PNAS
www.pnas.org

Supplementary Information for

The Effect of Deforestation on Access to Clean Drinking Water

Annie Mwayi Mapulanga and Hisahiro Naito

Corresponding author: Hisahiro Niato
E-mail: naito@dpipe.tsukuba.ac.jp

This PDF file includes:

References for SI reference citations

Background

Malawi is one of the poorest countries in the world, with GDP per capita of only 1,200 US dollars as of 2016. The CIA's *The World Factbook* ranks Malawi's GDP per capita as fifth from the bottom. Moreover, 50 percent of Malawi's population is considered poor (1). According to the 2008 Census, the population of Malawi is 13 million. In 1998, 85 percent of the population lived in rural areas.

In 1970, about 50 percent of Malawi's land was covered by forests (2). However, during the last 45 years, a substantial portion of forests was lost, with various estimates for the rates of deforestation. The Food and Agriculture Organization of the United Nations estimates that Malawi's forest area was 3,890 kha in 1990 and 3,237 kha in 2010, implying an annual deforestation rate of 0.9 percent (FAO, 2015). (3) estimate that 36 percent of the forest area was lost in the 27 years between 1972 and 2009, implying an annual deforestation rate of 1.63 percent. Our satellite images show that from 2000 to 2010, 14.2 percent of the forest was depleted, implying an annual deforestation rate of 1.51 percent.*

Table S1. Deforestation of Each Region during 1972–1992

	(1)	(2)	(3)	(4)	(5)	(6)
Region	Total Area of Each Region (Tha)	Forest Area in 1972 (Tha)	Forest Area in 1992 (Tha)	Ratio of Forest area in 1972	Ratio of Forest area in 1992	Difference (4)–(5)
Northern	2,713	1,507	470	55.6%	17.3%	38.2%
Central	3,562	1,488	777	41.8%	21.8%	20.0%
Southern	3,296	1,405	651	42.6%	19.7%	22.9%
National	9,571	4,400	1,898	46.0%	19.8%	26.1%

Source: Food and Agriculture Organization of the United Nations (2001). Calculated by the authors.
Note: Tha denotes thousand hectares.

Deforestation patterns are quite different across regions. Although 90 percent of the population lives in the central and southern regions of Malawi, the depletion rate of the forest in the northern region is much higher than that in the other regions. Table S1 shows the past pattern of deforestation in the three regions. In the northern region in 1972, although 56 percent of the land was covered by forests, this share decreased to 17.3 percent in 1992. In the southern region, 43 percent of the land was covered by forests, which decreased to 23 percent in 1992. Thus, for 20 years, in the northern region, the forest depleted by 69 percent [(56–17.3)/56=0.69], whereas in the central and southern regions, the forest depleted by 47 percent [(43–23)/43=0.47].

Table S2. Share of Population and Density by Region

	Share of Population (%)			Population Density		
Region	1987	1998	2008	1987	1998	2008
Northern	11.4	12.4	13.1	34	46	63
Central	38.9	40.9	42.1	87	114	155
Southern	49.6	46.6	48.8	125	146	185

Source: Malawi Population and Housing Census provided by the Minnesota Population Center (2018). Population density is the number of persons per square km.

Table S2 provides the population and population density of each region (6). These data sets indicate that, as of 1987, the share of the population in the northern region was about 10 percent, indicating that the density of the population in the southern region was four times higher than that in the northern region.

Table S3 shows the pattern of deforestation from 1990 to 2010. Tables S1, S2, and S3 imply that the

* In contrast, the Ministry of Natural Resources, Energy and Mining estimates an annual deforestation rate of 2.8 percent (4, 5).

Table S3. Ratio of the Forest Area of Each Region during 1990–2010

	Ratio of Forest Area			Difference	
	1990	2000	2010	1990–2000	2000–2010
Region	(1)	(2)	(3)	(1)–(2)	(2)–(3)
Northern	74.5%	53.1%	42.5%	21.4%	10.6%
Central	39.9%	35.9%	30.1%	4.0%	5.8%
Southern	45.2%	30.7%	29.6%	14.5%	1.1%
National	51.5%	39.0%	33.4%	12.5%	5.6%

Source: The authors' calculation based on the satellite images of land use for 1990, 2000, and 2010. The ratio of the forest area of each region is calculated by dividing the forest area in each region by the area of each region.

northern region experienced a higher deforestation rate although it had a relatively much smaller population during 1972–1992.

Water Access in Malawi

Regarding access to clean drinking water, we need to note that since Malawi's transition to a democratic government in 1994, the new government shifted the responsibility of providing safe drinking water to local communities, arguing that consumers should manage and finance the operation and maintenance of water supply. As a result, the government substantially reduced its own budget allocated to developing and maintaining a water system (8), and, in rural areas of Malawi, most water systems are developed and managed by local communities. This change implies that it is rare to see a large-scale water system in the rural areas of Malawi, as observed in developed countries, and households' access to water is likely to be affected by local forest and weather conditions.

Table 1 of the main text provides the share of each source of drinking water in rural populations based on the 1998 and 2008 censuses. In rural areas of Malawi, there are five types of clean drinking water sources: protected wells, boreholes/tubed wells, gravity-fed piped water, unprotected wells, and rivers/dams (9). Protected wells and tubed wells are dug by machines, and their structure prevents the surface water from flowing into the hole. The main difference between a tubed well and a protected well is that a tubed well has a tube inside of it, and a protected well has brick or stone lining in it. Since we do not have information on the depth of wells, they can be shallow or deep wells. Shallow wells have the risk of contamination. Since the structure and function of a tubed well and protected well are quite similar, we treat them as the same group in our analysis. A gravity-fed piped water system includes connected pipes from a water source to several taps in the village. As we explained above, building a water system is decentralized to local communities. Thus, it is rare to see the large-scale water systems common in developed countries in the rural areas of Malawi. An unprotected well is just a shallow hole that faces the risks of collapse and contamination during the rainy season. During the rainy season, the probability that contaminated surface water will flow into the hole is quite high.

Among populations living in rural areas in 1998, only 3.5 percent had a piped water system in their dwellings or yards. Only about 12 percent of the population had access to some type of piped water system, including a community tap, and 29 percent used unprotected wells as their source of drinking water.

Types of Forests in Malawi

Malawi's forest and woodland areas are dominated by so called miombo woodland (10). The mionbo woodland forms a large belt from Angola in the west to Mozambique and Tanzania in the east of central and south Africa. Miombo woodland consists of tropical forest, sub-tropical forest, savannas, and shrublands. Tropical forest and sub-tropical forest are dense and mildly-dense forests, respectively. In our analysis, we follow the FAO's definition of forest (11) for the classification of forest. FAO distinguishes forests from other wooded land. In the FAO's definition, savannas and shrublands are not included in the definition of forest. Thus, we exclude the savannas and shrublands from "forest" in our analysis and we classify dense and mildly-dense forests as "forest". Regarding the ability to retain water, our "forest" has almost equal ability, since we do not consider the savannas and shrublands as forest. Figure S1 shows a picture of a typical miombo forest, the Chimaliro Miombo Forest in Kasungu District, Central Region, Malawi.

Figure S1: Chimaliro Miombo Forest in Kasungu District, Central Region, Malawi

Data Sets

This study uses the satellite images of land cover and land use from 2000 and 2010, data of the Demographic Health Survey (DHS) of 2000 (16) and 2010 (17) with GPS location information, weather data, and population data. The satellite images are provided by the Ministry of Natural Resources, Energy and Mining of Malawi. DHS data sets are provided by the US Agency for International Development (USAID). For population data, we use the Gridded Population of the World (GPW) v4 (CIESIN) (18). This data set is constructed using micro-census data around the world and shows the population distribution for every 30 arc second of the earth. Regarding the population map of Malawi, GPW v4 uses the population information of each enumeration area of the census of Malawi, of which there are 12,641. Within each enumeration area, GPW v4 distributes the population equally. Although this uniform distribution assumption can be restrictive and data sets exist that use other information, such as settlement from satellite images, we believe that not using other surface information is desirable for the purpose of the regression. In addition, we find that the choice of the data set for the local population does not affect our estimation results. For temperature data, we use the temperature of the hottest (November) and the coldest (July) months from WorldClim (19). For rainfall data, we use the Climate Hazards Group InfraRed Precipitation with Station data 2.0 (20).

The National Statistics Office of the government of Malawi and USAID conduct the DHS, which is a nationally representative, cross-sectional geo-referenced survey. The DHS data sets report geographical coordinates of the primary sampling unit called a cluster. Each cluster is defined based on the census enumeration areas.[†] To apply the panel data analysis at the cluster level in rural households, we match the

[†] DHS changes each cluster's location within a 5-km distance to protect the confidentiality of survey respondents.

clusters in DHS 2000 in rural areas to clusters in DHS 2010, such that the distance between a cluster in DHS 2000 and a matched cluster in DHS 2010 is less than 5 km. After matching, 171 clusters are selected in both DHS 2000 and DHS 2010. The 171 clusters in DHS 2000 have 4,378 households, and the 171 clusters in DHS 2010 have 5,304 households.

The forest area ratio in each cluster is obtained from satellite images of land use and land cover maps from the Ministry of Natural Resources, Energy and Mining in Malawi. We create circles with 7.5-km radii using the cluster's GPS coordinate as the center of each circle and calculate the share of the forest area using the 2000 and 2010 satellite images. For robustness checks, we create circles with 12.5-km radii and calculate the ratio of the forest in each. Then, we conduct the same regression for robustness checks.

Regarding information on access to clean drinking water, DHS data have information on a household's source of drinking water. The DHS's final report classifies the source of drinking water into improved water sources and unimproved water sources. Following this DHS definition, we classify the sources of drinking water as clean drinking water or unsafe water if they are classified in the DHS as an improved or unimproved, respectively. More specifically, DHS 2000 classifies piped water into dwelling, piped water into yard, community tap, protected well, and borehole as improved sources of drinking water. It classifies unprotected wells, springs, rivers/streams/ponds, and rainwater as unimproved sources of drinking water. DHS 2010 classifies piped water into dwelling, piped water into yard, community tap, protected well, borehole, and protected spring as improved sources of drinking water, and it classifies unprotected well, unprotected spring, river/stream/pond, and rainwater as unimproved sources of drinking water. Our sensitivity checks for whether our results are sensitive to the classification of spring and rainwater show that our results are robust. Results of the sensitivity checks are available on request from the author.

Regarding the wealth information, we do not use the wealth index information of DHS data because the information set used to construct this wealth index includes the type of the source of drinking water, which is the dependent variable in our regression. To control the household wealth level, we directly use the information on floor material and ownership of a household asset, such as a bicycle and radio, which are used in the DHS data set to construct the wealth index.

In rural areas of Malawi, most households responded that their floor material is the ground or sand. Therefore, we code this variable as zero if the floor material is sand, ground, or dung. Otherwise, it is coded as one. Table S4 shows summary statistics of the main variables used in our regression analysis.

Table S4. Summary Statistics of the Main Variables

Years	2000		2010	
Variables	mean	sd	mean	sd
Clean drinking water dummy	0.619	0.486	0.825	0.380
Ratio of forest area	0.275	0.194	0.220	0.171
Log of rainfall	6.931	0.178	6.904	0.156
Log of population	10.73	0.659	10.94	0.631
Temperature of coldest month (July)	18.13	1.891	18.03	1.844
Temperature of hottest month (November)	25.01	2.048	24.93	1.986
Years of school head	4.019	3.674	5.005	3.868
Sex of head of household	0.708	0.455	0.709	0.454
Age of head of household	43.30	16.13	43.75	16.46
Number of household members	4.487	2.369	4.823	2.285
Good floor material	0.125	0.330	0.151	0.358
Radio ownership dummy	0.519	0.500	0.515	0.500
Bicycle ownership dummy	0.461	0.499	0.477	0.500
Latitude	14.12	1.926	14.11	1.861
Number of households	4,378		5,304	

Notes: The sample is the cluster-level panel data of the Demographic Health Survey (DHS) for 2000 and 2010. There are 171 clusters in each wave of DHS. The sample is restricted to rural clusters in 2000 and their paired clusters in 2010. The clean drinking water dummy is equal to 1 if a household has access to clean drinking water. To calculate the ratio of forest area, a circle with a 7.5-km radius is chosen. July is the hottest month of the year, and November is the coldest. The good floor material dummy is equal to 0 if the floor material is ground, sand, or dung. For other material, it is coded 1. Latitude is an absolute number and is always positive. The maximum and minimum values of the latitude are 17.12 and 9.67, respectively.

Explanation of the Model

In our analysis, we use a panel data set with the time fixed effect and cluster-fixed effect applying two-staged least squares estimation. This approach is sometimes called the fixed effect instrumental variable estimation (FEIV) and it is used extensively in the evaluation literature when a researcher want to analyze the causal effect of a variable on the outcome variable (21).

Our model consists of two equations:

$$\begin{aligned} \text{Water}_{ijt} &= \beta_0 + \beta_1 \text{RatioForest}_{jt} + \beta_2 X1_{ijt} \\ &+ \beta_3 X2_{j0} \times D_t + \alpha_{2t} + \alpha_{2j} + u_{ijt}, \end{aligned} \quad [1]$$

$$\begin{aligned} \text{RatioForest}_{jt} &= \gamma_0 + \gamma_1 \text{Latitude}_j \times D_t \\ &+ \gamma_3 X1_{ijt} + \gamma_4 X2_{j0} \times D_t + \alpha_{1t} + \alpha_{1j} + \varepsilon_{ijt}, \end{aligned} \quad [2]$$

where Water_{ijt} is a dummy variable indicating whether household i in cluster j at time t has access to clean drinking water. RatioForest_{jt} is the ratio of the area covered by the forest to the total area of the DHS cluster j at time t. α_{2t} is a time fixed effect, and α_{2j} is a cluster fixed effect. $X1_{ijt}$ is a vector of control variables that directly affect access to clean drinking water. $X2_{j0}$ is a vector of the cluster-level values at the initial period, and D_t is a time dummy. $X2_{j0} \times D_t$ allow clusters with different initial values to have different time trends. u_{ijt} is the error term that explains the variation of Water_{ijt}, which cannot be explained by the listed explanatory variables of equation (1). Since u_{ijt} might be correlated with RatioForest_{jt}, we cannot estimate β_1 of equation (1) consistently using ordinary least squares estimation (OLS). To estimate β_1 consistently, we introduce the first-stage equation (2) and apply two-staged least squares estimation (2SLS). α_{1t} is a time fixed effect, and α_{1j} is a cluster fixed effect. D_t is a time dummy, and Latitude_j is the latitude of cluster j. ε_{ijt} is the error term that captures the variation of RatioForest_{jt}, which cannot be explained by the list of explanatory variables in equation (2). The key assumption in the model in equations (1) and (2) is that the our instrumental variable $\text{Latitude}_j \times D_t$ is uncorrelated with u_{ijt}. This assumption implies $\text{Latitude}_j \times D_t$ does not affect the accessibility to clean drinking water other than through the ratio of forest area and other control variables, including the time fixed effect and cluster fixed effect. The time fixed effect controls the effect of variables that are constant across clusters but vary over two periods such as the national level time trend and macroeconomic shocks. The cluster fixed effect controls the effect of variables that are specific to each cluster but do not vary over two periods such as elevation and steepness of clusters.

To see the meaning of our model, for each pair of (t, j), we can take the average of both sides of equation (1) and subtract the equation with t=2000 from the equation with t=2010. Then, we have

$$\begin{aligned} \Delta\overline{\text{Water}}_j &= \beta_0 + \beta_1 \Delta\text{RatioForest}_j \\ &+ \beta_1 \Delta\overline{\text{X1}}_j + \beta_3 X2_{j0} + c_2 + (\overline{u}_{j,2010} - \overline{u}_{j,2000}), \end{aligned} \quad [3]$$

where $\Delta\overline{\text{Water}}_j = \overline{\text{Water}}_{j2010} - \overline{\text{Water}}_{j2000}$; $\Delta\text{RatioForest}_j = \text{RatioForest}_{j2010} - \text{RatioForest}_{j2000}$; $\Delta\overline{\text{X1}}_j = (\overline{\text{X1}}_{j2010} - \overline{\text{X1}}_{j2000})$; and c_2 is a new intercept. $\overline{\text{Water}}_{jt}$, $\overline{\text{X1}}_{jt}$, and $\overline{u}_{jt}$ are the cell average of Water_{jti} X1_{jti} and u_{jti} for each pair of (j,t).

Thus, equation (3) states that the difference of the average water accessibility between t=2010 and t=2000 is a function of the difference of the forest ratio between t=2010 and t=2000, the difference of the average of the household characteristics vector between t=2010 and t=2000, and the initial cluster-level characteristics in 2000. Note that since $\Delta\text{RatioForest}_j$ is correlated with $(\overline{u}_{j,2010} - \overline{u}_{j,2000})$, we cannot estimate equation (3) using the OLS.

Using the same procedure, we can rewrite equation (2) as follows:

$$\begin{aligned} \Delta\text{RatioForest}_j &= \gamma_1 \text{Latitude}_j \\ &+ \gamma_3 \Delta\overline{\text{X1}}_j + \gamma_4 X2_{j0} + c_{1.} + (\overline{\varepsilon}_{j,2010} - \overline{\varepsilon}_{j,2000},) \end{aligned} \quad [4]$$

where c_1 is a new intercept. Equation (4) states that the difference of the ratio of forest area between 2010 and 2000 can be a function of the latitude, the difference of the average of the household characteristics, and the initial cluster-level characteristics. From the assumption that $\text{Latitude}_j \times D_t$ is not correlated with u_{ijt}, Latitude_j is not correlated with $(\overline{u}_{j,2010} - \overline{u}_{j,2000})$. Thus, estimating equation (1) and (2) by 2SLS with the instrumental variable being $\text{Latitude}_j \times D_t$ is equivalent to estimating equations (3) and (4) by 2SLS with the instrumental variable being Latitude_j. Note that when we estimate the model, we still use equations (1) and (2) because controlling household characteristics will give higher statistical precision. However, even when we estimate (1) and (2), the identification mechanism comes from equations (3) and (4).

Equations (3) and (4) imply that we are essentially looking at the change in access to clean water and in the ratio of forest area during 2000–2010 across clusters with different latitudes while controlling the covariates. The key needed assumption is that *the latitude is not correlated with the change in access to clean drinking water other than through a change in the ratio of forest area and other control variables*. If latitude is correlated with the change in access to clean drinking water other than through a change in the ratio of forest area and change in control variables, we need to select a variable that captures such an effect. As a result, this becomes the standard for the selection of control variables.

Selection of Control Variables

Once we understand that equations (1) and (2) can be represented by equations (3) and (4), it is easy to identify which variable should be included as a control variable.

The key assumption for estimating (3) and (4) is that latitude is not correlated with the change in access to clean drinking water other than through a change in the ratio of forest area and other control variables. If latitude is correlated with change in access to clean drinking water other than through a change in the ratio of forest area and control variables, we need to select a variable that captures such an effect because our estimated coefficient of the ratio of forest area includes the effect from such a variable. Thus, this becomes the rule for the selection of control variables.

Histogram of the Difference of the Ratio of Forest between 2010 and 2000

A natural question of our 2SLS estimation is whether or not there is enough variation in the change of the ratio of forest area from 2000 to 2010. At the national level, the ratio of forest area decreases only 5 percentage points. However, at the local level, there is huge cross-sectional variation in the change of the ratio of forest area from 2000 to 2010 across clusters (Figure S2). This implies that it is reasonable to examine the variation in how those clusters experience changes in the availability of clean drinking water.

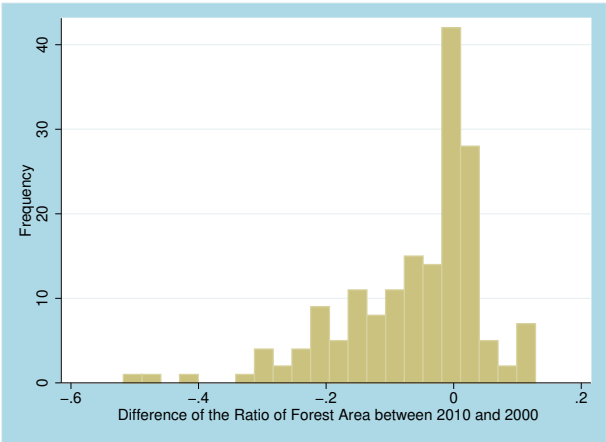

Figure S2: Histogram of the difference of the ratio of forest area between 2010 and 2000 for each cluster.

Standard Error Calculation

The standard error is calculated by assuming that the error terms are correlated within the cluster × year cell. In other words, we do not assume that the error terms in the same cluster are correlated over two periods. This is because the sum of the first-period cluster level error and the sum of the second-period cluster level error is always zero in the two-period fixed effect by construction. Note that, in the two-period fixed effect panel data, the famous result by (22) does not apply. The result of (22) applies only when the time period is greater than or equal to 3. (23) also makes a similar observation: for the T period difference-in-difference estimation, it is better to transform the T-period data into two period data and assume that the error term is not correlated over time to make the standard error calculation appropriate.

Additional Regression Results

Table S5 shows the estimation results of the OLS estimation. The estimated coefficients of Table S5 are one-tenth of the estimated coefficients of Table 2, and they are statistically insignificant. The exogeneity test rejects the null hypothesis that the ratio of forest is exogenous. Table S5 shows the importance of controlling the endogeneity of the ratio of forest area.

Table S5. Results of the Ordinary Least Squares Estimation
Effect of the Ratio of Forest Area (7.5-km radius)
on Access to Clean Drinking Water

Dependent variable	Clean drinking water dummy			
Variables	(1)	(2)	(3)	(4)
Ratio of forest area	0.0118	0.0227	0.0287	0.0373
	(0.157)	(0.155)	(0.154)	(0.154)
Log of rainfall	0.340***	0.331**	0.350***	0.337***
	(0.128)	(0.129)	(0.127)	(0.126)
Log of population		yes	yes	yes
Temperature		yes	yes	yes
Demographic characteristics			yes	yes
Household wealth				yes
R^2	0.264	0.266	0.272	0.274
N	9,682	9,682	9,682	9,682

Notes: Clustering robust standard errors are in parentheses, assuming that the error term is correlated within each cluster × year cell. All specifications include the cluster fixed effect, time fixed effect, and cluster-level initial values × time dummy in addition to the variables listed above. To calculate the size of the forest area ratio, a circle with a 7.5-km radius is chosen. *** p<0.01, ** p<0.05, and * p<0.1.

Table S6 shows the reduced-form regression where the dependent variable is the access to clean drinking water dummy. The OLS regression is applied by regressing the dependent variable on the instrumental variable and other control variables.

When we conduct the 2SLS estimation and examine the effect of the ratio of forest area on access to clean drinking water in Table 2, one natural question to our model is whether we might be picking up the effect of not only a change in the forest area but also a change in the ratio of cropland since the forest is cut due to an increase in population, it is possible that such a land is transformed to cropland. In addition, when the forest is transformed due to the pressure of increase in population, the land intensity might increase simultaneously. Since the size of the cropland and intensity of cropland can affect the water quality, our estimated coefficient might pick up the effect of not only a change in forest ratio on access to clean drinking water but also a change in cropland.

However, in our specification, such a possibility is unlikely. First, we control the effect of population increase as we include the log of population in the control variable. Since the key motivation for cropland expansion and an increase in cropland intensity is population increase, our estimate is unlikely to pick up the effect of a change in the ratio of cropland and an increase in cropland.

Table S6. Reduced-Form Regression of the Two-Stage Least Squares Estimation
Effect of the Latitude × Time Dummy
on Access to Clean Drinking Water

Dependent variable	Clean drinking water dummy			
Variables	(1)	(2)	(3)	(4)
Latitude	0.0292**	0.0292**	0.0281**	0.0284**
× time dummy	(0.0125)	(0.0125)	(0.0124)	(0.0123)
Log of rainfall	0.481***	0.481***	0.492***	0.480***
	(0.133)	(0.133)	(0.132)	(0.131)
Log of population		yes	yes	yes
Temperature		yes	yes	yes
Demographic characteristics			yes	yes
Household wealth				yes
R^2	0.268	0.268	0.273	0.276
N	9,682	9,682	9,682	9,682

Notes: The clustering robust standard errors are in parentheses, assuming that the error term is correlated within each cluster × year cell. All columns include the cluster-fixed effect, time fixed effect, log of rainfall, and cluster-level initial values × time dummy as control variables in addition to the variables listed above. *** p<0.01, ** p<0.05, and * p<0.1.

Table S7. Robustness Check (1) of Table 2
Controlling the Effect of Cropland and Cropland Intensity
Second-Stage Estimates of the Two-Stage Least Squares

Dependent variable	Clean drinking water dummy			
	(1)	(2)	(3)	(4)
Ratio of forest area	1.287**	1.195**	1.154**	1.147**
	(0.509)	(0.488)	(0.479)	(0.477)
Log of rainfall	0.544***	0.516***	0.526***	0.517***
	(0.148)	(0.144)	(0.142)	(0.142)
Log of population		yes	yes	yes
Temperature		yes	yes	yes
Demographic characteristics			yes	yes
Household wealth				yes
Ratio of cropland	yes	yes	yes	yes
Cropland intensity	yes	yes	yes	yes
R^2	0.254	0.257	0.264	0.266
Kleibergen-Paap rank	67.72	67.11	67.31	67.29
N	9,682	9,682	9,682	9,682

Notes: The estimated coefficients and standard errors of the second stage of the two-stage least squares estimation are displayed. The clustering robust standard errors are in parentheses, assuming that the error term is correlated within each cluster × year cell. All specifications include the time dummy, the cluster fixed effect log of rainfall, and the cluster-level initial values × time dummy, in addition to the variables listed above. All columns control the ratio of cropland the cropland intensity. The cropland intensity is measured by the ratio of the size of the perennial cropland over the size of the annual cropland. *** p<0.01, ** p<0.05, and * p<0.1.

To support our argument, in Table S7, we include the ratio of cropland and intensity of cropland in the control variables in all columns and run a 2SLS estimation. For calculating the ratio of cropland, we divide the total cropland area by the area of a circle with 7.5-km radius. For calculating the cropland intensity, we divide the area of perennial cropland by annual cropland. Table S7 shows that the estimated coefficients and their statistical significance do not change. In column (4) of Table S7, a one percent point increase in the cropland increases the probability of accessing clean drinking water by 1.1 percentage points.

In Table S8, we further control the effect of the wetland. When the forest is cut due to the pressure of increase in population, the size of the wetland is also changed. In this case, the estimated coefficient of Table 2 picks up the effect of not only a change in the ratio of forest area but also a change in the wetland. In Table S8, we include the ratio of the cropland, cropland intensity, and ratio of the wetland. Table S8

shows that the estimated coefficients of Table S8 are similar to the estimated coefficients of Table 2.

Table S8. Robustness Check (3) of Table 2
Controlling the Effect of Cropland, Cropland Intensity, and Wetland
in the Two-Stage Least Squares (2SLS): 2nd Stage Estimates of 2SLS

Dependent variable	Clean drinking water dummy			
	(1)	(2)	(3)	(4)
Ratio of forest area	1.419**	1.328**	1.288**	1.284**
	(0.559)	(0.535)	(0.525)	(0.523)
Log of rainfall	0.525***	0.496***	0.506***	0.497***
	(0.148)	(0.144)	(0.142)	(0.142)
Log of population		yes	yes	yes
Temperature		yes	yes	yes
Demographic characteristics			yes	yes
Household wealth				yes
Ratio of Cropland	yes	yes	yes	yes
Crop land Intensity	yes	yes	yes	yes
Ratio of wetland	yes	yes	yes	yes
R-squared	0.252	0.255	0.262	0.264
Kleibergen-Paap rank	61.28	61.44	61.61	61.60
N	9,682	9,682	9,682	9,682

Notes: The clustering robust standard errors are in parentheses, assuming that the error term is correlated within each cluster×year cell. All specifications include the time dummy, the cluster fixed effect log of rainfall, and the cluster-level initial values $\times$ time dummy, in addition to the variables listed above. All columns control the ratio of cropland, the cropland, and the intensity and ratio of the wetland. The cropland intensity is measured by the ratio of the size of the perennial cropland over the size of the annual cropland. *** p<0.01, ** p<0.05, and * p<0.1.

In Table S9, we examine whether or not our results are sensitive to the initial cluster-level characteristics. In Table S9, we drop the interaction term of the initial cluster-level characteristics and time dummy. Table S9 shows that the estimated coefficients of Table S9 are similar to the estimated coefficients of Table 2, which suggests the robustness of Table 2.

Table S9. Robustness Check (3) of Table 2
Effect of the Ratio of Forest Area on Access to Clean Water
without the Interaction Term

A. First-Stage Regression				
Dependent variable	Ratio of forest area			
	(1)	(2)	(3)	(4)
Latitude	0.0323***	0.0331***	0.0331***	0.0331***
×time dummy	(0.00399)	(0.00402)	(0.00401)	(0.00401)
Log of rainfall	-0.0844**	-0.0781*	-0.0782*	-0.0783*
	(0.0425)	(0.0424)	(0.0423)	(0.0423)
Kleibergen-Paap rank	65.65	67.84	68.08	68.20
R^2	0.948	0.948	0.948	0.948
B. Second-Stage Regression				
Dependent variable	Clean drinking water dummy			
	(1)	(2)	(3)	(4)
Ratio of forest area	0.977***	0.902**	0.903**	0.910**
	(0.376)	(0.365)	(0.362)	(0.360)
Log of rainfall	0.511***	0.477***	0.486***	0.480***
	(0.155)	(0.153)	(0.151)	(0.151)
R^2	0.252	0.256	0.261	0.264
C. Control Variables				
Log of population		yes	yes	yes
Temperature		yes	yes	yes
Demographic characteristics			yes	yes
Household wealth				yes
N	9,682	9,682	9,682	9,682

Notes: The clustering robust standard errors are in parentheses, assuming that the error term is correlated within each cluster × year cell. All specifications include the time dummy and the cluster fixed effect, in addition to the variables listed above. None of the columns above includes the cluster-level initial values × time dummy as control variables. *** p<0.01, ** p<0.05, and * p<0.1.

Table S10 displays the results of another robustness check using a large circle with a 12.5-km radius to calculate the forest area. Table S10 shows that the estimated coefficient of the ratio of forest area in Table S10 is slightly larger than the estimated coefficients in Table 2, but they are all positive and statistically significant. Table S10 shows that the results of Table 2 do not change even if we increase the radius of the circle of each cluster from 7.5 km to 12.5 km.

Table S10. Robustness Check (4) of Table 2
Effect of the Ratio of Forest Area (12.5-km radius) on Access to Clean Drinking Water in the Second-Stage Estimates of the Two-Stage Least Squares

Dependent variable	Clean drinking water dummy			
	(1)	(2)	(3)	(4)
Ratio of forest area	1.564**	1.409**	1.354*	1.368**
	(0.765)	(0.716)	(0.700)	(0.697)
Log of rainfall	0.705***	0.654***	0.659***	0.648***
	(0.210)	(0.199)	(0.195)	(0.195)
Log of population		yes	yes	yes
Temperature		yes	yes	yes
Demographic characteristics			yes	yes
Household wealth				yes
Kleibergen-Paap rank	32.24	33.90	34.05	34.12
R-squared	0.245	0.251	0.258	0.260
N	9,682	9,682	9,682	9,682

Notes: A circle with a 12.5-km radius from each cluster is used to calculate the ratio of the forest area. All notes of Table 2 apply.

Panel A of Table S11 shows the estimated coefficients of the instrumental variable in our first falsification test. Panel B of Table S11, which uses the radio ownership as a dependent variable, shows a similar pattern.

Panel C of Table S11 shows that the effect of the instrumental variable is again negative and statistically insignificant. Panels A, B and C of Table S11 suggest that it is unlikely that the southern region has a time trend of higher development. We can thus safely conclude that it is unlikely that a positive effect of the forest ratio on access to clean drinking water in our 2SLS estimation is the consequence of the southern region having a higher time trend of development.

Table S11. Falsification Tests in Reduced Form

	(1)	(2)	(3)	(4)
A. Dependent Variable: Access to the Electricity Dummy				
Latitude	-0.00180	-0.00183	-0.00247	-0.00287
× time dummy	(0.00349)	(0.00256)	(0.00242)	(0.00235)
R-squared	0.160	0.164	0.165	0.188
B. Dependent Variable: Radio Ownership Dummy				
Latitude	-0.0103	-0.0103	-0.0117*	-0.00834
× time dummy	(0.00906)	(0.00658)	(0.00660)	(0.00609)
R-squared	0.046	0.046	0.047	0.153
C. Dependent Variable: Good Floor Material Dummy				
Latitude	0.000342	0.000344	0.000170	-0.00119
× time dummy	(0.00612)	(0.00447)	(0.00449)	(0.00417)
R-squared	0.160	0.164	0.165	0.188
D. Control Variables				
Log of rainfall	yes	yes	yes	yes
Log of population		yes	yes	yes
Temperature			yes	yes
Demographic characteristics				yes
N	9,682	9,682	9,682	9,682

Notes. Clustering robust standard errors are in parentheses, assuming that the error term is correlated within each cluster × year cell. The above table shows the estimated coefficient of the instrumental variable in the reduced-form regression. All specifications include the time dummy and cluster fixed effect as control variables, in addition to the variables listed above. The dependent variable in Panel A is access to the electricity dummy. The dependent variable in Panel B is the radio ownership dummy.The dependent variable in Panel C is the good floor material dummy. *** p<0.01, ** p<0.05, and * p<0.1.

References

1. World Bank (2014) Malawi overview. *World Bank*.
2. Food and Agriculture Organization (2001) *Forestry Sector Outlook Studies—Counry Report: Malawi*. (United Nations).
3. Bone RA, Parks KE, Hudson MD, Tsirinzeni M, Willcock S (2017) Deforestation since independence: a quantitative assessment of four decades of land-cover change in malawi. *Southern Forests: a Journal of Forest Science* 79(4):269–275.
4. Ministry of Natural Resouce and Environment, Malawi (2011) Second national communication to the unfccc cop. (http://www.fao.org/3/a-i4808e.pdf). accessed Dec-03-2017.
5. Ministry of Natural Resouce and Environment, Malawi (2013) Malawi state of enviroment and outlook report. accessed Dec-03-2017.
6. National Statistical Office (2010) *Population and Housing Cenus 2008 Main Report*. (The Government of Malawi).
7. Minnesota Population Center (2018) *Integrated Public Use Microdata Series, International: Version 7.0*. (Minneapolis, MN: IPUMS).
8. Zuzani P, Ackim R, Kalulu K (2013) Sustainability of piped water supply schemes in rural malawi through community management. *Journal of Basic and Applied Scientific Research* 3(10):113–118.
9. Department of Lands, Valuation and Water (1993) *Gravity Fed Rural Piped Water Scheme:Rural Water Operator's Handbook*. (Government of Malawi).
10. Food and Agriculture Organization (2015) *Global Forest Resources Assessment 2015: Country Report Malawi*. (United Nations.).
11. Food and Agriculture Organization (2015) *Forest Resources Assessment 2015: Terms and Definitions*. (United Nations.).
12. Tole L (1998) Sources of deforestation in tropical developing countries. *Environmental Management* 22(1):19–33.

13. Minde I, Kowero G, Ngugi D, Luhanga J (2001) Agricultural land expansion and deforestation in malawi. *Forests, Trees and Livelihoods* 11(2):167–182.
14. Barbier EB (2004) Explaining agricultural land expansion and deforestation in developing countries. *American Journal of Agricultural Economics* 86(5):1347–1353.
15. Stickler MM, Huntington H, Haflett A, Petrova S, Bouvier I (2017) Does de facto forest tenure affect forest condition? community perceptions from zambia. *Forest Policy and Economics* 85:32–45.
16. Malawi. National Statistical Office and ICF Macro (Firm) (2001) *Malawi Demographic and Health Survey, 2000*. (National Statistical Office).
17. Malawi. National Statistical Office and ICF Macro (Firm) (2011) *Malawi Demographic and Health Survey, 2010*. (National Statistical Office).
18. Center for International Earth Science Information Network - CIESIN - Columbia University (2016) *Gridded Population of the World, Version 4 (GPWv4): Population Density Adjusted to Match 2015 Revision UN WPP Country Totals.* (Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC)).
19. Fick SE, Hijmans RJ (2017) Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37(12):4302–4315.
20. Funk C, et al. (2015) The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific data* 2:150066.
21. Wooldridge JM (2010) *Econometric analysis of cross section and panel data*. (MIT press).
22. Stock JH, Watson MW (2008) Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica* 76(1):155–174.

23. Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? *The Quarterly journal of economics* 119(1):249–275.