Avocado-Driven Deforestation in Michoacán, México
Supplementary Information Document

Eugenio Y. Arima
Audrey Denvir
Kenneth R. Young
Antonio González-Rodríguez
and Felipe García-Oliva

23 February 2022

Note

The following material constitutes the Supplementary Information (SI) Document pertaining to the Environmental Research Letters publication titled *Avocado-Driven Deforestation in Michoacán, Mexico*. This supporting document, once accessible for download on the ERL website, is currently unavailable, for unknown reasons.

SI 1 - Data Sources and Descriptive Statistics

Data on elevation (DEM), soil, land use, vegetation type, roads, and settlements were collected from Mexico's National Institute of Statistics and Geography (INEGI, in Spanish). Slope in percent was calculated from the DEM in ArcGIS Pro using the geodesic method. Soil classification at the group level (e.g. Andosol) were also included as categorical variables in the model.

INEGI land use and vegetation maps from 1992 and 2017 were used to isolate past and present spatial distribution of land use categories of particular interest. 2017 is used to represent present conditions since it is the most recent year from which INEGI land use and vegetation maps are available. Perennial agriculture layers from 2017 were cross-referenced with a visual assessment of high-resolution satellite imagery from Google Earth in order to create an avocado agriculture layer. Any avocado agriculture that was present in 2017, but not in the 1992 more general "all agriculture" layer, was deemed avocado expansion in that time period and used as the model dependent variable. Other land types from the 1992 maps were isolated and used to create binary (presence/absence) independent variable layers, e.g., oak forest, pine forest, etc. Similarly, vegetation type variables were created for 2017, not as model variables but for analysis of how projected land change would affect present vegetation coverage.

GIS vector files of roads and settlements (both rural and urban) were used to calculate raster files of Euclidean distance from these elements. The Phytosanitary Directory of Packing Houses from Mexico's National Service of Agrifood Health, Safety and Quality (SENASICA, in Spanish) was used to create a shapefile of all registered avocado packing house locations, and distance from

packing houses was calculated in ArcGIS Pro. Ultimately five distance variables were utilized: distance to roads, distance to avocado packing house, distance to cities, distance to all settlement (urban and rural), and distance from existing agriculture (calculated from 1992 agriculture land use, as designated by the INEGI maps described above).

Protected area information was obtained from the Global Forest Watch (Global Forest Watch 2019), and climatic data were obtained from the Centro de Ciencias de la Atmósfera at the Universidad Nacional Autónoma de México (Centro de Ciencias de la Atmósfera 2011). The climatic data included in the model were mean annual temperature and annual precipitation. These values are averaged by meteorological stations over the period of 1902-2011 from Mexico's National Meteorological Service.

Table S1. Variables' descriptive statistics and sources

Variable	Min	Max	Mean	SD	p50	Unit	Description
avoc[1]	0	1	0.0395	0.1948	0	binary	avocado agriculture in 2017
d2rds[2]	0	6,300	961.1251	876.8185	707.1068	meters	distance to roads
d2pack[3]	0	68,000	25,000	15,000.0	24,000	meters	distance to packing plant
d2cities[4]	0	38,000	9,400	6,200.0	7,800.0	meters	distance to city
d2local[4]	0	6,900	1,500	975.7836	1,200.0	meters	distance to local village
d2agri[5]	0	18,000	1,900	2,600.0	921.955	meters	distance to agricultural land use
elev[6]	360	3,900	1,900	603.3867	2,000.0	meters	MASL
slope[7]	0	306.5683	18.367	13.886	14.831	percent	terrain slope
precip[8]	651.3472	1,700	1,100	163.7795	1,100	mm	annual precipitation
temp[8]	7.1211	28.4926	17.9184	4.0455	16.9461	Celsius	average temperature
matto[9]	0	1	0.0679	0.2516	0	binary	matorral
oak[9]	0	1	0.1342	0.3409	0	binary	oak forest[9]
oakpine[9]	0	1	0.0685	0.2527	0	binary	oak-pine forest
oyamel[9]	0	1	0.0081	0.0899	0	binary	oyamel fir forest
pine[9]	0	1	0.0899	0.2861	0	binary	pine forest
pineoak[9]	0	1	0.3471	0.476	0	binary	pine-oak forest
selva[9]	0	1	0.1472	0.3543	0	binary	tropical forest
soilan[10]	0	1	0.3103	0.4626	0	binary	andosol
soilcm[10]	0	1	0.0252	0.1567	0	binary	cambisol
soillp[10]	0	1	0.1307	0.3371	0	binary	leptosol
soillv[10]	0	1	0.2961	0.4565	0	binary	luvisol
soilph[10]	0	1	0.062	0.2411	0	binary	phaeozem
soilrg[10]	0	1	0.0949	0.293	0	binary	regosol
soilvr[10]	0	1	0.0684	0.2524	0	binary	vertisol
protect[11]	0	1	0.0355	0.185	0	binary	protected areas
ejido[12]	0	1	0.4118	0.4922	0	binary	Communally managed land (ejidos and comunidades)

[1] File created from INEGI 2017 land use and vegetation data (described below) by visually assessing land identified as perennial agriculture and semi-perennial agriculture within study area.
[2] Roads vector file(s) downloaded on 12/8/20. Distance from roads layer calculated from roads layer.
[3] Packing plant location data from SENASICA
[4] Localities vector file (points) downloaded from INEGI on 12/8/20. Distance from Settlement layer calculated from localities layer (once for urban localities and once for all localities).
[5] Agriculture land use spatial data isolated from INEGI 1992 Conjunto de datos vectoriales de la carta de Uso del suelo y vegetación. Escala 1:250 000. Serie I. Continuo Nacional.
[6] Downloaded from INEGI on 12/17/2020.
[7] Calculate in ArcGIS Pro from the elevation layer with "Slope" function (Geodesic method).
[8] Downloaded from Centro de Ciencias de la Atmósfera, UNAM UNAM.
[9] Downloaded from INEGI, 2017: Conjunto de datos vectoriales de la carta de Uso del suelo y vegetación. Escala 1:250 000. Serie VI. Conjunto Nacional.
[10] Soil vector file downloaded from INEGI 12/8/2020; data rasterized and individual soil types isolated in individual raster file.
[11] Downloaded from Global Forest Watch of 8/12/20.
[12] Perimetrales de los núcleos agrarios certificados, downloaded from Registro Agrario Nacional on 10/24/21.

SI 2 - Spatial Allocation Model

In this model, avocado deforestation (y) in cell i belonging to a region j is represented as a binary outcome (1= deforested, 0=otherwise):

$$y_{ij} = \begin{cases} 0, & \text{if } y_{ij} < 0 \\ 1, & \text{if } y_{ij} > 0 \end{cases}$$

where, $y_{ij} = \mathbf{x}_{ij}\beta + \theta_j + \epsilon_{ij}$ and $\theta = \rho \sum_{k=1}^{m} w_{jk}\rho_k + \mu_j$. The vector $\mathbf{x_{ij}}$ contains the explanatory variables, β is the vector of parameters to be estimated; θ_j is the spatial effect on region j stemming from neighboring areas, k; w is a spatial weights matrix, and ρ is the spatial autocorrelation parameter also to be estimated. The idiosyncratic error e_{ij} is assumed to be normally distributed, conditional on θ; μ_j is also considered to be normally distributed, yielding a probit specification (for a formal description of the model, see Supporting Information document in Arima 2016; Smith and LeSage 2004).

The dependent variable is the deforestation attributed to avocado between 1992 and 2017. Avocado agriculture in 2017 was identified by cross-referencing areas categorized as perennial or semi-perennial agricultural land use by Mexico's National Institute of Statistics and Geography (INEGI) with a visual assessment of high-resolution Google Earth imagery of the study area. Areas determined to be avocado agriculture in 2017 that were located outside of the INEGI agricultural land use categorization in 1992 were determined to be new avocado expansion in that period. Water bodies, urban areas, other agriculture, and areas devoid of vegetation (e.g. mines) were masked out of the analysis. The model therefore estimates the probability of a given remaining cell being deforested for avocado in the future given prior avocado and other agricultural land uses.

The vector $\mathbf{x}$ of explanatory variables includes the relevant proximate drivers of deforestation frequently cited in the land change science literature (see section Data above). Table S1 describes each variable with respective source and descriptive statistics.

All digital GIS files were either in or converted to raster format, projected to Albers conic equal area projection, and resampled to 100 m cell resolution using the nearest neighbor algorithm. We masked out cells related to other land uses (e.g. crops, pastures, urban, and water bodies), yielding a total of 1,492,653 cells, which constitute the number of observations used in the regression model. Of those, 58,956 were of avocado plantations and the remaining cells are non-deforested. This mesoscale cell resolution of 100 m was chosen as a good compromise between the various scales of the data available. For example, the remote sensing work that produced the land use/cover map was based on Landsat imagery (30 m resolution) and on visual interpretation on Google Earth (< 5m), whereas other digital maps were at 1:250,000 to 1:10,000 scales. This cell size also conveniently represents one hectare on the ground.

The regional spatial autoregressive process was implemented following Arima (2016). The contiguous cells were labeled and assigned to regions formed by 10x10 neighboring cells, creating 21,860 regions.

This Bayesian statistical method uses Markov Chain Monte Carlo (MCMC) algorithms and the results presented here are based on the average of 500 valid draws after the first 500, used for convergence during the burn-in phase of the MCMC procedure, were discarded (Smith and LeSage 2004). This method is implemented in the function *semip_g* from LeSage's spatial econometric Matlab toolbox (LeSage 2015). The output of the model is a raster containing the predicted probability (ranging from 0-1) of each cell being converted to avocado.

SI 3 - Regression Results

Table S2. Bayesian probit regression model with regional spatial effects results.

Variable	Coefficient	Std Deviation	p-level
const	-4.376	0.707	0.000
d2rds	-5.106	0.396	0.000
d2pack	-0.345	0.023	0.000
d2cities	0.614	0.049	0.000
d2local	-3.666	0.683	0.000
d2agri92	-2.181	0.311	0.000
elev	2.851	0.272	0.000
elev2	-0.905	0.068	0.000
slope	-18.168	0.825	0.000
temp	-2.434	0.460	0.000
temp2	0.562	0.106	0.000
precip	2.925	0.450	0.000
precip2	-0.689	0.189	0.000
matto	-0.024	0.047	0.190
oak	-0.312	0.040	0.000
oakpine	-0.501	0.058	0.000
oyamel	0.151	0.138	0.202
pine	-0.279	0.037	0.000
pineoak	-0.414	0.028	0.000
selva	-0.352	0.042	0.000
soilan	1.308	0.109	0.000
soilcm	0.899	0.184	0.000
soillp	0.588	0.086	0.000
soillv	0.454	0.106	0.000
soilph	0.776	0.107	0.000
soilrg	0.594	0.111	0.000
soilvr	0.299	0.080	0.000
protect	-0.018	0.045	0.390
ejidos	-0.355	0.015	0.000
ρ	0.799	0.025	0.000

N = 1,492,653; Coefficients and standard deviations are the average and standard deviation respectively of 500 valid draws of a Monte Carlo Markov Chain process.

SI 4 - Monte Carlo Simulation Results

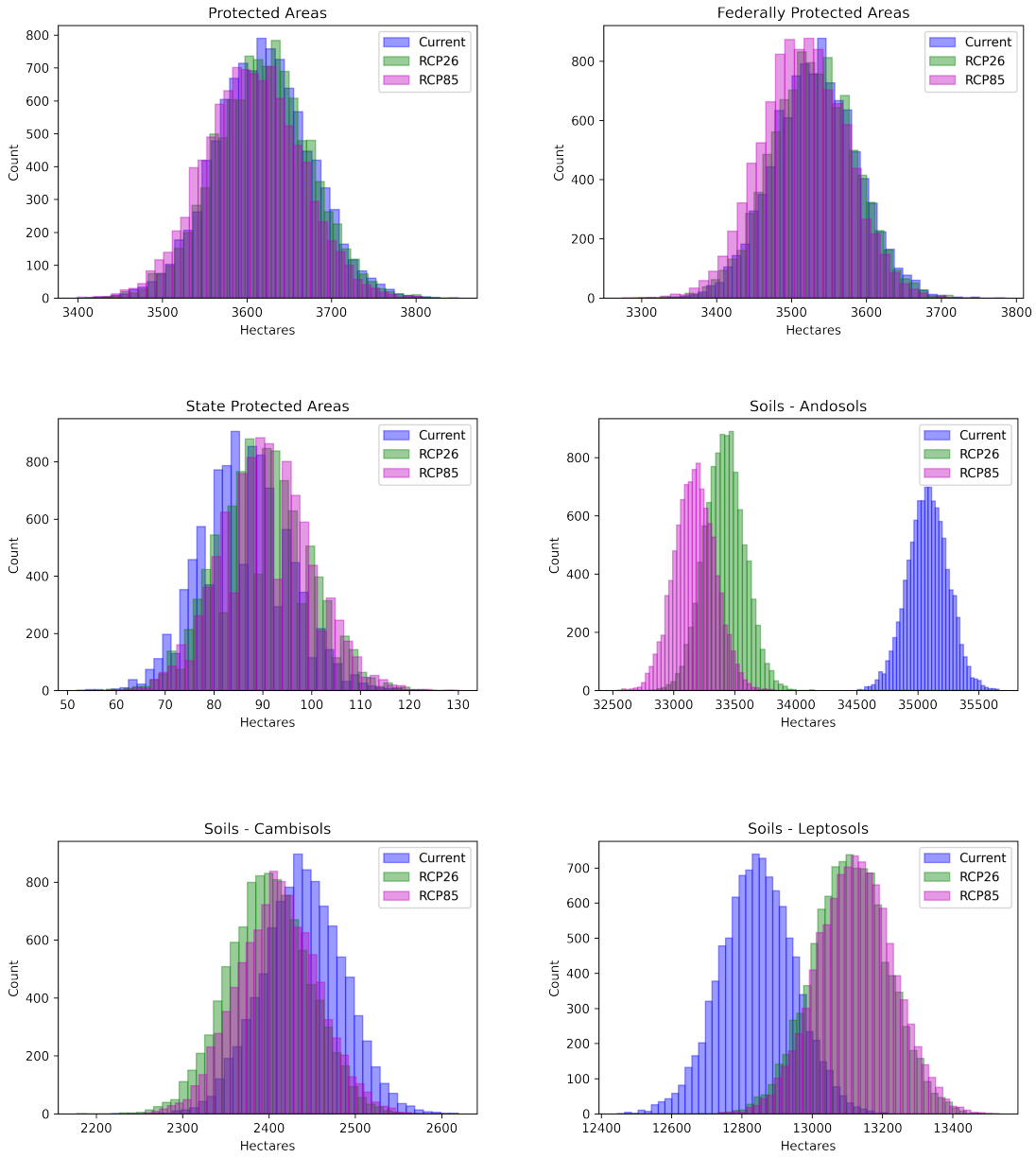

Figure S3. MC Simulations of Protected Area Encroachment and Soil Use for Avocado Expansion under Current Climate, RCP 2.6, and RCP 8.5 scenarios (Cont.)

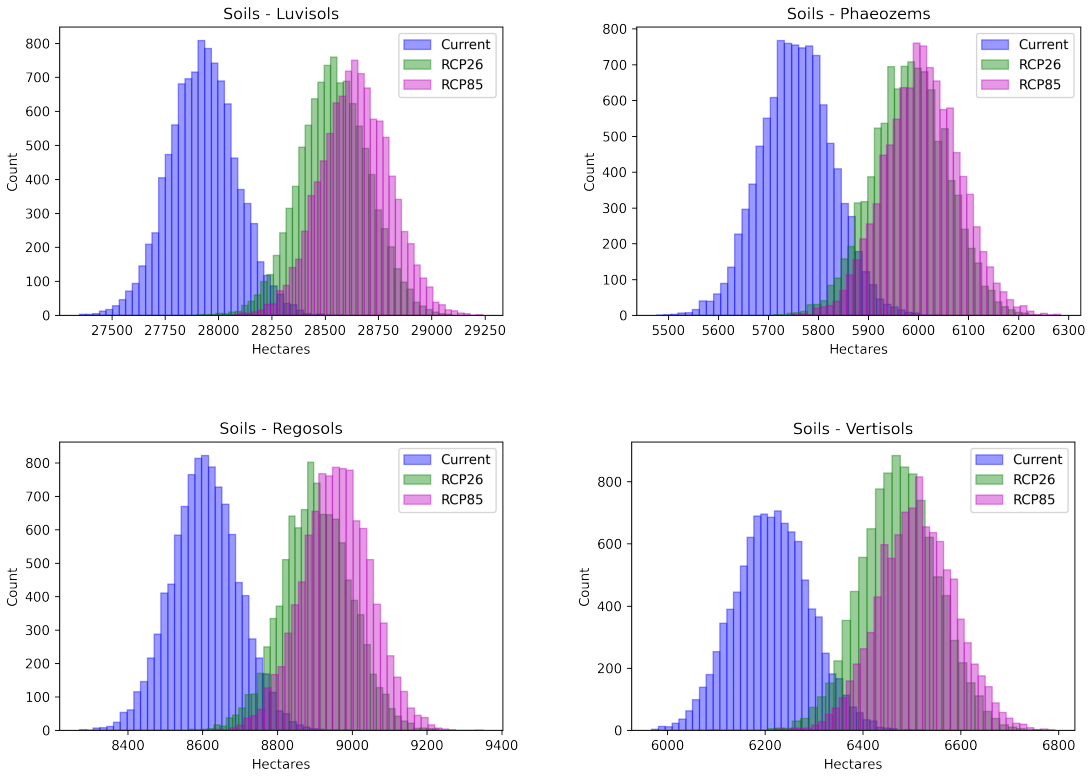

Figure S3. MC Simulations of Protected Area Encroachment and Soil Use for Avocado Expansion under Current Climate, RCP 2.6, and RCP 8.5 scenarios

Table S3a. Mean forest loss, by type, across 10,000 simulations under current climate.

Land Type	Area lost in km^2	Percent of land type lost from study area	Percent of land type lost from Michoacán
Oyamel fir forest	12.7	6.0	5.8
Cypress-mixed Forest	0.5	6.0	4.0
Mesophilic Montane Forest	12.1	9.6	9.0
Lowland Tropical Forest	197.6	6.2	1.1
Pine Forest	149.7	7.6	3.0
Pine-Oak Forest	273.9	7.7	4.5
Oak-Pine Forest	69.5	6.6	4.5
Oak Forest	116.9	6.1	3.2

The residual area lost is in the 'Other' category corresponding to the omitted variable in the regression.

Table S3b. Mean forest loss, by type, across 10,000 simulations under RCP 2.6.

Land Type	Area lost in km^2	Percent of land type lost from study area	Percent of land type lost from Michoacán
Oyamel fir forest	13.2	6.2	6.0
Cypress-mixed Forest	0.6	6.2	4.2
Mesophilic Montane Forest	11.5	9.1	8.5
Lowland Tropical Forest	206.7	6.4	1.1
Pine Forest	144.1	7.3	2.9
Pine-Oak Forest	265.1	7.4	4.4
Oak-Pine Forest	70.3	6.6	4.6
Oak Forest	121.8	6.3	3.3

The residual area lost is in the 'Other' category corresponding to the omitted variable in the regression.

Table S3c. Mean forest loss, by type, across 10,000 simulations under RCP 8.5.

Land Type	Area lost in km^2	Percent of land type lost from study area	Percent of land type lost from Michoacán
Oyamel fir forest	13.3	6.3	6.1
Cypress-mixed Forest	0.6	6.3	4.2
Mesophilic Montane Forest	11.3	9.0	8.4
Lowland Tropical Forest	207.7	6.5	1.1
Pine Forest	143.5	7.3	2.9
Pine-Oak Forest	264.1	7.4	4.4
Oak-Pine Forest	70.5	6.7	4.6
Oak Forest	122.5	6.4	3.4

The residual area lost is in the 'Other' category corresponding to the omitted variable in the regression.

SI 5 - Spatial Allocation Model Assessment

In order to examine location agreement/disagreement patterns, cells with highest probability of being avocado plantations according to our statistical model were cross-tabulated with actual avocado cells. The cells predicted to be avocado were selected by searching for a threshold probability p^* such that the sum of the number of cells j with estimated probability $\hat{p} \geq p^*$ equals the number of cells classified as avocado plantations in our land use maps. The calculated threshold probability p^* was found to be 0.4353. In other words, the number of cells with probability greater than 0.4353 equals 58,818 cells. This probit goodness-of-fit measure is preferred when one outcome is unbalanced in the sample, which is the case of the avocado belt in Michoacán, where only 3.9% of the cells are avocado, with the remaining classified as forests (Wooldridge 2010).

The model performance is very good (Table S4). It correctly predicts 99% of the forested cells but more importantly, it correctly predicts 81% of the avocado plantation cells, despite only 3.9% of the total cells in the raster being avocado. The overall accuracy is 98.5%. The model under predicts 11,194 cells and over predicts 11,246.

Table S4. Confusion Matrix

		Actual		
		0	**1**	**Total**
Predicted	**0**	1,422,589	11,194	1,433,783
	1	11,246	47,624	58,870
	Total	1,433,835	58,818	1,492,653

Spatially, both over (in yellow) and under predictions (in green) tend to be close to the correct predictions (in red). This is likely due to the strong spatial autocorrelation effect between regions captured by the coefficient ρ (0.79) in the regression. The gridded aspect of the predictions is also a reflection of the strong spatial effect of the 10×10 cells within each region. Figure S5 highlights such area in the western part of the avocado belt in Michoacán.

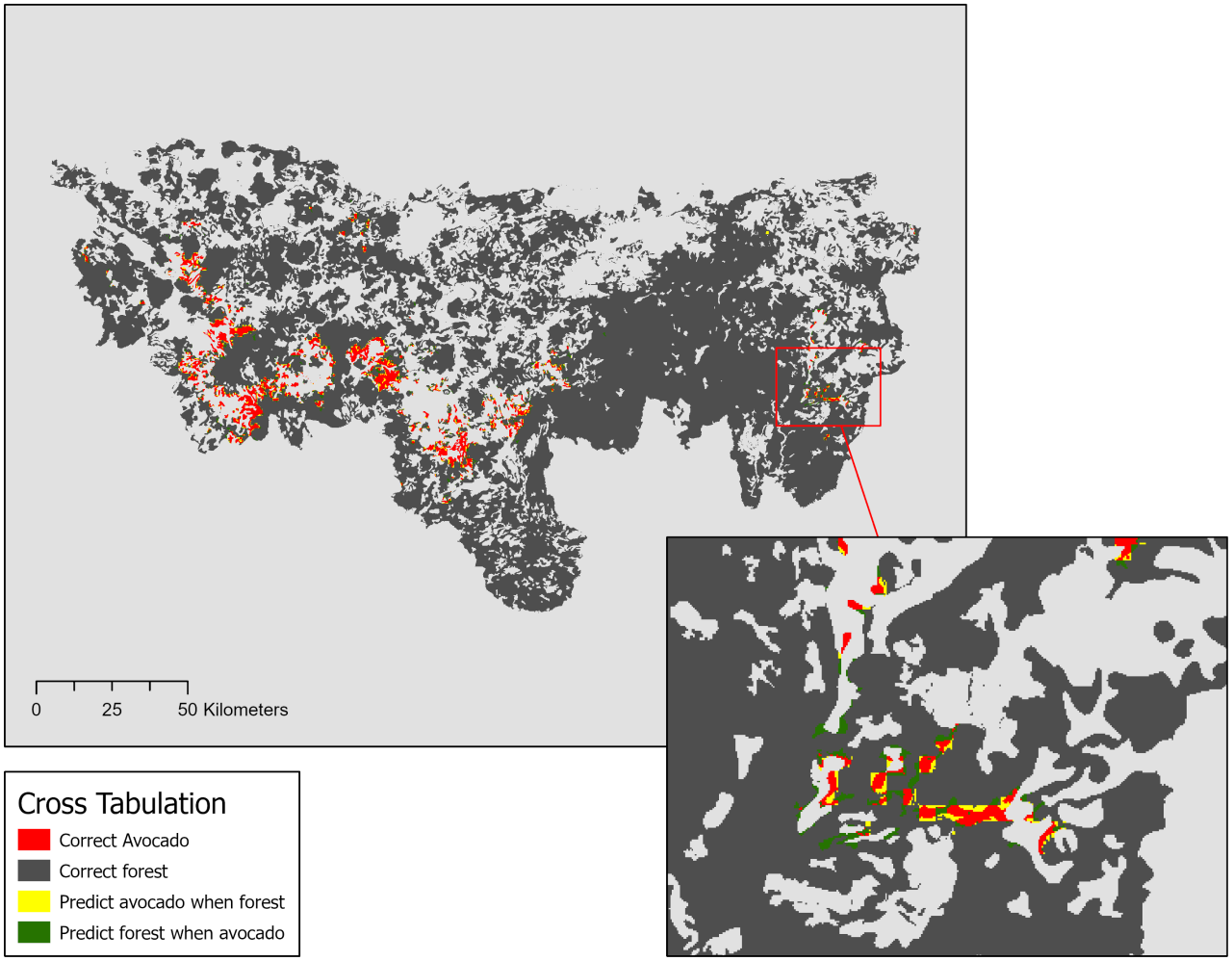

Figure S5. Cross tabulation of actual vs. predicted avocado, current climate.

SI 6 - Partial Effects Calculation

Unlike ordinary least squares where partial effects can be read directly from the estimated coefficients, probit models (and spatial probit models) require a more convoluted approach. To recall, the spatial probit model involves the definition of a latent variable y_{ij}:

$$y_{ij} = \begin{cases} 0, & \text{if } y_{ij} < 0 \\ 1, & \text{if } y_{ij} > 0 \end{cases}$$

where, $y_{ij} = \mathbf{x}_{ij}\beta + \theta_j + \epsilon_{ij}$ and $\theta = \rho \sum_{k=1}^{m} w_{jk}\rho_k + \mu_j$. The vector $\mathbf{x_{ij}}$ contains the explanatory variables, β is the vector of parameters to be estimated; θ_j is the spatial effect on region j stemming from neighboring areas, k; w is a spatial weights matrix, and ρ is the spatial autocorrelation parameter also to be estimated. The idiosyncratic error e_{ij} is assumed to be normally distributed, conditional on θ; μ_j is also considered to be normally distributed. Thus, the probability of observing avocado is given by $P(y = 1|\mathbf{x}_{ij}\beta, \theta) = P(y^* > 0|\mathbf{x}_{ij}\beta) = P(\epsilon > -\mathbf{x}\beta - \theta|\mathbf{x}\beta) = \Phi(\mathbf{x}\beta + \theta)$, where $\Phi(.)$ is the cumulative normal density function, yielding the spatial probit model.

The average partial effects (APEs) are calculated as follows. For a continuous variable c in level and quadratic form the APE_c is:

$$APE_c = N^{-1} \sum_{i=1}^{n} (\hat{\beta}_c + 2\hat{\beta}_{c^2} x_{ic}) \phi(x_i\hat{\beta} + \hat{\theta_j})$$

where ϕ is the standard normal density function, $\hat{\beta}_c$ is the estimated coefficient of the variable in level form, and $\hat{\beta}_{c^2}$ the estimated coefficient of the variable in quadratic form. APEs for a variable without its quadratic form are obtained by setting $\hat{\beta}_{c^2} = 0$. This APE formulation is a direct application of the chain rule of differentiation, observing that θ_j is not a function of x and that the standard normal is the derivative of the cumulative normal by definition.

For a binary variable b, the APE is defined as $N^{-1} \sum_{i=1}^{n} [\Phi(\mathbf{x_{ib^-}}\hat{\beta}_{\mathbf{ib^-}} + \hat{\beta}_b + \hat{\theta}_j] - \Phi(\mathbf{x_{ib^-}}\hat{\beta}_{\mathbf{ib^-}} + \hat{\theta}_j)]$, where the index b^- indicates that the variable b is not part of the vector and Φ is the cumulative normal.

Overall, the partial effects are very small in absolute terms (Table S6, second column), changing probabilities only at the third decimal place. However, it is important to highlight that the overall probability of observing avocado in the area is very small, only 0.039. If we take that as the benchmark of a naïve probability, then some changes are substantial relative to this benchmark (fourth column) as discussed in the article.

Table S6. Average marginal effects on probability resulting from change in value of variables - spatial probit model.

Variables	Partial Effect	Δvalue	% Naïve probability
d2rds	-0.00868	1 Km	-22%
d2pack	-0.00059	1 Km	-1%
d2cities	0.00104	1 Km	3%
d2local	-0.00623	1 Km	-16%
d2agri92	-0.00371	1 Km	-9%
elev	-0.00128	100 m	-3%
slope	-0.00031	1%	-1%
temp	-0.00099	1 Celsius	-3%
precip	0.00022	10 mm	1%
matto	-0.00041	Binary	-1%
oak	-0.00513	Binary	-13%
oakpine	-0.00814	Binary	-21%
oyamel	0.00260	Binary	7%
pine	-0.00464	Binary	-12%
pineoak	-0.00704	Binary	-18%
selva	-0.00580	Binary	-15%
soilan	0.02347	Binary	60%
soilcm	0.01676	Binary	43%
soillp	0.01039	Binary	26%
soillv	0.00790	Binary	20%
soilph	0.01427	Binary	36%
soilrg	0.01069	Binary	27%
soilvr	0.00525	Binary	13%
protect	-0.00030	Binary	-1%
ejido	-0.00602	Binary	-15%

SI 7 - Simulation of the impact of climate change on avocado probabilities

We used the estimated parameters from the spatial probit model (Table S2) and recalculated the avocado probability surfaces by incorporating future temperature and precipitation estimates under two climate change "representative concentration pathways" (RCP) RCP2.6 and RCP8.5 from the IPCC for the years between 2041-2060, which corresponds to our mid-century avocado projection. Let $\overline{BIO1}$ and $\overline{BIO2}$ be the vectors containing future annual average temperature and annual precipitation variables under climate change. Then, the new probability for cell i in region j under future climate change precipitation and temperature conditions ($\hat{prob}_{ij}$) is calculated as follows:

$$\hat{prob}_{ij} = \Phi(\hat{\alpha}_1\overline{BIO1} + \hat{\alpha}_2\overline{BIO2}^2 + \hat{\alpha}_3\overline{BIO12} + \hat{\alpha}_4\overline{BIO12}^2 + \mathbf{x^*}_{ij}\hat{\beta}^* + \hat{\rho}\sum_{k=1}^{m} w_{ij}\hat{\theta}_K)$$

where $\hat{\alpha}_1, \hat{\alpha}_2$, are the estimated coefficients for the effect of temperature (level and squared terms) and $\hat{\alpha}_3, \hat{\alpha}_4$, are the coefficients for precipitation (level and squared terms respectively). $mathbfx^*{}_{ij}\hat{\beta}^*$ denotes the matrix of the other explanatory variables and respective estimated coefficients but without the CURRENT temperature and precipitation variables.

The variable $\overline{BIO1}$ takes two different forms depending on the climate change scenario (RCP2.6 or RCP8.5) and is the average of the BIO1 WorldClim future climate data models from CMIP6:

$$\overline{BIO1} = n^{-1}\sum_{i=1}^{n} BIO1_{GCM_i}$$

where $BIO1_{GCM_i}$ is the BIO1 bioclimatic variable from each of the $i = 1, ..., 9$ GCM models for RCP2.6 or 8 models for the RCP8.5 available at the worldclim website.

Likewise, the future average precipitation across models for each scenario is:

$$\overline{BIO12} = n^{-1}\sum_{i=1}^{n} BIO12_{GCM_i}$$

For the RCP2.6 simulation, our models predict that an area of 334,776 ha will have higher probabilities of avocados when compared to the current scenario but the average increase in probability is very low (0.0007) with a maximum increase of 0.13. Over 1.16 million ha are projected to have lower probabilities, with an average reduction of -0.005 and by as much as -0.33.

Scenario RCP8.5 is less favorable for avocado expansion; only 238 k ha are predicted to have higher probabilities (average increase of 0.0005 with max increase of 0.09), whereas 1.25 million ha are projected to lower probabilities (average decrease of -0.007 by as much as -0.34).

Figure S8 below shows the change in probability for each climate change scenario with respect to the probabilities using current climatic data. In general, probabilities decrease where we currently observe or predict expansion of avocados in the central-western portion of the belt. Increases in probability occur in the eastern side of the region near the Monarch Reserve for both RCPs. Larger increases in probability are observed under RCP2.6 (Figure S6 - panel A) than under RCP8.5 (panel B) due to more pronounced changes in precipitation and milder increases in temperatures for the former scenario.

The correlation matrix (Figure S9) shows that changes in estimated probabilities are positively correlated with changes in precipitation (0.300 and 0.322), meaning that places with future higher

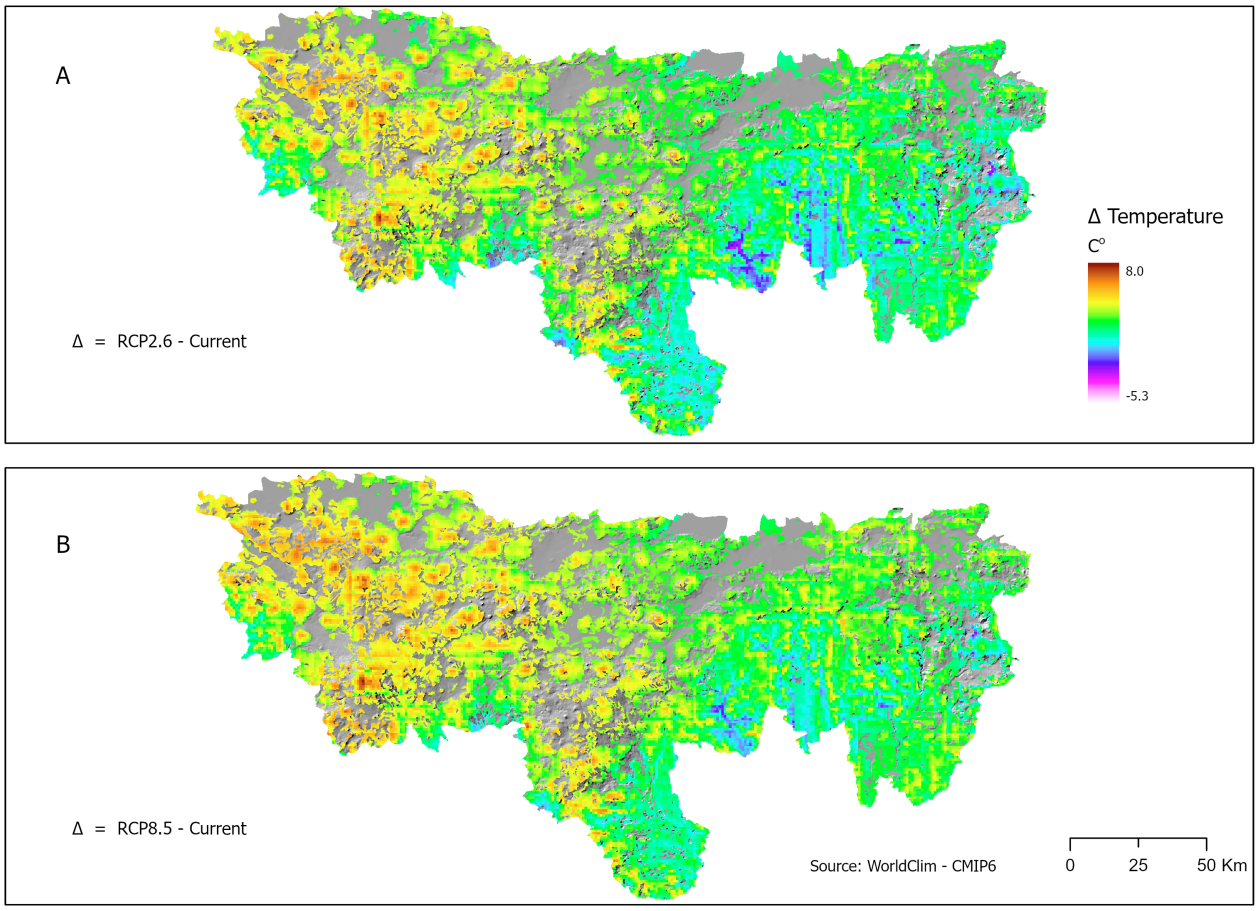

Figure S6. Change in temperature between RCP scenarios 2.6 (A) and 8.5 (B) with respect to current temperatures.

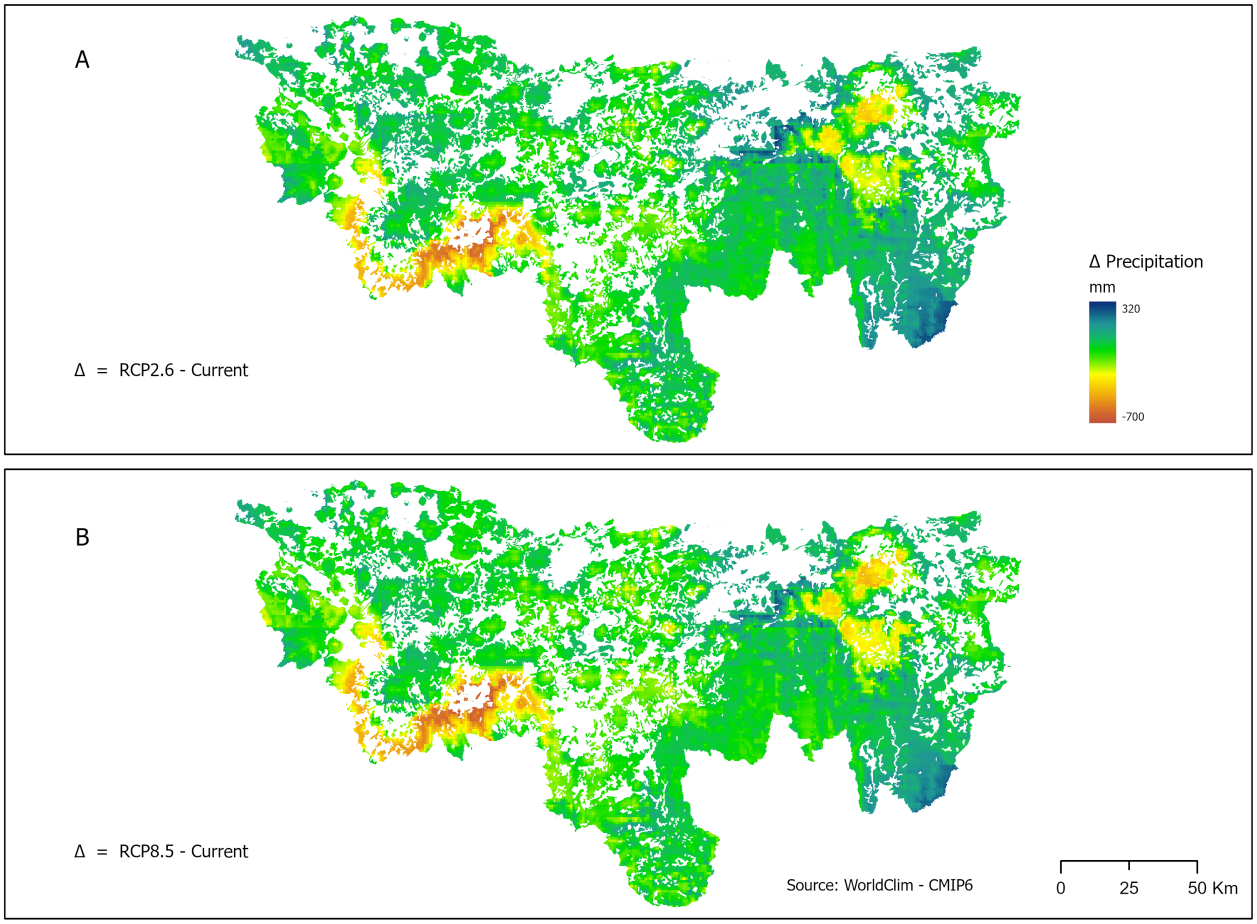

Figure S7. Change in precipitation between RCP scenarios 2.6 (A) and 8.5 (B) with respect to current precipitation.

precipitation than current levels will have higher probabilities of becoming avocados. Changes in temperature are negatively correlated with changes in probability (-0.170 and -0.169), i.e. places that will become hotter (positive change in temperature) will have lower probabilities of avocado.

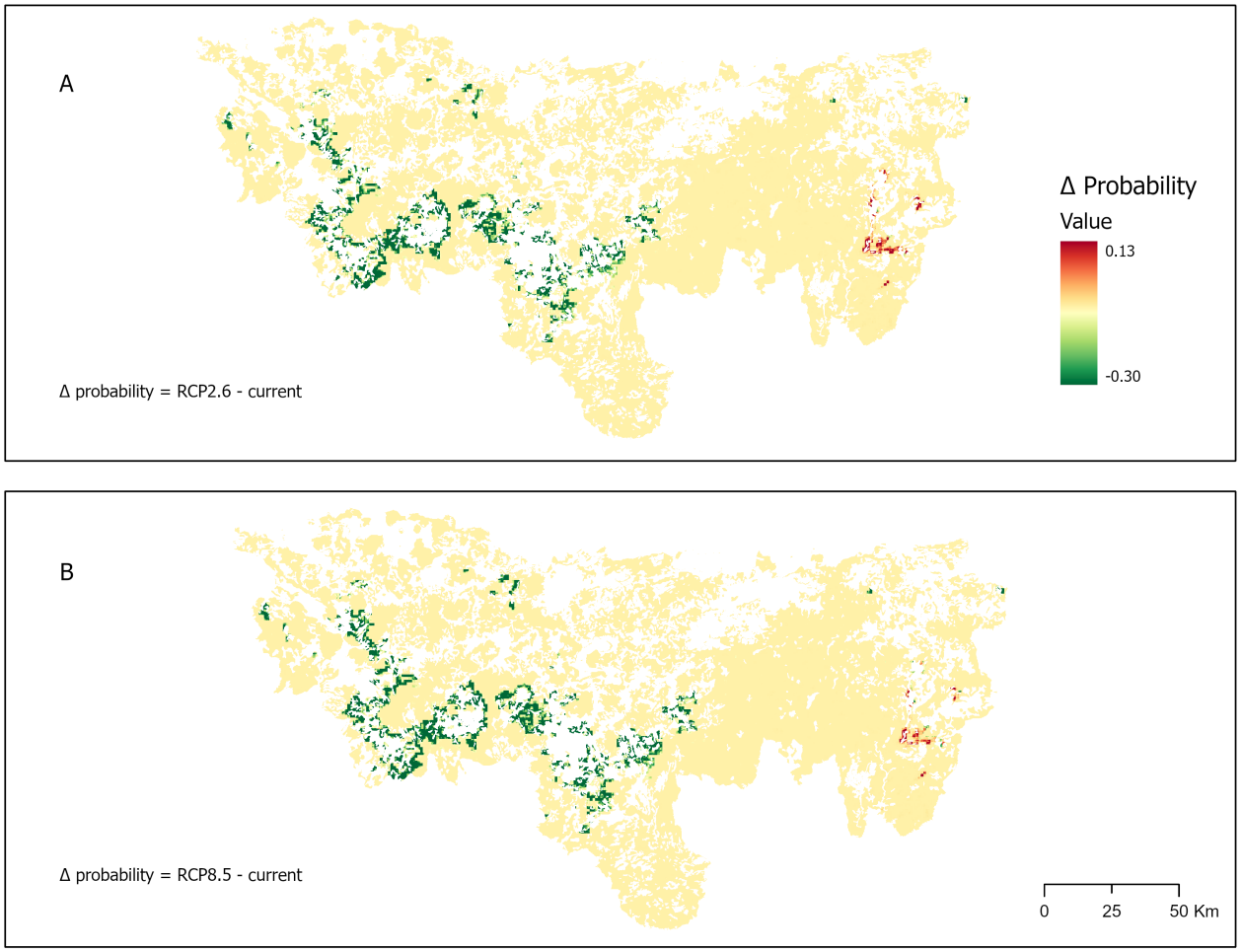

Figure S8. Simulated changes in avocado probabilities between RCP scenarios 2.6 (A) and 8.5 (B) with respect to current probabilities.

SI 8 - Correlation matrix between all independent variables

Multicollinearity among independent variables can lead to larger error variances but those problems are attenuated here due to the large number of observations used in the regression (n= 1,492,653;

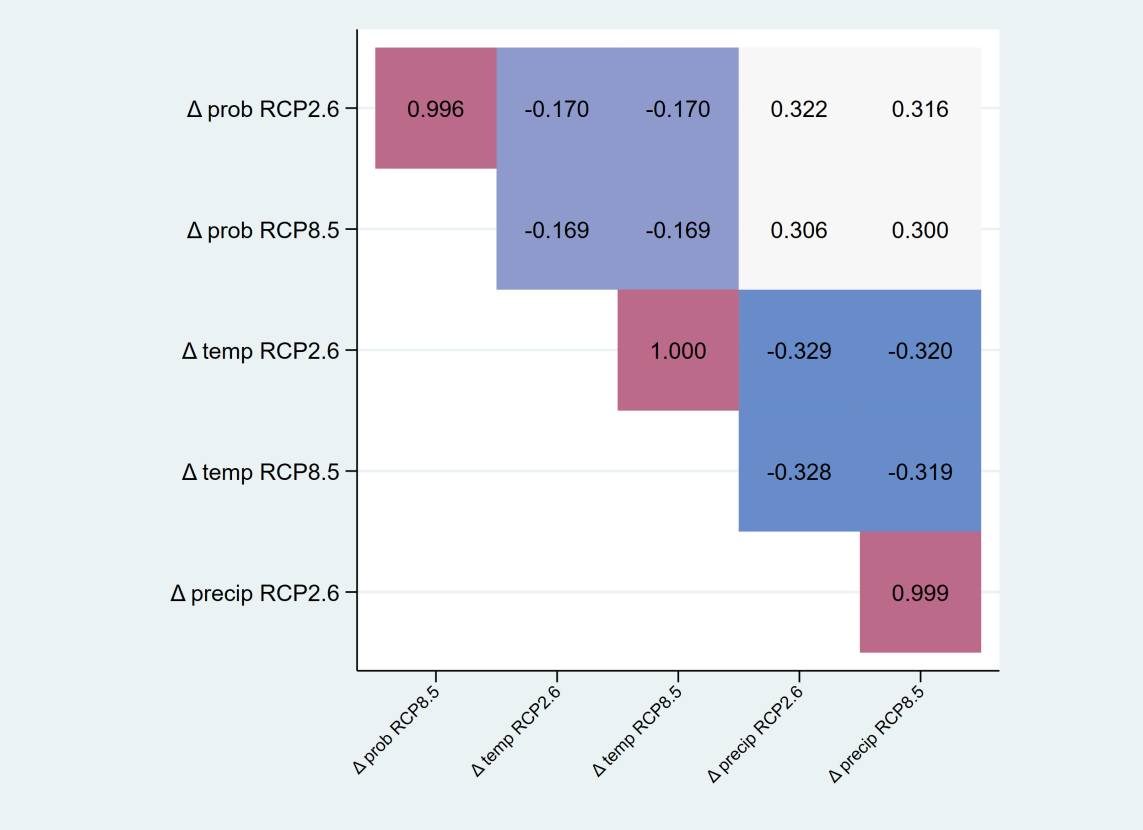

Figure S9. Correlation matrix between changes in probability, changes in precipitation, and changes in temperature between current and future climate change scenarios.

Goldberger 1991). Figure S10 shows the correlation matrix between all independent variables. As expected, the level terms for precipitation, temperature, and elevation are (positively) highly correlated with their respective squared terms, as well as elevation and presence of selva (tropical forest) and elevation with andosols (negative). Of interest, are the distance variables, particularly distance to cities (d2cities), which showed a positive coefficient in both spatial and standard (non-spatial) probit estimations (see S9 below). This variable is positively correlated with distance to 1992 agriculture and temperature, and negatively correlated with elevation. Nonetheless, the linear projection of d2cities onto the other variables only explains 47% (R^2) of the total variance of the former (see regression below where d2cities is the regressand and the other variables the regressors).

```
      Source |       SS           df       MS      Number of obs   = 1,492,653
-------------+----------------------------------   F(5, 1492647)   >  99999.00
       Model |  270797.985         5  54159.5971   Prob > F        =    0.0000
    Residual |  308384.705 1,492,647  .206602569   R-squared       =    0.4676
-------------+----------------------------------   Adj R-squared   =    0.4676
       Total |   579182.69 1,492,652  .388022587   Root MSE        =    .45454

------------------------------------------------------------------------------
    d2cities | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      d2agri |   1.095456   .0015422   710.34   0.000     1.092434    1.098479
        elev |   .1449463   .0117503    12.34   0.000     .1219162    .1679764
       elev2 |   -.050334   .0029312   -17.17   0.000    -.0560792   -.0445889
        temp |  -2.286985   .0261463   -87.47   0.000    -2.338231    -2.23574
       temp2 |   .7279842   .0068104   106.89   0.000     .7146362    .7413323
       _cons |   2.297204   .0228246   100.65   0.000     2.252468    2.341939
------------------------------------------------------------------------------
```

As a robustness check, we ran a linear probability model (OLS non-spatial version) without the distance to agriculture variable and the coefficient for distance to cities remained positive (0.01156) and statistically significant (std. err. 0.0024). We ran a second model with the further removal of temperature (both level and squared terms) and the positive and statistically significant coefficient for distance to cities still held. We opted therefore to maintain the variable distance to cities in the main model reported in the manuscript. The results suggest that the probability of avocadoes does indeed increase as we move away from cities, conditional on the other covariates. A possible explanation could be related to prices of land near cities, which are usually higher due to competing alternative uses (e.g. residential, chacras for vegetables and dairy) and more fragmented (i.e. smaller areas) thus increasing the costs for avocado production.

SI 9 - Non-spatial (standard) probit model

The non-spatial (standard) probit model without the regional spatially correlated effect returns estimated coefficients that are similar in magnitude and sign to the Bayesian spatial probit model (Table S7 below). Average marginal effects are also similar (Table S8), with the exception of leptosols (soillp) which showed a larger positive marginal effect. The greatest advantage of the spatial model when compared to the standard (non-spatial) model is the much superior performance

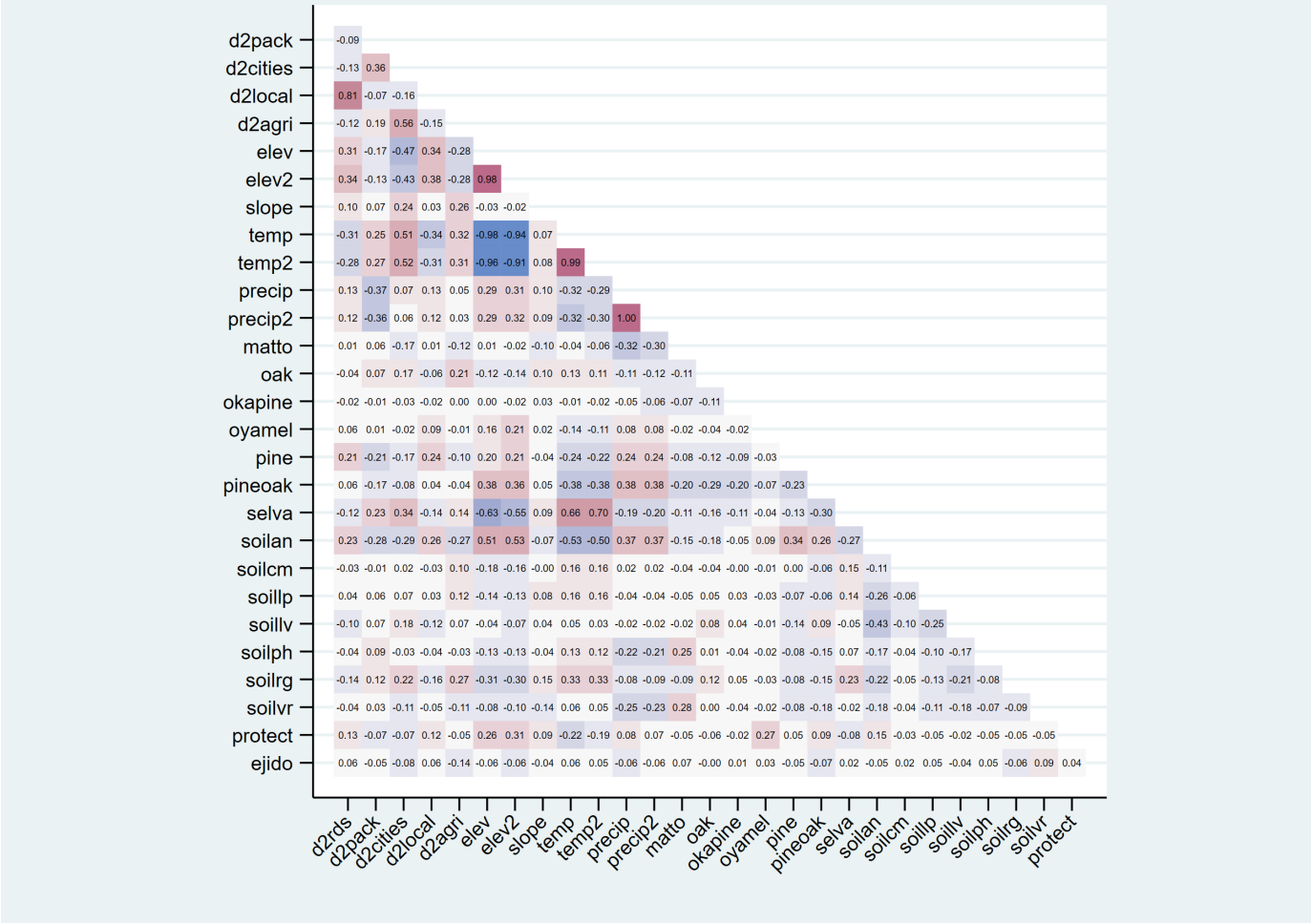

Figure S10. Heatmap correlation matrix between independent variables. Note: Precipitation and its squared term are shown to be perfectly collinear due to rounding.

in terms of spatial accuracy. The non-spatial model predicts 49% of the avocado pixels correctly (Table S9) whereas the spatial model accuracy (Table S4) is much better at around 81%. The non-spatial model completely misses the avocado areas to the east near the Monarch Reserve and other non-continuous avocado areas outside the core avocado area in the central part of Michocán (Figure S11).

Table S7. Non-spatial probit regression results

Variable	Coefficient	Std. Error	p-level
const	0.4853	0.2618	0.064
d2rds	-1.7565	0.0582	0.000
d2pack	-0.3346	0.0041	0.000
d2cities	0.3801	0.0077	0.000
d2local	-1.8504	0.0488	0.000
d2agri	-3.1936	0.0427	0.000
elev	12.8482	0.2165	0.000
elev2	-3.6496	0.0553	0.000
slope	-7.6175	0.2623	0.000
temp	-16.8206	0.3299	0.000
temp2	4.3961	0.0930	0.000
precip	5.1005	0.1948	0.000
precip2	-1.7729	0.0815	0.000
matto	-0.4682	0.0214	0.000
oak	-0.5628	0.0172	0.000
oakpine	-0.1996	0.0132	0.000
pine	-0.0858	0.0103	0.000
pineoak	0.0503	0.0085	0.000
selva	0.1381	0.0204	0.000
soilan	0.8167	0.0172	0.000
soilcm	-0.0433	0.0262	0.098
soillp	0.5833	0.0196	0.000
soillv	-0.0366	0.0178	0.039
soilph	0.4882	0.0214	0.000
soilrg	-0.1800	0.0262	0.000
protect	0.3569	0.0145	0.000
ejido	-0.0996	0.0055	0.000

Note: variables 'oyamel' and 'soilvr' were dropped from the non-spatial regression due to non-concavity issues during the likelihood maximization procedure.

Table S8. Average marginal effects on probability resulting from change in value of variables – non-spatial probit model

Variables	Partial Effect	Δ value	% Naïve probability
d2rds	-0.09529	1 Km	-23.7%
d2pack	-0.01815	1 Km	-4.7%
d2cities	0.020623	1 Km	5.3%
d2local	-0.10039	1 Km	-26.1%
d2agri	-0.17325	1 Km	-45.7%
elev	-0.0925	100 m	-22.7%
slope	-0.41326	1%	-10.5%
temp	-0.12417	1 Celsius	-31.6%
precip	0.052335	10 mm	1.4%
matto	-0.02016	Binary	-41.4%
oak	-0.02354	Binary	-52.0%
okapine	-0.00992	Binary	-13.7%
pine	-0.00452	Binary	1.2%
pineoak	0.002731	Binary	20.2%
selva	0.007997	Binary	34.7%
soilan	0.046667	Binary	131.8%
soilcm	-0.0023	Binary	5.1%
soillp	0.04004	Binary	116.4%
soillv	-0.00196	Binary	4.5%
soilph	0.033399	Binary	89.0%
soilrg	-0.00892	Binary	-17.7%
protect	0.023067	Binary	57.4%
ejido	-0.00535	Binary	-12.9%

Table S9. Cross-tabulation between predicted vs. actual avocado, non-spatial probit

		Actual		
		0	**1**	**Total**
Predicted	**0**	1,403,896	29,887	1,433,783
	1	29,939	28,931	58,870
	Total	1,433,835	58,818	1,492,653

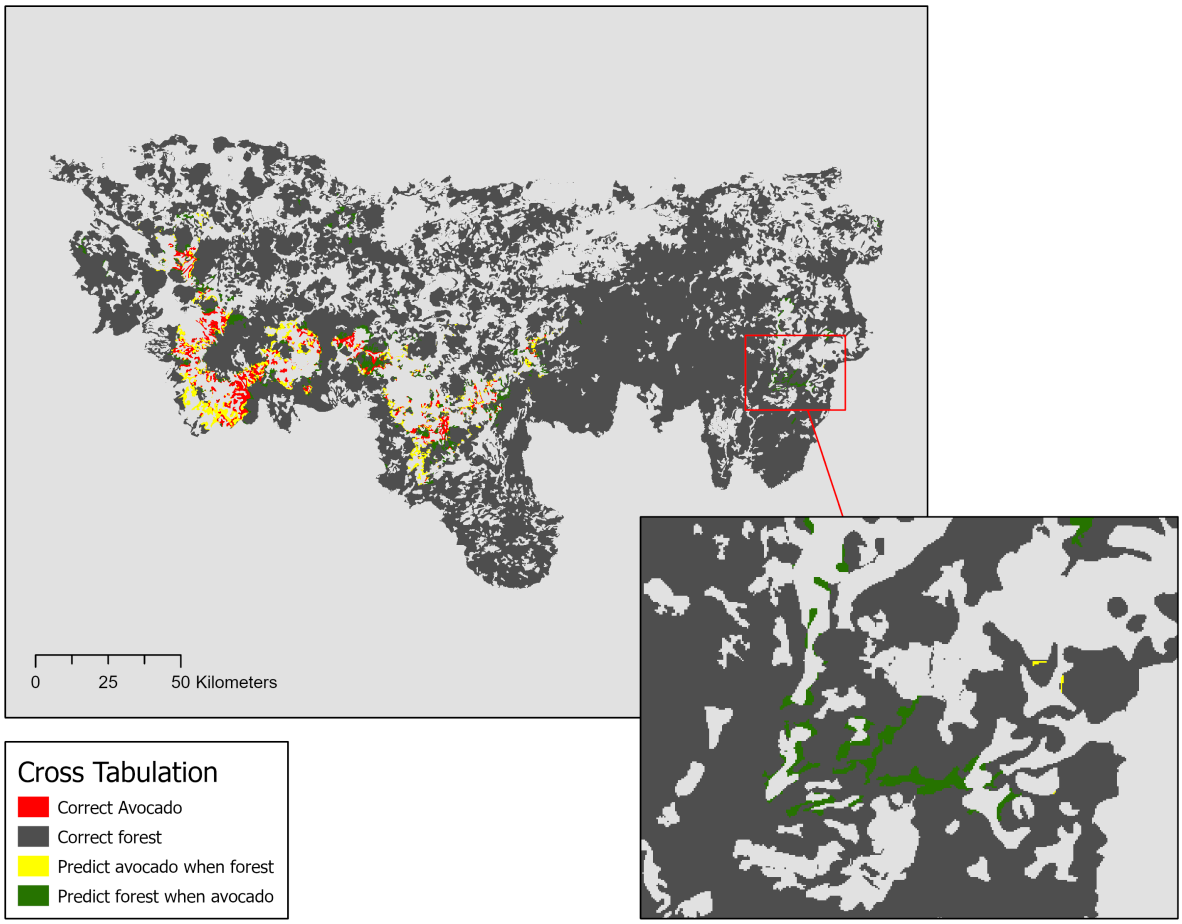

Figure S11. Cross tabulation of actual vs. predicted avocado, non-spatial probit model

S10 - Heterogeneity of avocado presence/absence among categorical variables.

	matto	
avoc	0	1	Total
0	1,332,734	100,963	1,433,697
1	58,552	404	58,956
Total	1,391,286	101,367	1,492,653

	oak	
avoc	0	1	Total
0	1,234,259	199,438	1,433,697
1	58,082	874	58,956
Total	1,292,341	200,312	1,492,653

	oakpine	
avoc	0	1	Total
0	1,334,101	99,596	1,433,697
1	56,261	2,695	58,956
Total	1,390,362	102,291	1,492,653

	oyamel	
avoc	0	1	Total
0	1,421,536	12,161	1,433,697
1	58,956	0	58,956
Total	1,480,492	12,161	1,492,653

	pine	
avoc	0	1	Total
0	1,311,302	122,395	1,433,697
1	47,108	11,848	58,956
Total	1,358,410	134,243	1,492,653

	pineoak		
avoc	0	1	Total
0	950,689	483,008	1,433,697
1	23,937	35,019	58,956
Total	974,626	518,027	1,492,653

	selva		
avoc	0	1	Total
0	1,215,197	218,500	1,433,697
1	57,782	1,174	58,956
Total	1,272,979	219,674	1,492,653

	soilan		
avoc	0	1	Total
0	1,016,109	417,588	1,433,697
1	13,415	45,541	58,956
Total	1,029,524	463,129	1,492,653

	soilcm		
avoc	0	1	Total
0	1,396,643	37,054	1,433,697
1	58,409	547	58,956
Total	1,455,052	37,601	1,492,653

	soillp		
avoc	0	1	Total
0	1,242,969	190,728	1,433,697
1	54,530	4,426	58,956
Total	1,297,499	195,154	1,492,653

avoc	soillv 0	1	Total
0	997,466	436,231	1,433,697
1	53,238	5,718	58,956
Total	1,050,704	441,949	1,492,653

avoc	soilph 0	1	Total
0	1,342,492	91,205	1,433,697
1	57,653	1,303	58,956
Total	1,400,145	92,508	1,492,653

avoc	soilrg 0	1	Total
0	1,292,752	140,945	1,433,697
1	58,300	656	58,956
Total	1,351,052	141,601	1,492,653

avoc	soilvr 0	1	Total
0	1,332,410	101,287	1,433,697
1	58,191	765	58,956
Total	1,390,601	102,052	1,492,653

avoc	protect 0	1	Total
0	1,382,804	50,893	1,433,697
1	56,865	2,091	58,956
Total	1,439,669	52,984	1,492,653

avoc	ejido 0	1	Total
0	837,780	595,917	1,433,697
1	40,150	18,806	58,956
Total	877,930	614,723	1,492,653

References

Goldberger, Arthur Stanley (1991). *A Course in Econometrics*. Harvard University Press. 430 pp. ISBN: 978-0-674-17544-0.

Smith, T. E. and James P LeSage (2004). "A Bayesian probit model with spatial dependencies". In: *Advances in Econometrics: spatial and spatiotemporal econometrics*. Ed. by James P LeSage and Pace Kelley. Vol. 18. Oxford: Elsevier, pp. 127–160.

Wooldridge, J. (2010). *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press. 752 pp.

LeSage, James P (2015). *Econometrics Toolbox*. Issue: Last Accessed: July 3 2015 Volume: 2015.

Arima, E. Y. (Mar. 24, 2016). "A Spatial Probit Econometric Model of Land Change: The Case of Infrastructure Development in Western Amazonia, Peru". In: *PLOS ONE* 11.3. Publisher: Public Library of Science, e0152058. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0152058. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152058 (visited on 04/25/2025).