Supplementary Methods and Results

1 Overview

This document provides supplementary information on methods and results from the different components of the analysis described in the main text. The code used to conduct the analysis is presented after sections 1-X.

2 Reference data accuracy

The accuracy of the reference vector dataset was assessed by Estes *et al.* (2016), and described further here. The assessment was undertaken using a sub-sample of 609 1 km^2 grid cells, which were selected using a weighted randomized sampling scheme. Weights were derived from a logistic regression model of cropland occurrence probability in South Africa, with 1 corresponding to the lowest quartile of probability, and 4 the highest. The accuracy assessment included all cropland classes mapped within the reference dataset, which ranged from communal/smallholder fields to commercial row crops to orchards and other types of horticulture. A visual accuracy assessment was conducted within each 1 km^2 grid cell, wherein each cell was divided into 25 smaller cells of 200 X 200 m (4 ha), and then the proportion of each cell overlapping with cropfields visible in underlying, high resolution Google Maps imagery was then calculated to the nearest 5% of coverage, using a finer 20-cell mesh overlaid on each sub-cell within which field presence/absence was recorded. The same procedure was then performed to assess coverage by reference map polygons, as well as the intersections and differences between cells determined to be occupied by i) visual assessment and ii) by the digitized polygons. These intersections and differences were used to calculate the area of true positives, false positives, true negatives, and false negatives within each 1 km^2 grid cell, which were then used to calculate accuracy measures. In this case, accuracy was assessed using two landcover classes: cropland and non-cropland, thus a two class confusion matrix was constructed (Table 1) according to remote sensing classification accuracy guidelines set out by Olofsson *et al.* (2014). In this case, the reference data were the visually interpreted crop field presences/absences, and the map data were the field boundary vectors being assessed (i.e. the data which provide the reference maps in this study), the class "crop" refers to crop field presence and "non-crop" to crop field absence. The total area of vectorized crop fields was used to calculate the proportion W of South Africa's total area mapped as crop fields, and of the areas "mapped" (by omission) as having other cover types. These were then used to weight the proportions of each mapped class i corresponding to each reference class j by the total area A mapped for class i. The total accuracy (0.97, or 97%) was calculated by summing the diagonal (bold in Table 1), as well as the producer's accuracy P for each reference class j, and the user's accuracy U for each mapped class i.

Table 1: Confusion matrix for the assessment of the hand-digitized crop field boundaries used to generate reference cropland cover percentage values. Here *Reference* denotes a visual assessment of crop field presence in high resolution imagery within a sample of 609 sites, while *Map* refers to the vectorized crop field boundaries. The assessment had two classes: crop fields and non-crop fields, and proportions corresponding to these two categories are the proportions of areas determined by class agreements and disagreement in the 609 sites (see text above), which are weighted by the proportion W of each class' total mapped area A_i in South Africa. P and U provide the Producer's and User's accuracies, respectively.

		Reference (j)				
		Crop	Non-crop	W_i	A_i (ha)	U_i
Map (i)	Crop	**0.108**	0.007	0.114	14018567	0.943
	Non-crop	0.021	**0.865**	0.886	108541733	0.977
	Total	0.128	0.872	1.000	122560300	
	P_j	0.840	0.992			

3 Cropland map error analysis

We examined the impact of several sources of uncertainty on our results. The first is the potential temporal mismatch between the reference dataset and the landcover products we were testing. We tested this by examining two versions of the reference dataset, the initial version created in 2007, and the updated version from 2011 used in the main analysis. The 2011 version shows 3% more cropland area than the 2007 version. To examine the effect of the over mapping bias and accuracy measures, we converted the 2007 vector maps to gridded cropland percentage and subtracted each test map, and then compared the differences in bias and mean absolute error (MAE) values calculated between both (Table 2). The largest difference was 1.63%, between the 1 km SA-LC residuals, which means that the overestimation bias by SA-LC was actually greater relative to the 2007 version of the reference map. Except for the corresponding MAE value, all other difference were <1%.

Table 2: Differences in the bias and mean absolute errors values resulting from cropland percentage residuals calculated using 1) the 2011 reference map and 2) the 2007 reference map, with 2007 values subtracted from 2011 values. Differences from each test map and each scale of aggregation are shown.

	Resolution	SA-LC	GlobCover	MODIS	GeoWiki
1	1 km	1.63	-0.80	0.12	0.64
2	5 km	0.99	-0.38	0.40	0.80
3	10 km	0.89	-0.27	0.44	0.77
4	25 km	0.83	-0.12	0.47	0.76
5	50 km	0.79	0.04	0.47	0.73
6	100 km	0.76	0.19	0.54	0.72
7	1 km	1.30	-0.77	-0.10	0.27
8	5 km	-0.22	-0.44	-0.08	0.12
9	10 km	-0.48	-0.34	-0.08	0.12
10	25 km	-0.61	-0.20	-0.08	0.20
11	50 km	-0.67	-0.09	-0.36	0.22
12	100 km	-0.70	0.04	-0.24	0.19

As a further exploration of uncertainty, we included the full range of variability that resulted from a) using either the 2007 or 2011 map to calculate bias and accuracy, and b) from different levels of cropland percentages used when converting MODIS and GlobCover mixed cropland classes into

cropland percentage maps, following Fritz *et al.* (2015). We pooled the residuals from across each of these permutations, for each reference-test map combination, and examined how their overall mean values change with aggregation scale (Fig. 1), and also assessed the total distribution of errors within different class of cropland density, determined by dividing the 2011 1 km reference map into 20 different zones, or bins, of cropland density, ranging from 0-5% cropland cover up to 95-100% cover. We then calculated summary statistics from the pooled residuals and absolute values of residuals within each of these bins (Fig. 2).

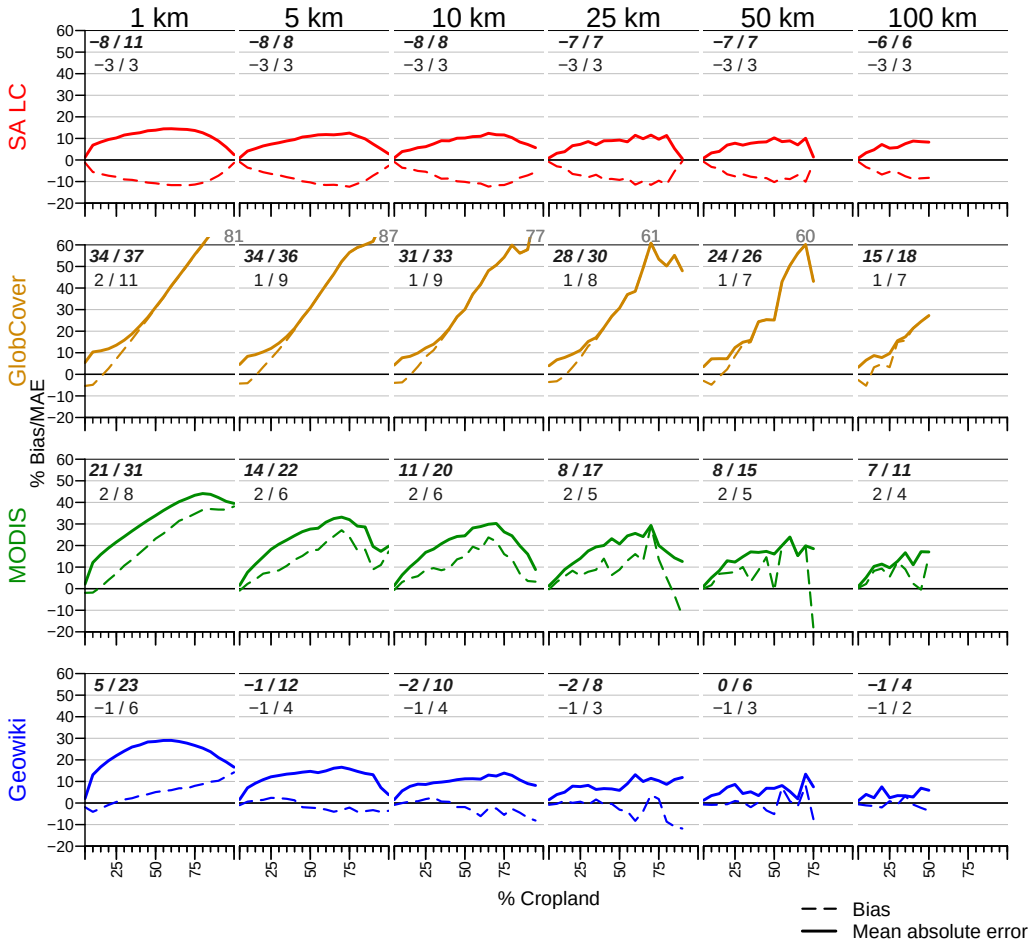

Figure 1: Biases and mean absolute errors (MAE) for each of the cropland maps as a function of cropland density (calculated using the 2011 reference maps) and aggregation scales. Rows present biases by map product, columns by aggregation scale. Dash lines indicate bias at each level of cropland density, calculated in bins spanning 5% of density (e.g. 0-5% cropland cover, 5-10%, etc.), while solid lines indicate the mean absolute error. The black numbers in each plot area present the overall means of bias/MAE for each sensor-scale combination. The bin-wise and overall mean statistics were calculated from pooled map errors calculated from differences between the 2007 reference map and each cropland map (including all three variants–high, medium, and low–of the MODIS and GlobCover-derived cropland maps), and the 2011 reference map and each cropland map.

We also calculated how different methods of calculating the bias and accuracy statistics impacted out findings, using three different methods. First, we simply took the straight averages across the entire country, which substantially understates bias and accuracy because cropland covers only 10% of the country, and all landcover products successfully discriminate the non-cropped regions. We also compared the average error metrics extracted from within just the agricultural regions of the maps, defined here as the union of areas having >0.5% cropland in the reference and each test map. This also tends to understate bias, because the area having <5% cropland is much larger than areas having

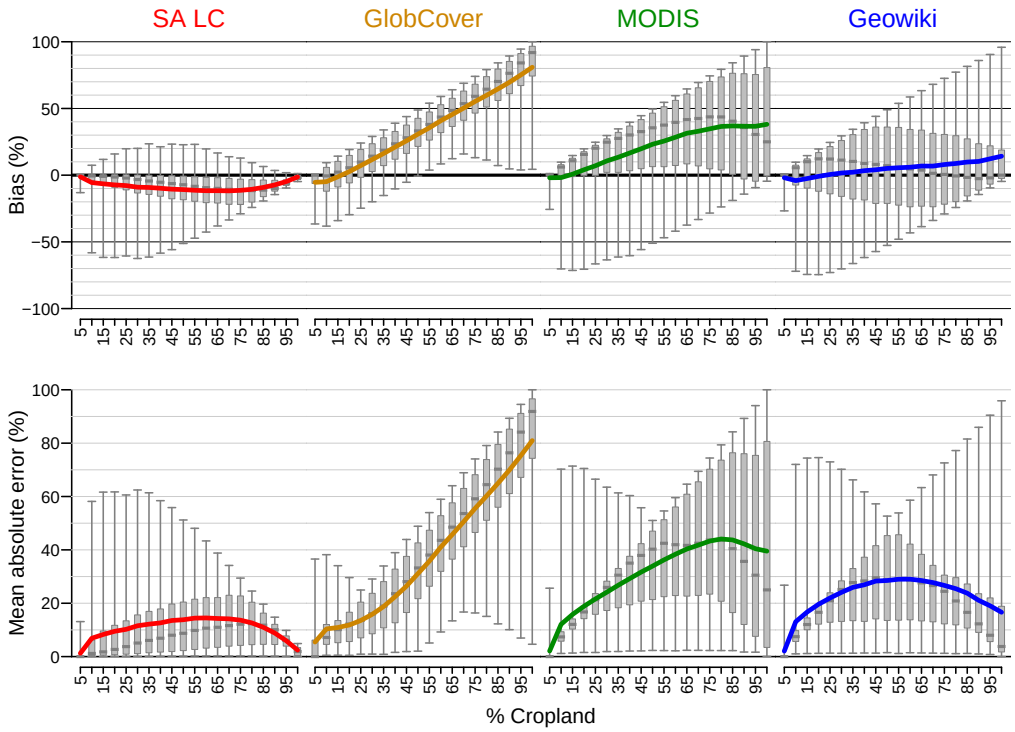

Figure 2: Biases and mean absolute errors (MAE) for each of the cropland maps at 1 km resolution, as a function of cropland density. Colored lines (color-coded to map product name) show the bias/MAE at each level of cropland density, calculated in bins spanning 5% (e.g. 0-5% cropland cover, 5-10%, etc.). Box plots show the variability of bias in each bin (whiskers = 2.5 and 97.5 percentiles, box the inter-quartile, and grey bar in box the median). Biases are presented in the top row, and MAEs in the bottom row. Statistics were calculated from pooled map errors calculated from differences between the 2007 reference map and each cropland map (including all three variants–high, medium, and low–of the MODIS and GlobCover-derived cropland maps), and the 2011 reference map and each cropland map.

higher densities of cropland (see Fig. 4 frequencies of cells per five percentile bins of cropland density). Finally, we calculated a density-independent metrics, which is similar to the density weighted mean calculated for our main analysis, which was calculated by averaging the mean bias and MAE values within different levels of cropland cover (divided in increments of 5%). The results of these three alternate metrics are shown in Figure 3.

We used magisterial district boundaries for South Africa (Fig. 5) to extract absolute residual values for agricultural areas (>0.5% cropland cover) and 2011 reference cropland density values, the means of which respectively served as the response and predictor in the generalized additive model (Wood, 2001) analysis (main text).

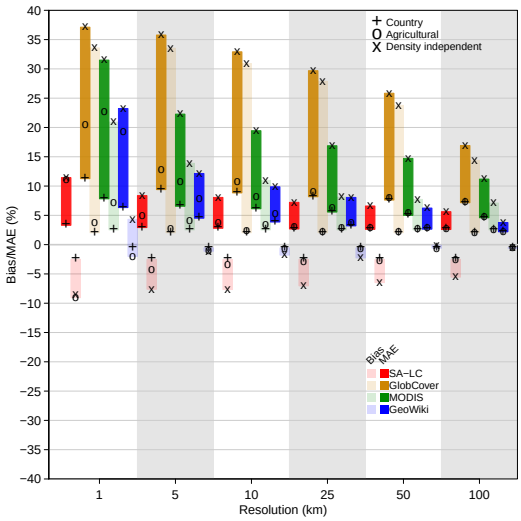

Figure 3: A comparison of three alternate methods for calculating bias and accuracy (mean absolute error): as a straight average across the entire country, averaged over agricultural areas only (the union of areas defined as having >0.5% cropland cover in the reference map and each test map), and independent of cropland density, wherein the mean bias/MAE values for each of 20 cropland cover classes (representing 5% increments of cover 0% to 100% defined by the reference map) were calculated and then averaged.

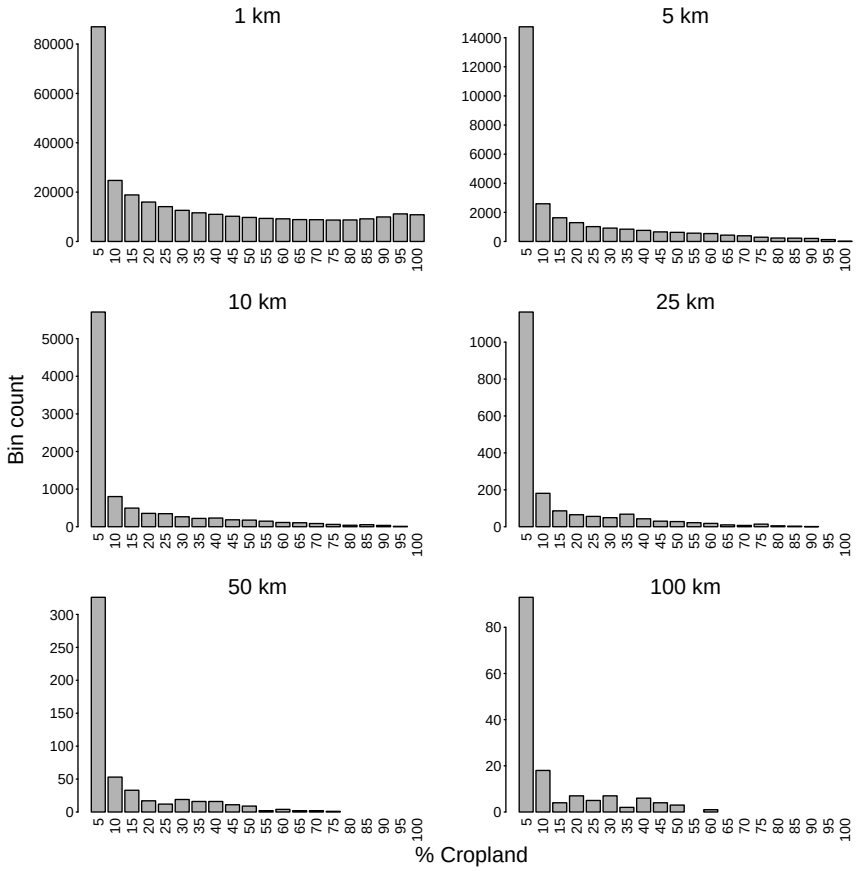

Figure 4: Number of cells within each cropland density bin at each scale of aggregation, where bins represent 5% increments of cropland cover (values on x-axis provide the upper limit of each bin). Bin values were based on the 2011 reference map, excluding areas with <0.5% cropland.

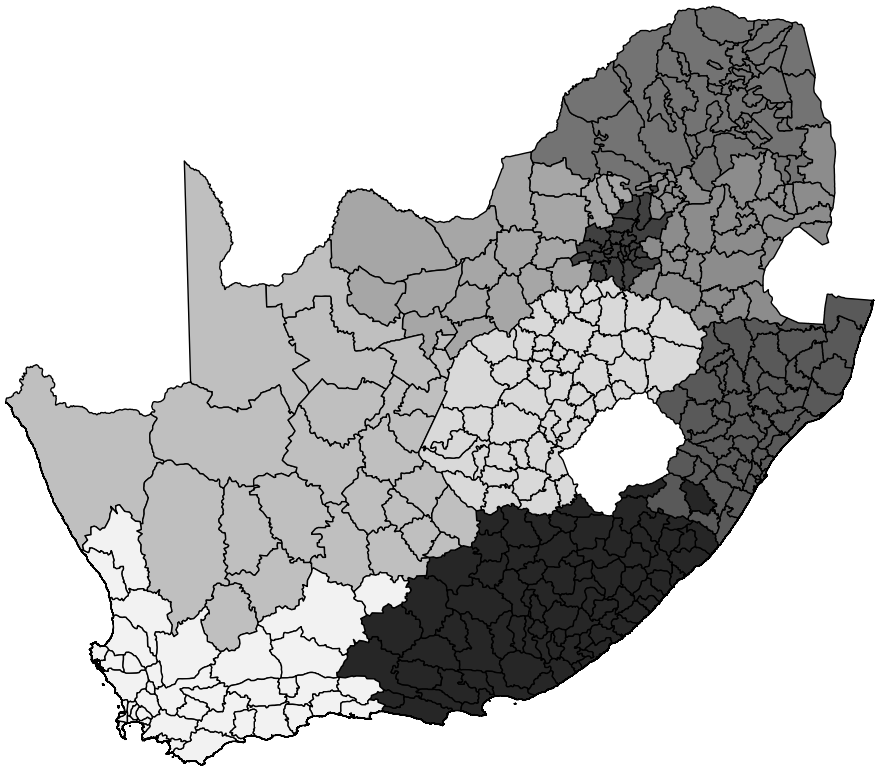

Figure 5: South Africa's magisterial districts.

4 Carbon analysis

To calculate carbon stocks using the method of Ruesch & Gibbs (2008), we simplified their landcover-specific carbon density classes into six categories, based on the similarity of their class types and the fact that they had the same assigned carbon densities. Other classes were dropped because of their low level of, or lack of, occurrence. The class adjustments were as follows:

- Classes 1, 3, 6, 8 corresponding to broadleaf and mixed forests were reduced to a single forest class.
- Classes 4 and 5 (needleaf forests) were dropped.
- Classes 9, 10, 17 (secondary forests, forest/cropland mosaic) where merged to a single secondary forest class.
- Classes 20-23 (water, snow, ice, built-up areas) were dropped.
- Class 19 (bare areas) was dropped.

All other classes (cropland, shrubland, spare vegetation) were retained. Each of these classes has a specific carbon density according to the ecofloristic zone it is found in, of which there are 10 for Africa (Zones 6-9 and 10-15; see Ruesch & Gibbs, 2008). We calculated the area of each ecofloristic zone using their polygon boundary maps[1]. For each of the simplified cover classes, we then calculated the mean ecofloristic zone carbon density value, weighted by ecofloristic zone area. We used these values to generate the different carbon maps, as described in the main text.

Maps of the mean residual differences between the reference map and each of the 5 derived carbon maps for each test map are shown in Figure 6. Table 3 provides the bias and mean absolute errors for the different carbon maps and how they vary with scale.

[1] available from cdiac.ornl.gov/ftp/global_carbon/ecofloristic_zones.zip

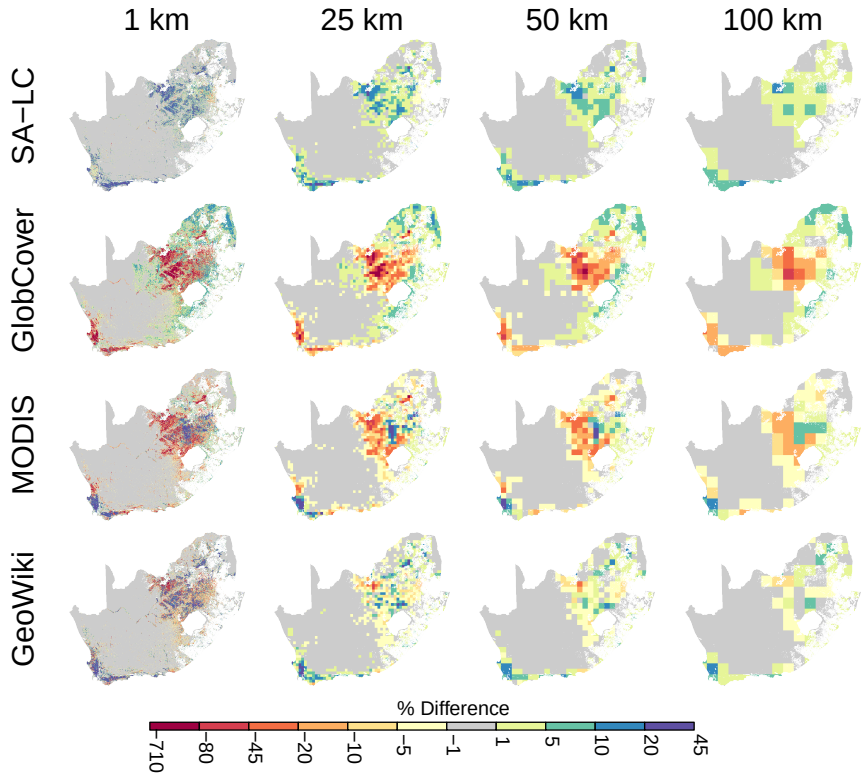

Figure 6: Spatial patterns of error (averaged across all five different possible cover types adjacent to cropland) in carbon stock estimates.

Table 3: Biases and mean absolute errors, weighted by reference cropland density, for each of the test maps across aggregation scales and each possible landcover type sharing the pixel with cropland.

Metric	Map	Cover	1 km	5 km	10 km	25 km	50 km	100 km
Bias	SA-LC	All	10.9	9.6	8.2	6.5	5.0	4.2
Bias	GlobCover	All	-123.4	-47.6	-35.9	-24.8	-17.4	-12.3
Bias	MODIS	All	-66.0	-17.6	-12.0	-8.3	-6.2	-4.1
Bias	GeoWiki	All	-20.4	2.1	2.3	1.3	0.3	0.5
Bias	SA-LC	Forest	22.7	19.7	16.9	13.3	10.4	9.0
Bias	GlobCover	Forest	-276.2	-98.3	-73.3	-50.2	-35.5	-25.4
Bias	MODIS	Forest	-146.5	-36.1	-24.5	-17.0	-12.9	-8.8
Bias	GeoWiki	Forest	-46.1	4.3	4.6	2.7	0.6	1.0
Bias	SA-LC	Secondary	18.4	16.7	14.6	11.8	9.5	8.2
Bias	GlobCover	Secondary	-186.3	-79.3	-61.2	-43.8	-31.7	-23.2
Bias	MODIS	Secondary	-101.0	-30.6	-21.5	-15.2	-11.7	-8.0
Bias	GeoWiki	Secondary	-30.5	3.4	3.7	2.2	0.6	0.9
Bias	SA-LC	Shrubland	17.9	16.4	14.3	11.6	9.4	8.1
Bias	GlobCover	Shrubland	-178.2	-77.1	-59.8	-42.9	-31.2	-22.9
Bias	MODIS	Shrubland	-96.8	-29.9	-21.1	-15.0	-11.5	-7.9
Bias	GeoWiki	Shrubland	-29.2	3.3	3.6	2.2	0.6	0.9
Bias	SA-LC	Grassland	0.3	0.3	0.3	0.3	0.2	0.2
Bias	GlobCover	Grassland	-1.9	-1.2	-1.1	-0.9	-0.8	-0.6
Bias	MODIS	Grassland	-1.1	-0.6	-0.5	-0.4	-0.3	-0.2
Bias	GeoWiki	Grassland	-0.3	0.0	0.1	0.0	0.0	0.0
Bias	SA-LC	Sparse	-4.6	-5.2	-5.1	-4.8	-4.6	-4.4

Bias	GlobCover	Sparse	25.4	18.1	16.1	13.9	12.2	10.5
Bias	MODIS	Sparse	15.4	9.1	7.6	6.3	5.5	4.4
Bias	GeoWiki	Sparse	4.0	-0.3	-0.6	-0.5	-0.3	-0.4
MAE	SA-LC	All	19.2	12.5	10.7	8.6	6.9	6.0
MAE	GlobCover	All	134.9	56.2	43.8	31.9	23.9	18.2
MAE	MODIS	All	84.8	33.2	26.2	19.9	14.9	11.4
MAE	GeoWiki	All	47.3	17.9	12.8	8.8	5.8	3.9
MAE	SA-LC	Forest	34.8	21.0	17.5	13.6	10.6	9.1
MAE	GlobCover	Forest	278.2	100.3	75.3	52.4	37.6	27.7
MAE	MODIS	Forest	168.6	56.2	42.9	31.6	22.9	17.1
MAE	GeoWiki	Forest	90.5	29.9	20.9	14.0	8.9	5.8
MAE	SA-LC	Secondary	27.4	17.9	15.2	12.1	9.6	8.3
MAE	GlobCover	Secondary	188.1	81.1	63.1	45.7	33.8	25.3
MAE	MODIS	Secondary	118.9	47.6	37.3	28.1	20.8	15.7
MAE	GeoWiki	Secondary	66.6	25.5	18.1	12.4	8.0	5.4
MAE	SA-LC	Shrubland	26.6	17.6	14.9	11.9	9.5	8.2
MAE	GlobCover	Shrubland	179.9	79.0	61.7	44.9	33.2	24.9
MAE	MODIS	Shrubland	114.2	46.6	36.6	27.7	20.5	15.5
MAE	GeoWiki	Shrubland	64.3	24.9	17.8	12.2	7.9	5.3
MAE	SA-LC	Grassland	0.4	0.4	0.3	0.3	0.2	0.2
MAE	GlobCover	Grassland	1.9	1.3	1.1	1.0	0.8	0.7
MAE	MODIS	Grassland	1.4	0.9	0.8	0.7	0.6	0.5
MAE	GeoWiki	Grassland	0.9	0.5	0.4	0.3	0.2	0.2
MAE	SA-LC	Sparse	6.7	5.8	5.4	4.9	4.7	4.4
MAE	GlobCover	Sparse	26.4	19.6	17.7	15.7	14.0	12.3
MAE	MODIS	Sparse	20.7	14.9	13.3	11.5	9.7	8.3
MAE	GeoWiki	Sparse	14.1	8.4	6.6	5.1	3.9	2.9

5 Gridded maize yield and production

For our crop yield and production analysis, we followed several steps to create the gridded yield and crop production maps. First, we calibrated cropland percentage maps so that they matched total cropland areas reported for coarser administrative units, following methods set out by Ramankutty *et al.* (2008):

1. We extracted the 2011 reference percentages within each of South Africa's 9 provinces (the same units used by Ramankutty *et al.*, 2008), converted these to proportions and summed them to calculate the "reported" cropland areas for each province. We did the same to calculate provincial cropland areas for each test map;

2. We divided the "reported" provincial cropland area estimates by provincial areas to calculate *rpcf*, the reference provincial cropland fractions, and then did the same with test map provincial cropland areas to calculate *tpcf*, the test map provincial cropland fractions;

3. A province-specific correction factor *upcf* (*rpcf/tpcf*) was then calculated and applied to each testmap cropland fraction (*tcf*) in each pixel (x, y):

$$\mathrm{tcfa}_{x,y} = \mathrm{upcf}_{x,y} * \mathrm{tcf}_{x,y} \tag{1}$$

   Yielding *tcfa*, the calibrated test map cropland fraction.

The differences between the reference map cropland fraction and *tfca* are shown in Figure 7 (where they were adjusted back to percentages), while Table 4 provides the corresponding bias and accuracy values (MAE).

We then used *tcfa*, per Monfreda *et al.* (2008), to disaggregate magisterial district-level reported maize yields and harvested areas (Stats SA, 2007), according to the following steps:

1. We first disaggregated maize yields using the following formula:

$$\mathrm{fcrop}_{x,y} = \mathrm{tfca}_{x,y} \frac{\mathrm{cmda}}{\mathrm{tmda}} \tag{2}$$

   Where *fcrop* is the fraction of each cell in each district harvested for maize, *cmda* is the district-reported harvested area for maize, and *tmda* is the total area of the magisterial district;

2. The reported maize yield for each magisterial district was then assigned to each pixel in *fcrop* in the district having non-zero maize harvested areas;

3. Production estimates were calculated by multiplying yield and harvested areas.

The maps of differences between reference-derived yield and production maps and those based on the test maps are shown in Figures 8 and 9, where the values were normalized to the country mean yield or production (calculated from the reference map). Table 5 presents the bias and accuracy values for each reference-test comparison of gridded yield and production estimates, giving both the cropland density-weighted variants of bias and accuracy, and for comparison the same values calculated as straight averages within agricultural pixels (i.e. pixels in the reference or test map having .0.5% cropland). In the former variant, yields show relatively much less bias than in the latter, where large yield biases were driven by the discrepancies in the low cropland density regions in the center of the country. In contrast, the density-weighted measures reveal large production biases, whereas the unweighted agricultural region measures show no bias in production estimates. The density-weighted variant shows that error scales with cropland cover, as seen in Figure 2 in the main text.

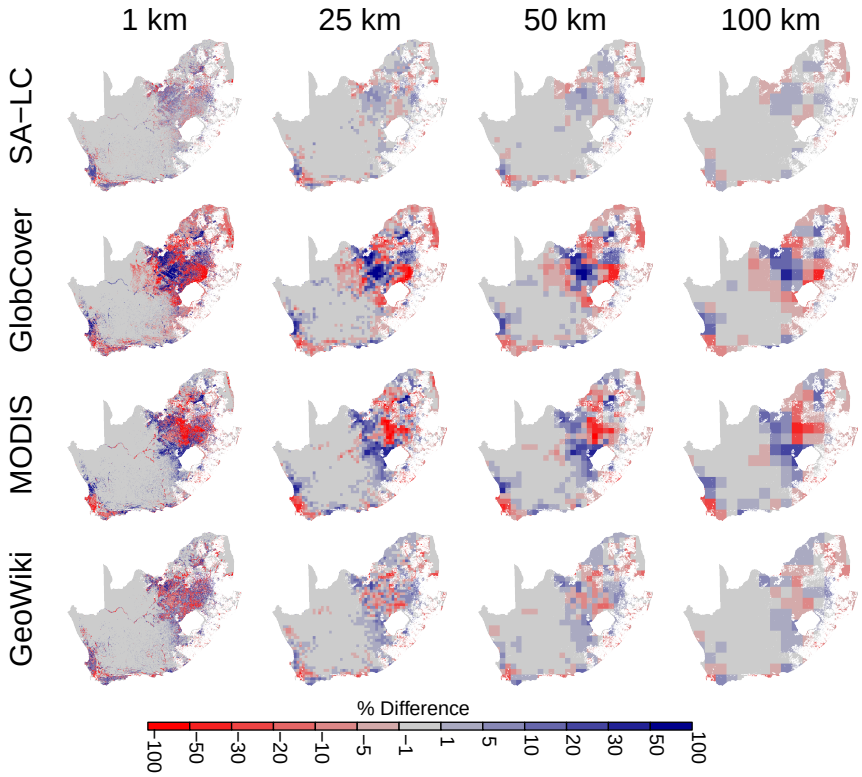

Figure 7: Errors in cropland maps adjusted using provincial cropland area statistics.

Table 4: Bias and mean absolute errors (MAE) in statistically constrained cropland maps across aggregation scales, weighted by density of cropland cover in the reference map.

Metric	Map	1 km	5 km	10 km	25 km	50 km	100 km
Bias	GeoWiki	9.7	1.1	0.6	0.4	0.5	0.1
Bias	GlobCover	34.5	18.3	14.5	10.6	7.6	4.6
Bias	MODIS	17.8	5.5	3.2	1.3	0.1	-1.3
Bias	SA-LC	6.6	2.7	2.1	1.6	1.1	0.6
Accuracy	GeoWiki	23.8	12.6	9.4	6.8	4.8	3.0
Accuracy	GlobCover	42.3	27.3	23.3	18.8	15.6	11.2
Accuracy	MODIS	33.8	21.5	18.4	15.3	12.7	10.6
Accuracy	SA-LC	11.4	6.0	4.7	3.7	2.8	1.9

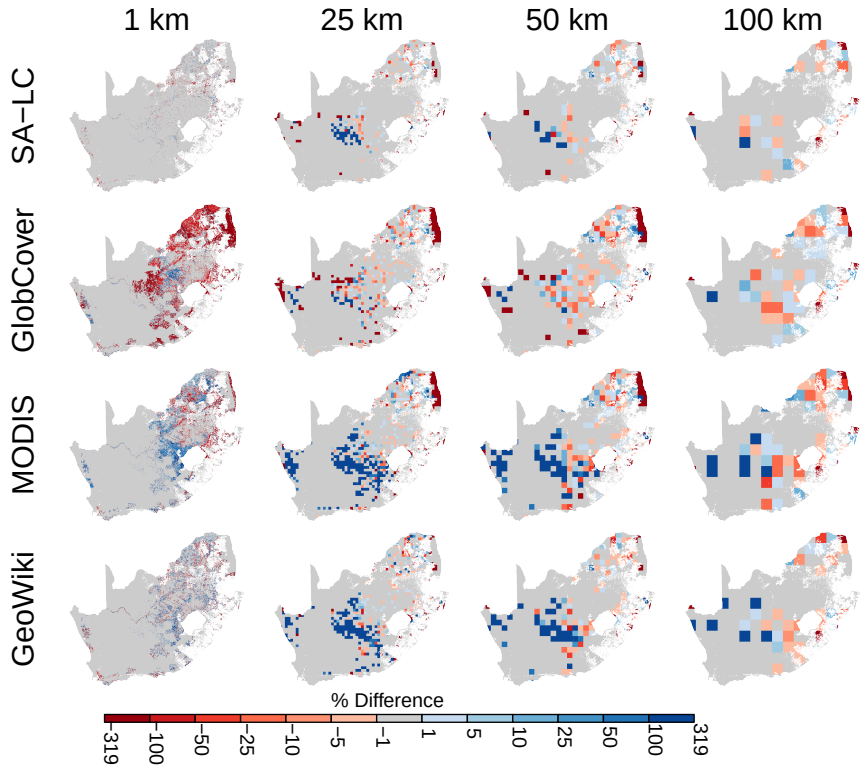

Figure 8: Errors (normalized to the reference-derived country mean) in disaggregated maize yield estimates.

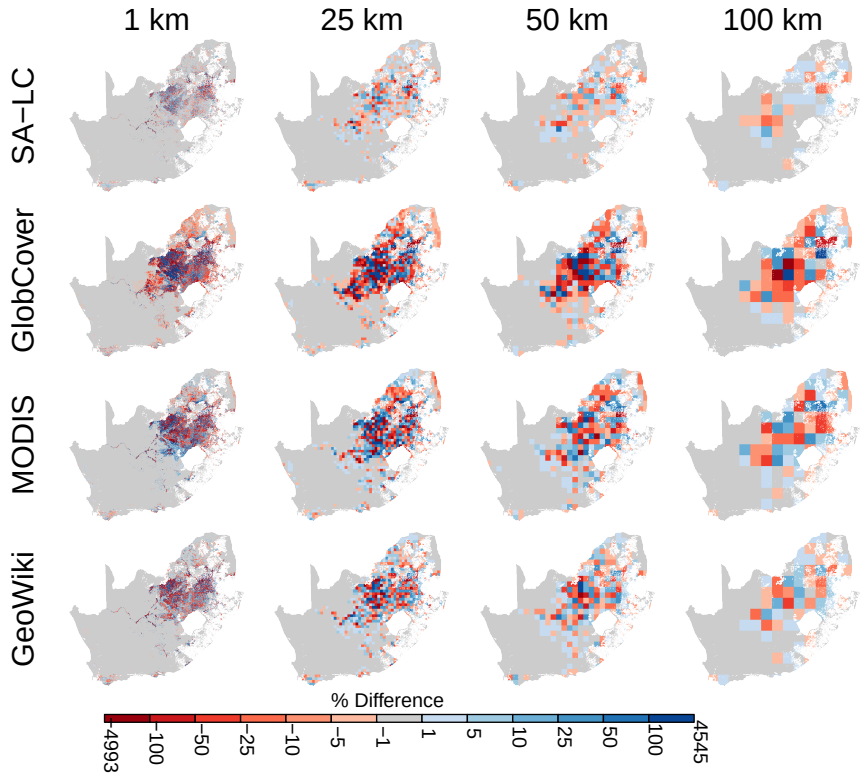

Figure 9: Errors (normalized to reference-derived country mean) production estimates calculated from disaggregated maize yield and harvested area estimates.

Table 5: Biases and mean absolute errors (MAE) in disaggregated maize yield and production (calculated from disaggregated yield and harvested area estimates) maps. Values for both density-weighted and agricultural areas bias and accuracy are presented. Bias and MAE were normalized to their respective mean values calculated from reference maps.

Region	Metric	Map	Variable	1 km	5 km	10 km	25 km	50 km	100 km
Density	Bias	SA-LC	Yield	1.2	0.3	0.0	0.0	0.0	-0.3
Density	Bias	GlobCover	Yield	9.8	0.9	0.0	-0.6	-0.6	-0.6
Density	Bias	MODIS	Yield	19.6	8.9	5.7	3.0	1.5	-0.6
Density	Bias	GeoWiki	Yield	8.0	3.0	1.5	0.6	0.3	-0.6
Density	Bias	SA-LC	Production	6.9	1.6	0.5	-0.2	-0.1	-0.1
Density	Bias	GlobCover	Production	60.5	50.2	43.7	35.1	23.3	12.5
Density	Bias	MODIS	Production	21.9	6.0	1.8	-1.8	-0.9	-0.5
Density	Bias	GeoWiki	Production	12.7	-3.3	-4.6	-3.8	-0.5	-0.9
Density	MAE	SA-LC	Yield	1.2	0.3	0.3	0.3	0.6	0.9
Density	MAE	GlobCover	Yield	9.8	1.2	0.9	1.5	1.8	1.8
Density	MAE	MODIS	Yield	19.6	9.2	6.2	4.5	3.9	2.4
Density	MAE	GeoWiki	Yield	8.0	3.3	1.8	1.8	1.2	1.2
Density	MAE	SA-LC	Production	19.0	14.3	11.8	8.9	5.1	2.3
Density	MAE	GlobCover	Production	95.6	102.0	100.4	88.1	65.8	46.4
Density	MAE	MODIS	Production	66.8	62.0	58.4	46.4	25.7	14.6
Density	MAE	GeoWiki	Production	47.3	43.6	37.6	29.3	19.4	7.9
Agricultural	Bias	SA-LC	Yield	-5.1	-0.3	3.0	3.6	3.6	1.5
Agricultural	Bias	GlobCover	Yield	-58.0	-36.0	-22.3	-11.9	-8.9	-1.5
Agricultural	Bias	MODIS	Yield	5.1	21.4	29.2	26.8	20.5	11.6

Agricultural	Bias	GeoWiki	Yield	2.4	24.4	29.5	25.3	21.4	9.8
Agricultural	Bias	SA-LC	Production	0.0	-0.1	-0.1	-0.1	-0.0	0.0
Agricultural	Bias	GlobCover	Production	0.0	-0.1	0.0	0.1	0.3	0.3
Agricultural	Bias	MODIS	Production	0.0	-0.1	-0.1	-0.1	0.0	-0.1
Agricultural	Bias	GeoWiki	Production	0.0	0.1	0.0	0.0	0.1	0.1
Agricultural	MAE	SA-LC	Yield	15.5	16.7	19.9	15.8	12.2	6.8
Agricultural	MAE	GlobCover	Yield	71.7	48.2	38.1	23.5	17.3	6.2
Agricultural	MAE	MODIS	Yield	55.9	51.2	50.9	44.9	38.4	20.8
Agricultural	MAE	GeoWiki	Yield	41.1	41.1	40.5	35.1	28.6	14.6
Agricultural	MAE	SA-LC	Production	19.7	11.3	8.6	5.5	3.3	1.9
Agricultural	MAE	GlobCover	Production	55.7	55.5	52.5	42.2	28.1	17.3
Agricultural	MAE	MODIS	Production	56.0	41.3	35.6	24.9	14.1	8.4
Agricultural	MAE	GeoWiki	Production	43.7	30.2	23.5	15.3	9.3	4.0

6 Evapotranspiration analysis

A number of the variables used by the Variable Infiltration Capacity (VIC; Liang *et al.*, 1994) model are linked to a 0.25° resolution, AVHRR-derived landcover map. These include seasonal leaf area index (LAI) phenologies (Fig. 10), as well as other properties such as plant rooting depth and infiltration rates. Landcover properties therefore impact the model's simulation of water balance. To test how landcover map error impacts VIC simulations, we created five new versions of VIC's native landcover scheme, one based on the reference cropland map, and the other four on each of the test maps. For each version, we first reprojected the relevant 25 km cropland map to the geographic coordinate system used by VIC and resampled it to 0.25°. We then adjusted VIC's landcover scheme by replacing its existing cropland percentages with those from our map, and then proportionally adjusted the remaining cover types in each cell to accommodate the changed cropland amounts. Since LAI is strongly linked to rainfall seasonality in South Africa, which varies between winter (May-August) rainfall in the west and southwest of the country and summer (October-March) rainfall in the east and northeast, we assigned LAI curves that peaked during the winter months to the cover types in the western half of the country, and those peaking in the summer months to the eastern half (Fig. 10).

After making these adjustments, we conducted five different VIC simulations, one for each of the adjusted landcover schemes. The model was run at a daily time step for 28 years, from 1981-2008, from which monthly total evapotranspiration values were extracted. We calculated from these results the average annual total ET, the maximum and minimum monthly ET observed throughout the entire time series, and the mean ET during the month which on average had the highest ET during the time series. We then found the differences between the reference and test map variants of each of these ET variables, and calculated their bias and MAE values (Fig. 11). The results are similar across the four variables, and we present the difference maps for the mean annual ET comparison in the main text.

7 Agent-based model assessment

7.1 Maize production

The agent-based model simulates household food production using a look-up table that links maize yields to several different variables: planting date; cultivar (open-pollinated or hybrid); soil properties, and weather. The look-up table itself was based on a series of yield simulations conducted by the DSSAT cropping system model run over 31 years (1979-2010) for a location in the southern Province of Zambia, corresponding to the region where household survey data were collected. The model

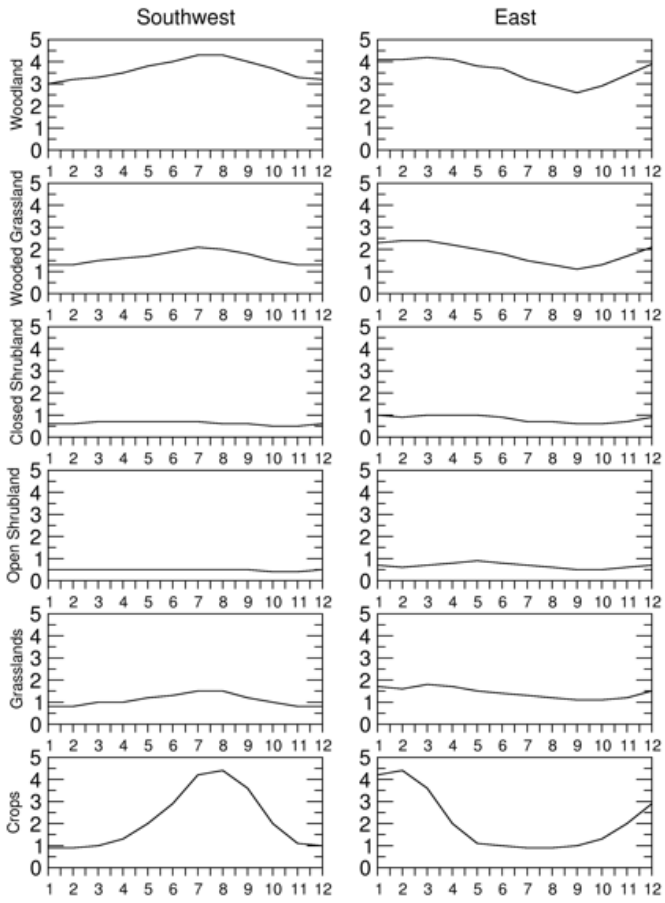

Figure 10: Seasonal LAI curves for different cover types in VIC's landcover scheme, showing the phenologies used for the winter-rainfall portion of the country in the west and southwest (left column), and for the summer rainfall region in the country's easter to northeast (right column).

was run for three dominant soil series[2], using two open-pollinated and two hybrid cultivars (each pair representing a short- and medium-length growing season variety) whose coefficients were obtained from the inputs files for the PSIMs modeling platform (Elliott *et al.*, 2014). Each cultivar was simulated under planting dates ranging from October 15th until January 30th, varying by 15 day increments, with row spacing fixed at 90 cm and planting density at 3.7 plants m^2, and 5 kg ha^{-1} applied at planting. Models were run using weather data extracted from a bias-corrected version of the Princeton Global Meteorological forcing dataset (Chaney *et al.*, 2014; Estes *et al.*, 2014; Sheffield *et al.*, 2006).

7.2 District selection

For our analysis of how landcover map error impacts agent-based model results, we selected four districts in South Africa which had similar climatic characteristics (∼800 mm annual rainfall) to the region in Zambia where the crop model simulations and household survey data were collected. These were four contiguous districts along the western border of Lesotho, Clocolan, Ficksburg, Fouriesburg, and Marquard (Fig. 12, top left), which had between 29-45% of their areas covered by cropland, according to the 2011 reference map (Fig. 12, top right).

[2]extracted from a gridded version of the ISRIC-WISE database, available from http://dssat.net/649

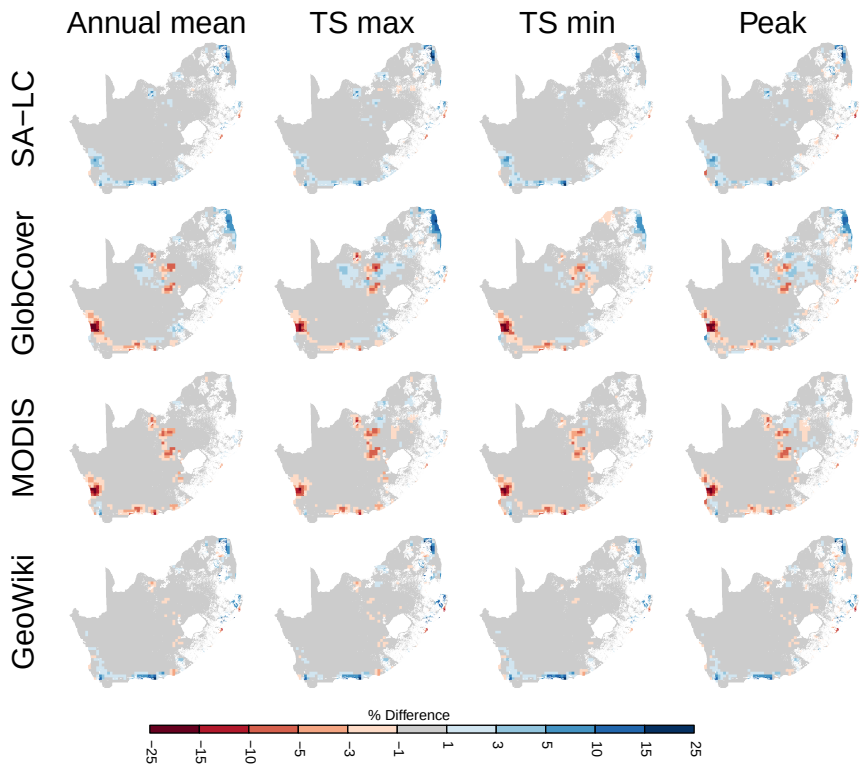

Figure 11: Maps of error between reference- and test map-derived ET estimates simulated by the VIC model. Four different variants of ET were assessed: annual mean (left column, also presented in Fig. 2 of the main text), the maximum (second column, TS max) and minimum (third column, TS min) ET values observed during the 30 year times, and mean ET during the month in which ET peaks (right column, Peak).

7.3 Model description

The following summary follows the Overview, Design and Details protocol (Grimm *et al.*, 2006; Polhill *et al.*, 2008) designed as a standardized method for describing individual-based models and agent-based models.

7.3.1 Purpose

We developed an agent-based simulation to explore the intra-seasonal dynamics during the maize production of smallholder farmers under climate changes with an emphasis on household-level heterogeneity.

7.3.2 State variables and scales

The primary analytical components in the model are actors (households) and cells (land). We designed a greedy algorithm that allocates land to households based on a heuristic rule that households tend to live near each other to form communities like villages. The simulation runs on four districts in South Africa. For each district, a set of synthetic household agents is created from the land cover data of that district and the farmer register in Monze District of Zambia. We extracted the frequencies of cultivated area from the farmer register using bins (in hectare) of 0-1, 1-2, 2-3, 3-4, 4-5, and 5-10. Because each of the land cells is 1 hectare (ha), we considered the households in the same bin having the same integer value of cultivated area. For example, households in the bin (0, 1] all have 1 ha of cropland, and households in the bin (5, 10] all have 8 ha of cropland. From this distribution of cultivated area (Figure 13), we calculated its mean value as:

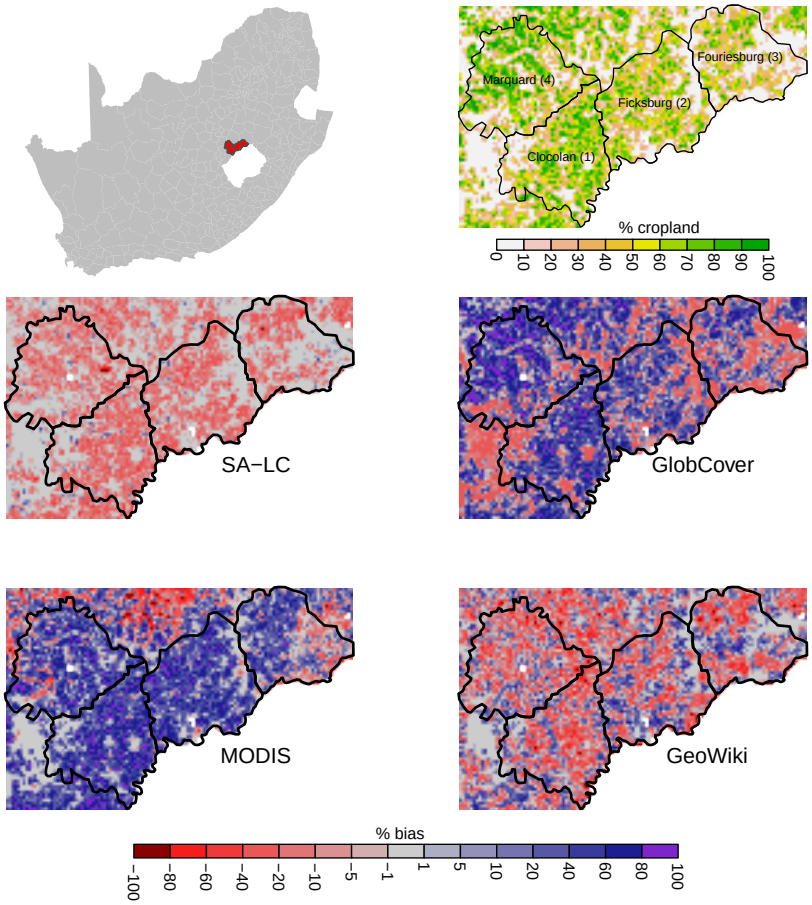

Figure 12: The location of the four selected magisterial districts (top left) used in evaluating agent allocation bias, the reference levels of cropland cover within those districts (top right), and the difference in cropland percentage between the reference and each of the four cropland maps (lower four panels).

$$
\begin{aligned}
&\textit{Mean value of cultivated area} \\
&= \textit{area of } 1^{st} \textit{ bin} * 1 + \textit{area of } 2^{nd} \textit{bin} * 2 + \textit{area of } 3^{rd} \textit{ bin} * 3 + \textit{area of } 4^{th} \textit{ bin} * 4 \\
&\quad + \textit{area of } 5^{th} \textit{ bin} * 5 + + \textit{area of } 6^{th} \textit{ bin} * (6+10)/2 \\
&= 0.3308 * 1 + 0.3591 * 2 + 0.1619 * 3 + 0.0835 * 4 + 0.0335 * 5 + 0.0062 * 5 * 8 \\
&= 2.2854 \textit{ (ha)}
\end{aligned}
$$

We then used the mean value of cultivated area to divide the total area of cropland in each district to get the number of households (Table 6).

Table 6: Number of households created and their total cultivated area.

District	Cropland area (ha)	Calculated number of households	Actual number of household agents	Total cultivated area of household agents
clocolan	54413.3	23810	23810	54410
ficksburg	51363.93	22475	22475	51367
fouriesburg	29402.99	12866	12865	29400
marquard	55633.12	24343	24341	55615

Four key household initial characteristics define agents: (1) cultivated area (used primarily for

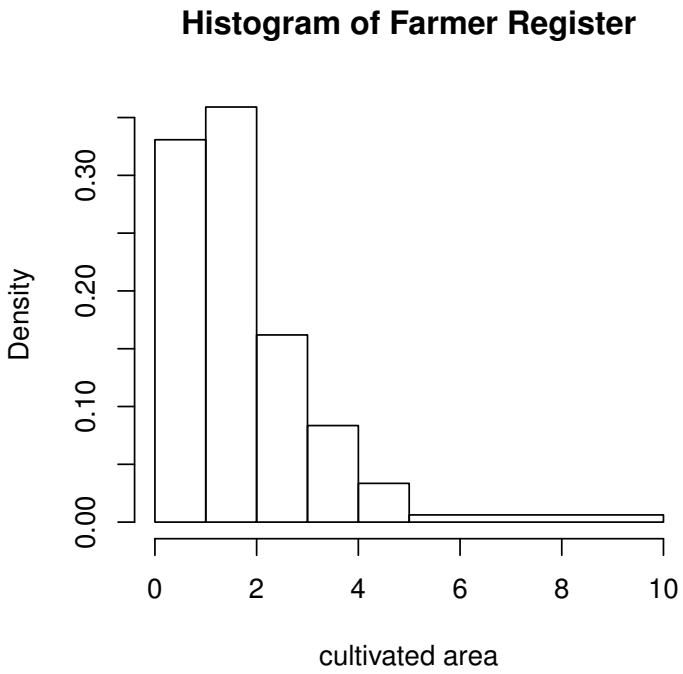

Figure 13: Histogram of cultivated area in the farmer register.

maize production), (2) day of planting maize, (3) soil properties of land cells, and (4) seed type of maize. To determine (1), the household agents are assigned 1 ha, 2 ha, 3 ha, 4 ha, 5ha, and 8 ha, according to the frequencies we extracted from the farmer register previously. The results in Table 1 show a good matching between the total area of cropland from the land cover data and the total cultivated area of household agents. As for (2), (3), and (4), we create a distribution from our household survey that is shown in Figure 14. In addition to the initial characteristics, each household has two internal variables: household ID and harvest amount. Each land cell has the following variables: household ID (if assigned), land cover type (cropland or not), and revenue (if harvested). The harvest amount of a household is the sum of revenue of cropland cells managed by that household.

The model runs biweekly for the growing season from October 2007 to April 2008, with all household harvested on April 1st in 2008. The four districts in South Africa are represented at a 100m spatial resolution. We took the gridded cropland reference map at 1 km and disaggregated it into 100m land cells that are either cropland or not.

7.3.3 Process overview and scheduling

We developed an algorithm shown in Algorithm 1, to allocate the cropland cells to households. Our land allocation algorithm first chooses a number of seed households (HHs) in the procedure ALLOCATE_HH and invokes ALLOCATE_MANY_FARMLAND to randomly assign unallocated cropland cells/patches to them. Then the algorithm randomly selects an unallocated cropland cell/patch that is adjacent to some already allocated cell/patch, and allocates it to the next household using ALLOCATE_MANY_FARMLAND. In this way, the households should be located close to each other to form communities. This is repeated until there is no unallocated land or there is no household that hasn't been assigned any land. ALLOCATE_MANY_FARMLAND repeatedly invokes ALLOCATE_FARMLAND until required amount of cropland has been allocated to a household. Comments are denoted by right-pointing triangles.

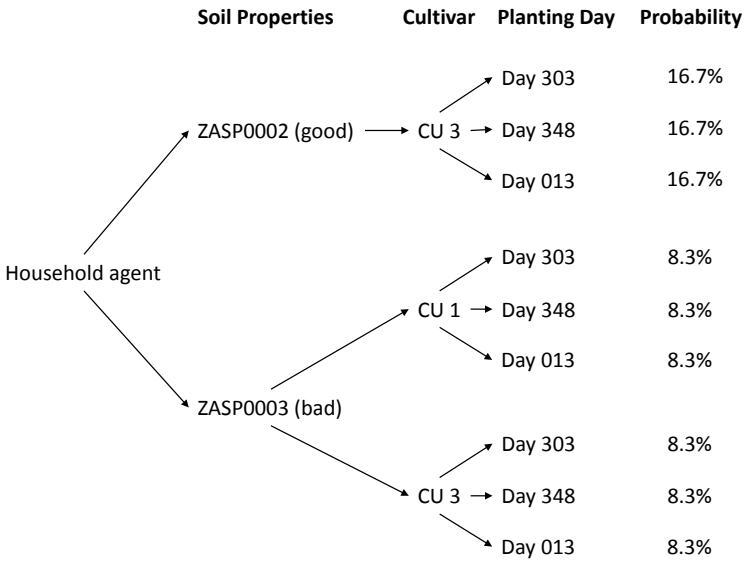

Figure 14: Distribution of soil properties, maize type (cultivar), and planting day. The soil properties, cultivar, and planting day can be used to find the yield in the look-up table created from the DSSAT cropping system model. CU 1 indicates a local maize type and CU 3 indicates a hybrid one. Day 303 means the 303$^{\text{rd}}$ day of the year.

Once a cropland cell is allocated to a household agent, its soil property is determined by the characteristics of the household. The simulation runs biweekly and when the planting day of a household arrives, that household will plant maize on all of its cropland cells. When April 1st arrives, all households will harvest their cropland cells where the yield is calculated using the look-up table from the DSSAT cropping system.

7.4 Design concepts

7.4.1 Emergence

Cropland cells are allocated to households with seeding and geospatial proximity searching. The resulting household communities define the scope of labor sharing activities.

7.4.2 Adaptation

7.4.3 Fitness/objectives

The objective of the model is to evaluate the impact of climate variability on food production, one component of food security. An important aspect of climate change impact research in agricultural systems is the ability of farmers to adapt to changing climate conditions. This means farmers must have some signal detection mechanism and then some process by which they modify their behavior based on their own experiences, or experiences they learn from others. However, in the version of the model used in this paper we reduced the overall complexity to focus primarily on the impact of crop bias in land cover data on model initialization (or rather sensitive the allocation of land to agents is to different land cover datasets). In particular, we do not employ agent-interactions or learning in the version of the model used here.

7.4.4 Prediction

The look-up table assumes no deviation from optimal yields for the specified planting scenarios. In other words, households do not take into account the possibility of frost, pests, or other disturbances

that affect yields in the real world.

7.4.5 Interaction

Households interact with the cropland cells through planting and harvesting. Because labor sharing is turned off in the model there is no formal agent-to-agent interaction in the model.

7.4.6 Sensing

Households do not sense the actions of other agents or environmental conditions in the version of the model used for this manuscript.

7.4.7 Model calibration

The model was calibrated during two phases: land allocation and food production. After the cropland cells are allocated to households, the model is considered well calibrated when all households were allocated their appropriate area of cropland, and all cropland was allocated. The land allocation algorithm starts with an initial value of maximum searching scope. The model is then run iteratively, increasing the value of maximum searching scope if there are households not allocated any cropland and cropland unallocated. When harvesting is finished at April 1st, the total production is compared against the district level census data, and the average yield is compared against the post-harvest survey data.

Algorithm 1 Algorithm to allocate cropland patches (cells) to households.

procedure ALLOCATE_FARMLAND(H, P) ▷ H: household, P: patch
  $A \leftarrow$ the area of farmland needed by H
  if $A >$ the area of P **then** ▷ P is fully occupied by H
    $occupiedRatio(P) \leftarrow 1$
  else
    $occupiedRatio(P) \leftarrow (A - 1)$ ▷ P is partially occupied by H
  end if
  $N \leftarrow$ neighbor farmland (in radius r) of P ▷ r is a global parameter of allocation radius
  $status(N) \leftarrow$ tentative seed patches
end procedure

procedure ALLOCATE_MANY_FARMLAND(H, P) ▷ H: household, P: patch
  Invoke allocate_farmland(H, P)
  repeat
    $searchRadius \leftarrow 700m$ ▷ starting from a threshold value
    $UP \leftarrow$ a randomly selected unoccupied farmland within $searchRadius$ of P
    Invoke allocate_farmland(H, UP)
    if $A >$ the area of P **then**
      $searchRadius \leftarrow (searchRadius + 100m)$
    end if
  until H is assigned enough farmland $\vee$ $searchRadius == s$ ▷ s is a global parameter of the maximum search radius
end procedure

procedure ALLOCATE_HH
  $i \leftarrow 1$ ▷ the id of current household to be allocated
  repeat
    $SH \leftarrow$ the ith household
    $status(SH) \leftarrow$ seed household
    $SP \leftarrow$ a randomly selected patch
    Invoke allocate_many_farmland(SH, SP)
    $i \leftarrow (i + 1)$
  until $i == numSeed$ $\vee$ there is no unoccupied land ▷ $numSeed$ is a global parameter of the total number of seed households created during initialization
  repeat
    $SH \leftarrow$ the ith household
    $TSP \leftarrow$ a randomly selected patch so that $status(TSP) ==$ tentative seed patch $\wedge$ $occupiedRatio(TSP) == 0$
    Invoke allocate_many_farmland(SH, TSP)
    $i \leftarrow (i + 1)$
  until $i == numHHs$ $\vee$ there is no unoccupied land ▷ $numHHs$ is the total number of households
end procedure

References

Chaney NW, Sheffield J, Villarini G, Wood EF (2014) Spatial Analysis of Trends in Climatic Extremes with a High Resolution Gridded Daily Meteorological Data Set over Sub-Saharan Africa. *Journal of Climate*.

Elliott J, Kelly D, Chryssanthacopoulos J, *et al.* (2014) The parallel system for integrating impact models and sectors (pSIMS). *Environmental Modelling & Software*, **62**, 509–516.

Estes LD, Chaney NW, Herrera-Estrada J, Sheffield J, Caylor KK, Wood EF (2014) Changing water availability during the African maize-growing season, 19792010. *Environmental Research Letters*, **9**, 075005.

Estes LD, McRitchie D, Choi J, *et al.* (2016) A platform for crowdsourcing the creation of representative, accurate landcover maps. *Environmental Modelling & Software*, **80**, 41–53.

Fritz S, See L, McCallum I, *et al.* (2015) Mapping global cropland and field size. *Global Change Biology*, **21**, 1980–1992.

Grimm V, Berger U, Bastiansen F, *et al.* (2006) A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, **198**, 115–126.

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**, 14415.

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.

Olofsson P, Foody GM, Herold M, Stehman SV, Woodcock CE, Wulder MA (2014) Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, **148**, 42–57.

Polhill JG, Parker DC, Brown DG, Grimm V (2008) Using the ODD protocol for comparing three agent-based social simulation models of land use change. *Journal of Artificial Societies and Social Simulation*, **11**, 1–25.

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.

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*.

Sheffield J, Goteti G, Wood EF (2006) Development of a 50-Year High-Resolution Global Dataset of Meteorological Forcings for Land Surface Modeling. *Journal of Climate*, **19**, 3088–3111.

Stats SA (2007) Commercial Census of Agriculture, South Africa. http://www.statssa.gov.za/agriculture/default.asp

Wood SN (2001) mgcv: GAMs and Generalized Ridge Regression for R. *R News*, **1**, 20–25.