Quantifying the impacts of bias in landcover data on global change analyses

Lyndon Estes [*], Peng Chen [†], Stephanie Debats [*], Tom Evans [†], Fanie Ferreira [‡], Gabrielle Ragazzo [*], Justin Sheffield [*] and Kelly Caylor [*]

[*]Princeton University, Princeton, NJ USA,[†]Indiana University, Bloomington, IN USA, and [‡]GeoTerraImage, Pretoria, RSA

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Blah blah.

landcover | bias | remote sensing | agriculture | crop yield | harvested area | carbon | agent-based model | landscape

Abbreviations: GTI, GeoTerraImage; SSA, sub-Saharan Africa

The nature and distribution of landcover is a fundamental determinant of many environmental and social processes that drive or are affected by global change (1), such as agricultural production and food security (2–4), carbon cycling (5, 6), biodiversity loss (7, 8), or demographic changes (9). Landcover maps are therefore critical for understanding the nature and impact of such changes (10), and they need to be accurate at the finest scales at which the underlying processes operate. For example, agricultural productivity and nutrient loadings can vary greatly between neighboring fields, and field sizes are often <2 hectares in regions where smallholder farming still dominates (11, 12). To understand agriculturally driven processes, it is thus necessary to accurately delineate fields at their smallest grain size, and to do so at regional to global scales to have a consistent set of maps.

Landcover data can only be developed with satellite imaging, but often the average size class of the cover type of interest is smaller than the sensor resolution, or spectrally indistinct from other neighboring covers, which propagates classification error (10, 13, 14). The result is that landcover datasets are generally inaccurate at finer scales and greatly differ between one another, particularly in those parts of the world undergoing the most rapid land use changes, where the aforementioned sources of bias tend to be most pronounced (15–17).

These errors are well-known (16, 17, 10, 18, 19), and there are a variety of efforts underway to improve landcover maps, particularly for agriculture (20, 14). What is less known is the degree to which these errors bias measurements built upon the distributional and areal information in landcover. An impediment to this understanding is that the errors are hard to quantify because spatially extensive reference data are not available for most regions of the world–particularly over Africa and other developing regions. Errors assessment therefore typically rely on a small number of ground truth points or survey data aggregated to political boundaries. For this reason, we have a better understanding of the biases between landcover datasets or in relation to country-level statistics (16, 17, 21) than we do of how error changes over spatial gradients or as a function of aggregation scale.

Being unable to fully quantify the errors in landcover maps of course makes it difficult, if not impossible, to quantify their impact on downstream analyses. There has been some work examining how such error influences climate simulations (22), agricultural land use patterns (23), and carbon flux (24) and human population estimates (9), but these either use simulated landcover errors (22) or compare relevant differences in estimates between different satellite-derived landcover maps (9, 24). The exception is (23), who use a high quality, ground-collected reference map detailing farm land use parcels in central Belgium, but the number of sites and region were both fairly restricted, and the parcels were not spatially contiguous.

There is thus an urgent need to more precisely quantify landcover map errors and how these vary over large regions, particularly for the regions where landcover is changing most rapidly yet is most poorly known. We address this need in this study, using a unique, high accuracy agricultural landcover map for South Africa to quantify the errors in several latest generation landcover maps that are broadly used in global change studies. We use this information to examine how i) landcover properties and related classification schemes influence error, ii) how these errors change with aggregation scale, with the specific goal of determining "safe" scales for making area-based calculations, and 3) how these errors propagate through several different forms of downstream analyses that broadly represent the global change research focus areas, including biogeochemical and land use change studies, food security assessments, land surface hydrology and climatology, and human geography.

Study area and landcover data

Our study focused on South Africa, which comprises nearly 6% of sub-Saharan Africa's (SSA) landmass, and has a large, diverse agricultural sector, ranging from large commercial operations to smallholder farms (25, 26). This diversity suggests that the country's agricultural landcover spans the range of types that are found throughout the rest of SSA.

The South African government commissioned a whole country cropland boundary map in order to stratifying the annual aerial crop type census used to calculate harvested area estimates (27). The map was made by trained workers who visually interpreted high resolution satellite imagery and manually digitized field boundaries following a standardized mapping protocol. The resulting vectorized field maps, which were made in 2007 and updated in 2011, provide a unique, high accuracy reference dataset of both crop field distribution and size classes. We converted the vector data into a rasterized estimated of cropland percentage at 1 km resolution (henceforth the "reference map"), which was 97% accurate in distinguishing cropped from non-cropped areas.

Reserved for Publication Footnotes

We compared our reference percent cropland estimates to those created from four satellite-derived landcover datasets. We obtained South Africa's 30 m resolution National Landcover map (SA-LC) for 2009 (28), the 500 m resolution MODIS Landcover for 2011 (29, 30), the 300 m resolution GlobCover 2009 (31), and the new 1 km Geo-wiki hybrid-fusion cropland map for Africa (18). We chose these particular datasets because they are nearly contemporaneous with our reference data, and represent the major types of landcover products used by researchers: SA-LC typifies the higher resolution, Landsat-derived maps that are developed individually for many countries (32), MODIS and GlobCover are widely used global-scale products (33, 34), while Geo-Wiki incorporates the first three datasets and is the current state of the art for agricultural landcover maps. We extracted the cropland classes from the first three datasets and converted these to 1 km resolution percent cropland estimates (hereafter simply the "landcover maps"), resulting in 4 maps to compare to our reference cropland map (the "reference map").

Quantifying Error

We used these maps to first quantify error in cropland area estimates. We calculated error as the difference between the reference and landcover maps at different scales of aggregation (1 to 100 km), in order to estimate bias and how it varies with scale. Next, we assessed how bias correlates with the amount of cropland cover in agricultural landscapes, to gain insight into how landscape patterns may affect error.

We undertook five further analyses to investigate how map error can impact assessments founded on landcover maps. These include first-order analyses, in which values for a variable of interest are mapped to particular cover type(s), and second-order analyses, in which a process model draws on the cover types' values to calculate an output value. We created four datasets to represent second order analyses. The first was a series of maps of vegetated carbon stocks created following the methodology of Ruesch and Gibbs' (35). The second was constrained cropland percentage maps, which, following Ramankutty et al (36) were adjusted so that their total cropland areas matched provincial-level reported cropland totals. Using these adjusted cropland percentage maps, we disaggregated district-reported maize harvested area and yields (following 37) . We then compared differences between total carbon stock estimates calculated from the reference map with those from the four cropland maps, and again examined how these differences changed as a function of aggregation scale. We made the same comparisons for total maize harvested area, average yield, and total production.

For the second-order analyses, we examined how cropland cover errors influence 25 km resolution monthly evapotranspiration estimates produced using the Variable Infiltration Capacity (38) land surface hydrology model. For this example, we used the cropland maps to adjust the seasonally varying, landcover-specific leaf area index (LAI) values that VIC uses to partition water vapor fluxes into their evaporative and transpirative components. In the second example, we examined how these errors can impact the parameterization of an agent-based food security model (39). Spatially-explicit, agent-based models are frequently employed in land change science, and require an initialization step to assign landscape resources to model agents (e.g. 40–42). In this case, we used the cropland maps to allocate farmland to agents representing individual households in political districts, with the model assigning each household its initial cropland holdings using a function that considers total district cropland and how much cropland is near the agent's location.

Results

Bias and its correlates. We created the 1 km reference and landcover maps, removing all croplands marked as communal or smallholder farmland in the reference vector maps (individual fields were not mapped with the same precision) or as plantation forestry (SI), and then aggregated each map to 5, 10, 25, 50, and 100 km resolutions. We then subtracted each landcover map from the reference map at each scale of aggregation (Fig. 1).

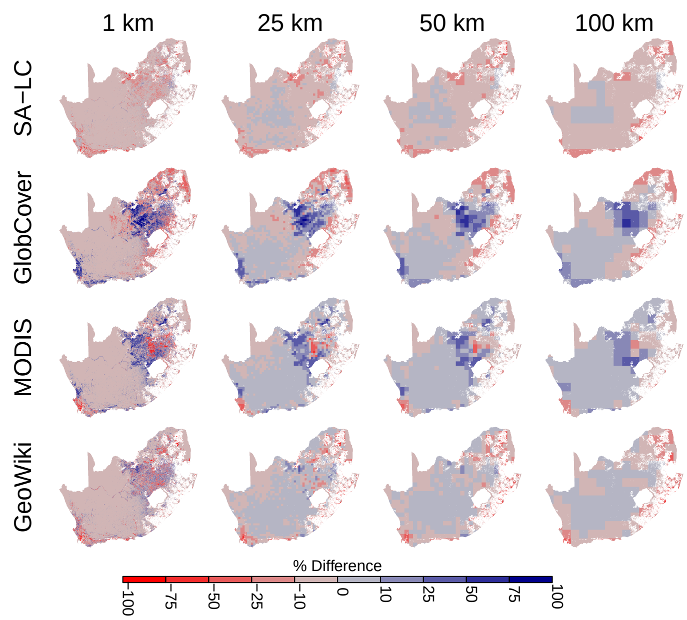

Fig. 1. Differences in percent cropland estimates between the reference map and each of the four landcover maps. Rows indicate the landcover map being assessed (by subtraction from the reference map), while columns refer to resolution of aggregation. White indicates areas with no data where communal farmlands or plantation forests were removed.

The spatial patterns of bias are most pronounced in the MODIS and GlobCover maps, with both substantially underestimating the amount of cropland in most regions of the country, particularly in the central part of the country where the bulk of the country's cropland is found (blue areas in Fig. 1). The mean of these differences, which is the landcover map's bias, are 21% and 34% at 1 km for MODIS and GlobCover, respectively (Fig. S1), meaning that each map underestimates cropland by that amount at this resolution. MODIS bias drops to 8% at 50 km of aggregation, whereas GlobCover bias is still 24% at 100 km (Fig. S1).

The SA-LC and GeoWiki datasets show much spatial variability and overall bias.

Error as a function of cropland density.

Potential bias in harvested areas, yield, and production estimates.

Potential bias in estimates of carbon stocks.

Potential bias in harvested areas, yield, and production estimates.

Impacts on evapotranspiration estimates.

Initialization errors in spatial agent-based models.

Discussion

Blather.

More blather. Materials and Methods

Methods. Perhaps it is right **SI Materials and Methods**.

Digital RCD Analysis.

Appendix: App 1

Appendix

This is an example of an appendix without a title.

ACKNOWLEDGMENTS. I thank everyone tearfully.

References

1. Lambin EF (1997) Modelling and monitoring land-cover change processes in tropical regions. *Progress in Physical Geography* 21(3):375–393.
2. Lark TJ, Salmon JM, Gibbs HK (2015) Cropland expansion outpaces agricultural and biofuel policies in the United States. *Environmental Research Letters* 10(4):044003.
3. Wright CK, Wimberly MC (2013) Recent land use change in the Western Corn Belt threatens grasslands and wetlands. *Proceedings of the National Academy of Sciences* 110(10):4134–4139.
4. Licker R et al. (2010) Mind the gap: how do climate and agricultural management explain the yield gap of croplands around the world? *Global Ecology and Biogeography* 19(6):769–782.
5. Asner GP et al. (2010) High-resolution forest carbon stocks and emissions in the Amazon. *Proceedings of the National Academy of Sciences* 107(38):16738–16742.
6. Gaveau DLA et al. (2014) Major atmospheric emissions from peat fires in Southeast Asia during non-drought years: evidence from the 2013 Sumatran fires. *Scientific Reports* 4.
7. Newbold T et al. (2015) Global effects of land use on local terrestrial biodiversity. *Nature* 520(7545):45–50.
8. Luoto M, Virkkala R, Heikkinen RK, Rainio K (2004) Predicting bird species richness using remote sensing in boreal agricultural-forest mosaics. *Ecological Applications* 14(6):1946–1962.
9. Linard C, Gilbert M, Tatem AJ (2010) Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal* 76(5):525–538.
10. See L et al. (2015) Improved global cropland data as an essential ingredient for food security. *Global Food Security* 4:37–45.
11. Jain M, Mondal P, DeFries RS, Small C, Galford GL (2013) Mapping cropping intensity of smallholder farms: A comparison of methods using multiple sensors. *Remote Sensing of Environment* 134:210–223.
12. Debats S, Luo D, Estes L, Fuchs T, Caylor K (year?) A generalized computer vision approach to mapping agricultural fields in Sub-Saharan Africa. *Remote Sensing of Environment*.
13. Lobell DB (2013) The use of satellite data for crop yield gap analysis. *Field Crops Research* 143:56–64.
14. Estes L et al. (2015) DIYlandcover: Crowdsourcing the creation of systematic, accurate landcover maps. *PeerJ PrePrints* 3:e1266.
15. Estes LD et al. (2013) Projected climate impacts to South African maize and wheat production in 2055: a comparison of empirical and mechanistic modeling approaches. *Global Change Biology* 19(12):3762–3774.
16. Fritz S, See L, Rembold F (2010) Comparison of global and regional land cover maps with statistical information for the agricultural domain in Africa. *International Journal of Remote Sensing* 31(9):2237–2256.
17. Fritz S et al. (2011) Cropland for sub-Saharan Africa: A synergistic approach using five land cover data sets. *Geophysical Research Letters* 38:L04404.
18. Fritz S et al. (2015) Mapping global cropland and field size. *Global Change Biology* 21(5):1980–1992.
19. Verburg PH, Neumann K, Nol L (2011) Challenges in using land use and land cover data for global change studies. *Global Change Biology* 17(2):974–989.
20. Fritz S et al. (2012) Geo-Wiki: An online platform for improving global land cover. *Environmental Modelling & Software* 31:110–123.
21. Kaptu Tchuent AT, Roujean JL, De Jong SM (2011) Comparison and relative quality assessment of the GLC2000, GLOBCOVER, MODIS and ECOCLIMAP land cover data sets at the African continental scale. *International Journal of Applied Earth Observation and Geoinformation* 13(2):207–219.
22. Ge J et al. (2007) Impacts of land use/cover classification accuracy on regional climate simulations. *Journal of Geophysical Research: Atmospheres* 112(D5):D05107.
23. Schmit C, Rounsevell MDA, La Jeunesse I (2006) The limitations of spatial land use data in environmental analysis. *Environmental Science & Policy* 9(2):174–188.
24. Quaife T et al. (2008) Impact of land cover uncertainties on estimates of biospheric carbon fluxes. *Global Biogeochemical Cycles* 22(4):GB4016.
25. Hardy M, Dziba L, Kilian W, Tolmay J (2011) Rainfed Farming Systems in South Africa in *Rainfed Farming Systems*, eds. Tow P, Cooper I, Partridge I, Birch C. (Springer Netherlands), pp. 395–432.
26. Estes LD et al. (2014) Using changes in agricultural utility to quantify future climate-induced risk to conservation. *Conservation Biology* 28(2):427–437.
27. Fourie A (2009) Better Crop Estimates in South Africa. *ArcUser Online* (1).
28. SANBI (2009) National Landcover 2009, (South African National Biodiversity Institute; National Department of Environmental Affairs and Tourism, Pretoria, South Africa), Technical report.
29. DAAC) LPDAACL (2011) MODIS MCD12q1 Land Cover Type Yearly L3 Global 500 m SIN Grid. Version 5.01, (NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota), Technical report.
30. Friedl MA et al. (2010) MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment* 114(1):168–182.
31. Arino O et al. (2012) *Global land cover map for 2009 (GlobCover 2009)*. (European Space Agency & Universit Catholique de Louvain).
32. Fry J, Coan M, Homer C, Meyer D, Wickham J (2009) Completion of the National Land Cover Database (NLCD) 1992-2001 Land Cover Change Retrofit Product, (U.S. Geological Survey), USGS Numbered Series 2008-1379.
33. Gross D et al. (2013) Monitoring land cover changes in African protected areas in the 21st century. *Ecological Informatics* 14:31–37.
34. Shackelford GE, Steward PR, German RN, Sait SM, Benton TG (2015) Conservation planning in agricultural land-

scapes: hotspots of conflict between agriculture and nature. *Diversity and Distributions* 21(3):357–367.
35. Ruesch A, Gibbs HK (2008) New IPCC Tier-1 global biomass carbon map for the year 2000. *Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory, Oak Ridge, Tennessee. Available online at: http://cdiac. ornl. gov/epubs/ndp/global_carbon/carbon_documentation. html.*
36. Ramankutty N, Evan AT, Monfreda C, Foley JA (2008) Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Global Biogeochemical Cycles* 22:19 PP.
37. Monfreda C, Ramankutty N, Foley JA (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:GB1022.
38. Liang X, Lettenmaier DP, Wood EF, Burges SJ (1994) A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research* 99(D7):14415.
39. Chen P, Plale B, Evans T (2013) *Dependency Provenance in Agent Based Modeling.* pp. 180–187.
40. Manson SM, Evans T (2007) Agent-based modeling of deforestation in southern Yucatn, Mexico, and reforestation in the Midwest United States. *Proceedings of the National Academy of Sciences* 104(52):20678–20683.
41. Evans TP, Kelley H (2004) Multi-scale analysis of a household level agent-based model of landcover change. *Journal of Environmental Management* 72(1-2):57–72.
42. Kelley H, Evans T (2011) The relative influences of landowner and landscape heterogeneity in an agent-based model of land-use. *Ecological Economics* 70(6):1075–1087.