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 Land-cover 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 crop-land 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 first 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.

Bias

Landcover bias. 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), as horticulture, or plantation forestry (SI), and calculated the total cropland extent in the remaining area (1,081,000 km^2, or 90% of South Africa). The 2011 reference map showed a cropland area of 104,310 km^2, which the SA-LC and GeoWiki maps overestimated by 26 and 5.8%, respectively, and GlobCover and MODIS underestimated by 21 and 26.1%.

We then aggregated each map to 5, 10, 25, 50, and 100 km resolutions, and subtracted the four landcover maps from the reference map at each scale of aggregation to asses the spatial patterns of bias (Fig. 1).

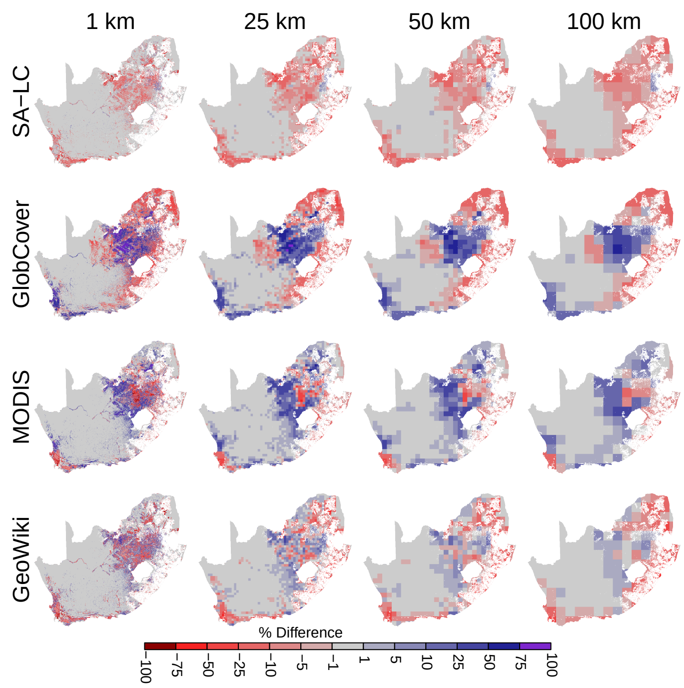

Fig. 1. Differences in percent cropland estimates between the reference map and each of the four landcover maps. Rows indicate the landcover map being assessed (by subtraction from the reference map), while columns refer to resolution of aggregation. White indicates areas with no data where communal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

The most pronounced extreme biases are in the MODIS and GlobCover maps, with both underestimating cropland extent by 10-75% in the center of the country (blue areas in Fig. 1, and the dominant production region), and overestimating along the eastern to northern margins (red areas in Fig. 1). The average biases for GlobCover and MODIS are 21% and 34% at 1 km, meaning that each map underestimates cropland by that amount at this resolution. MODIS mean bias drops to 8% at 50 km of aggregation, whereas GlobCover bias is still 24% at 100 km (SI Appendix, Fig. S1).

The SA-LC map uniformly overestimates cropland throughout the country (Fig. 1), but the magnitude of mean bias is much lower at -8% at 1 km to -6% at 100 km (Fig. S1). The GeoWiki map has a very heterogeneous pattern of bias, having a slight underestimation tendency (5% mean bias) at 1 km, which changes to a smaller overestimation bias (-2 %) at 100 km (Fig. S1).

The role of landscape characteristics in shaping bias. To investigate the degree to which landscape features influence landcover map bias, we extract all pixels in agricultural areas (>0.5% cropland) of the 1 km reference map using the the boundaries of 354 magisterial districts (South Africa's finest administrative unit, which average 3,445 km^2 in size; SI Appendix, Fig. S3), and averaged the values of these pixels within each district. These provided an estimate of the intensity of cropping within the farmed parts of the landscape (with districts defining the landscape scale), as well as the degree of mixing between cropland and other land covers. We then extracted the cropland map biases for the same locations, and calculated the mean absolute bias within each district. Absolute bias describes the magnitude of bias but not its direction, and is more informative about the likelihood of the map being biased at any given point in the landscape than actual bias, as positive and negative biases can cancel each other out within small areas.

A generalized additive model fit to district-level absolute bias (log-transformed) shows that absolute bias peaks at 50% cropland cover for all but the GlobCover map (which continued to increase with cropland cover), and is lowest when the landscape is dominated either by cropland or other cover types (Fig. 2). In other words, bias is highest when cropland cover is mixed evenly with other cover types.

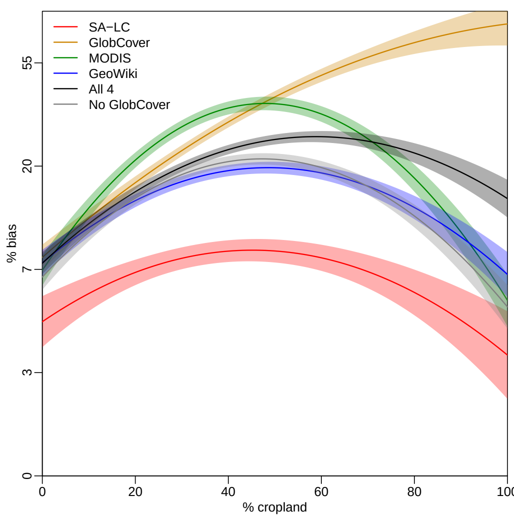

Fig. 2. The relationship between the mean absolute bias in cropland maps and cropland cover within agricultural areas (reference map pixels having >0.5% cropland), averaged within the boundaries of magisterial districts (n = 345), as fit with a generalized additive model. Prediction curves are color-coded to the different cropland maps, with the solid line indicating predicted absolute bias, and the lighter shading the standard error of the coefficients. Models were also fit to the mean absolute biases across all four maps (black curve), and from all maps exlcuding GlobCover (grey surve).

The impact of bias on calculating carbon stocks. We used the data of (35) to calculate carbon densities for African forests, shrublands, croplands, grasslands, and sparse habitats. We assigned the cropland carbon values to map cells in proportion to their cropland cover. For the non-cropland proportions, we assigned the carbon value from each of the other types, such that we created four different carbon maps for each landcover map at each aggregation scale (Fig. S4), which allowed us to test how carbon estimates vary as a function of i) bias in the cropland estimates and ii) the characteristics of adjacent cover types.

At the 1 km scale, the differences between country-wide total carbon stocks made using any of the cropland maps and the reference were within +/-3%, regardless of which cover type was adjacent to cropland (Table S1), because differences between total stock estimates are diluted by the large area (∼50-70%) of the country having no cropland. Comparing the carbon stocks from each map's agricultural area reveals much greater differences (Table S1). SA-LC overestimates carbon stocks by just 2% when the adjacent cover type is forest, and up to 15% when it is spare cover. MODIS ranges from negligible overestimates in denser carbon classes (forest, secondary forest, and shrublands) to 8-13% underestimates for the grassland and sparse classes. GeoWiki underestimates for all types, from <1% for sparse cover to 8% for forest. GlobCover shows the largest total carbon stock bias, with over-estimates ranging from for sparse lands to 162% for forest. The magnitude of this bias is due to false positives–GlobCover identified cropland in nearly 50% of pixels, compared to 30% for the other three cropland maps.

The spatial patterns of carbon bias (Fig. S4) are a reflection of the cropland biases (Fig. 1). Where cropland is underestimated and the surrounding cover type is of higher carbon density than cropland, carbon density is overestimated. For lower density (grassland and sparse vegetation), there may be slight underestimates. These tendencies are reflected in the average of each map's spatial biases over agricultural areas (Fig. 3). For example, MODIS and GlobCover overestimate carbon by ∼50% on average at 1 km resolution when forest is the cropland-adjacent cover (stars in semi-transparent green and gold bars, Fig. 3; Table S2). When it is sparse vegetation (open circles in Fig. 3), MODIS underestimates by nearly 3% at all scales, whereas maps that overestimate cropland (e.g. SA-LC, semi-transparent red) overestimate carbon density for this cover. Overall, GeoWiki has the lowest average spatial biases, for all cover types and all resolutions. It's maximum bias is a 12% overestimate at 1 km for forest cover, thereafter all biases are less than 10%, and converge to +/-2% above 1 km aggregation (Fig. 3, Table S2). All maps' biases are within +/-10% bias after aggregation to 25 km.

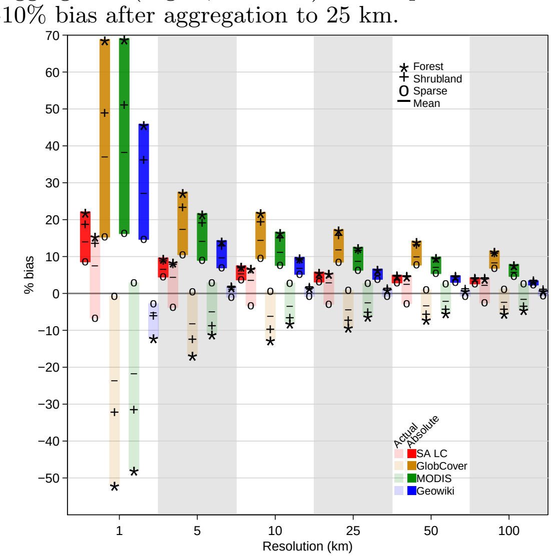

Fig. 3. The average spatial biases in carbon density estimates based on cropland maps, expressed as a percent difference relative to the reference map. The values of actual bias (incorporating bias direction and magnitude) and absolute bias (magnitude only) are presented in semi-transparent and solid colors, respectively, with colors denoting the specific cropland map and symbols indicating which cover type was used to calculate cropland adjacent carbon density. The bar represents the mean biases calculated across each of the 5 cover types. Shrubland and grassland bias values are near zero and not shown for display clarity.

The patterns of mean absolute spatial bias (solid colored bars in Fig. 3) generally follow that of actual bias, but with somewhat higher magnitudes and a few important differences. The most notable is that Geo-Wiki has fairly large absolute bias at 1 km, averaging 27% across across cover types (line in solid blue bar, Fig. 3, Table S2), which is close to the 36-37% for GlobCover and MODIS. SA-LC has the lowest absolute bias across scales, averaging (across cover types) from 14% at 1 km to 3% at 100 km. (Fig. 3, Table S2). The increase in Geo-Wiki's absolute spatial bias relative to SA-LC's can be attributed to the spatially heterogeneous nature of its cropland biases (Fig. 1) compared to the other three cropland maps.

Bias in harvested areas, yield, and production estimates. We used the reference dataset to calculate total cropland area within South Africa's nine provinces, and then adjusted the the four cropland maps so that they each summed to the same area within each province (per 36). Despite this statistical constraint, the updated cropland maps still had substantial spatial biases similar in pattern (Fig. S5) to those in the unadjusted maps (Fig. 1), and we evaluated how these biases affected crop production estimates. We derived maize yield and harvested area estimates by disaggregating district-level (n = 354, mean area = 3,445 km^2) agricultural census data (43) following Monfreda *et al.* (37), and used that to calculate spatial biases in yield and production estimates at each aggregation scale.

The yields disaggregated onto the cropland maps were markedly different to those on the reference map, particularly in the lower density cropland areas in the center of the country, where GlobCover overestimated yields and MODIS and GeoWiki (to a lesser extent) underestimated them at 1 km resolution (Fig. S6). However, the mean spatial bias in yields at this resolution was pronounced only for GlobCover, equivalent to nearly 60% of the mean reference yield of 3.4 tons ha^{-1}, compared to +/-5% for the other maps (Fig. 5). Interestingly, GeoWiki and MODIS biases increased with aggregation, peaking at 10 km where both datasets underestimated yields by 30%, thereafter declining to 10% at 100 km, whereas GlobCover's mean yield bias declined linearly with aggregation (Fig. 5).

In contrast to yields, production errors were unbiased in aggregate (Fig. 5). The statistical constraints on harvested and cropland areas led to the canceling out of spatial errors in production estimates, which is evident in the checkerboard-like pattern in maps of production biases (Figure S7). However, this pattern of bias means that the average absolute biases in production estimates are large, between 40 to 55% for GeoWiki, MODIS, and GlobCover at 1 km, and remaining generally high (<10%) for these three datasets even up to 50-100 km of aggregation (Fig. 5). SA-LC production biases were lowest across all spatial scales (20% at 1 km to 2% in 100 km).

Absolute mean biases in yield were also substantial, and generally 10-15% larger than production biases across all spatial scales, except for GlobCover where absolute production biases exceeded yield bias at 5-100 km of aggregation.

Bias in evapotranspiration estimates. Compared to carbon and crop related examples, actual and absolute biases in evapotranspiration calculated using the VIC model were small and averaged to less than +/-1%. There are several hotspots of discrepancy, however, which are evident in maps of spatial biases (Fig. 6). The most pronounced of these are the 5-15% overestimates by VIC initialized with MODIS and GlobCover in the center of the country, and along the southern and western coasts, particularly the latter where the ET overestimates reach 25%. These correspond primarily to margins of major crop production regions–in the center is the westernmost boundary of the summer rainfall growing region, where maize is the primary crop, while the bias hotspot up the west coast marks one of the two boundaries of the winter rainfall region where wheat farming dominates(25). These two boundaries correspond respectively to approximately 400 and 200 mm of growing season rainfall.

SA-LC and GeoWiki also produced biased ET estimates along the southern and western coasts, but here the tendency was to underestimate ET, while biases in the center of the

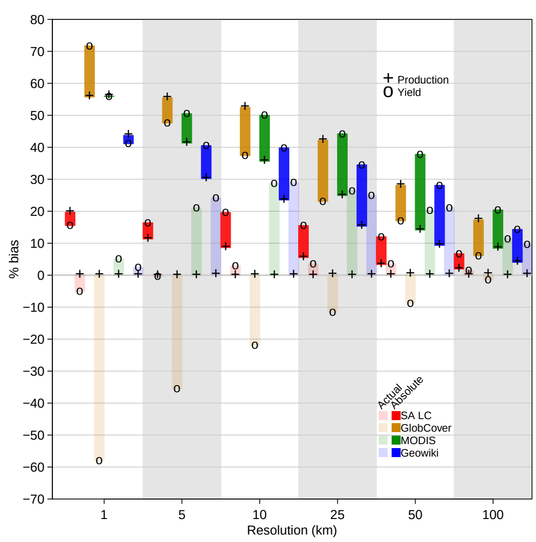

Fig. 4. The mean spatial biases in disaggregated maize yield estimates, and the bias in total crop production estimates calculated from these yields and constrained crop area estimates. The values of actual bias (incorporating bias direction and magnitude) and absolute bias (magnitude only) are presented in semi-transparent and solid colors, respectively, with colors denoting the specific cropland map and symbols indicating whether the bias is related to one of two yield aggregation methods and associated crop production estimates.

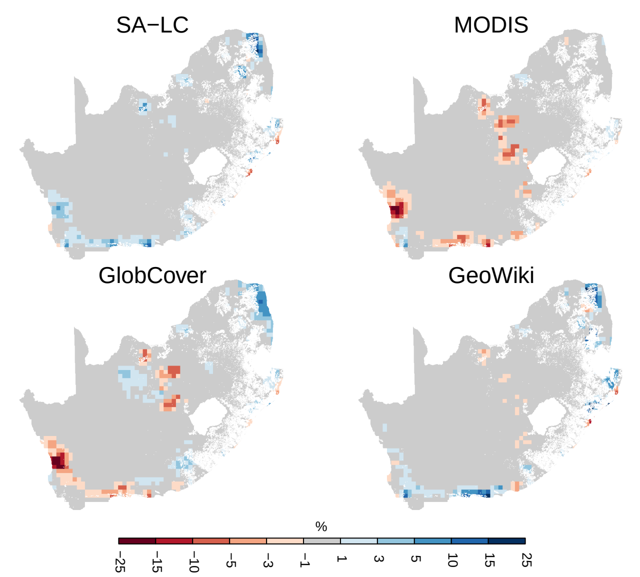

Fig. 5. Differences in annual mean evapotranspiration estimates from 29-year runs of the VIC land surface hydrology model when initialized with LAI response curves derived from the reference map, versus those from the four cropland maps.

country were either negligible to absent. All but MODIS underestimated ET by 5-15% in the northern tip of the country.

Initialization biases in agent-based models. We used an agent-based model of food security that represents the interactions between hundreds of individual farming households over multiple seasons (39). The model is thus computationally intensive, and, like many spatial ABMs, run over smaller geographic domains (e.g. districts, rather than an entire country) and at higher spatial resolutions (10s to 100s of meters) that are needed to represent the different land units of single farmers. To match these computational characteristics, we selected four contiguous magisterial districts (ranging from 1,040-1,343 km^2, Fig. S8) in the eastern part of the country, having between 28-45% of their areas devoted to cropland, according to the reference map.

We disaggregated the cropland percentages in all maps to binary cropland/non-cropland cover types with 100 m resolution, which matches the typical field size (1 ha) for smallholder farmers in household survey data (collected in Zambia) used in developing the agent-based model (39). These surveys found mean household crop field area to be 2 ha, which we divided into reference cropland areas to estimate the total households within each district. We then initialized the model by assigning each household agent two cropland pixels. In order to emulate the natural groupings of communities, the model only assigns a household fields that are within 1.5 km of other agents' fields, and if those pixels have not already been allocated to another agent. The model thus iteratively grows "communities" until all households are assigned cropland, or all available cropland is allocated.

We used each of the five maps to initialize the model, and assessed how bias in cropland cover estimates impacted the initialization process, in terms of how many agents were not assigned fields, and how many fields were left unallocated. For the first metric, there was a one-to-one relationship between the percent by which cropland area underestimated and the percent of households left without and farmland (Fig. 6, left panel), the most extreme examples occurring when MODIS cropland was used to initialize the model in districts 1 and 2, where ∼85% of agents could not be allocated farmland because MODIS underestimated cropland by that amount ∼85%. No agents were left without fields where maps overestimated cropland extent (GeoWiki, SA-LC), but in these cases the percent of cropland left unallocated matched the size of the overestimate (e.g. ∼20% for SA-LC, Fig. 6 right panel). Interestingly, the overall relationship between land allocation and cropland bias was U-shaped, as the model also failed to give land to agents when cropland was underestimated by 50% or more (Fig. 6, right panel), as exemplified by the MODIS result for districts 1 and 2, where 7-12% of cropland was left unallocated despite the fact that only 85% of agents had no land. This curious relationship occurred because cropland tends to cluster, and when it is underestimated, the size of these clusters is small, resulting in islands of cropland that fall outside of the search radius (which is constrained by an absolute distance and the proximity of other agents) within which cropland is sought when agents are seeded onto the landscape.

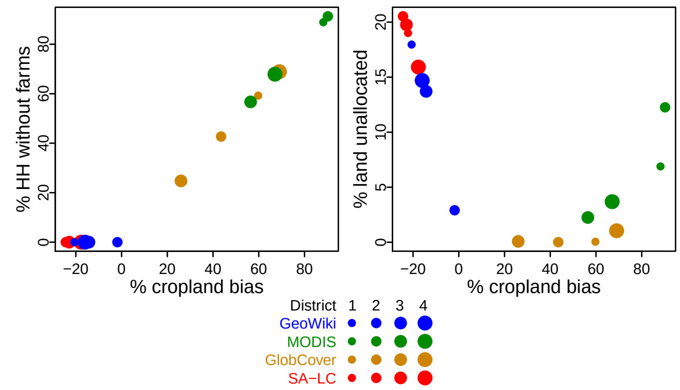

Fig. 6. Biases in agent-based model initialization relative to cropland map biases, measured in terms of the percent of households having no cropland allocated (left) and the percent of cropland left unallocated (right). Dot sizes correspond to district numbers, colors represent the landcover map.

Discussion

Some points to make thrown out now because out of time. Others will be added. Please add any important implications/considerations you see from the results.

Some points from my notes last year:

- Bias as a function of scale
  1. At 1 km resolution all landcover products are still fairly biased.
  2. Bias drops to acceptable levels quickly for geowiki–at 5X5 km, mean bias is just 1% (overestimated). The absolute bias for this dataset is 10% or lower from 10X10 km resolution and coarser.
  3. The SA dataset's bias is fairly consistent but low across all levels of aggregation, amounting to no more than an 8% overestimate of cropland with absolute bias of similar magnitude.
  4. MODIS and GlobCover biases (mostly of underestimation) do not dissipate until the higher levels of aggregation. MODIS's actual bias (under-estimation) falls below 10% at 20 km resolution, but the absolute bias remains above 10% until more than a 100-fold aggregation is done (>100 km resolution). For GlobCover, it is still too high.
- Bias as a function of cropland cover
  1. Classification algorithms are thus more error-prone where landcover is mixed/heterogenous.
  2. The exception to this lies in the GlobCover dataset, where bias primarily increases as a function of cropland cover. The reason for this is that GlobCover's cropland classes do not provide for 100% cropland cover, so aggregation tends to exacerbate underestimates.
  3. Thus caution is needed when aggregating a mixed pixel class.
     - An example illustrats this: take 4 1 km pixels, 2 of which are 100% cropland, 2 of which are other cover types. Imagine a landcover product classifies 3 of these as cropland (2 correct, 1 an error of commission), using a cropland class that is defined as 50%

cropland. Aggregating the actual fraction by a factor of 4 will result in a new 4 km pixel having 50% cropland, whereas aggregating the landcover product's pixels will give just 38% cropland, even when factoring in the incorrect classification.)

- Bias as a function of method
  1. Higher resolution and ensemble-based approaches have less bias.
  2. geowiki represents a fusion of multiple coarse resolution data sources that has undergone extensive validation using a crowdsourcing approach
  3. the SALC dataset is based on 30 m landsat data, but incorporates a range of ancillary data and expert judgement
  4. MODIS and GlobCover data are effectively single source/single algorithm.
  5. **Newer points begin here**
  6. Statistically constrained constrained landcover estimation approaches provide accurate area-based inferences when aggregated. But spatial errors are still high, as seen with GeoWiki and production/yield estimates. Using these to identify yield gaps at specific map locations is inappropriate, or even for a larger location if it does not coincide with the geographic boundaries of the statistical unit.
  7. Constrained estimates are also dependent on the accuracy of the statistics.
- Fix above to have section on bias for global change studies
  1. Scales at which it is safe to estimate values of say carbon stocks.
  2. Above point about bias in disaggregated yield estimates - no point mapping these out. A new approach might be to take these statistically reported yields and then combine them with satellite data to estimate yield variability within the district. That way would have meaningful reason for disaggregating yields, and would be pegged to real yield values, which would help minimize errors in remote sensing of yields.
  3. Something on ET - doesn't seem to matter much, but land-atmosphere interactions can make these discrepancies meaningful, particularly since biases occur in arid areas where a lot of irrigation happens–can cause significant impacts on regional climate. etc. etc. Also we didn't change out land cover types, and the vegetation in SA around the cropland will have reasonably similar LAI and ET responses (I think), thus impact more muted than it might be elsewhere (e.g. in forested landscapes).
  4. Agent-based models. Tom, Peng, something of significance/implications of this, please
- Will need a section on way forward for data, etc. Key role of accurate landcover, particularly agricultural. New methods, vectorized field boundaries seem to be highly valuable, Mapping Africa, Stephanie's paper, Geo-Wiki, etc are the way ahead.

Mixed landscapes increase the chances of omission and commission errors by increasing the number of cover classes, or because such landscapes are less spectrally distinct (14)

More blather. Materials and Methods

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

Describe weighted mean bias reasons.

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 landscapes: 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.
43. Africa SS (2007) Commercial Census of Agriculture, South Africa.