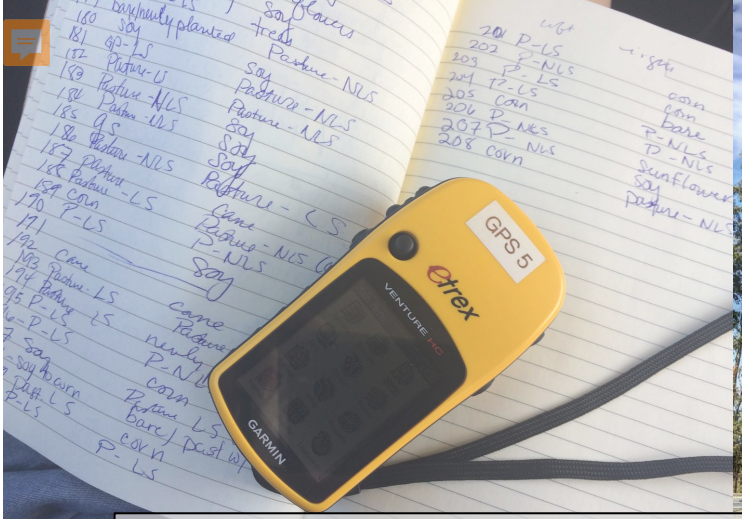



A Landsat-based Global Land Cover Training Dataset from 1985 to 2019

Stanimirova R., Tarrio, K., Turlej, K., McAvoy, K., Stonebrook S., Hu, K-T., Arevalo, P., Zhang, Y., Bullock, E., Woodcock, C.E., Loveland, T.R., Olofsson, P., Zhu, Z., Barber, C., Friedl, M.A.



Photo credit: Radost Stanimirova, Yingtong Zhang

BOSTON UNIVERSITY

NASA

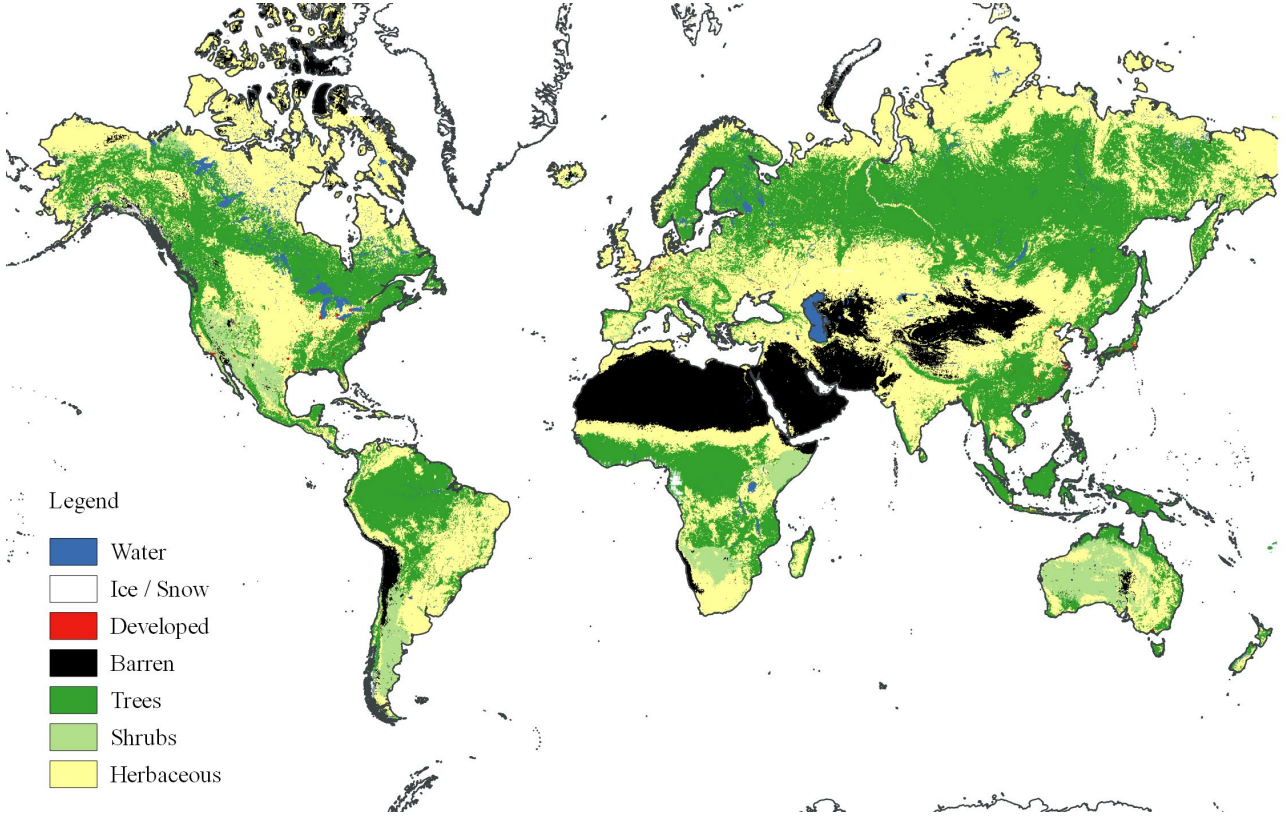
International Collaborators

- Australia
- Brazil
- Mexico
- Botswana
- Laos
- Ireland
- Mozambique
- Russia
- Bangladesh
- Germany
- India
- Togo

Motivation

- Machine learning (ML) models require large, spatially explicit training datasets.
- Collecting high quality training data is costly and labor-intensive.

GLanCE: Global Land Cover and Estimation

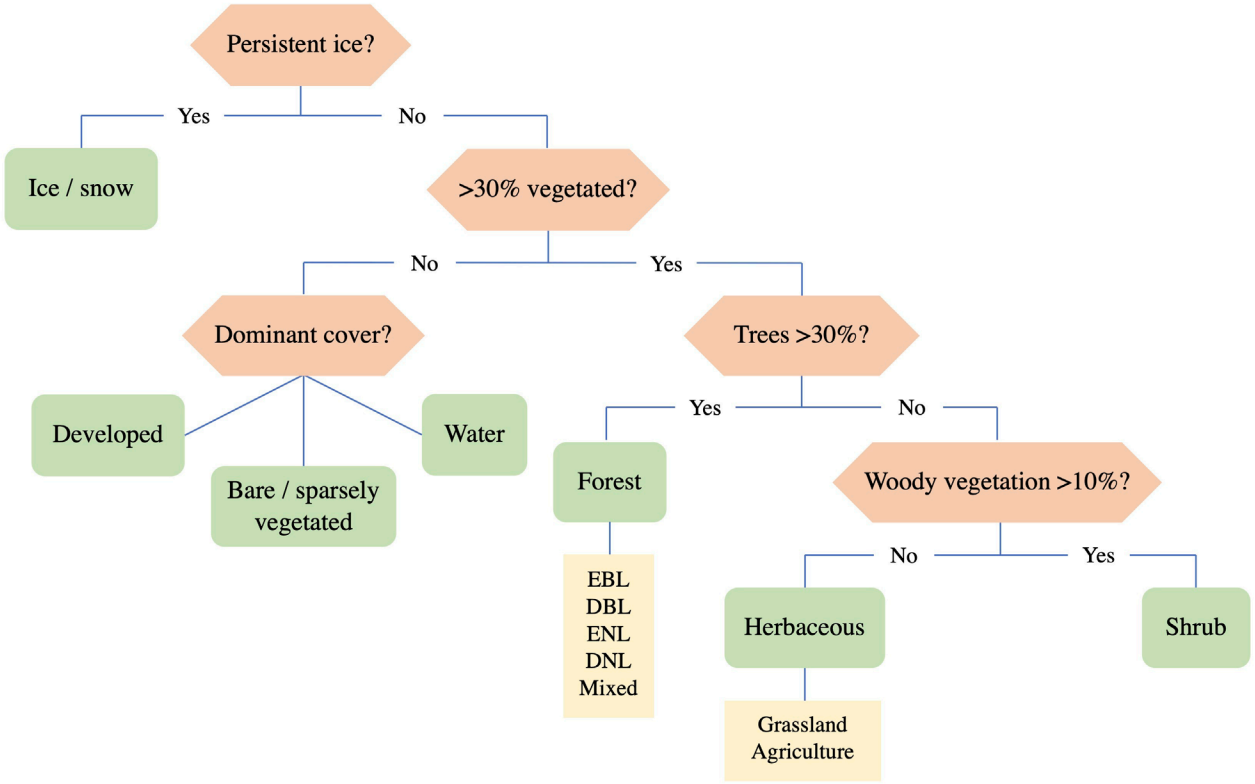


Friedl et al., 2022

Objectives

- Leverage GEE and ML to ensure training data quality and ecoregion representation.
- Represent abrupt change and gradual land cover transitions in the training dataset.
- Build a community resource.

GLanCE land cover legend



Level 1	Level 2
Ice/snow	Ice/snow
Water	Water
Developed	Developed
Barren / sparsely vegetated	Soil
	Rock
	Beach/Sand
Trees	Deciduous
	Evergreen
	Mixed
Shrub	Shrub
Herbaceous	Grassland
	Agriculture
	Moss/lichen

Friedl et al., 2022

Data fields

Leaf type

% impervious

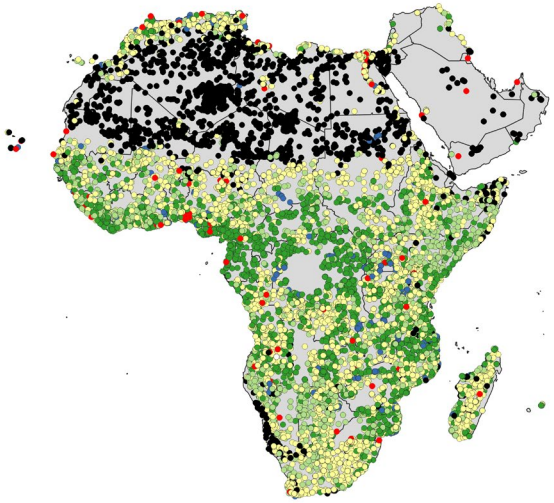
Edges

Vegetation density

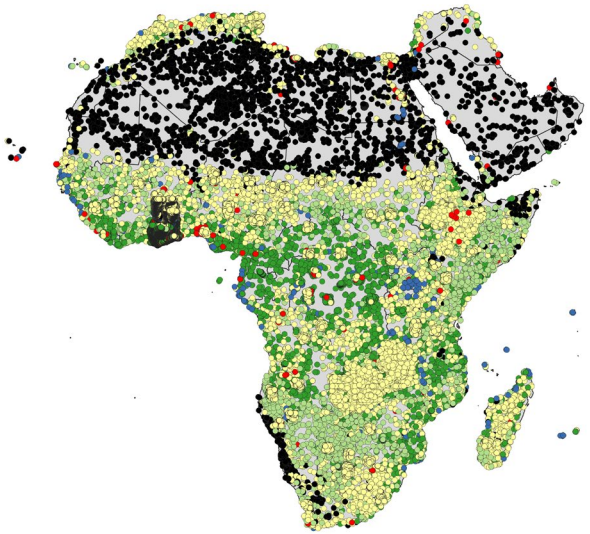
- Leaf type
- % impervious
- Forest edge vs interior
- Vegetation density (%)
- Change
- Confidence LC label

Training data set collection and curation

In-house (n=15,016)

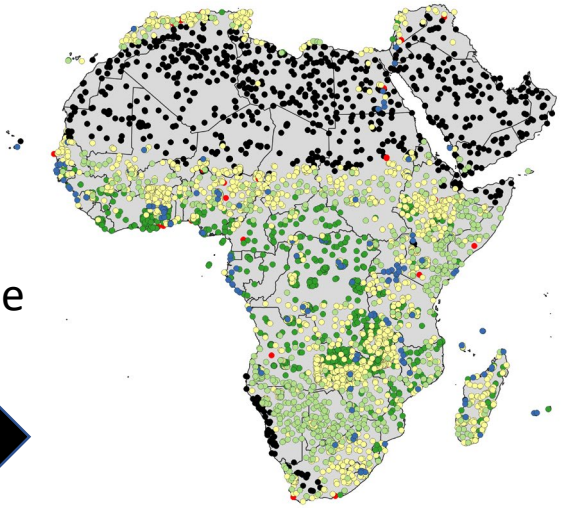


Supplementary ALL (n=268,009)



Sample down

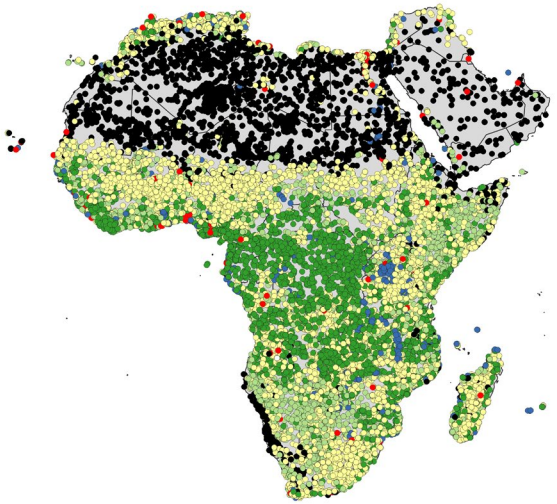
Supplementary sample (n=5661)



We supplement training data collected in-house with publicly available and collaborator contributed ancillary data

Training data set collection and curation

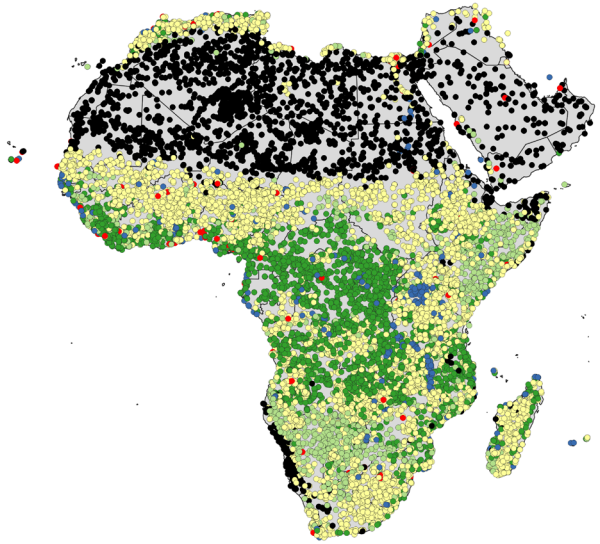
Combined (n=22,210)



Filter

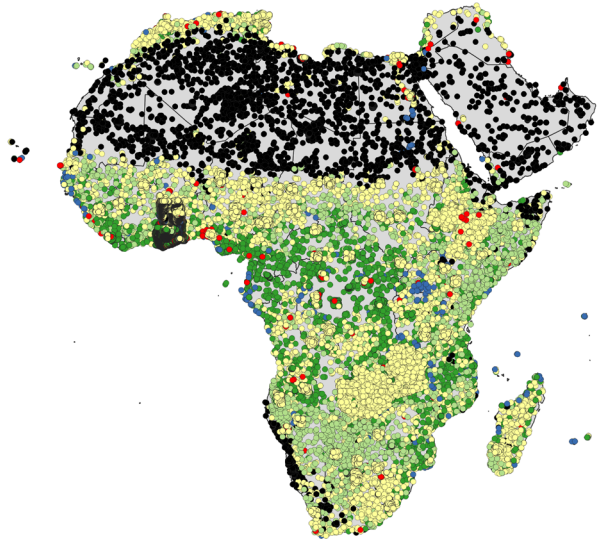
Product #1

Filtered (n=18,394)



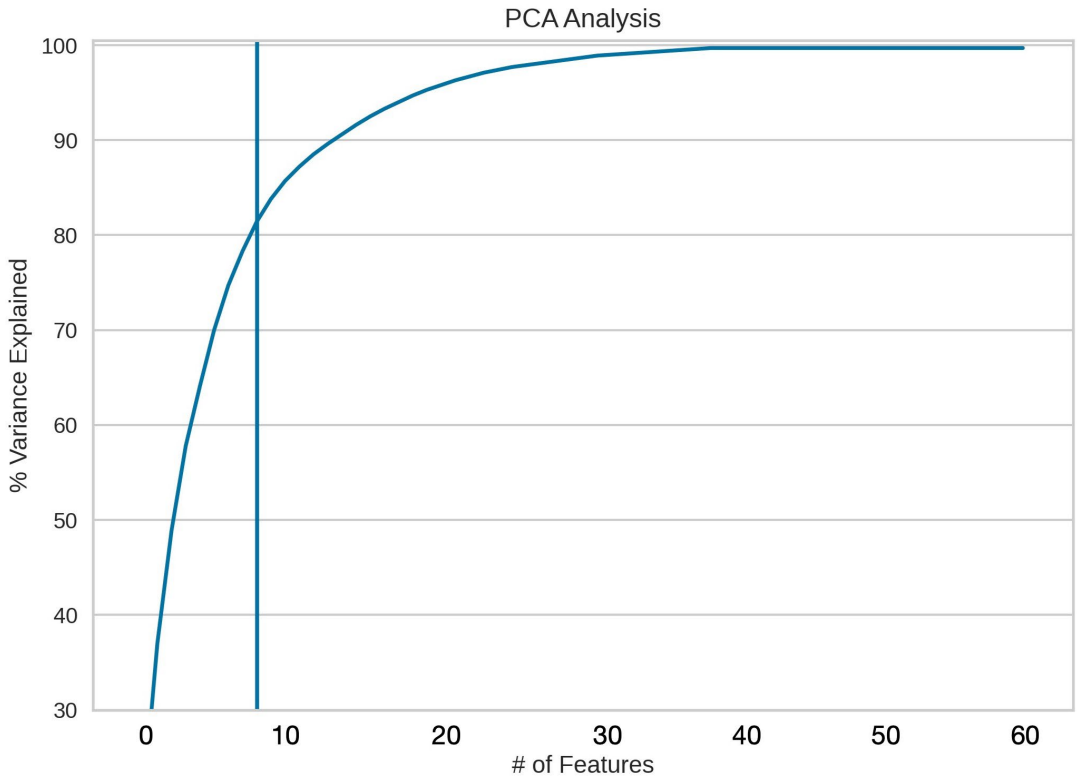
Product #2

Full dataset (n=283,025)

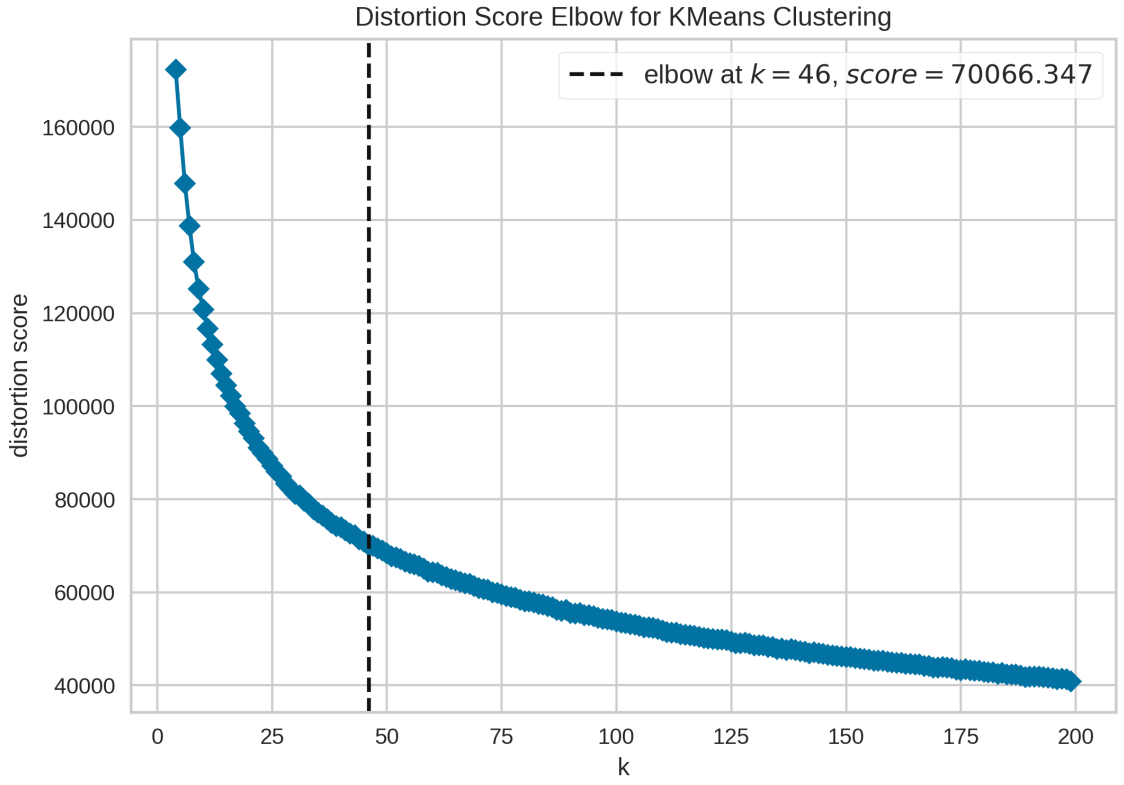


In-house data: Cluster-based sampling

Principal Component Analysis



K-means clustering

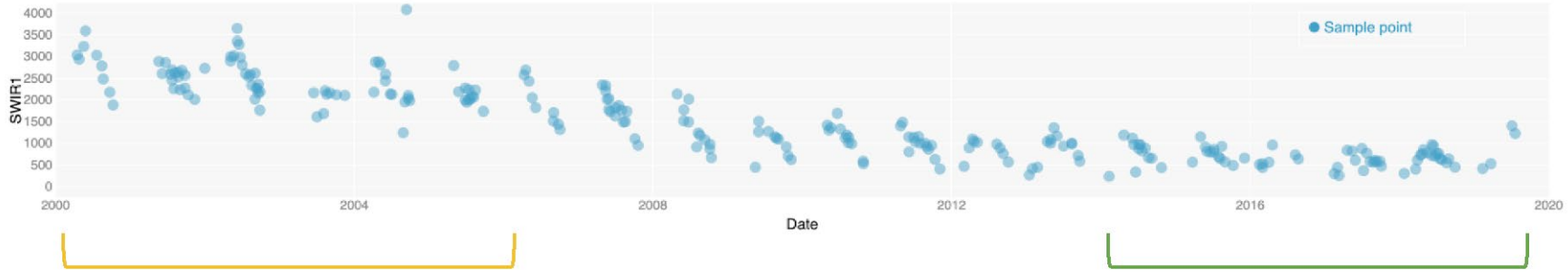


In-house data: Data collection

Pixel is definitely grass-like from 1997 - 2007

Pixel is definitely forest from 2013 - onward

Pixel gradually becomes forested from 2007-2013 (grass > shrub > tree succession), but hard to tell *when* (+ no high res. imagery to confirm)

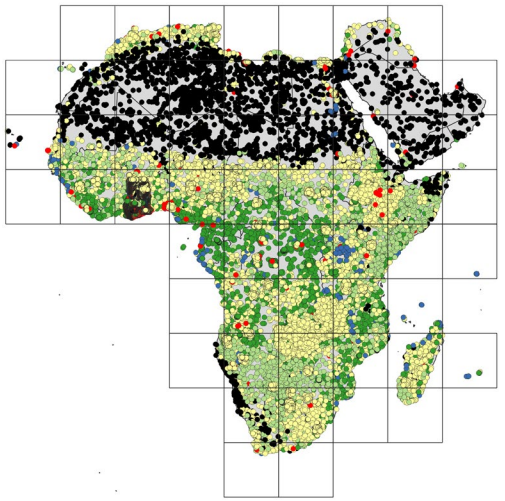


Transitional herbaceous (regrowth)

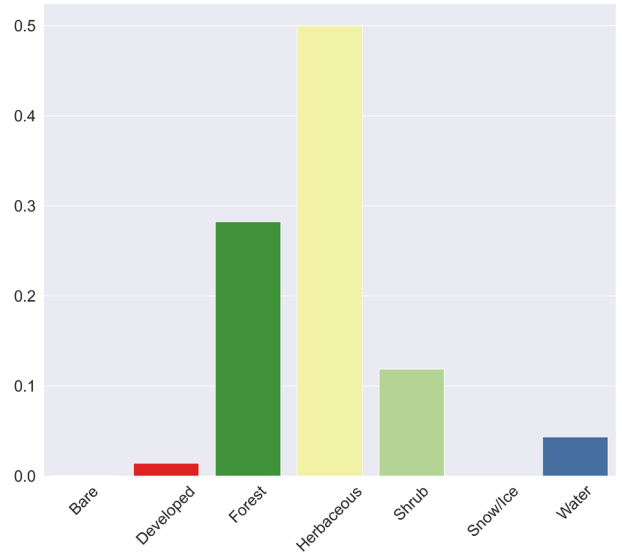
Transitional forest (regrowth)

Supplementary data: Sampling

Sample down supplementary

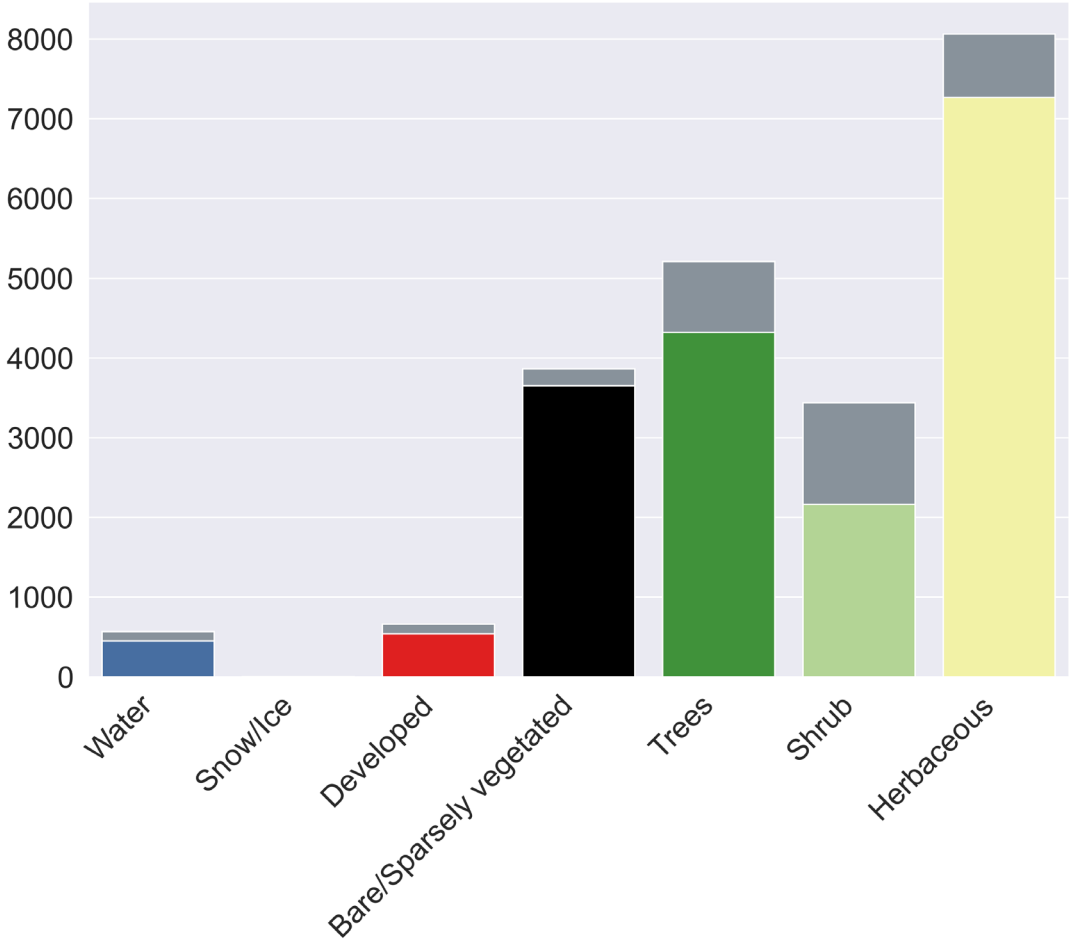


Proportional distribution



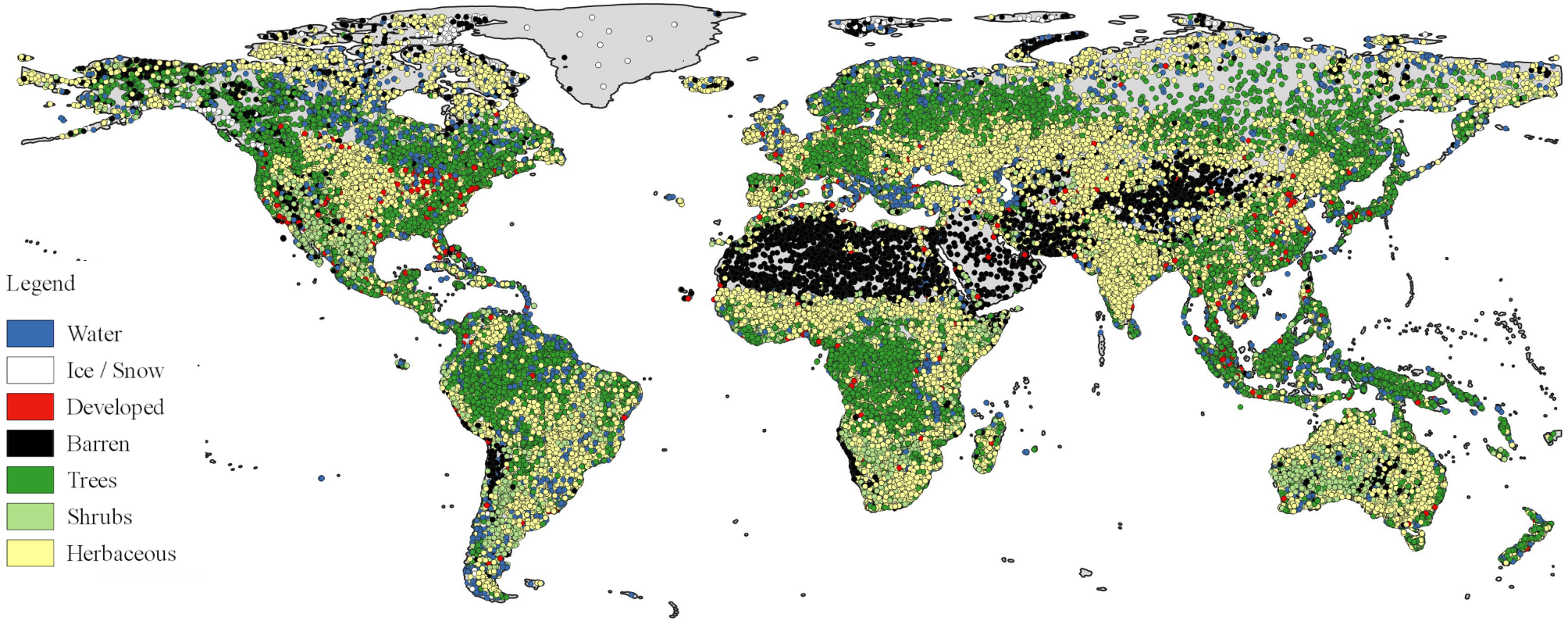
1. Add samples proportionally to underlying distribution per grid
2. Add samples spatially distributed
3. Identify candidate MODIS LC pixels as training samples based on an approach by Zhang & Roy (2017)

Filtering



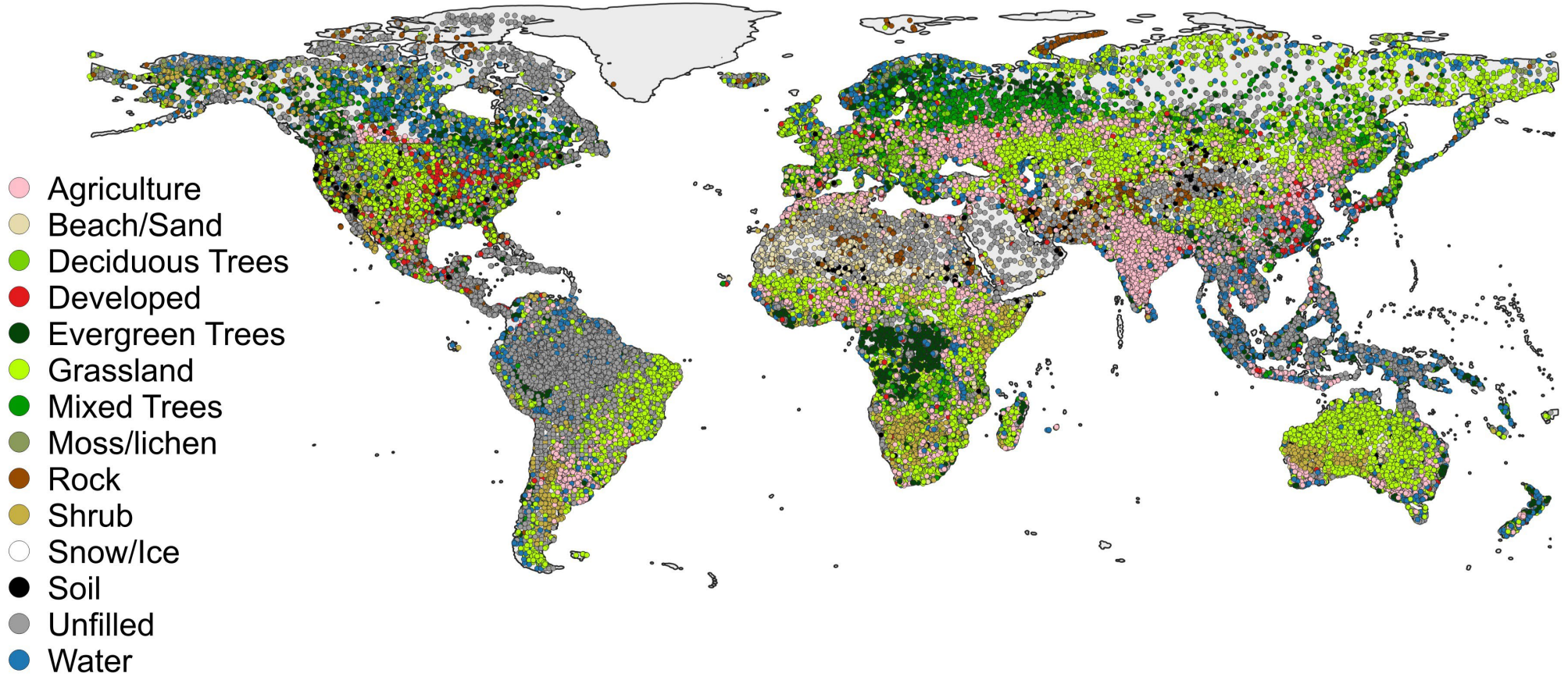
- Used a cross validation procedure to remove poorly labeled samples (Brodley & Friedl, 1999)
- With this procedure we remove ~15% of training data

Level 1 sample (n=129,948)



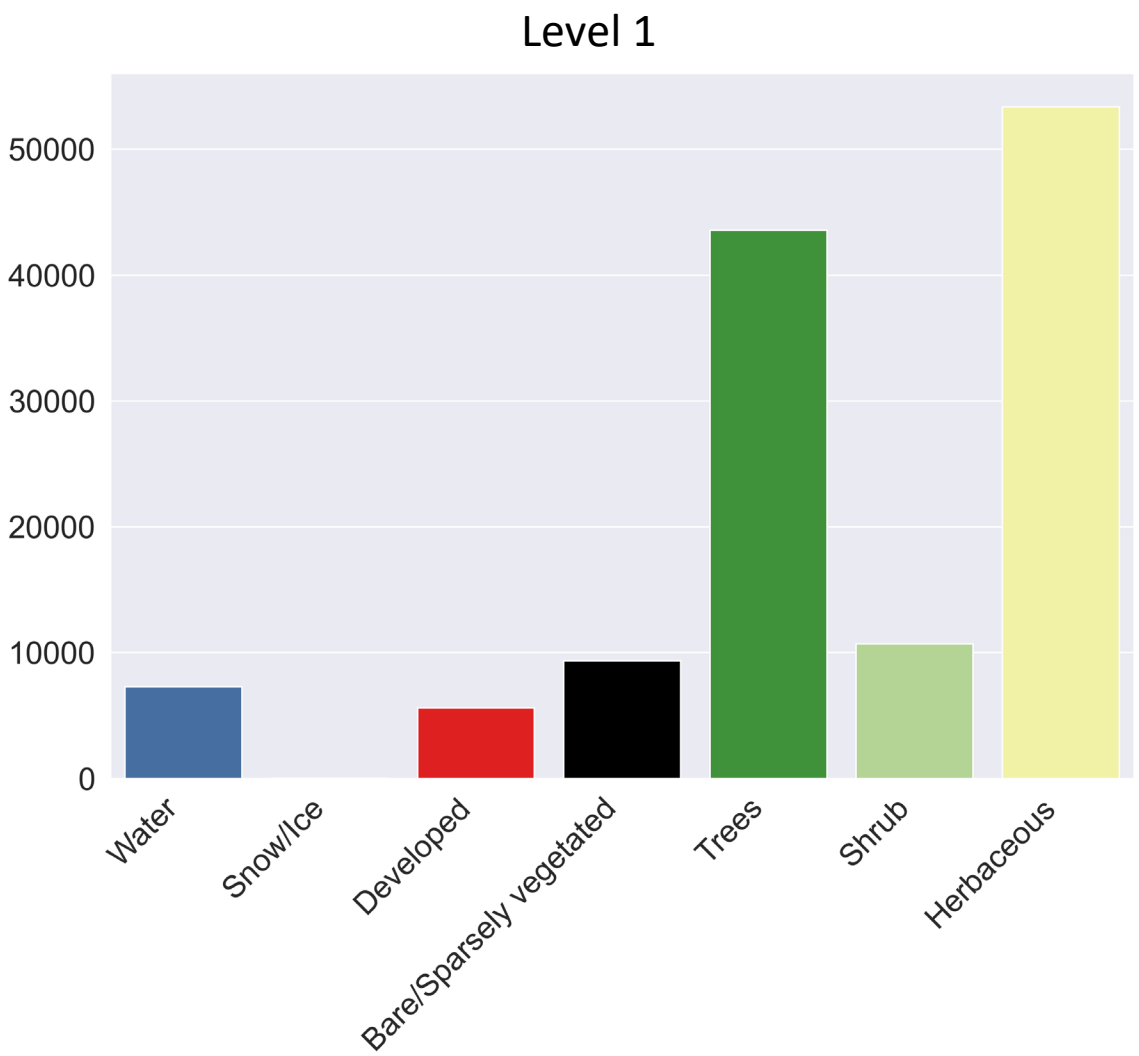
Out of 2,033,832 total points available to us.

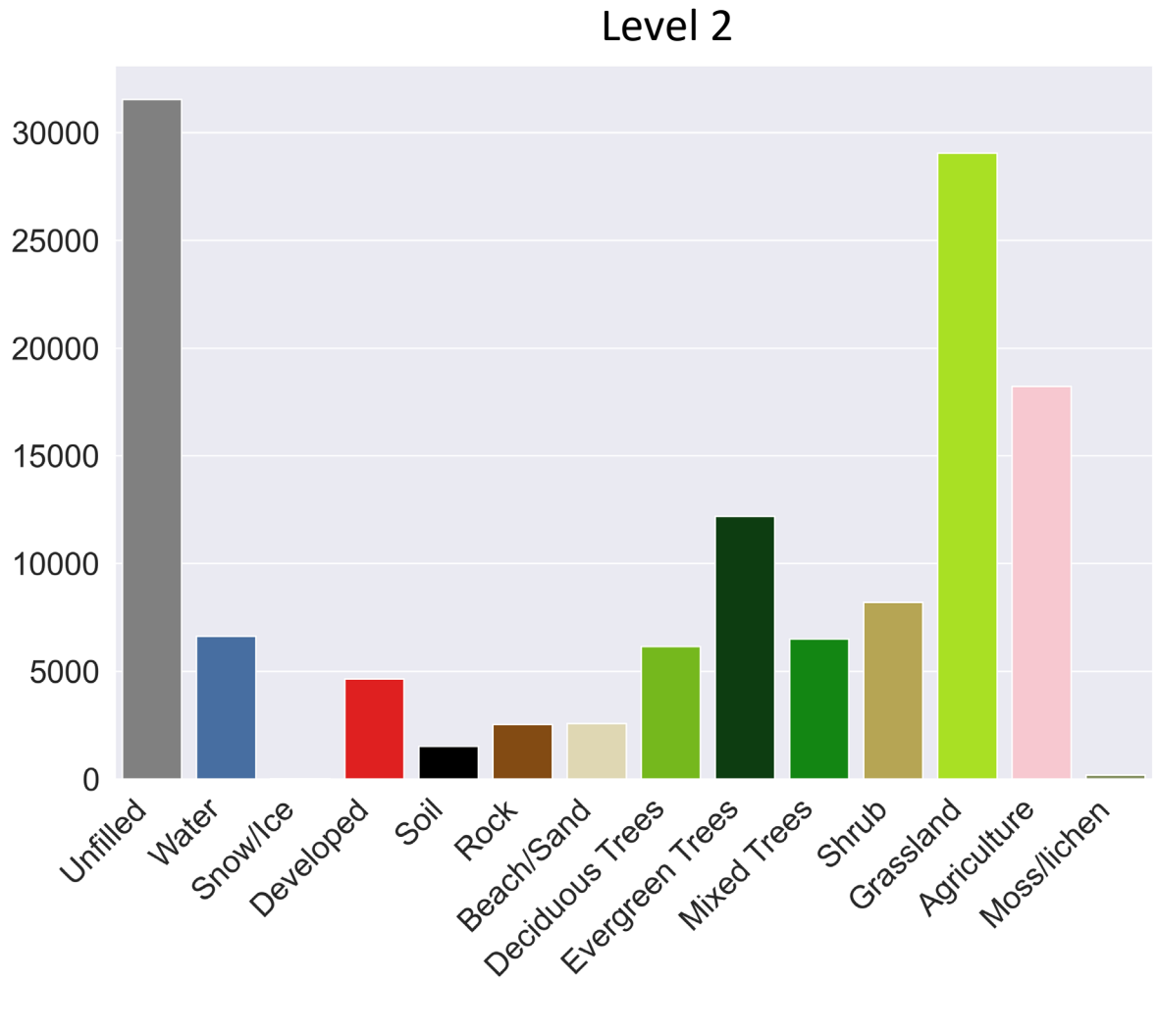
Level 2 sample (n=129,948)



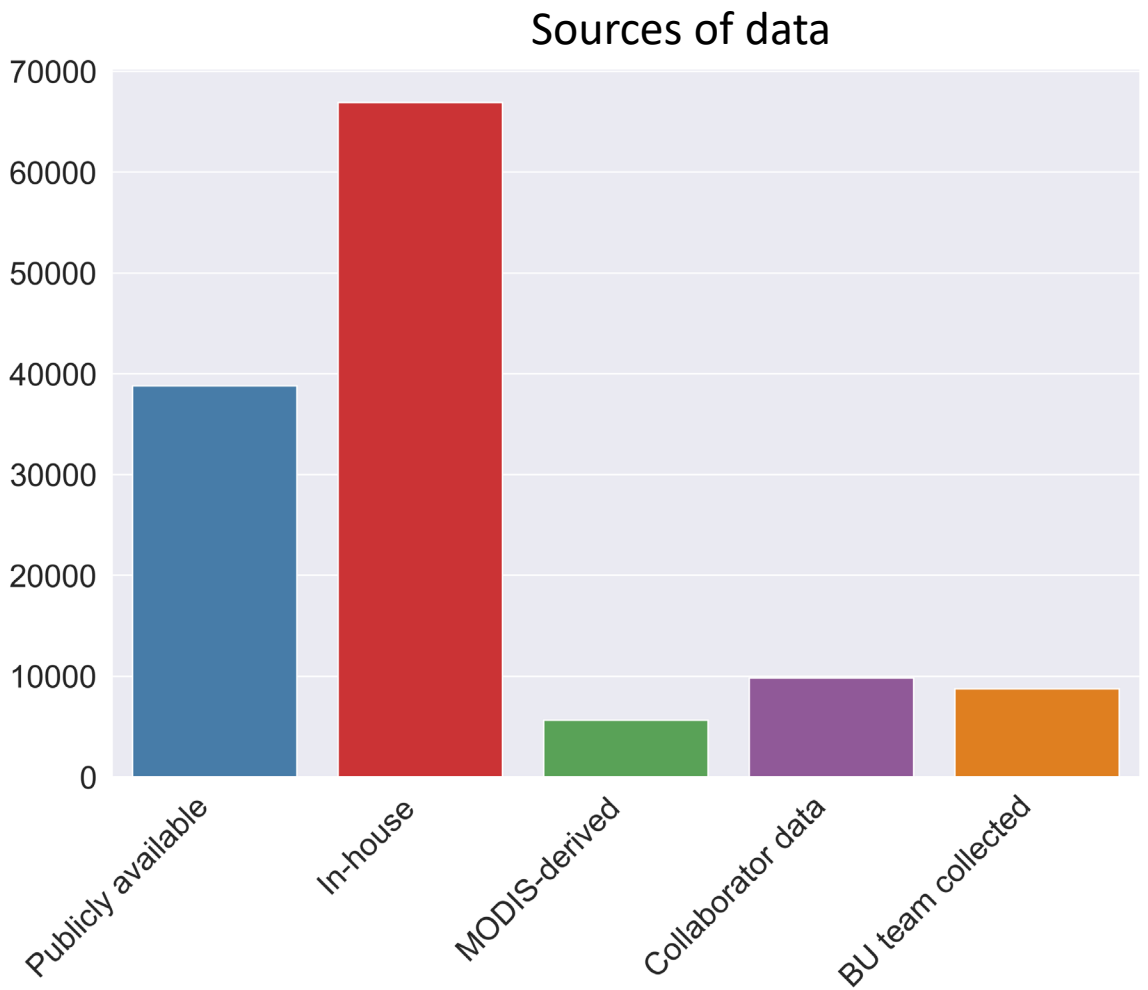
Out of 2,033,832 total points available to us.

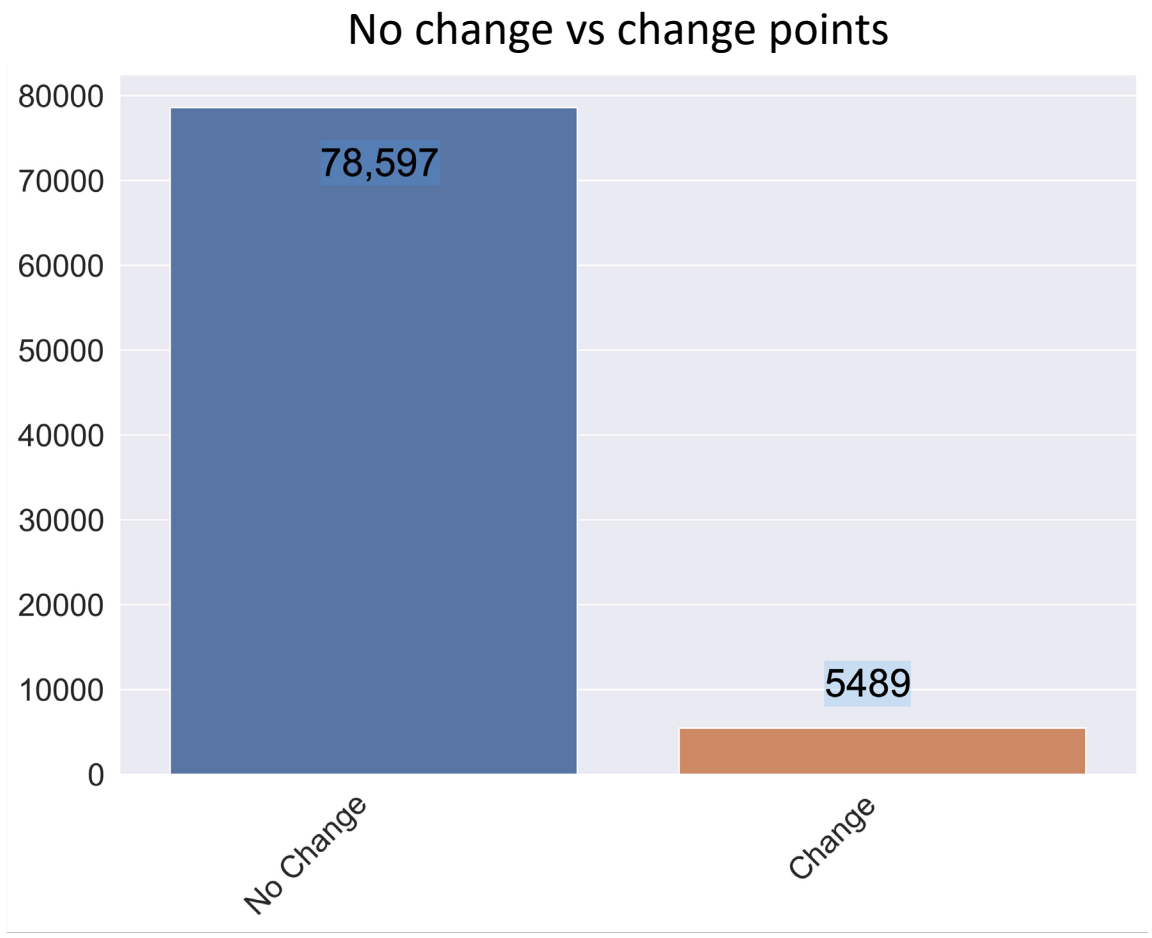
Land cover level 1 & 2 distribution





Dataset characteristics





Conclusions

- Global, geographically diverse land cover training dataset
- Will be released as open access via DAAC
  - Sampled down dataset: 129,948 points
  - Full dataset: 2,033,832 points
- Want to contribute your own dataset? Reach out!

Radost Stanimirova
Postdoctoral Researcher
Email: rkstan@bu.edu
Website: rkstan.github.io
Twitter: https://twitter.com/radost_stan

BOSTON UNIVERSITY

NASA

References

Class definitions

Level 1	Level 2	Description
Ice/snow	Ice/snow	Land areas where snow and ice cover is greater than 50% throughout the year.
Water	Water	Areas covered with water throughout the year: streams, canals, lakes, reservoirs, oceans.
Developed	Developed	Areas of intensive use; land covered with structures, including any land functionally related to developed/built-up activity.
Barren / sparsely vegetated	Barren / sparsely vegetated	Land comprised of natural occurrences of soils, sand, or rocks where less than 10% of the area is vegetated.
Forest	/	Land where tree cover is greater than 30%. Note that cleared trees (i.e., clear-cuts) are mapped according to *current* cover (e.g., barren/sparsely vegetated, shrubs, or grasses).
	Deciduous broadleaf forest	*definition
	Evergreen broadleaf forest	*definition
	Deciduous needleleaf forest	*definition
	Evergreen needleleaf forest	*definition
	Mixed forest	*definition
	Forested wetland*	*definition
Shrub	Shrub	Land dominated by shrubs. Total vegetation cover exceeds 10% , shrub cover is greater than 10% and tree cover is less than 30%.
	Shrub wetland*	*definition
Herbaceous	/	Land dominated by herbaceous plants. Total vegetation cover exceeds 10%, tree cover is less than 30%, and shrubs comprise less than 10% of the area.
	Grassland	*definition
	Agriculture	*definition

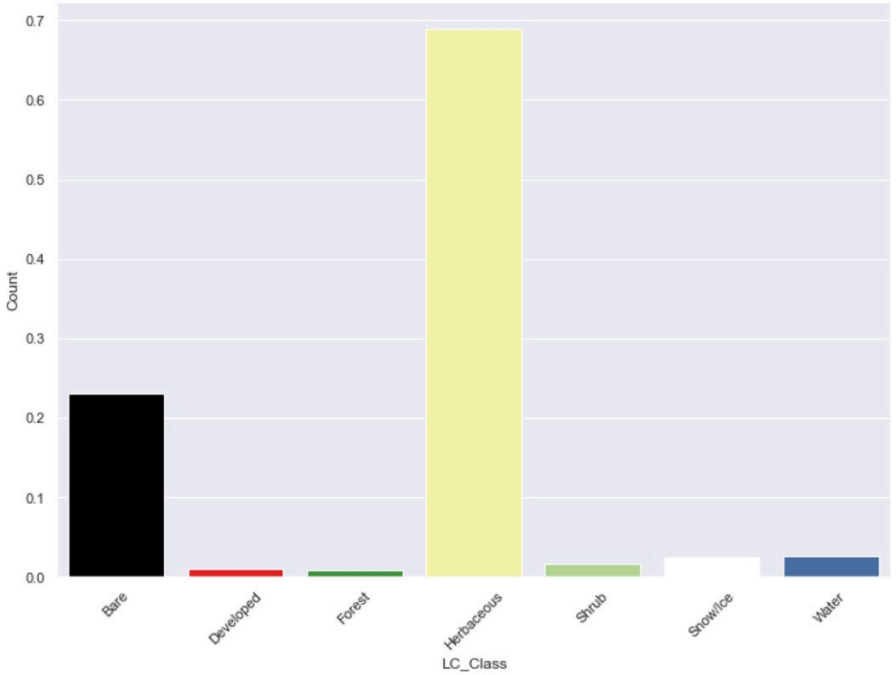
Data fields

Column	Notes
Latitude	
Longitude	
Start Year	1985-2019
End Year	1985-2019
Level 1 Land Cover	** See table
Level 2 Land Cover	** See table
Leaf Type	Broadleaf, Needleleaf, Mixed
Impervious Percent	Low (0%-30%), Medium (30%-60%), High (60%-100%)
Location	Interior, Exterior
Vegetation Density	Sparse (0%-30%), Open (30%-60%), Closed (60%-100%)
Vegetation Modifier	Cropland, Plantation, Wetland, Riparian/Flood, Mangrove, Trees/Shrub Present
Segment Type	Stable, Transitional
Change	No change, Change
Confidence Land Cover Label	1 (lowest) - 3 (highest)
Level1 Ecoregion	Based on World Wildlife Fund ecoregions
Level2 Ecoregion	Only for North America based on EPA ecoregions
Continent	North America, South America, Africa, Europe, Asia, Oceania
Continent Code	1, 2, 3, 4 ,5, 6
Dataset	STEP, Clustering, LCMAP, ABoVE, MapBiomas, Training_augment, MODIS_algo, GeoWiki, RadEarth, Collaborator data, BU Team Collected, GLC30, LUCAS
Dataset_Code	1, 2, 3, 4, 5, 999, 700, 701, 702, 703, 704, 705, 706
Glance_ID	unique ID for each sample
ID	ID for unique combination of latitude and longitude

Dataset	Spatial extent	Years	Number of samples	Original source
STEP	Global	2000-2019*		Sulla-Menasche et al. (2019)
Clustering	Global	1997-2020*		Turlej et. al (in prep)
GeoWiki	Global	2011, 2012	16,543	Fritz et al. (2017)
LandCoverNet	Global	2018	662	Alemohammad et al. (2020)
GLC30	Global	2015		Liu et al. (2019)
ABoVE	Canada and Alaska	1984-2014*	9,073	Wang et al. (2019)
LCMAP	Conterminous United States	1985-2017*	17,476	Stehman et al. (2021)
MapBiomas	South America (northern)	1985-2020	20,119	Souza et al. (2020)
Team collected	South America (southern)	2017, 2018	1,667	Graesser et al. (2022); Stanimirova et al. (2022)
	Colombia	2001-2016*	776	Arevalo et al. (2019)
	West Africa	2001-2020*	**	Tarrio et al. (in prep)
	Georgia (country)	2000-2020*		Chen et al. (2021)
	Laos	2000-2020*		Chen et al. (in prep)
Collaborator collected	Zambia	2008		Contact Eric
	Ethiopia	2001-2018		Contact Sylvia
	Ghana	2017	994	Contact Foster

Calculate Target Proportional Distribution

- Use 4 global land cover datasets
  - MODIS LC (500 m)
  - ESA World Cover (10 m)
  - Copernicus Global Land cover (100 m)
  - ESRI (10 m)
- Crosswalk each legend to GLANCE key



- Calculate proportional distribution of each dataset, in each grid, at native resolution
- Calculate median of all proportional distributions