

Supplementary Methods and Results

1 1 Overview

2 This document provides supplementary information on methods and results from the different com-
3 ponents of the analysis described in the main text. The code used to conduct the analysis is presented
4 after section 7.

5 2 Reference data accuracy

6 The accuracy of the reference vector dataset was assessed by Estes *et al.* (2016), and described further
7 here. The assessment was undertaken using a sub-sample of 609 1 km² grid cells, which were selected
8 using a weighted randomized sampling scheme. Weights were derived from a logistic regression
9 model of cropland occurrence probability in South Africa, with 1 corresponding to the lowest quartile
10 of probability, and 4 the highest. The accuracy assessment included all cropland classes mapped within
11 the reference dataset, which ranged from communal/smallholder fields to commercial row crops to
12 orchards and other types of horticulture. A visual accuracy assessment was conducted within each 1
13 km² grid cell, wherein each cell was divided into 25 smaller cells of 200 X 200 m (4 ha), and then
14 the proportion of each cell overlapping with cropfields visible in underlying, high resolution Google
15 Maps imagery was then calculated to the nearest 5% of coverage, using a finer 20-cell mesh overlaid
16 on each sub-cell within which field presence/absence was recorded. The same procedure was then
17 performed to assess coverage by reference map polygons, as well as the intersections and differences
18 between cells determined to be occupied by i) visual assessment and ii) by the digitized polygons.
19 These intersections and differences were used to calculate the area of true positives, false positives,
20 true negatives, and false negatives within each 1 km² grid cell, which were then used to calculate
21 accuracy measures. In this case, accuracy was assessed using two land cover classes: cropland and
22 non-cropland, thus a two class confusion matrix was constructed (Table 1) according to remote sensing
23 classification accuracy guidelines set out by Olofsson *et al.* (2014). In this case, the reference data
24 were the visually interpreted crop field presences/absences, and the map data were the field boundary
25 vectors being assessed (i.e. the data which provide the reference maps in this study), the class “crop”

26 refers to crop field presence and “non-crop” to crop field absence. The total area of vectorized crop
 27 fields was used to calculate the proportion W of South Africa’s total area mapped as crop fields, and
 28 of the areas “mapped” (by omission) as having other cover types. These were then used to weight the
 29 proportions of each mapped class i corresponding to each reference class j by the total area A mapped
 30 for class i . The total accuracy (0.97, or 97%) was calculated by summing the diagonal (bold in Table
 31 1), as well as the producer’s accuracy P for each reference class j , and the user’s accuracy U for each
 32 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)		W_i	A_i (ha)	U_i
Map (i)	Crop	Crop	Non-crop			
	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			

33 3 Input land cover data

34 The land cover data used in this analysis are illustrated in Figure 1, which shows their original form
 35 prior to being converted to 1 km cropland percentages over a small region of South Africa.

36 4 Cropland map error analysis

37 We examined the impact of several sources of uncertainty on our results. The first is the potential
 38 temporal mismatch between the reference dataset and the land cover products we were testing. We
 39 tested this by examining two versions of the reference dataset, the initial version created in 2007, and
 40 the updated version from 2011 used in the main analysis. The 2011 version shows 3% more cropland
 41 area than the 2007 version. To examine the effect of the over mapping bias and accuracy measures,

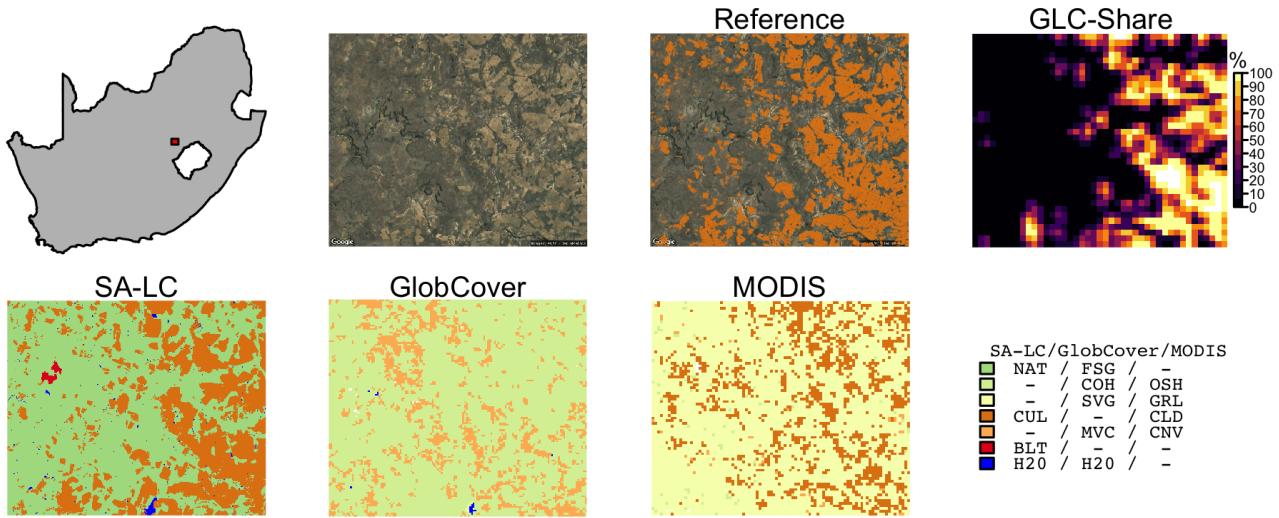


Figure 1: The original form of the reference dataset (field boundary polygons) and land cover datasets evaluated in this study, illustrated in relation to an underlying high resolution satellite image (from Google Maps, second map from left on top row) selected from a sub-region (red area in top left map) of South Africa. The original cover types (excluding a few classes having <10 pixels in the selected location) comprising the SA-LC, GlobCover, and MODIS land cover datasets are shown (GLC-Share is provided as cropland percentages), with the legend at the right providing the corresponding acronym for each class. For SA-LC, these classes are natural vegetation (NAT), cultivation (CUL), built (BLT), and water (H2O). For GlobCover, the classes are mosaic forest or shrubland (50-70%) / grassland (20-50%) (FSG), closed to open herbaceous vegetation (COH), sparse vegetation (SVG), mosaic vegetation (50-70%) / cropland (20-50%) (MVC), and water (H2O). For MODIS, the classes are open shrubland (OSH), grassland (GRL), cropland (CLD), and cropland/natural vegetation mosaic (CNV)

42 we converted the 2007 vector maps to gridded cropland percentage and subtracted each test map, and
 43 then compared the differences in bias and mean absolute error (MAE) values calculated between both
 44 (Table 2). The largest difference was 1.63%, between the 1 km SA-LC residuals, which means that
 45 the overestimation bias by SA-LC was actually greater relative to the 2007 version of the reference
 46 map. Except for the corresponding MAE value, all other difference were <1%.

47 As a further exploration of uncertainty, we included the full range of variability that resulted from
 48 a) using either the 2007 or 2011 map to calculate bias and accuracy, and b) from different levels of
 49 cropland percentages used when converting MODIS and GlobCover mixed cropland classes into crop-
 50 land percentage maps, following Fritz *et al.* (2015). We pooled the residuals from across each of these
 51 permutations, for each reference-test map combination, and examined how their overall mean values
 52 change with aggregation scale (Fig. 2), and also assessed the total distribution of errors within differ-
 53 ent class of cropland density, determined by dividing the 2011 1 km reference map into 20 different
 54 zones, or bins, of cropland density, ranging from 0-5% cropland cover up to 95-100% cover. We then
 55 calculated summary statistics from the pooled residuals and absolute values of residuals within each

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

56 of these bins (Fig. 3).

57 We also calculated how different methods of calculating the bias and accuracy statistics impacted
 58 out findings, using three different methods. First, we simply took the straight averages across the
 59 entire country, which substantially understates bias and accuracy because cropland covers only 10%
 60 of the country, and all land cover products successfully discriminate the non-cropped regions. We also
 61 compared the average error metrics extracted from within just the agricultural regions of the maps,
 62 defined here as the union of areas having >0.5% cropland in the reference and each test map. This also
 63 tends to understate bias, because the area having <5% cropland is much larger than areas having higher
 64 densities of cropland (see Fig. 5 for frequencies of cells per five percentile bins of cropland density).
 65 Finally, we calculated a density-independent metric, which produces values similar to the density-
 66 weighted means presented in our main analysis, which was calculated by averaging the mean bias and
 67 MAE values within different levels of cropland cover (binned into increments of 5%, excluding areas
 68 with <0.5% cropland). The results of these three alternate metrics are shown in Figure 4.

69 We used magisterial district boundaries for South Africa (Fig. 6) to extract absolute residual values
 70 for agricultural areas (>0.5% cropland cover) and 2011 reference cropland density values, the means
 71 of which respectively served as the response and predictor in the generalized additive model (Wood,
 72 2001) analysis (main text).

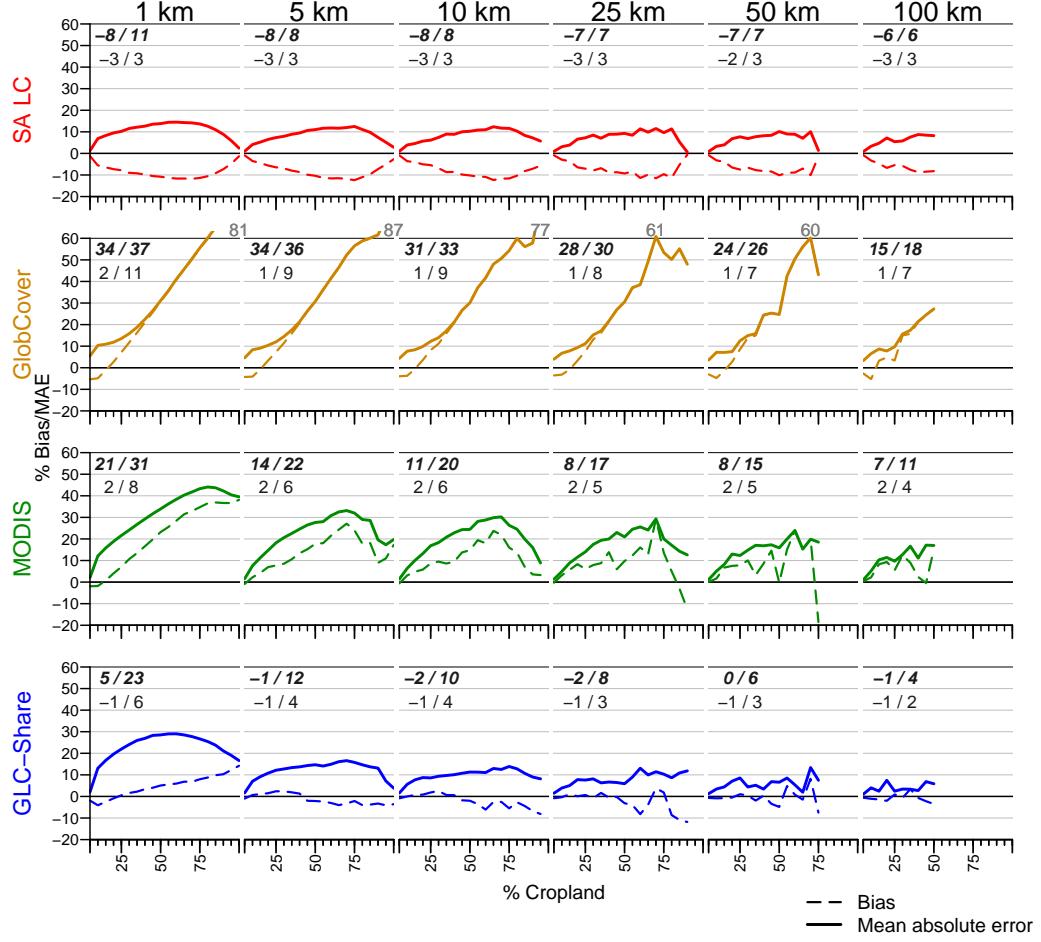


Figure 2: Biases and mean absolute errors (MAE) for each of the cropland maps as a function of cropland density (calculated using the 2011 reference maps) and aggregation scales. Rows present biases by map product, columns by aggregation scale. Dash lines indicate bias at each level of cropland density, 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; bold numbers provide the mean of bin-wise means of bias/MAE, while non-bold numbers represent the mean of bin-wise means weighted by the number of cells contributing to each bin. 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.

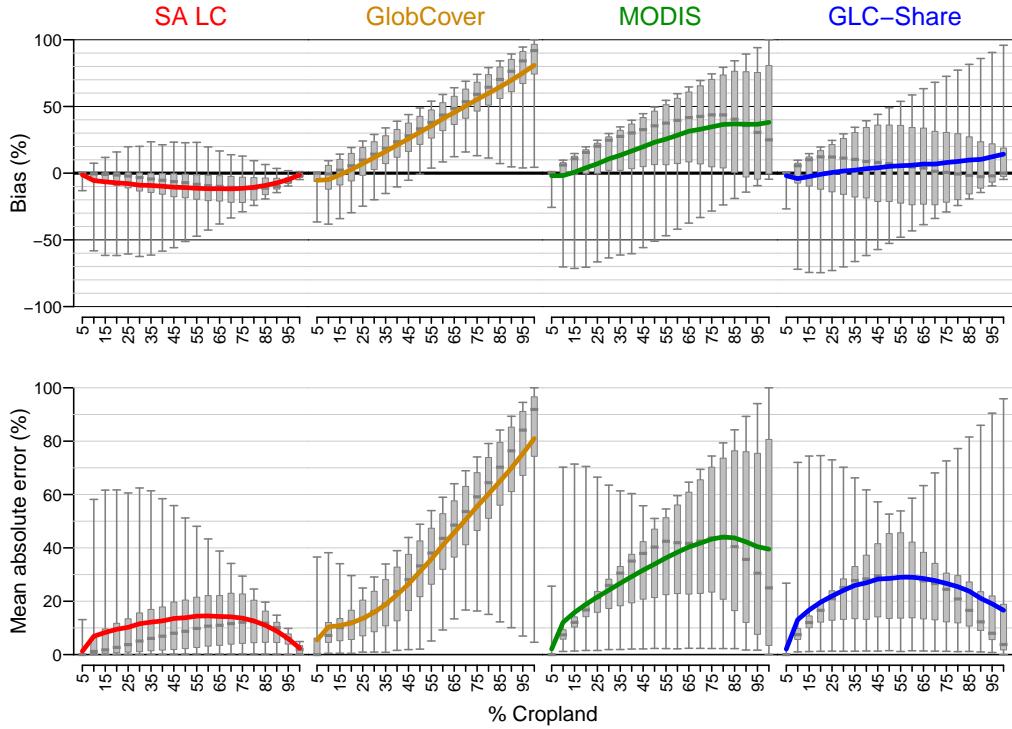


Figure 3: Biases and mean absolute errors (MAE) for each of the cropland maps at 1 km resolution, as a function of cropland density. Colored lines (color-coded to map product name) show the bias/MAE at each level of cropland density, calculated in bins spanning 5% (e.g. 0-5% cropland cover, 5-10%, etc.). 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.

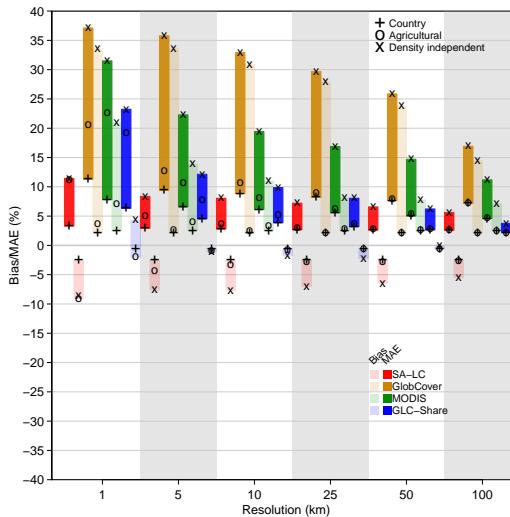


Figure 4: A comparison of three alternate methods for calculating bias and accuracy (mean absolute error): as a straight average across the entire country, averaged over agricultural areas only (the union of areas defined as having >0.5% cropland cover in the reference map and each test map), and independent 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.

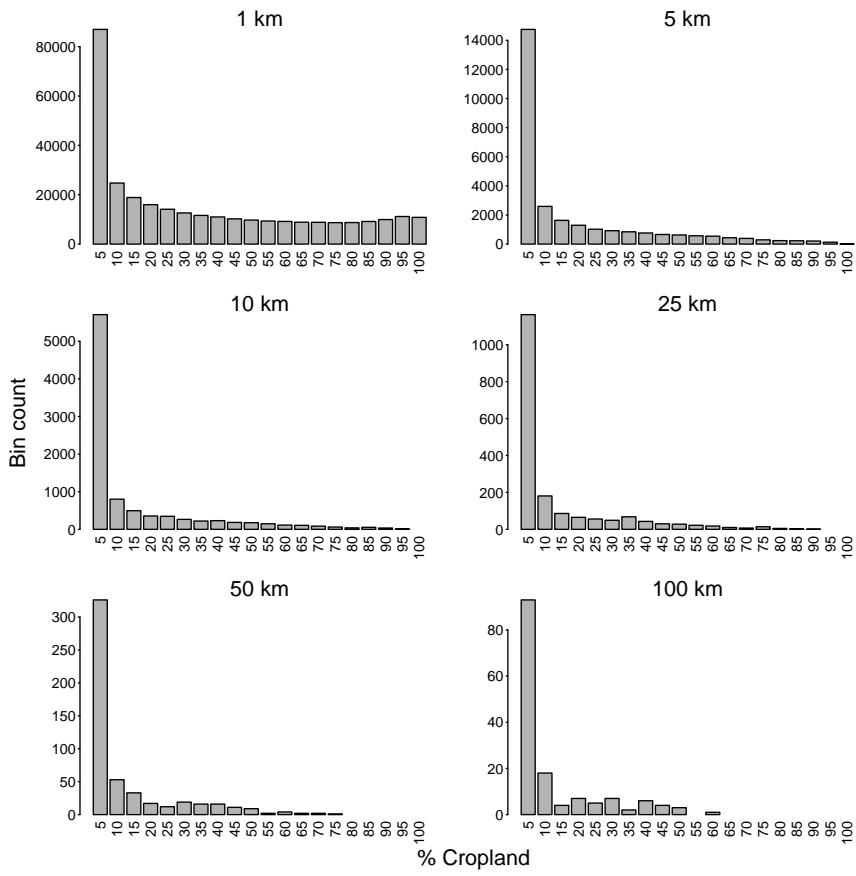


Figure 5: Number of cells within each cropland density bin at each scale of aggregation, where bins represent 5% increments of cropland cover (values on x-axis provide the upper limit of each bin). Bin values were based on the 2011 reference map, excluding areas with <0.5% cropland.

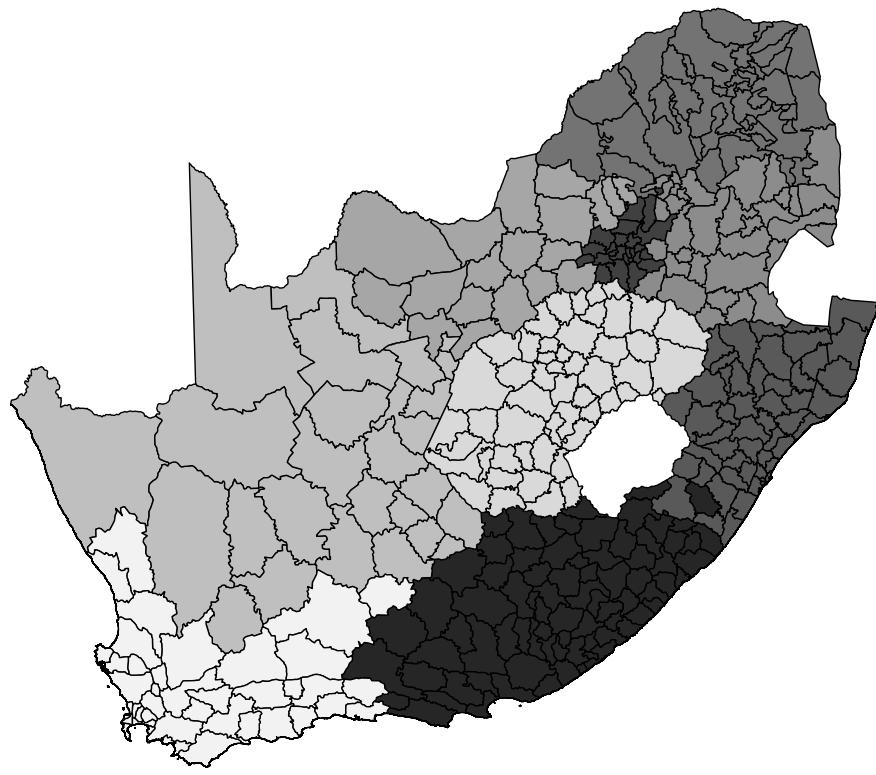


Figure 6: South Africa's magisterial districts.

73 **5 Carbon analysis**

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

- 78 • Classes 1, 3, 6, 8 corresponding to broadleaf and mixed forests were reduced to a single forest
79 class.
- 80 • Classes 4 and 5 (needleleaf forests) were dropped.
- 81 • Classes 9, 10, 17 (secondary forests, forest/cropland mosaic) were merged to a single sec-
82 ondary forest class.
- 83 • Classes 20-23 (water, snow, ice, built-up areas) were dropped.
- 84 • Class 19 (bare areas) was dropped.

85 All other classes (cropland, shrubland, spare vegetation) were retained. Each of these classes has
86 a specific carbon density according to the ecofloristic zone it is found in, of which there are 10 for
87 Africa (Zones 6-9 and 10-15; see Ruesch & Gibbs, 2008). We calculated the area of each ecofloristic
88 zone using their polygon boundary maps¹. For each of the simplified cover classes, we then calculated
89 the mean ecofloristic zone carbon density value, weighted by ecofloristic zone area. We used these
90 values to generate the different carbon maps, as described in the main text.

91 Maps of the mean residual differences between the reference map and each of the 5 derived carbon
92 maps for each test map are shown in Figure 7. Table 3 provides the bias and mean absolute errors for
93 the different carbon maps and how they vary with scale.

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

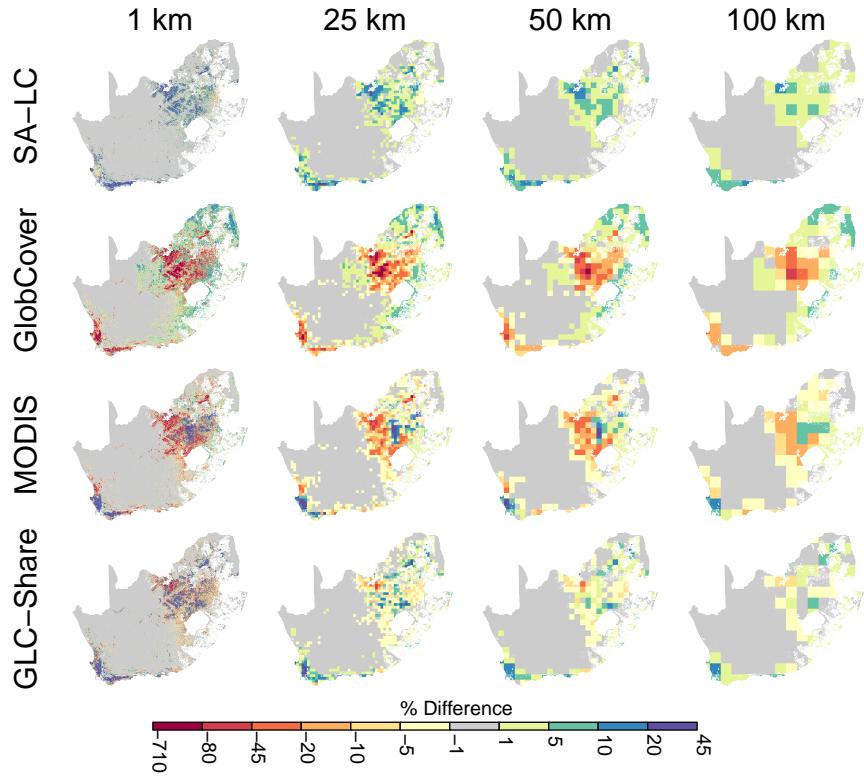


Figure 7: Spatial patterns of error (averaged across all five different possible cover types adjacent to cropland) in carbon stock estimates.

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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	Sparse	14.1	8.4	6.6	5.1	3.9	2.9

94 **6 Gridded maize yield and production**

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

- 99 1. We extracted the 2011 reference percentages within each of South Africa's 9 provinces (the
100 same units used by Ramankutty *et al.*, 2008), converted these to proportions and summed them
101 to calculate the “reported” cropland areas for each province. We did the same to calculate
102 provincial cropland areas for each test map;
- 103 2. We divided the “reported” provincial cropland area estimates by provincial areas to calculate
104 $rpcf$, the reference provincial cropland fractions, and then did the same with test map provincial
105 cropland areas to calculate $tpcf$, the test map provincial cropland fractions;
- 106 3. A province-specific correction factor $upcf$ ($rpcf/tpcf$) was then calculated and applied to each
107 testmap cropland fraction (tcf) in each pixel (x, y):

$$tcfa_{x,y} = upcf_{x,y} * tcf_{x,y} \quad (1)$$

108 Yielding $tcfa$, the calibrated test map cropland fraction.

109 The differences between the reference map cropland fraction and $tfca$ are shown in Figure 8 (where
110 they were adjusted back to percentages), while Table 4 provides the corresponding bias and accuracy
111 values (MAE).

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

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

$$fcrop_{x,y} = tfca_{x,y} \frac{cmda}{tmida} \quad (2)$$

115 Where $fcrop$ is the fraction of each cell in each district harvested for maize, $cnda$ is the district-
116 reported harvested area for maize, and $tmda$ is the total area of the magisterial district;

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

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

120 The maps of differences between reference-derived yield and production maps and those based on
121 the test maps are shown in Figures 9 and 10, where the values were normalized to the country mean
122 yield or production (calculated from the reference map). Table 5 presents the bias and accuracy values
123 for each reference-test comparison of gridded yield and production estimates, giving both the cropland
124 density-weighted variants of bias and accuracy, and for comparison the same values calculated as
125 straight averages within agricultural pixels (i.e. pixels in the reference or test map having .05%
126 cropland). In the former variant, yields show relatively much less bias than in the latter, where large
127 yield biases were driven by the discrepancies in the low cropland density regions in the center of
128 the country. In contrast, the density-weighted measures reveal large production biases, whereas the
129 unweighted agricultural region measures show no bias in production estimates. The density-weighted
130 variant shows that error scales with cropland cover, as seen in Figure 2 in the main text.

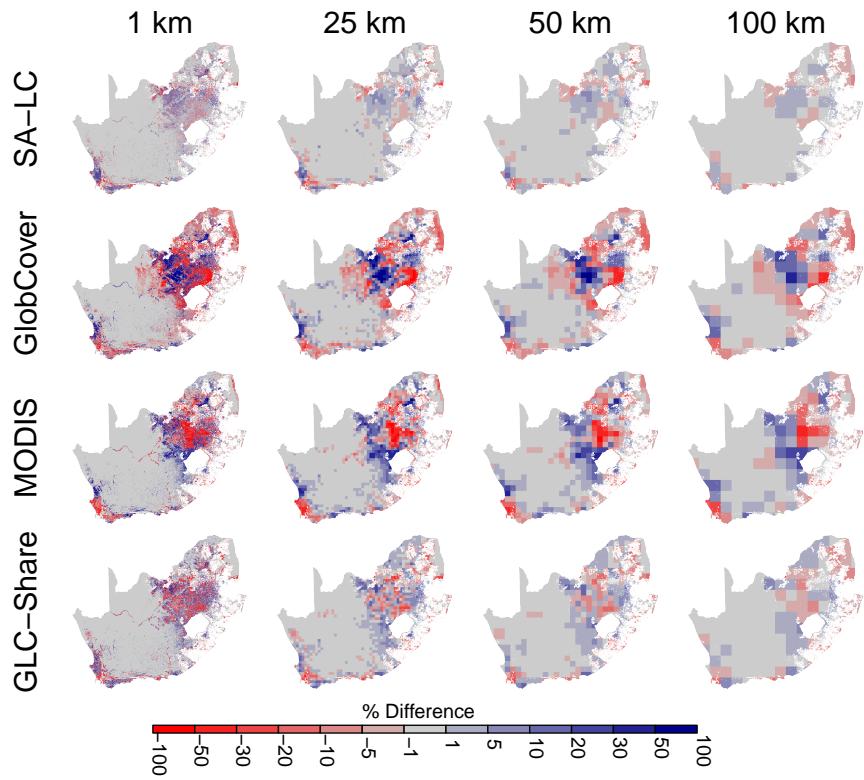


Figure 8: Errors in cropland maps adjusted using provincial cropland area statistics.

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	GLC-Share	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	GLC-Share	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

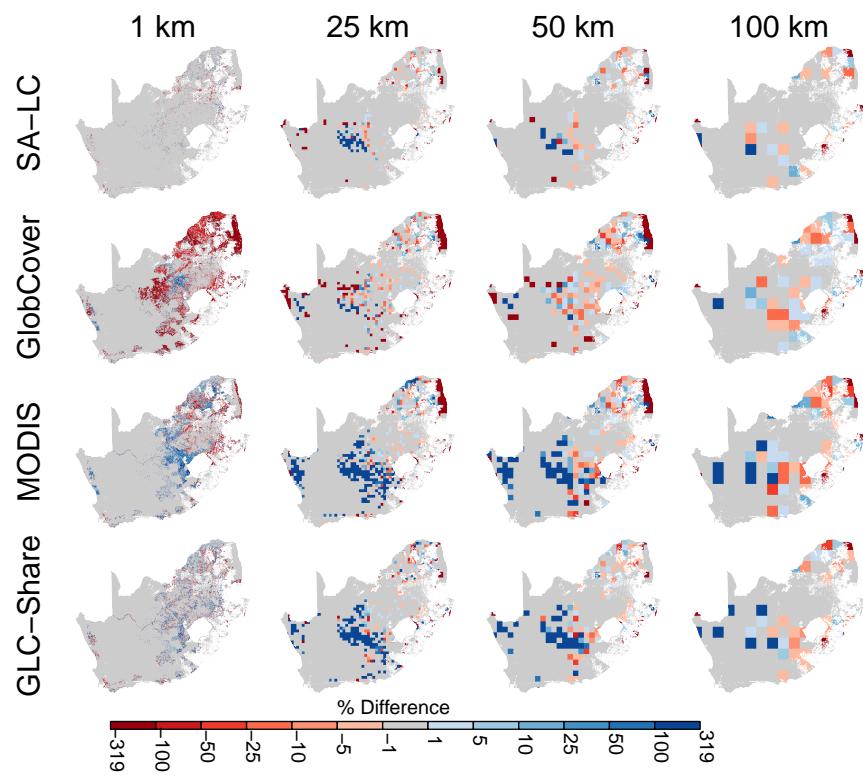


Figure 9: Errors (normalized to the reference-derived country mean) in disaggregated maize yield estimates.

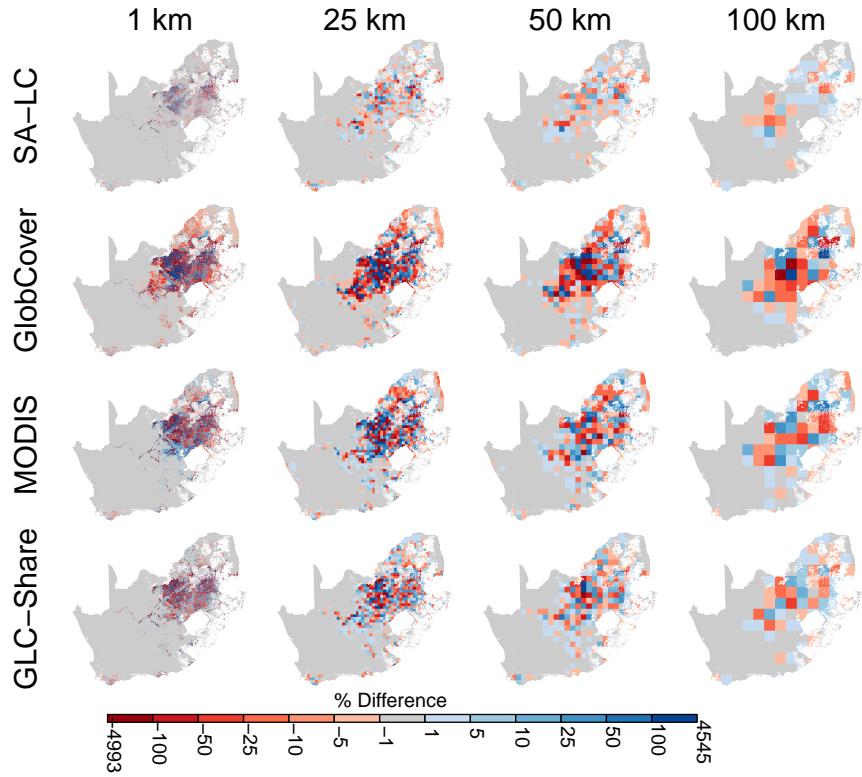


Figure 10: Errors (normalized to reference-derived country mean) production estimates calculated from disaggregated maize yield and harvested area estimates.

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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	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	GLC-Share	Production	43.7	30.2	23.5	15.3	9.3	4.0

131 7 Evapotranspiration analysis

132 A number of the variables used by the Variable Infiltration Capacity (VIC; Liang *et al.*, 1994) model
 133 are linked to a 0.25° resolution, AVHRR-derived land cover map. These include seasonal leaf area in-
 134 dex (LAI) phenologies (Fig. 11), as well as other properties such as plant rooting depth and infiltration

135 rates. Land cover and cover properties therefore impact the model’s simulation of water balance. To
136 test how land cover map error impacts VIC simulations, we created five new versions of VIC’s native
137 land cover scheme, one based on the reference cropland map, and the other four on each of the test
138 maps. For each version, we first reprojected the relevant 25 km cropland map to the geographic coor-
139 dinate system used by VIC and resampled it to 0.25° . We then adjusted VIC’s land cover scheme by
140 replacing its existing cropland percentages with those from our map, and then proportionally adjusted
141 the remaining cover types in each cell to accommodate the changed cropland amounts. Since LAI
142 is strongly linked to rainfall seasonality in South Africa, which varies between winter (May-August)
143 rainfall in the west and southwest of the country and summer (October-March) rainfall in the east
144 and northeast, we assigned LAI curves that peaked during the winter months to the cover types in the
145 western half of the country, and those peaking in the summer months to the eastern half (Fig. 11).

146 After making these adjustments, we conducted five different VIC simulations, one for each of the
147 adjusted land cover schemes. The model was run at a daily time step for 28 years, from 1981-2008,
148 from which monthly total evapotranspiration values were extracted. We calculated from these results
149 the average annual total ET, the maximum and minimum monthly ET observed throughout the entire
150 time series, and the mean ET during the month which on average had the highest ET during the time
151 series. We then found the differences between the reference and test map variants of each of these ET
152 variables (Fig. 12), and calculated their bias and MAE values. The results are similar across the four
153 variables, and we present the difference maps for the mean annual ET comparison in the main text.

154 8 Agent-based model assessment

155 8.1 Survey statistics

156 The frequency distributions of household cropland holdings were derived from survey data obtained
157 from Zambia (no equivalent data were available for South Africa). We used these statistics to calculate
158 the “true” number of households per district, as well as household cropland area distributions. The
159 average field size in these studies was 1 ha, which determined the resolution of the base cropland maps
160 (100 m).

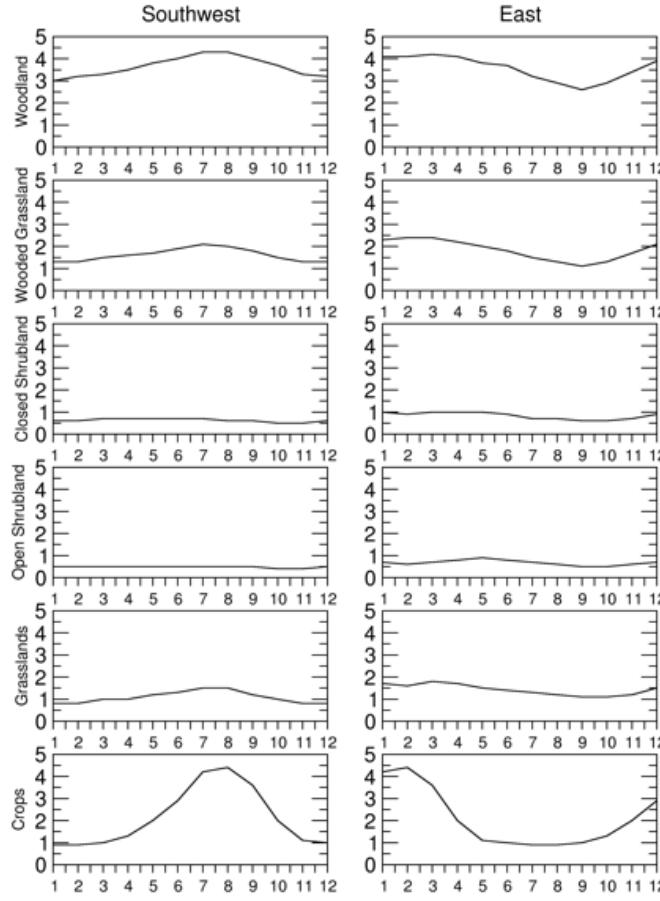


Figure 11: Seasonal LAI curves for different cover types in VIC's land cover scheme, showing the phenologies used for the winter-rainfall portion of the country in the west and southwest (left column), and for the summer rainfall region in the country's easter to northeast (right column).

161 8.2 Maize production

162 The agent-based model simulates household food production using a look-up table that links maize
 163 yields to several different variables: planting date; cultivar (open-pollinated or hybrid); soil proper-
 164 ties, and weather. The look-up table itself was based on a series of yield simulations conducted by the
 165 DSSAT cropping system model run over 31 years (1979-2010) for a location in the southern Province
 166 of Zambia, corresponding to the region where household survey data were collected. The model was
 167 run for three dominant soil series², using two open-pollinated and two hybrid cultivars (each pair rep-
 168 resenting a short- and medium-length growing season variety) whose coefficients were obtained from
 169 the inputs files for the PSIMs modeling platform (Elliott *et al.*, 2014). Each cultivar was simulated un-

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

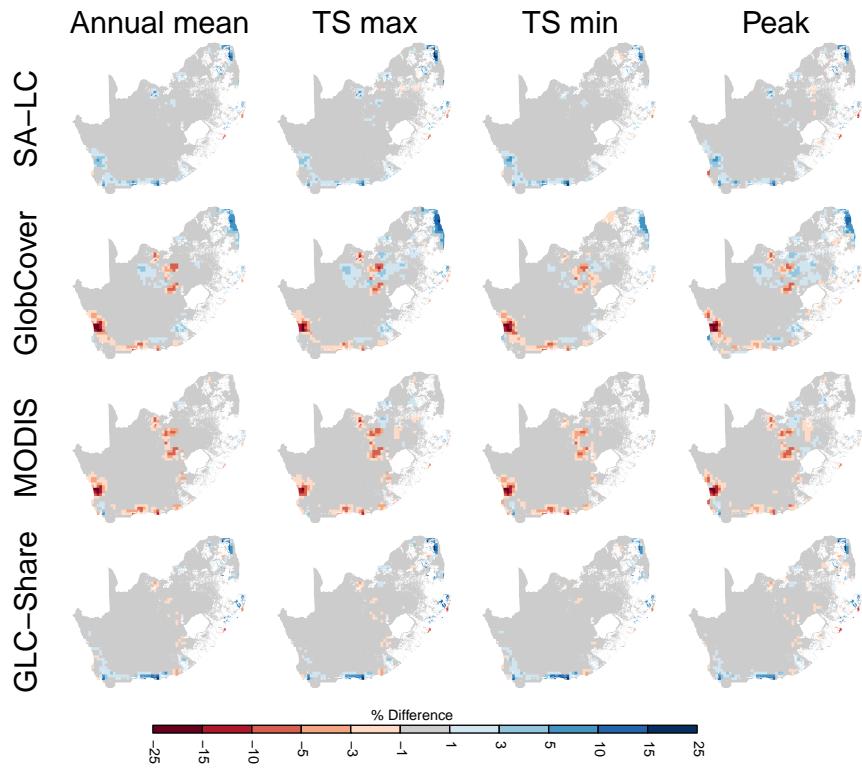


Figure 12: Maps of error between reference- and test map-derived ET estimates simulated by the VIC model. Four different variants of ET were assessed: annual mean (left column, also presented in Fig. 2 of the main text), the maximum (second column, TS max) and minimum (third column, TS min) ET values observed during the 30 year times, and mean ET during the month in which ET peaks (right column, Peak).

170 der planting dates ranging from October 15th until January 30th, varying by 15 day increments, with
 171 row spacing fixed at 90 cm and planting density at 3.7 plants m², and 5 kg ha⁻¹ applied at planting.
 172 Models were run using weather data extracted from a bias-corrected version of the Princeton Global
 173 Meteorological forcing dataset (Chaney *et al.*, 2014; Estes *et al.*, 2014; Sheffield *et al.*, 2006).

174 8.3 District selection

175 For our analysis of how land cover map error impacts agent-based model results, we selected four
 176 districts in South Africa which had similar climatic characteristics (~800 mm annual rainfall) to the
 177 region in Zambia where the crop model simulations and household survey data were collected. These
 178 were four contiguous districts along the western border of Lesotho, Clocolan, Ficksburg, Fouriesburg,
 179 and Marquard (Fig. 13, top left), which had between 29-45% of their areas covered by cropland,
 180 according to the 2011 reference map (Fig. 13, top right).

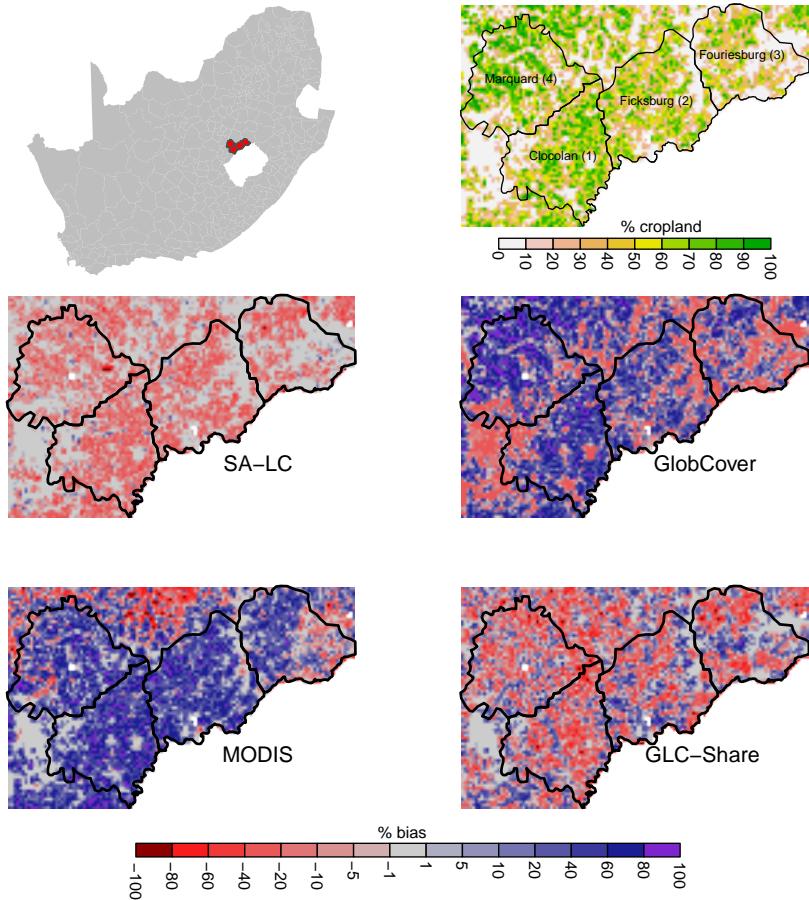


Figure 13: The location of the four selected magisterial districts (top left) used in evaluating agent allocation bias, the reference levels of cropland cover within those districts (top right), and the difference in cropland percentage between the reference and each of the four cropland maps (lower four panels).

181 8.4 Model description

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

185 8.4.1 Purpose

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

189 **8.4.2 State variables and scales**

190 The primary analytical components in the model are actors (households) and cells (land). We designed
 191 a greedy algorithm that allocates land to households based on a heuristic rule that households tend to
 192 live near each other to form communities like villages. The simulation runs on four districts in South
 193 Africa. For each district, a set of synthetic household agents is created from the land cover data of
 194 that district and the farmer register in Monze District of Zambia. We extracted the frequencies of
 195 cultivated area from the farmer register using bins (in hectare) of 0-1, 1-2, 2-3, 3-4, 4-5, and 5-10.
 196 Because each of the land cells is 1 hectare (ha), we considered the households in the same bin having
 197 the same integer value of cultivated area. For example, households in the bin (0, 1] all have 1 ha
 198 of cropland, and households in the bin (5, 10] all have 8 ha of cropland. From this distribution of
 199 cultivated area (Figure 14), we calculated its mean value as:

Mean value of cultivated area

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

200 We then used the mean value of cultivated area to divide the total area of cropland in each district
 201 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
cloolan	54413.3	23810	23810	54410
ficksburg	51363.93	22475	22475	51367
fouriesburg	29402.99	12866	12865	29400
marquard	55633.12	24343	24341	55615

202 Four key household initial characteristics define agents: (1) cultivated area (used primarily for
 203 maize production), (2) day of planting maize, (3) soil properties of land cells, and (4) seed type of

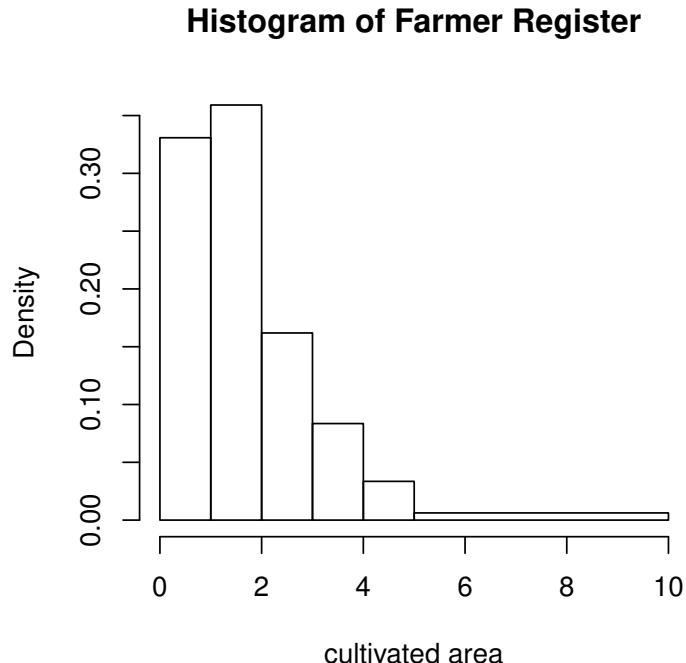


Figure 14: Histogram of cultivated area in the farmer register.

204 maize. To determine (1), the household agents are assigned 1 ha, 2 ha, 3 ha, 4 ha, 5ha, and 8 ha,
 205 according to the frequencies we extracted from the farmer register previously. The results in Table
 206 1 show a good matching between the total area of cropland from the land cover data and the total
 207 cultivated area of household agents. As for (2), (3), and (4), we create a distribution from our house-
 208 hold survey that is shown in Figure 15. In addition to the initial characteristics, each household has
 209 two internal variables: household ID and harvest amount. Each land cell has the following variables:
 210 household ID (if assigned), land cover type (cropland or not), and revenue (if harvested). The harvest
 211 amount of a household is the sum of revenue of cropland cells managed by that household.

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

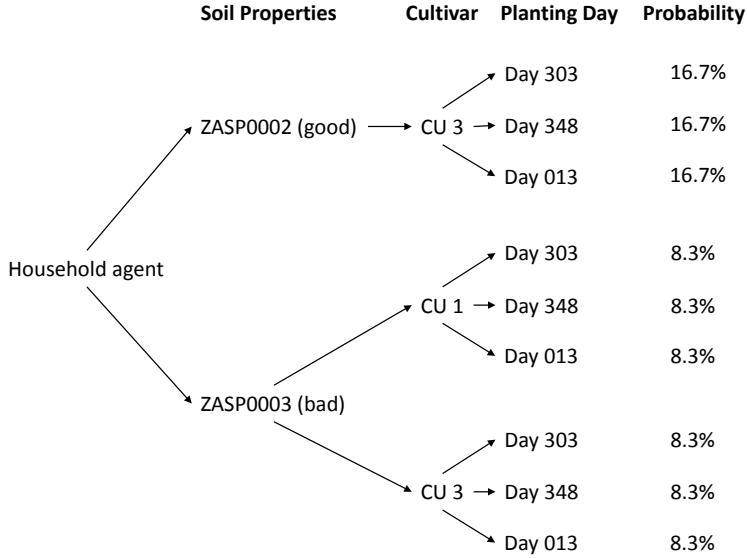


Figure 15: Distribution of soil properties, maize type (cultivar), and planting day. The soil properties, cultivar, and planting day can be used to find the yield in the look-up table created from the DSSAT cropping system model. CU 1 indicates a local maize type and CU 3 indicates a hybrid one. Day 303 means the 303rd day of the year.

216 8.4.3 Process overview and scheduling

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

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

232 **8.5 Design concepts**

233 **8.5.1 Emergence**

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

236 **8.5.2 Adaptation**

237 **8.5.3 Fitness/objectives**

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

247 **8.5.4 Prediction**

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

251 **8.5.5 Interaction**

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

254 **8.5.6 Sensing**

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

257 **8.5.7 Model calibration**

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

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

```
1: procedure allocate_farmland( $H, P$ ) ▷  $H$ : household,  $P$ : patch
2:    $A \leftarrow$  the area of farmland needed by  $H$ 
3:   if  $A >$  the area of  $P$  then ▷  $P$  is fully occupied by  $H$ 
4:      $occupiedRatio(P) \leftarrow 1$ 
5:   else
6:      $occupiedRatio(P) \leftarrow (A - 1)$  ▷  $P$  is partially occupied by  $H$ 
7:   end if
8:    $N \leftarrow$  neighbor farmland (in radius  $r$ ) of  $P$  ▷  $r$  is a global parameter of allocation radius
9:    $status(N) \leftarrow$  tentative seed patches
10:  end procedure

11: procedure allocate_many_farmland( $H, P$ ) ▷  $H$ : household,  $P$ : patch
12:   Invoke allocate_farmland( $H, P$ )
13:   repeat
14:      $searchRadius \leftarrow 700m$  ▷ starting from a threshold value
15:      $UP \leftarrow$  a randomly selected unoccupied farmland within  $searchRadius$  of  $P$ 
16:     Invoke allocate_farmland( $H, UP$ )
17:     if  $A >$  the area of  $P$  then
18:        $searchRadius \leftarrow (searchRadius + 100m)$ 
19:     end if
20:   until  $H$  is assigned enough farmland  $\vee$   $searchRadius == s$  ▷  $s$  is a global parameter of the maximum search radius
21:  end procedure

22: procedure allocate_hh
23:    $i \leftarrow 1$  ▷ the id of current household to be allocated
24:   repeat
25:      $SH \leftarrow$  the  $i$ th household
26:      $status(SH) \leftarrow$  seed household
27:      $SP \leftarrow$  a randomly selected patch
28:     Invoke allocate_many_farmland( $SH, SP$ )
29:      $i \leftarrow (i + 1)$ 
30:   until  $i == numSeed \vee$  there is no unoccupied land ▷  $numSeed$  is a global parameter of the total number of seed households created during initialization
31:   repeat
32:      $SH \leftarrow$  the  $i$ th household
33:      $TSP \leftarrow$  a randomly selected patch so that  $status(TSP) ==$  tentative seed patch  $\wedge$   $occupiedRatio(TSP) == 0$ 
34:     Invoke allocate_many_farmland( $SH, TSP$ )
35:      $i \leftarrow (i + 1)$ 
36:   until  $i == numHHS \vee$  there is no unoccupied land ▷  $numHHS$  is the total number of households
37:  end procedure
```

266 References

- 267 Chaney NW, Sheffield J, Villarini G, Wood EF (2014) Spatial Analysis of Trends in Climatic Extremes with a High
268 Resolution Gridded Daily Meteorological Data Set over Sub-Saharan Africa. *Journal of Climate*.
- 269 Elliott J, Kelly D, Chryssanthacopoulos J, *et al.* (2014) The parallel system for integrating impact models and sectors
270 (pSIMS). *Environmental Modelling & Software*, **62**, 509–516.
- 271 Estes LD, Chaney NW, Herrera-Estrada J, Sheffield J, Taylor KK, Wood EF (2014) Changing water availability during
272 the African maize-growing season, 1979–2010. *Environmental Research Letters*, **9**, 075005.
- 273 Estes LD, McRitchie D, Choi J, *et al.* (2016) A platform for crowdsourcing the creation of representative, accurate land-
274 cover maps. *Environmental Modelling & Software*, **80**, 41–53.
- 275 Fritz S, See L, McCallum I, *et al.* (2015) Mapping global cropland and field size. *Global Change Biology*, **21**, 1980–1992.
- 276 Grimm V, Berger U, Bastiansen F, *et al.* (2006) A standard protocol for describing individual-based and agent-based
277 models. *Ecological Modelling*, **198**, 115–126.
- 278 Liang X, Lettenmaier DP, Wood EF, Burges SJ (1994) A simple hydrologically based model of land surface water and
279 energy fluxes for general circulation models. *Journal of Geophysical Research*, **99**, 14415.
- 280 Monfreda C, Ramankutty N, Foley JA (2008) Farming the planet: 2. Geographic distribution of crop areas, yields, physi-
281 ological types, and net primary production in the year 2000. *Global Biogeochemical Cycles*, **22**, GB1022.
- 282 Olofsson P, Foody GM, Herold M, Stehman SV, Woodcock CE, Wulder MA (2014) Good practices for estimating area
283 and assessing accuracy of land change. *Remote Sensing of Environment*, **148**, 42–57.
- 284 Polhill JG, Parker DC, Brown DG, Grimm V (2008) Using the ODD protocol for comparing three agent-based social
285 simulation models of land use change. *Journal of Artificial Societies and Social Simulation*, **11**, 1–25.
- 286 Ramankutty N, Evan AT, Monfreda C, Foley JA (2008) Farming the planet: 1. Geographic distribution of global agricul-
287 tural lands in the year 2000. *Global Biogeochemical Cycles*, **22**, 19 PP.
- 288 Ruesch A, Gibbs HK (2008) New IPCC Tier-1 global biomass carbon map for the year 2000. *Carbon Dioxide Information
289 Analysis Center (CDIAC), Oak Ridge National Laboratory, Oak Ridge, Tennessee. Available online at: [http://cdiac.
ornl.gov/epubs/ndp/global_carbon/carbon_documentation.html](http://cdiac.
290 ornl.gov/epubs/ndp/global_carbon/carbon_documentation.html)*.
- 291 Sheffield J, Goteti G, Wood EF (2006) Development of a 50-Year High-Resolution Global Dataset of Meteorological
292 Forcings for Land Surface Modeling. *Journal of Climate*, **19**, 3088–3111.
- 293 Wood SN (2001) mgcv: GAMs and Generalized Ridge Regression for R. *R News*, **1**, 20–25.