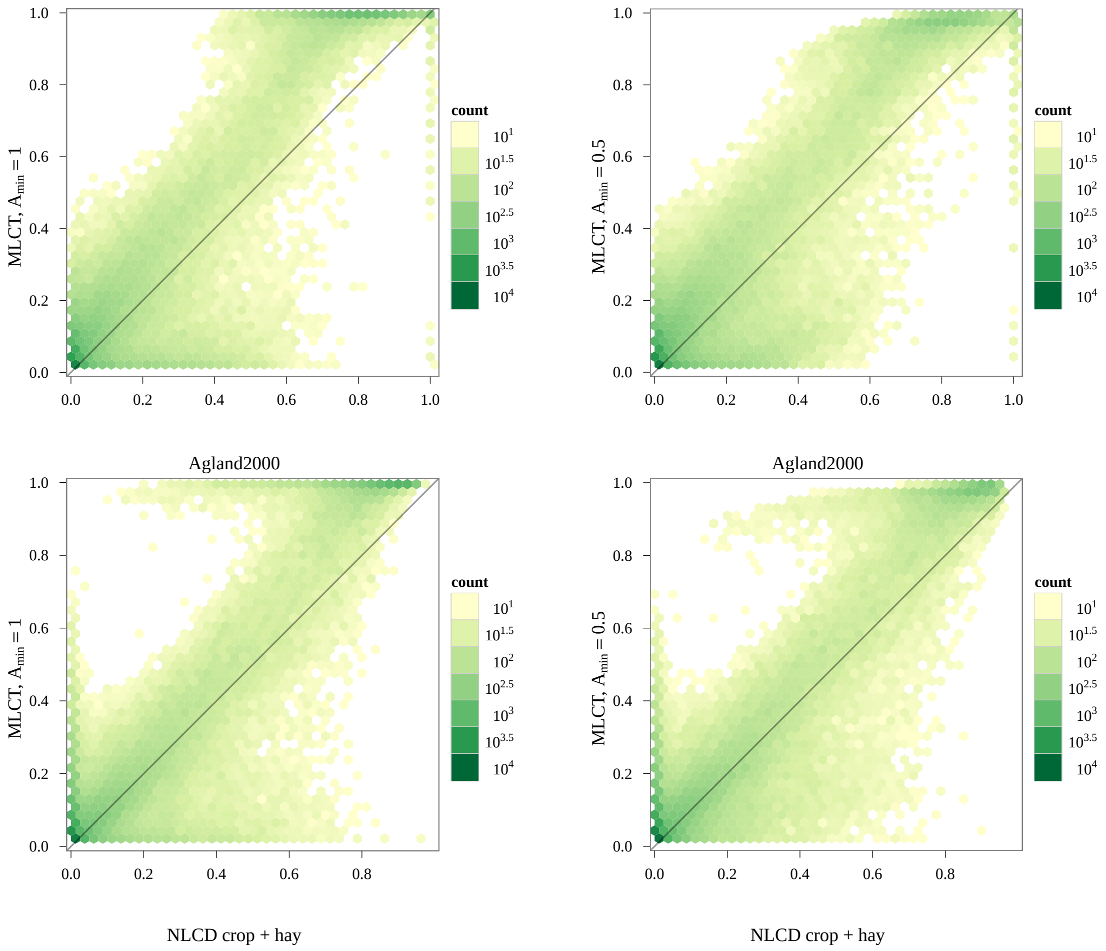


Figure 10: Hexbin plot of MLCT crop $A_{min} = 1.0$ (left) and $A_{min} = 0.5$ (right), both with mosaic removed, versus Agland2000 crop (top) and NLCD cropland plus hay/pasture (bottom).

Figure 10 plots the crop fractions of MLCT with A_{min}=1 and 0.5 versus the Agland2000 crop and NLCD 'crop+hay' layers. As one would expect there is an overall correlation among these variables, especially given that Agland2000 provides prior probabilities to the MLCT classification (and that an early version of MLCT, BU-MODIS provides a key input to the construction of Agland2000). It is clear that the MLCT primary class exhibits a positive bias overall, although a subset that is negatively biased is also apparent for low values of the Agland2000 crop fraction in the interval $[0.1, 0.5]$. Also of particular note is the drastic decrease in correlation when Aglands2000 reaches 1.0 relative to the stronger relationship over the interval $[0.8, 1.0)$. There appears to be something peculiar about the Agland2000 allocation procedure that drives the crop fraction to its maximum in areas where the remote sensing data clearly resists such a characterization. One possible explanation is systematic errors in the agricultural census data that drive the Agland2000 algorithm, forcing unrealistically high concentrations in order to satisfy the algorithm's constraints.

Figure 10 also shows the correlation of MLCT vs. the NLCD crop and hay/pasture layers. We see clearly the positive bias of MLCT relative to NLCD in general, and especially over the interval $[0.8, 1.0]$. These are the cells in which MLCT sees near complete cropland, whereas NLCD sees sometimes substantial mixing with other classes, which we presume to be largely due to small features missed by MLCT. A curious feature arises here as well, similar to the decorrelation where Agland2000 showed 100% crop, there is a decorrelation where NLCD shows close to exactly 0% crop. The origin of this feature is not clear to us, but we suspect a classification error in MLCT.

We expect that setting A_{min}=1 will produce a maximum overall bias and attendant error by assigning entire pixels to the cropland class and not allowing for the possibility of mixed covers. The results on Table 6 indicate that A_{min}=0.5 is more representative of the distribution of cropland as compared to both NLCD and Agland2000 because, although the total area indicated is higher according to Table 5, there is less error on a cell-by-cell basis indicating that it does a better job of representing the spatial distribution than A_{min}=1.0. It reduces the RMSE against Agland2000 from 0.180 to 0.165, and against NLCD from 0.175 to 0.161. This is reflected in the structure revealed by Figure 10 where fewer cells in the MLCT data are assigned 100% crop because of the secondary class. Additionally, where crop was included in a secondary class it also caused cells of near-zero value for MLCT to lift away from the x-axis. The uncorrelated observations for Agland2000 equal to 1.0 and NLCD equal to zero are still present, however.

This result persuades us that considering the secondary class with $A_{min} = 0.5$ is indeed the correct approach. From this point forward we will consider only the statistics derived from setting A_{min}=0.5 for the aggregation of the MLCT data due to this improved fit with Agland2000 and NLCD cropland and its full consideration of all information imparted by the MLCT data.

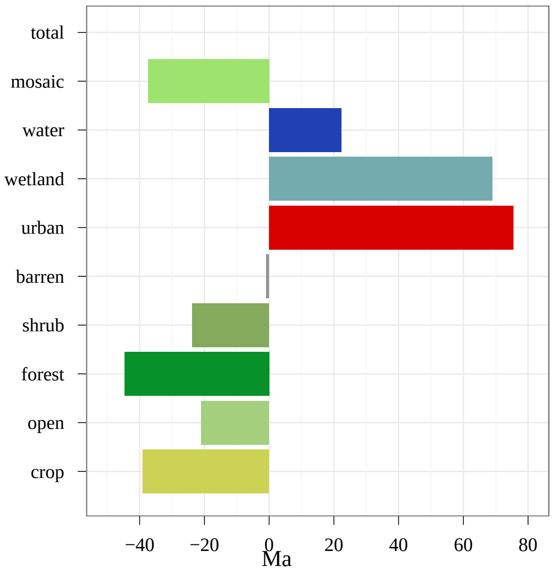
4.4 Evaluation of NLCD Offsets

We can perform the same error calculations as above to assess whether adding NLCD offsets improves overall cropland accuracy. Table 6 gives the RMSEs for the various intermediate products. We see that each step in the algorithm improves the overall error of the data product against both Agland2000 and NLCD.

We next examine the effect on the total areas for all classes. Figure 11 shows the totals by class of the offsets that result from this calculation. The item labeled "total" appears blank because a value of zero is plotted there indicating that area was conserved in this operation, which is to say that area subtracted from one class was reallocated to another. As expected, the most significant offset was for the urban class, representing the low-density infrastructure outside of concentrations of development large enough and dense enough to be identified in the MLCT classification. Water and wetland fractions were also increased to bring the total areas of those classes in line with NLCD.

However, the most important outcome with respect to our stated objective of bringing total cropland areas in line with the total from NLCD and Aglands2000 is the reduction of crop areas by 39 Ma (15.8 Mha) and mosaic areas by 37.4 Ma (15.2 Mha). This total reduction of 57.8 Ma (23.4 Mha) of the final crop class after mosaic decomposition brings the national cropland area for PEEL_0 into close agreement with NLCD, MLU, and Agland2000, 437.6 Ma vs. 449.2, 441.3, and 446.5 Ma respectively. See Table 5 and Figure 12 for a detailed comparison of the effects of the offset. Figure 13 shows two additional hexbin plots using the offset-adjusted MLCT fractions, which is the PEEL_0 data set short of disaggregating the cropland area into the crop sub-classes. The six RMSE values given in Table 6 correspond to the six hexbin plots in Figures 10 and 13.

The significance of this result is that it is not conditioned by the desired cropland area estimate, rather it shows a convergence in these estimates by selectively incorporating information about other classes from NLCD. Figure 20 shows the PEEL$_0$ data set with a single cropland class.



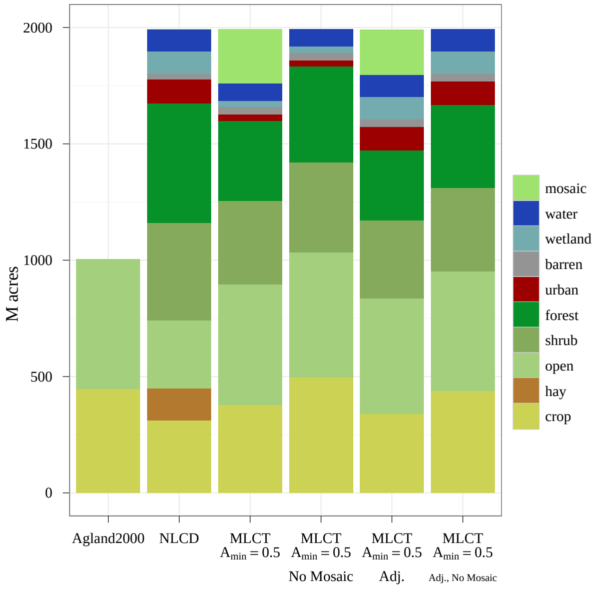


Figure 11: Total offsets calculated from NLCD.

Figure 12: Total acreages after NLCD adjustment.

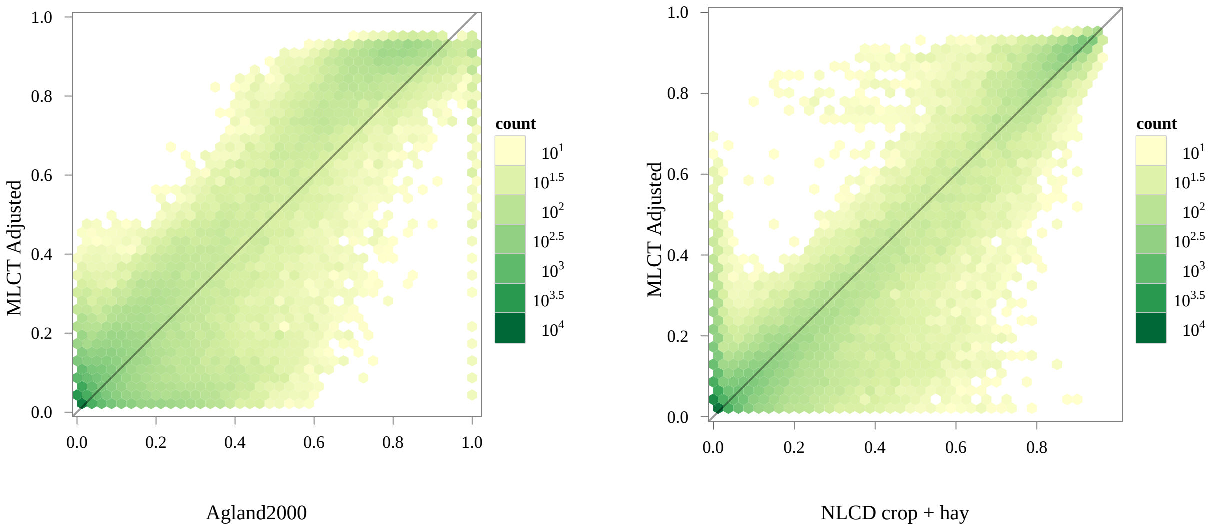


Figure 13: Hexbin plot of MLCT adjusted crop versus Agland2000 cropland (left) and versus NLCD crop plus hay/pasture (right).

4.5 Potential Further Offsets and Adjustments

There is little consensus among the datasets regarding the relative coverage of forest, shrub, and open land, or indeed the definitions of these classes. Thus, we find it impractical to compare among them. MLCT shows 1220.3 Ma of land in one of these three classes, just slightly less than the final amount (1228.6 Ma) in PEEL$_0$, the additional area from mosaic decomposition being roughly cancelled by the reductions from the WWU offsets. This number compares favorably to the 1224.5 Ma distributed among these three classes in NLCD and the 1241.3 Ma distributed among forest and open in MLU, which lacks a class analogous to "shrub".

We attribute these differences to inconsistent class definitions due to the fairly continuous transition from open to shrub and from shrub to forest. For instance, NLCD defines forest land to be any land with greater than 25% coverage by a tree canopy that is generally greater than six meters tall. Shrubland in NLCD is similarly defined as any area with a shrub canopy (defined as generally less than 5 or 6 meters high) typically covering greater than 20% of the pixel. Open herbaceous land is thus roughly everything else that is not subject to intensive management (besides wetlands), meaning the pixel is at least 80% open grassland. This definition is rather strict by comparison to the definitions of open grasslands used in MLU and MLCT, which leads to substantial disagreement among these datasets and is simply the result of class definitions.

Thus far we have refrained from any direct adjustment of the cropland layer in MLCT, and have shown that, given a principled approach and understanding of the differences between these datasets, it is possible to bring the USDA MLU, Agland2000, NLCD, and MLCT estimates of cropland largely into agreement at the national level and in terms of their distributions. If we accept the premise that higher resolution datasets, such as NLCD, or datasets that include some regional census information, such as Agland2000, are likely to contain more accurate representations of cultivated and managed lands, then it follows that these data should contain additional information about the spatial distribution of cropland within a given region that can be accommodated by a similar offset mechanism applied directly to the cropland data layer.

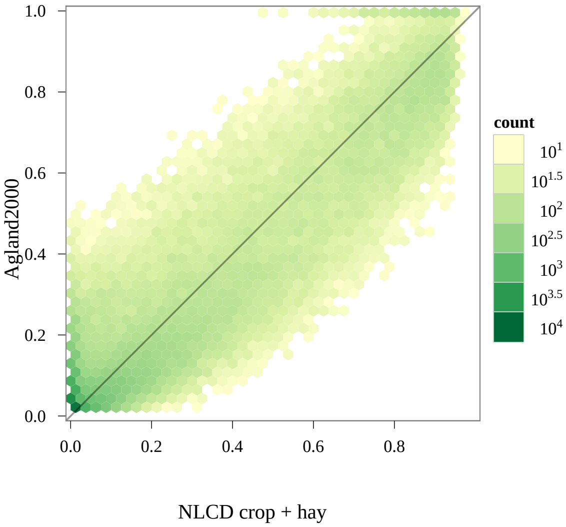


Figure 14: Hexbin plot of NLCD cropland plus hay/pasture versus Agland2000 crop.

Figure 14 shows the hexbin plot of the NLCD 'crop+hay' vs. Agland2000 cropland. The RMSE across the cUSA between these layers is 0.129, which implies that there is roughly as much variation in the spatial distribution of cropland between these two products as there is between each of them and our improved MLCT product. We thus see little principled motivation to choose one or the other crop layer as a target for further tuning. We will explore these distinctions and potential improvements along these lines in future work.

5 Computational Tools and Environment

Because our goal is to develop the most accurate and consistent complete global LULC data set, we focus on developing open-source algorithms, to be released with the dataset upon publication, that are easily extensible and can accommodate new data and techniques as they become available. We refer the reader to *Best* [2011] for further details on the software environment and reproducible research framework under which this work was conducted. We note this analysis would not have been possible without the excellent `raster` package for `R` developed by *Hijmans and van Etten* [2011] and other contributors. We consider this interface to geospatial raster data sets in the `R` statistical analysis environment an important contribution to spatial analysis and a laudable accomplishment because it unleashes the power of a sophisticated, popular, open-source, free software programming language for statistical operations on large geospatial raster data sets.

6 Conclusions

We have described and validated a new algorithm for combining a wide range of land cover data from a number of sources, scales, and extents to construct hybrid data products with improved accuracy and fidelity. Because many of the most interesting steps in our algorithm depend on high quality, high resolution land cover data that is not yet available on a global basis, we have focussed our description of the algorithm only on the US. Even so, the positive results that we have obtained in the US lead us to expect that our procedures for aggregating MLCT to a coarser resolution using a statistical interpretation of sub-pixel cover and distinguishing likely components of the mosaic class, can also help to more accurately characterize the distribution of global land cover, especially cropland.

Because we have constructed the portions of the algorithm that require high resolution/high value data as offsets to the initial cover representation, we can extend the algorithm globally, calculating, estimating, or modeling these offsets region by region where possible, and ignoring them where information is unavailable. Using this algorithm, we have constructed a customized data product, PEEL_0, designed to initialize the PEEL model, but we stress that the target of this work is is more than just a static land cover data set; it is an algorithmic framework and open source code base for producing global or regional land cover data products with customized cover representations, which integrates high value information at multiple scales to improve and validate land cover characterizations.

In the absence of high-resolution data on rural development, the low-density portion of the "urban" class that falls below MLCT's detection threshold, we propose that it might be possible to model the over-estimation factor of the MLCT cropland class. This factor would be defined as the ratio of total area encompassed by the MLCT cropland classification to acreage actually under cultivation and could potentially be modeled as a function of classification confidence and secondary class using the data described and produced here as a training set. The null hypothesis in the formulation of such a model is that there is enough diversity among agricultural landscapes in our cUSA study area to adequately characterize agricultural landscapes around the world in this regard. Similarly it might be possible to directly model the “urban” percentage below the MLCT detection threshold as a function of population density and agricultural productivity, identifying said threshold in the process. There is a clear dependency between these offsets in agriculturally productive regions so modeling them in conjunction somehow may be constructive. We expect that global offsets for the water and wetland classes will be harder to obtain without corresponding proxy statistics with which to formulate a model but perhaps we can expect greater availability of spatially explicit catalogs of ecological services and sensitive/protected areas in the near future that would close these gaps in the available information. These directions will be pursued in future work.

References

Bartholomé, E., and A. S. Belward (2005), GLC2000: a new approach to global land cover mapping from earth observation data., *International Journal of Remote Sensing*, *26*(9), 1959 – 1977.

Best, N. (2011), Synthesis of a complete land use/land cover data set for the conterminous united states emphasizing accuracy in area and distribution of agricultural activity, Master’s thesis, Northeastern Illionois University.

Biradar, C. M., et al. (2009), A global map of rainfed cropland areas (GMRCA) at the end of last millennium using remote sensing, *International Journal of Applied Earth Observation and Geoinformation*, *11*(2), 114 – 129, doi:DOI:10.1016/j.jag.2008.11.002.

Fisher, P. F., A. J. Comber, and R. Wadsworth (2005), *Re-presenting GIS*, chap. Land Use and Land Cover: Contradiction or Complement, pp. 85–98, John Wiley & Sons Ltd.

Friedl, M. A. (2002), Global land cover mapping from MODIS: algorithms and early results, *Remote Sensing of Environment*, *83*(1-2), 287–302.

Friedl, M. A., D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang (2010), MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets, *Remote Sensing of Environment*, *114*(1), 168–182, doi:10.1016/j.rse.2009.08.016.

Hansen, M., R. DeFries, J. R. G. Townshend, and R. Sohlberg (2000), Global land cover classification at 1 km resolution using a decision tree classifier, *Int J Rem Sens*, *21*, 1331–1365.

Hijmans, R., N. Garcia, and J. Wieczorek (2009), Global administrative areas (gadm) database.

Hijmans, R. J., and J. van Etten (2011), *raster: Geographic analysis and modeling with raster data*, r package version 1.8-12.

Homer, C., C. Huang, L. Yang, B. Wylie, and M. Coan (2004), Development of a 2001 National Land-Cover Database for the United States, *Photogrammetric Engineering Remote Sensing*, *70*(7), 829–840.

Homer, C., et al. (2007), Completion of the 2001 National Land Cover Database for the Conterminous United States, *Photogrammetric Engineering and Remote Sensing*, *73*(4), 337–341.

Jones, R., R. Leemans, L. Mearns, N. Nakicenovic, A. Pittock, S. emenov, and J. kea (2001), *IPCC AR-3 Working Group 2*, chap. 3. Developing and Applying Scenarios, pp. 147–190, Cambridge University Press.

LP DAAC (2008), Modis land cover type (MLCT, MCD12Q1 v005), `https://lpdaac.usgs.gov/lpdaac/products/modis_products_table/land_cover/yearly_l3_global_500_m/mcd12q1`, these data are distributed by the Land Processes Distributed Active Archive Center (LP DAAC), located at the U.S. Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center (lpdaac.usgs.gov).

Lubowski, R. N., M. Vesterby, S. Bucholtz, A. Baez, and M. J. Roberts (2006), Major Uses of Land in the United States, 2002. (Economic information bulletin; no. 14), *Tech. rep.*, United States Department of Agriculture, Economic Research Service.

Matsuoka, Y., M. Kainuma, and T. Morita (1995), Scenario analysis of global warming using the Asian Pacific Integrated Model (AIM), *Energy Policy*, *23*(4-5), 357–371, doi:10.1016/0301-4215(95)90160-9.

Monfreda, C., N. Ramankutty, and J. A. Foley (2008), Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000, *Global Biogeochemical Cycles*, *22*(1), 1–19, doi:10.1029/2007GB002947.

Ramankutty, N., A. T. Evan, C. Monfreda, and J. A. Foley (2008), Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000, *Global Biogeochemical Cycles*, *22*(1), 101,029/, doi:10.1029/2007GB002952.

Sellers, P. J., et al. (1997), Modeling the exchanges of energy, water, and carbon between continents and the atmosphere, *Science*, *275*(5299), 502–509, doi:10.1126/science.275.5299.502.

Thenkabail, P., et al. (2008), A Global Irrigated Area Map (GIAM) Using Remote Sensing at the End of the Last Millennium.

van Vuuren, D. P., B. Eickhout, P. L. Lucas, and M. G. J. den Elzen (2006), Long-Term Multi-Gas Scenarios to Stabilise Radiative Forcing Exploring Costs and Benefits Within an Integrated Assessment Framework, *Energy Journal*, *3*(Special Issue #3), 201–234.

van Vuuren, D. P., M. G. J. den Elzen, P. L. Lucas, B. Eickhout, B. J. Strengers, B. van Ruijven, S. Wonink, and R. van Houdt (2007), Stabilizing greenhouse gas concentrations at low levels: an assessment of reduction strategies and costs, *Climatic Change*, *81*(2), 119–159, doi:10.1007/s10584-006-9172-9.

You, L., and S. Wood (2006), An entropy approach to spatial disaggregation of agricultural production, *Agricultural Systems*, *90*(1-3), 329–347, doi:10.1016/j.agsy.2006.01.008.

You, L., S. Wood, and U. Wood-Sichra (2006), Generating global crop distribution maps: from census to grid, in *Selected paper at IAEA 2006 Conference at Brisbane, Australia*, 202, pp. 1–16.

A cUSA maps of data sets

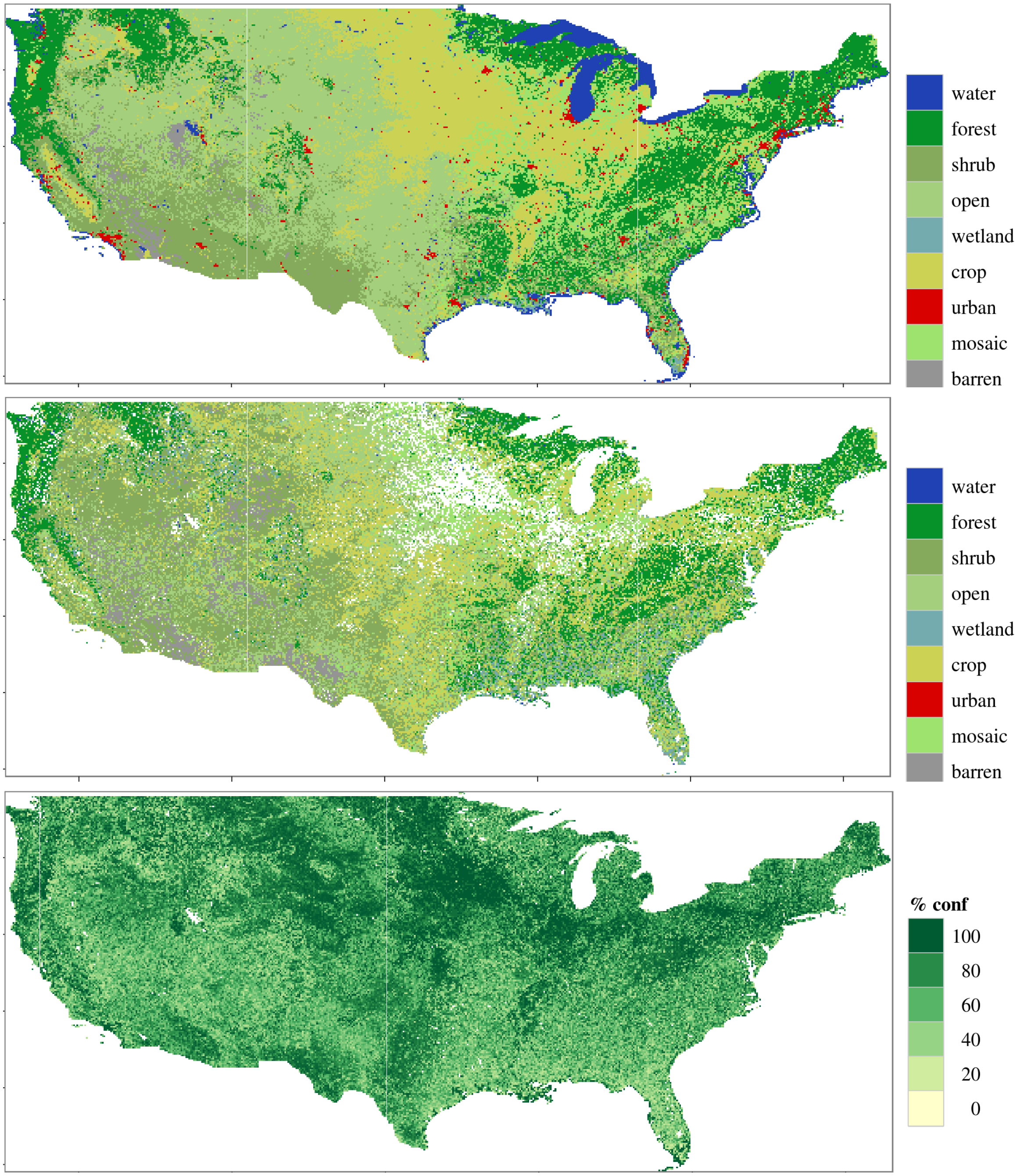


Figure 15: MLCT primary reclassified cover (top), secondary reclassified cover (middle), and primary cover classification confidence(bottom) for the cUSA.

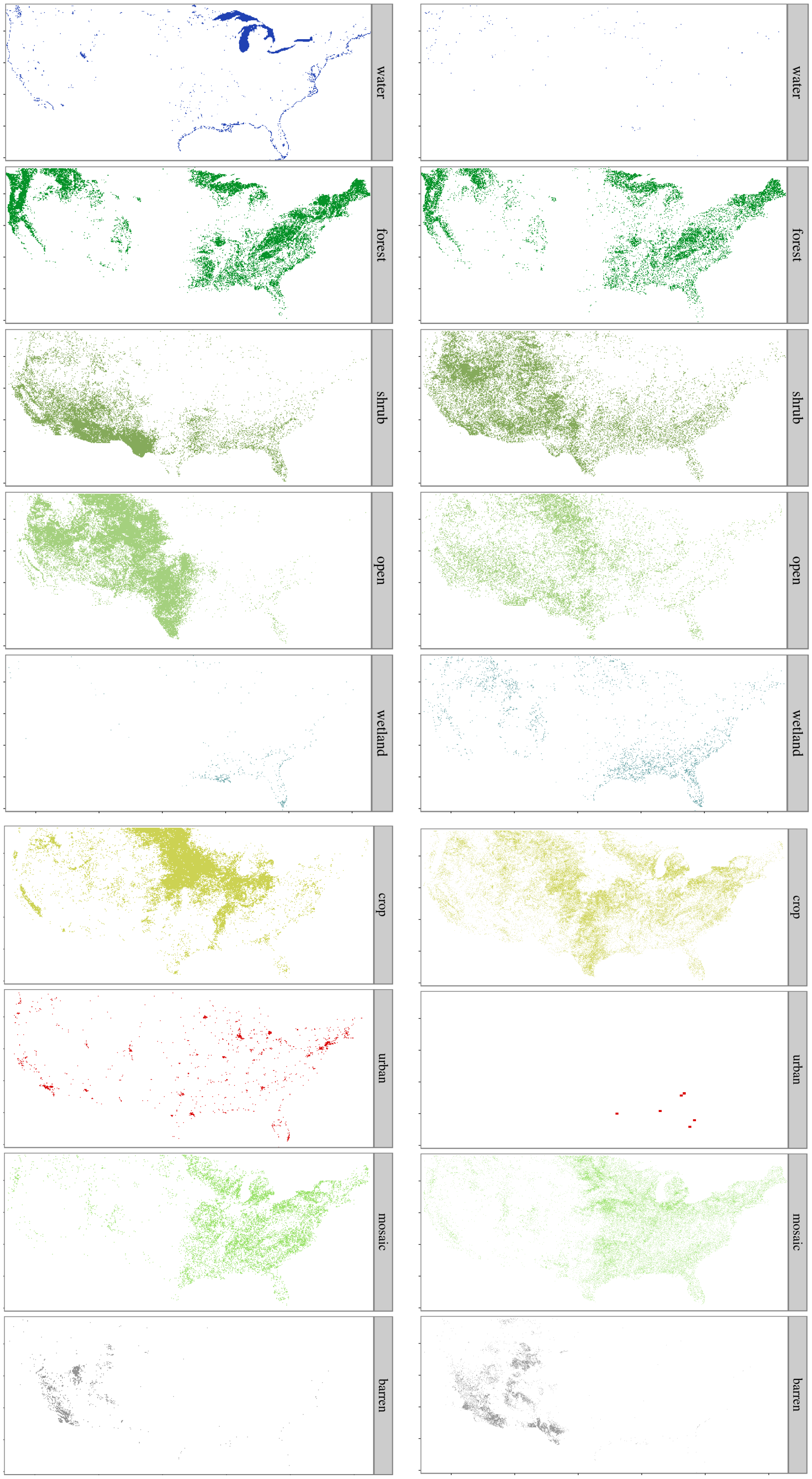


Figure 16: MLCT primary (left) and secondary covers (right).

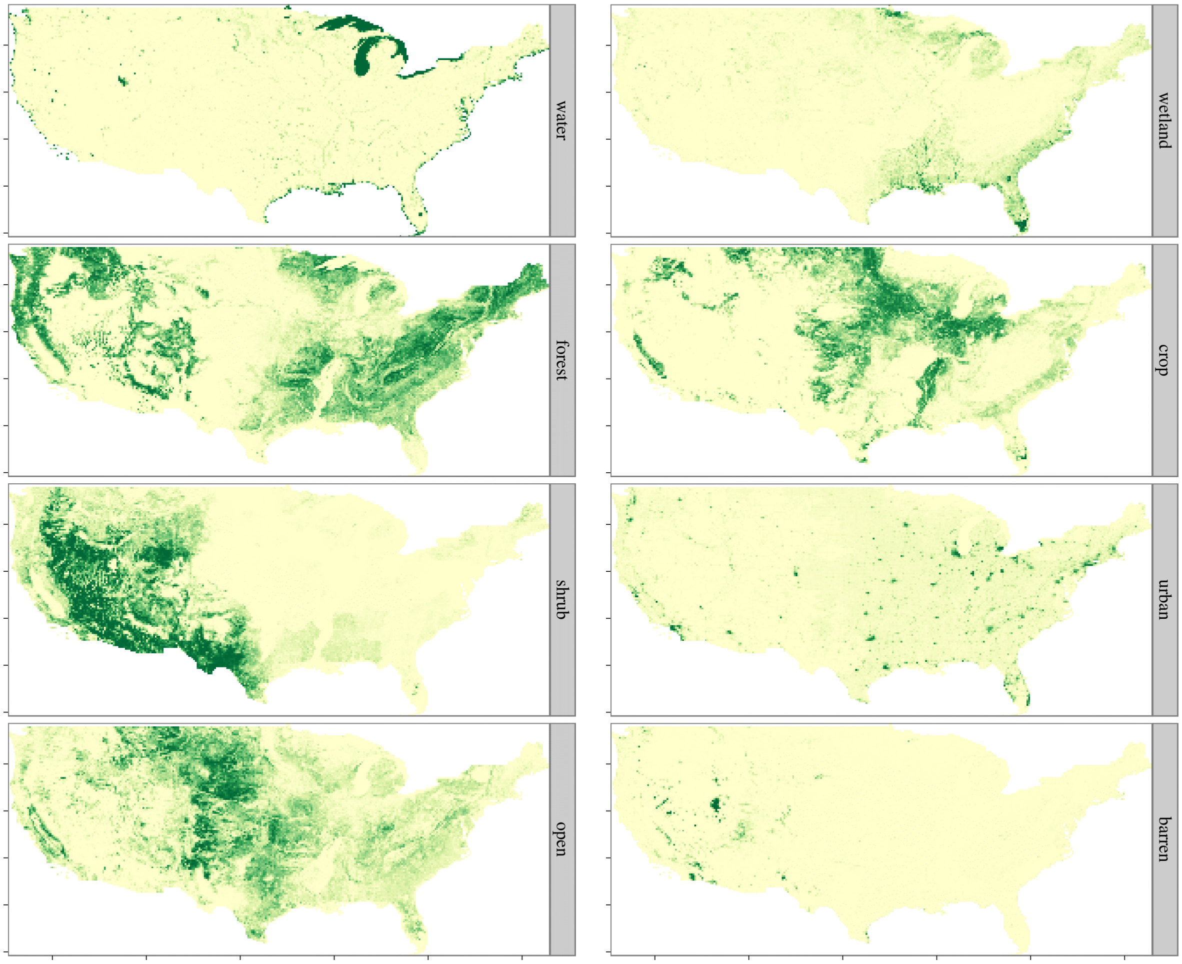


Figure 17: NLCD aggregated cover fractions.

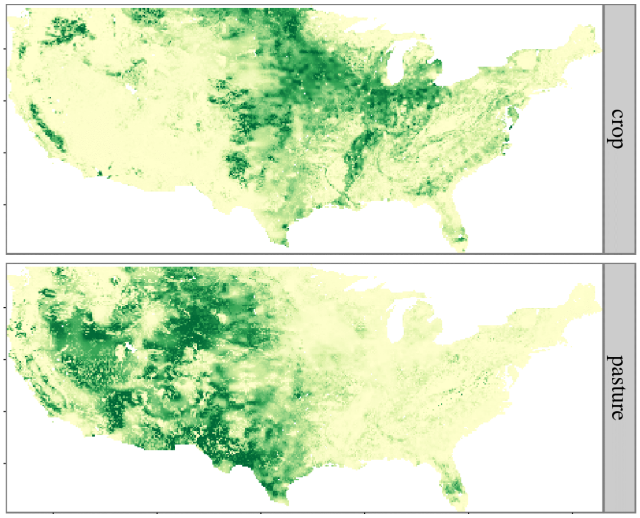


Figure 18: Agland2000 distribution in cUSA study area.

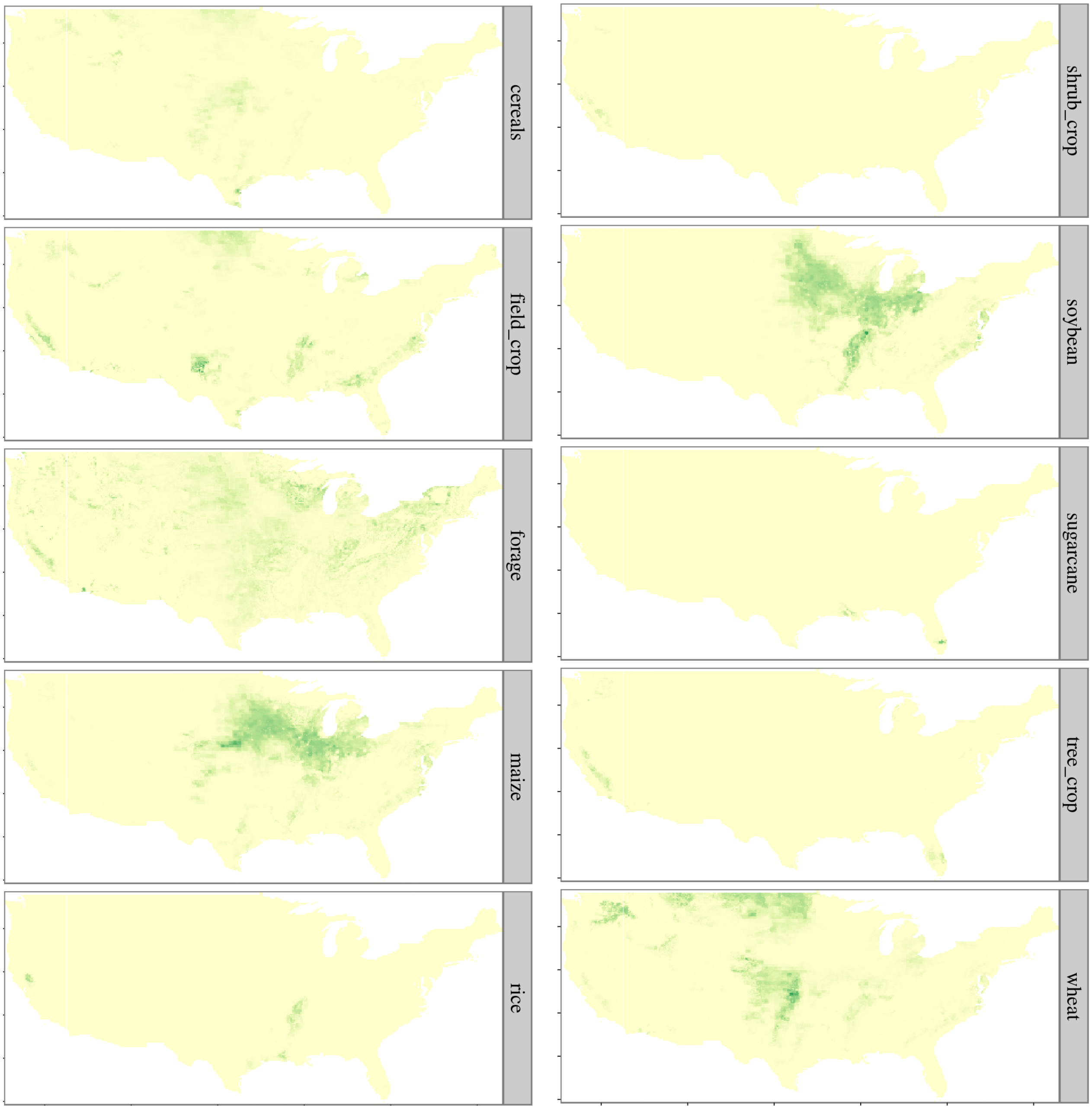


Figure 19: 175Crops2000 category maps.

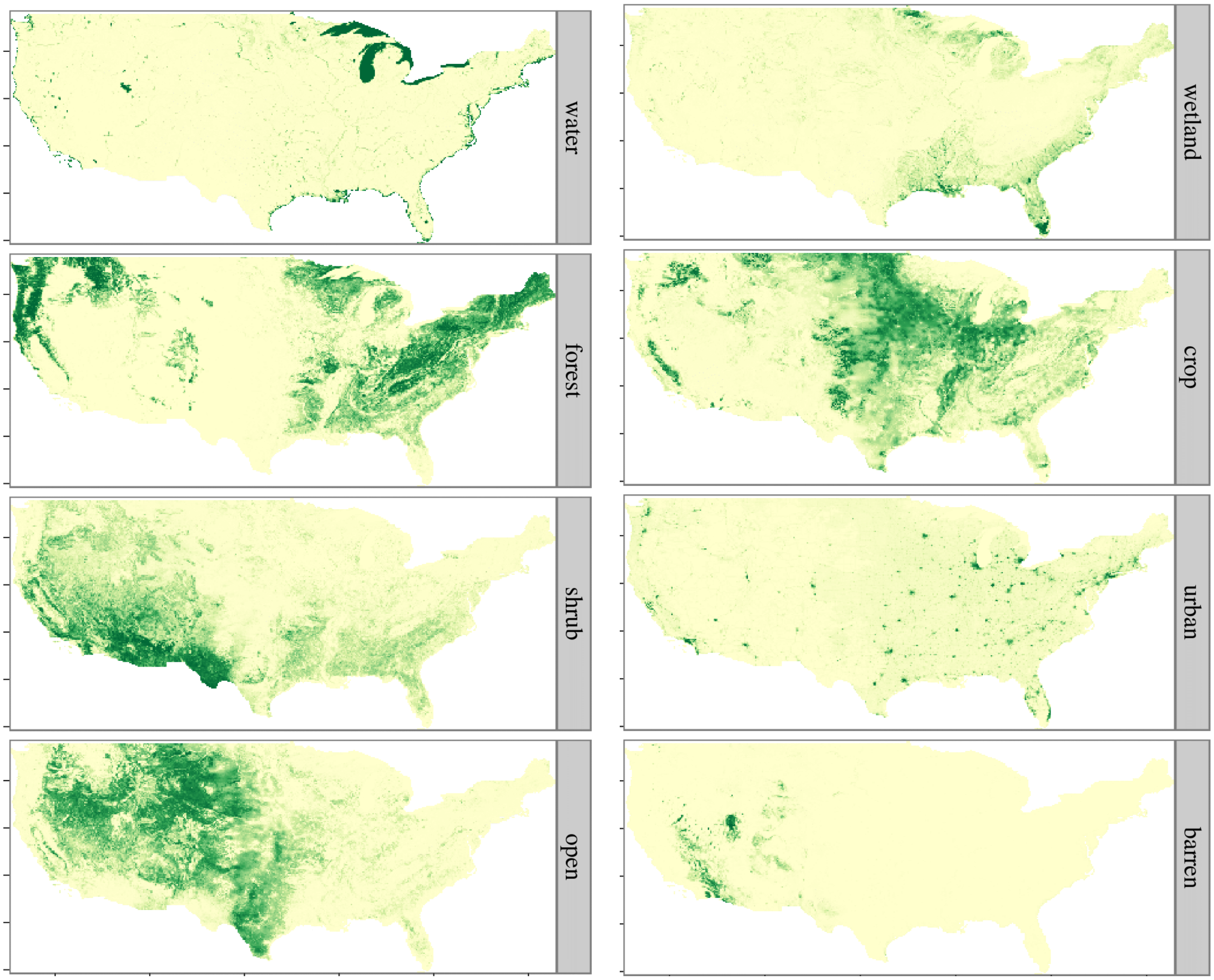


Figure 20: Final PEEL_0 maps.

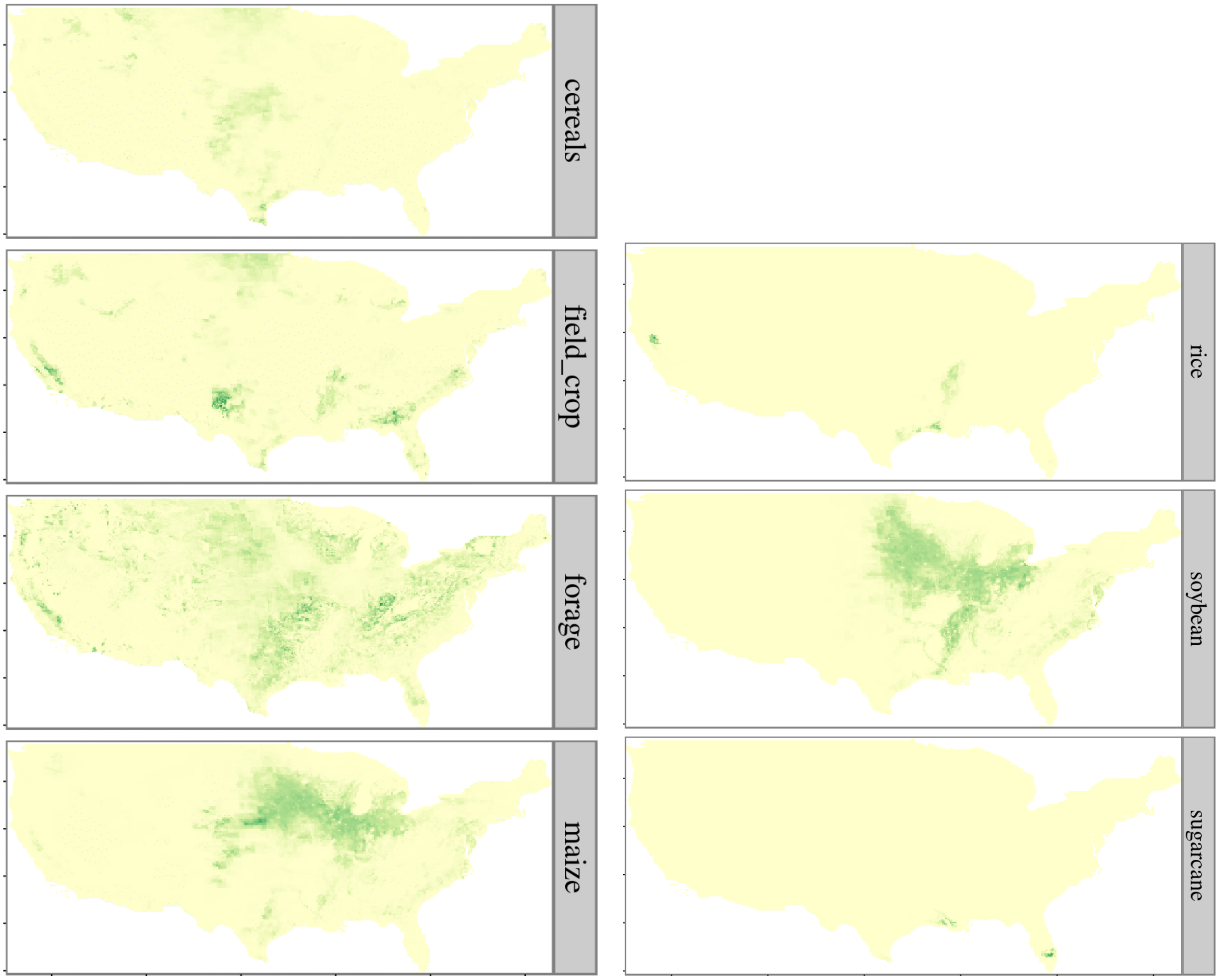


Figure 21: Normalized fractions for crop sub-classes.