Large-Scale Irrigation and Commercialization of Agriculture in South Africa

Preliminary Draft

Tereza Varejkova*

November 29, 2023

Abstract

I estimate the impacts of irrigation canals on annual crop productivity and the structure of the agricultural sector in South Africa. I use remotely sensed measures of crop yields and a novel land cover classification dataset in a regression discontinuity framework with relative elevation to the nearest canal as a running variable. I find that canals increase yields of two major crops (maize and wheat) and also lead to expansion of commercial farming. On the other hand, subsistence farmers with access to irrigation do not expand area under production. Furthermore, they experience lower yields relative to their counterparts in the control areas. Subsistence farmers are mostly concentrated in the former homelands, and therefore, large-scale irrigation exacerbates the post-apartheid spatial inequalities in South Africa.

1 Introduction

Structural transformation has long been discussed as a driver of economic development (Lewis, 1954; Kuznets, 1957; Ranis and Fei, 1961). While developed economies tend to be characterized by low shares of agricultural employment, these shares remain high in

*PhD Candidate, University of Maryland, College Park, Department of Economics. E-mail: tvarejko@umd.edu. I am extremely grateful to my advisors Sebastian Galiani, Guido Kuersteiner, and Maureen Cropper for their guidance and support. I am also grateful to Taryn Dinkelman, Ken Leonard, Cory Smith, Mamidipudi Ramakrishna Sharan, and Mahima Vasishth for invaluable feedback and suggestions that I received during my presentations at outside seminars. Finally, I would like to thank Rashid Hassan (University of Pretoria) and Marion Delport (BFAP South Africa) for useful comments on the previous version of this paper.

developing countries, particularly in sub-Saharan Africa. Furthermore, agricultural sectors in high-income countries consist of large-scale, commercial farming, whereas many sub-Saharan economies still rely on subsistence agriculture. For instance, in Africa, 80% of farmers operate on land with a total area of less than 2 ha (Lowder, Skoet and Raney, 2016). Transition from small-scale, subsistence farming towards large-scale, commercial agriculture can play an important role in the process of structural transformation by providing cheap and abundant food supplies to the growing manufacturing and services sectors (Suri and Udry, 2022; Collier and Dercon, 2014). Therefore, it is important to understand what factors contribute to commercialization of agriculture in Africa.

This paper examines how access to surface irrigation in South Africa affects agricultural productivity and commercialization of agriculture. Tatlhego et al. (2022) address a similar question and evaluate the correlation between an increase in irrigated area and farm size in seven countries, including South Africa. They find that the average farm size increased near large dams. Although this study is insightful, it does not identify a causal relationship. It simply evaluates a pre-post relationship by comparing the average farm size before and after dam construction. However, it is possible that large-scale farms were also established nearby, outside the command areas of these dams. In this paper I aim to establish a causal link between access to irrigation and rise in commercial agriculture. My identification strategy exploits the fact that irrigation is gravity-based. The basic principle behind gravity-based irrigation is that water flows downhill, following the slope of the land. A large dam is used to fill a network of irrigation canals, which are designed to carry water to the fields. Therefore, land that lies below a canal can be easily irrigated, whereas land above the canal is less likely to be irrigated.[1] I use this fact to evaluate the impact of irrigation canals using a regression discontinuity design (RDD) with relative elevation to the nearest canal

[1] It is still possible to irrigate land above canals by pumping water uphill, but it is very costly as it requires significant amounts of electricity.

as the running variable. Areas that lie topographically below a canal are defined as treated, while areas that lie above a canal are defined as control. A similar approach was previously applied in the context of India (Asher et al., 2022). An alternative running variable would be the distance to the boundary of a dam's command area. (Blakeslee et al., 2023; Jones et al., 2022).

The RDD method with relative elevation to the nearest canal as the running variable ideally requires a unit of analysis that is highly spatially disaggregated. For instance, Asher et al. (2022) use village-level data and the elevation of each village is determined as the 5^{th} percentile of the pixel distribution within each polygon. This approach is prone to measurement error since the village fields can be both above and below the nearest canal, but they will all be assigned the same treatment status. In contrast, I am able to assign the treatment status at the field or even sub-field level, since I use outcome variables derived from raster datasets with resolution of 30 meters and 20 meters.

As a proxy for agricultural productivity, I use the Enhanced Vegetation Index (EVI) derived from Landsat 8 satellite imagery, available at a 30 meters resolution. EVI can perform better than the other commonly used measure of crop productivity, Normalized Difference Vegetation Index (NDVI), especially in contexts with dense vegetation such as irrigated agriculture (Wardlow and Egbert, 2010). I focus on two major annual crops grown in South Africa — wheat and maize — and I construct separate productivity measures for each crop based on the crop calendar. Indicators for land use are derived from the South African National Land Cover 2018 (SANLC 2018) dataset and the South African National Land Cover 1990/2020 Change (SANLC 1990/2020). SANLC is a 73-class land cover classification dataset published by the South African Department of Rural Development and Land Reform (DRDLR) and generated by automated mapping models based on high-resolution satellite imagery. I use SANLC in two ways. First, I derive an indicator on whether the land is agricultural or not in order to apply a “crop mask" when I estimate the effect of irrigation canals on crop yields. This ensures credibility of the EVI measure as a proxy for yields by removing any noise from non-crop vegetation. Second, I construct indicator variables for land use outcomes, in particular, whether the land is fallow or used for growing annual crops, and whether there was an expansion of commercial or subsistence farming between 1990 and 2020. To define the treatment variable, I collect GPS coordinates of 144 irrigation canals from South Africa’s Department of Water and Sanitation (DWS). The unit of analysis is a 30 meters by 30 meters grid cell and the treatment status of these grid cells is determined based on their elevation derived from the ALOS Global Digital Surface Model which is also available at a 30 meters resolution.

The RD results indicate that irrigation canals increase agricultural production both at the intensive and the extensive margins, that is, irrigation increases both crop yields and the extent of cultivated area. However, only commercial farmers benefit from access to irrigation. I find that expansion of commercial cropland is more likely to have happened in areas below the canals, but the same does not hold for the expansion of subsistence cropland. Furthermore, I find that subsistence farmers in areas below the canals experience worse crop yields than farmers in areas above the canals. This is an unexpected result given the South African government’s emphasis on investment in irrigation schemes as a way to enhance food security among smallholder farmers (Sinyolo, Mudhara and Wale, 2014). One possible concern regarding the validity of my results is the use of data for only one year, which was furthermore a drought year. To address this issue of “temporal external validity", I extend the analysis to a longer time period (2013–2019) and use the UMD GLAD Global Cropland dataset (Potapov et al., 2022) to generate a new “crop mask". This dataset does not allow me to distinguish between commercial and subsistence cropland, but given that most subsistence farmers are confined to the former homelands, I can estimate heterogeneous treatment effects for homelands versus non-homelands and use it as a proxy for the differential impact on commercial and subsistence farmers.[2]

This paper contributes to the literature on impacts of large-scale irrigation in both developed and developing countries (Duflo and Pande, 2007; Zaveri, Russ and Damania, 2020; Dillon and Fishman, 2019; Hornbeck and Keskin, 2014; Strobl and Strobl, 2011; Olmstead and Sigman, 2015). Most of the previous work relies on an instrumental variable (IV) method that exploits exogeneity of geographical variables in order to predict suitability of places for dam construction. These studies have found evidence that irrigation increases crop yields, decreases dependency on rainfall, and reduces poverty rates. Recently, regression discontinuity designs have been introduced as a new empirical strategy for assessing the impact of irrigation dams an canals (Asher et al., 2022; Blakeslee et al., 2023; Jones et al., 2022; Hagerty, 2021). This approach leads to more valid estimates of treatment effects since RDD relies on weaker assumptions than IV.

This paper also expands the evidence on irrigation specifically in South Africa. Blanc and Strobl (2014) compare the performance of large versus small dams[3]. While they find that large dams have unequal distributional impacts with a positive effect on cropland downstream and a negative effect in the direct vicinity of a large dam, they also find that large dams substantially increase the positive impact of small dams. Mettetal (2019) examines the adverse environmental impacts of small, medium, and large dams. She finds that through reduced water access and increased pollution, irrigation dams increase infant mortality by 10-20 percent. Both of these previous studies employ an IV approach. This

[2] The former homelands were places where Black South African population was forcibly confined to during the apartheid. Despite the efforts of the post-apartheid governments, these areas still lag behind in terms of economic performance. Some of the irrigation canals and their command areas fall within the territory of these former homelands, and therefore, there is variation in access to irrigation even in the former homelands.

[3] Large and small dams are differentiated based on their height. Dams with a height between 5 m and 12 m are considered small, dams with a height between 12 m and 3 m are medium, and dams with height over 30 m are considered large.

is the first paper that uses RDD approach in the context of South Africa. A contribution of this paper is also the focus on the heterogeneity based on the type of farmer (commercial versus subsistence) and the former homelands status.

Finally, this paper also contributes to the literature on causal effects of agricultural productivity growth. In an early work, Foster and Rosenzweig (2004) estimate the effects of yield improvements associated with the Green Revolution in India and find that agricultural productivity increases do not spur growth of the non-agricultural sector. On the other hand, Gollin, Hansen and Wingender (2021) perform a cross-country analysis of the adoption of high-yielding varieties (HYVs) and find that the associated higher yields led to an increase in income and slower population growth. Bustos, Caprettini and Ponticelli (2016) study the adoption of yield-enhancing genetically engineered soybeans in Brazil and also find evidence of an increased industrial growth. While this paper does not speak to the issue of industrialization *per se*, I still contribute to this literature through examining the impact on the structure of the agricultural sector, which can be a prelude to a wider structural transformation.

The paper is organized as follows. In Section 2, I provide some background information on irrigation in South Africa. In Section 3, I describe the data. Section 4 presents the empirical strategy and discusses the identifying assumptions. In Section 5, I present the main results and in Section 6, I conclude.

2 Irrigation in South Africa

South Africa is a country plagued by water scarcity, yet its agriculture is heavily reliant on water. In the colonial times, the British rulers planned to establish the area as an agricultural colony, and therefore, the construction of large-scale irrigation projects started already in the pre-apartheid era as a means to alleviate the prevalent water issues (Bablin, 2021).

Nowadays, South Africa has a well-developed agricultural sector despite water scarcity, which is made possible by irrigation. Irrigation systems cover 1.3 million hectares or 7.2% of arable land (Dennis and Nell, 2002) exploited by both commercial and small-scale, subsistence farmers. This irrigated area generates 30% of the country's crop production. Irrigation is beneficial to farmers in three ways: it increases yields, allows planting multiple crops per year, and expands potentially productive area. There are four main types of irrigation: flood, sprinkler, center pivot, micro-drip, and micro-spray irrigation.[4] Flood irrigation is the least efficient type, which can lead to water loss, but it is the most viable option for smallholder farmers for whom the more technologically advanced systems might be too costly. On the other hand, center pivot irrigation is the most capital intensive option (Lichtenberg, 1989).

Irrigation in South Africa comes mainly from surface water. Groundwater irrigation is implemented on 1% of cultivated land (Altchenko and Villholth, 2015), whereas 7.2% of cultivated land is irrigated in total. This implies that approximately 14% of crops are irrigated with groundwater and 80% are irrigated with surface water. Surface water is stored in 200 large dams, as well as in many medium and small dams, and it is distributed through a network of primary and secondary canals that span over 8,000 kilometers. Figure A–1 in the Appendix shows the number of large dams constructed per decade. The first large dam with the main purpose of irrigation was built in 1913 and the majority of dams were constructed by the beginning of the 1990s. Only six large dams were completed in the 2000s.

South Africa is divided into nineteen water management areas. As per the National Water Act (1998), each water management area is supposed to be under the supervision of a

[4] https://southafrica.co.za/irrigation.html. Accessed on 30 November 2023.

catchment management agency (CMA). The roles assigned to CMAs include management of water resources by deciding on water allocations, water use licensing and monitoring of compliance, support to marginalized farmers, and monitoring of water quality and pollution. There are significant administrative delays in establishing CMAs. Only two CMAs are implemented to date, namely, the Breede-Gouritz and the Inkomati-Usutu CMAs. Payments for water use are collected by the Department of Water and Sanitation, although smallholder farmers who grow food for subsistence are exempted from these fees (Chipfupa and Wale, 2019).

3 Data

3.1 Data sources

Outcome variables. I estimate the effect of canals on two categories of outcomes: agricultural productivity and land use. Due to lack of other data sources, I use remote sensing data to derive outcome variables in both categories. The main advantage of remote sensing data is its high granularity (Donaldson and Storeygard, 2016), which allows for a precise definition of treated and untreated areas. A more traditional source of data on agricultural outcomes would be a survey. However, due to concerns of data privacy, most surveys are anonymized before they are made public, which means that it is impossible to identify which village each observation comes from. For instance, the lowest spatial disaggregation level at which the South Africa's Census of Agricultural Households identifies observations is the level of municipality. For a spatial regression discontinuity design, a higher spatial resolution is necessary. A disadvantage of using remotely sensed data, as opposed to an agricultural survey, is the inability to account for variables that would allow for a more comprehensive analysis (input costs, plot area, etc.).

Agricultural yields are proxied by Enhanced Vegetation Index (EVI) derived from Landsat 8 satellite imagery[5] which is available at a 30 meters resolution. EVI is calculated from the Near-Infra-Red (NIR), Red and Blue image bands of each scene and ranges in value from -1 to 1, where negative values indicate non-vegetated areas such as water or barren land and positive values indicate dense vegetation. EVI improves on the Normalized Difference Vegetation Index (NDVI), which is another commonly used satellite-derived proxy for agricultural yields, by adjusting for canopy background and reducing atmosphere influences (Huete et al., 2002). It is therefore more suitable for areas with high amounts of biomass, such as irrigated land. Both EVI and NDVI have been shown to be reliable proxies for agricultural yields in various geographical contexts (Lobell et al., 2020; Burke and Lobell, 2017; Asher and Novosad, 2020). South Africa has two agricultural seasons and I define each season's measure of productivity as the maximum value of EVI over the growing and harvesting stages.[6] The main crop of the summer season is maize which is grown and harvested from mid-November until mid-May. The main crop of the winter season is wheat which is grown and harvested from mid-June until the end of November.[7] South Africa also has two distinct climates in terms of when most of the rainfall occurs. West of the country is characterized by winter rainfall with smaller amounts of total precipitation, whereas the East is characterized by summer rainfall with larger amounts of precipitation.[8] This translates into distinct growing seasons for maize in the East and in the West. To ensure comparability of climate relevant to agricultural production, I focus my analysis on the Eastern part of the country.[9]

[5] Landsat 8 Collection 1 Tier 1 8-Day EVI Composite. Courtesy of the U.S. Geological Survey.

[6] I am not concerned about contamination by non-crop vegetation because I restrict the analysis to agricultural land by applying a crop mask as discussed in the following paragraph.

[7] See Figure A–3 in the Appendix.

[8] Climate Change Knowledge Portal. South Africa. Current Climate. Climatology. https://climateknowledgeportal.worldbank.org/country/south-africa/climate-data-historical. [Accessed on 31 October 2023.]

[9] I focus on areas that lie east of the 25^{th} meridian. This includes the provinces Limpopo, Mpumalanga,

Data on land use comes from the South Africa National Land Cover 2018 (SANLC 2018) and the South Africa National Land Cover 1990/2020 Change (SANLC 1990/2020) released by the South African Department of Rural Development and Land Reform (DRDLR). SANLC 2018 is a raster dataset available at the resolution of 20 meters, generated from automated mapping models using Sentinel 2 satellite imagery for the period of 1 January 2018 to 31 December 2018. SANLC 1990/2020 is based on a comparison of SANLC 1990 and SANLC 2020 datasets.[10] I use the land cover data in two ways. First, I create a crop mask and apply it to the data when analyzing agricultural productivity. This allows me to eliminate concerns about contamination of the productivity measure by non-crop vegetation. Second, I generate key outcomes to study the effect on agricultural production at the extensive margin and on land use change (expansion of land farmed by commercial farmers versus subsistence farmers).

Treatment status. To determine the treatment status I collect GPS coordinates of 144 canals from the South Africa's Department of Water and Sanitation[11] and elevation data from ALOS Digital Surface Model (Tadono et al., 2014), available at a 30 meters resolution. First, I determine the elevation relative to the sea level for each of the 144 irrigation canals. Next, I determine the elevation of each grid cell in my dataset relative to the nearest canal by taking a difference between the elevation of the canal and the elevation of the given grid cell. Areas with positive relative elevation are defined as treated, whereas areas with negative relative elevation are considered as control.

Covariates. I gather several geophysical covariates to perform balance tests and verify the validity of RD estimates. I also include these covariates into my main specification in

KwaZulu-Natal, Free State, Gauteng, Northwest, and parts of Northern Cape and Eastern Cape.

[10] The 72-class SANLC 1990 was generated from the Landsat 5 imagery for the period of 1989–1991 and is available at the resolution of 30 meters. The 73-class SANLC 2020 was generated from the Sentinel 2 imagery for the period of 1 January 2020 to 31 December 2020 and is available at the resolution of 20 meters.

[11] As explained in the previous footnote, I only focus on canals that lie east of the 25th meridian.

order to improve precision. I extract daily precipitation values from the Climate Hazards Center InfraRed Precipitation with Station (CHIRPS) dataset (Funk et al., 2015). The data is available at the resolution of 5,566 meters. I construct a precipitation measure as the total annual precipitation in each cell and then I average this value over the years 2014–2018. Next, I extract the daily land surface temperature at the resolution of 1,000 meters from the MODIS Terra Land Surface Temperature dataset (Wan, Hook and Hulley, 2021). I construct a temperature measure as the monthly maximum temperature averaged over the the years of 2014–2018. The terrain ruggedness index (TRI) comes from Nunn and Puga (2012). It is a measure of topographic variability within a given area that quantifies the variation in elevation between neighboring cells in a digital elevation model. Finally, I calculate distance to the nearest river using WWF HydroSHEDS dataset (Grill et al., 2019).

Homelands. A shapefile with boundaries of of former homelands is obtained from the Department of Agriculture, Land Reform and Rural Development (DALRRD).

3.2 Summary statistics

The unit of analysis is a 30 meters by 30 meters grid cell. The grid cells are obtained by converting a geospatial raster dataset with all the above described variables into a delimited text format. The conversion is limited to areas within 10 km of each canal. This results in 56,465,817 observations. The regression discontinuity (RD) sample further restricts the dataset to observations that are within 50 meters of relative elevation to the nearest canal and excluding a 3 meter donut hole. This results in 27,340,027 observations, out of which 6,363,358 grid cells are classified as agricultural land. Summary statistics for the full sample and the RD sample are presented in Table 1.

Restricting the sample to a narrow bandwidth of 50 meters relative elevation does not result in large differences compared to the full sample, although the land that lies within

50 meters of relative elevation to the nearest canal is more likely to be an agricultural field. Furthermore, columns (2)–(5) present summary statistics disaggregated by the treatment and former homelands status. Treated areas (those lying below canals) have significantly higher values of EVI, they are more likely to consist of agricultural land and grow commercial irrigated annual crops. They are less likely to grow commercial rainfed annuals, be cultivated by subsistence farmers, or lie fallow. There are also significant differences in terms of geophysical covariates. Treated areas have slightly lower maximum monthly temperature (36.2 degrees versus 35.4 degrees) and slightly higher mean annual precipitation (523 mm versus 555 mm). They are also on average closer to the nearest canal (5,283 m versus 5,086 m) and to the nearest river (1,120 m versus 840 m) and have lower elevation above the sea level (1,101 m versus 1,048 m). Although these differences are significant, they do not pose a concern for the RD analysis since the RD assumptions demand an absence of discontinuity at the threshold. The test of these assumptions is presented in Table 2.

Former homelands cover 9% of the RD sample. They tend to be slightly more agricultural than non-homelands (23.2% versus 24.4%), but consist almost exclusively of land farmed by subsistence farmers (68%) or fallow land (30%). Only 1% of homelands agricultural land is exploited by commercial farmers. In terms of geophysical characteristics, homelands tend to have higher maximum monthly temperature (35.9 degrees versus 36.5 degrees), higher mean annual rainfall (528 mm versus 579 mm), they are on average further away from the canals (5,129 m versus 6,154 m) and slightly closer to rivers (1,035 m versus 1,28 m). They also lie at a much lower elevation above the sea level (1,109 m versus 846 m).

4 Empirical Strategy

4.1 Previous literature

The main challenge in estimating the impact of irrigation canals is their endogenous placement. For instance, in order to maximize the return on investment, irrigation dams (and the corresponding network of irrigation canals) might be placed in areas with better agricultural potential. Alternatively, irrigation dams might be constructed in politically favored places, which could be correlated with the provision of other goods and services correlated with agricultural yields. A related issue is the lack of historical data for the period before canals were constructed, which prevents the use of panel data methods. Most canals were build in the 1960s and 1970s and the Landsat 8 EVI measures are available only from 2013.[12]

The previous literature employed two main strategies to deal with this endogeneity issue. Early studies relied on an instrumental variable approach. Duflo and Pande (2007) use river gradient which reflects the geographic suitability to predict the distribution of irrigation dams across districts in India. A similar approach was previously employed also in the context of Sub-Saharan Africa. In addition to the river gradient, Strobl and Strobl (2011) distinguish between ephemeral and perennial rivers to predict the distribution of dams across the continent. Ephemeral rivers are considered less suitable for dam construction.

The second empirical strategy is regression discontinuity (RD) design and has been made possible only recently with the availability of high-resolution data on crop yields. The running variable is either distance to the command area boundary (Blakeslee et al., 2023; Jones et al., 2022; Hagerty, 2021) or relative elevation to the nearest canal (Asher

[12] Lansat 7 NDVI measures go a bit further back in time. They are available from 2000.

et al., 2022). The main advantage of an RDD is that it produces causal estimates under weaker assumptions than IV approach. The only assumption needed for internal validity of RD estimates is continuity at the threshold of all the observable and unobservable characteristics that could be correlated with the outcome. On the other hand, IV method requires the assumptions of relevance, exogeneity, and exclusion restriction, which are more likely to be violated. For instance, the instrument used in Duflo and Pande (2007) and subsequent studies (river gradient) might be affecting yields directly and not only through the channel of increasing the probability of dam presence. Lower river gradient makes the area more suitable for construction of a dam. However, lower river gradient is also mechanically correlated with the land gradient, which makes the area more suitable for agriculture and also leads to higher yields. This would imply a violation of the exclusion restriction. Furthermore, the IV approach only allows to estimate the local average treatment effect (LATE), that is, the impact of dams that were built because of favorable geophysical characteristics (the so-called compliers). This approach cannot say anything about dams being built for other reasons, which also makes the RDD more attractive.[13]

In this paper I use an RDD with relative elevation to the nearest canal as the running variable. A visual representation of the elevation-based RDD is shown in Figure A–2 in the Appendix. My main specification consists of regressing outcome variables on a treatment indicator and a linear function of the running variable while allowing for different slopes in the treatment and the control groups.

4.2 Regression Discontinuity Design

My empirical strategy relies on the fact that surface irrigation is gravity-based. Water from large irrigation dams is distributed through a network of main and secondary canals that

[13] Although it should be noted that RDD also yields local effects in the sense that the effects are only estimated for observations that are within the selected bandwidth of the running variable.

carry water from uphill areas toward downhill areas. It is technically possible to pump water uphill from nearby canals, and therefore, the running variable does not perfectly determine the treatment. However, land below canals still has a higher probability of being irrigated, which allows for a fuzzy RD design. Figure 1 shows the probability of land being irrigated as a function of the relative elevation to the nearest canal both for the agricultural and full samples.

The main fuzzy RD specification estimates the local average treatment effect (LATE) of irrigation canals on the agricultural productivity and land use outcomes for areas just below the canals. Following (Imbens and Lemieux, 2008; Gelman and Imbens, 2019), I regress each outcome on the treatment indicator (whether an observation lies below the canal) while controlling linearly for the running variable (relative elevation to the nearest canal) separately on each side of the threshold:

$$\begin{aligned} Y_{id} =& \beta_0 + \beta_1 Treat_{id} + \beta_2 Rel\_Elev_{id} + \beta_3 Rel\_Elev_{id} \times Treat_{id} \\ & + \beta_4 X_{id} + \mu_d + \epsilon_{id} \end{aligned} \tag{1}$$

Y_{id} is the outcome of interest in grid cell i and district d. $Treat_{id}$ is the treatment indicator that is equal to one when a grid cell lies below a canal, which is determined based on its elevation relative to the nearest canal. The running variable Rel_Elev_{id} is calculated as the elevation of the nearest canal minus the elevation of a given grid cell.[14] Observations with positive values of Rel_Elev_{id} lie below a canal (treatment group) and observations with negative values lie above a canal (control group). I interact the running

[14] Elevation of each grid cell is determined using the ALOS Digital Surface Model data. The dataset is available at the resolution of 30 meters, which is the same as the resolution of the EVI outcome variable and similar to the resolution of the land use outcome variables (20 meters). The values from the ALOS raster data are thus simply matched to the raster data containing the outcomes. This considerably improves accuracy of the analysis. For instance, Asher et al. (2022) need to determine the elevation of each polygon (village) in their dataset. They do so by taking the 5th percentile of the elevation distribution of the pixels constituting the polygon. However, it is likely that some parts of these polygons (villages) lie above the canal even though they are coded as treated.

variable Rel_Elev_{id} with the treatment dummy to allow for different slopes at each side of the threshold.

X_{id} refers to three geophysical covariates that I control for in order to improve precision of the RD estimates: average maximum monthly temperature, average total annual precipitation, terrain ruggedness index, distance to the nearest canal, and distance to the nearest river. Finally, I include district fixed effects μ_d to account for unobserved, time-invariant characteristics that may vary across districts, such as differences in agricultural policy, institutions or infrastructure, that may affect the outcome variable.

Standard errors are clustered at the canal level since it is the unit at which the treatment is assigned (Abadie et al., 2023). This also allows for arbitrary spatial correlation in places around each canal. To ensure comparability of treatment and control units, I restrict the sample to grid cells that lie within 10 kilometers of distance and 50 meters of relative elevation to the nearest canal. There might also be some ambiguity in the treatment status for areas that are very close to the threshold of zero relative elevation to the nearest canal. There are two reasons for this ambiguity. First, there might be some measurement error, either in the implied elevation of a canal, or in the elevation of a given grid cell, and this measurement error affects the determination of the treatment status for places very close to the threshold. Second, as mentioned earlier, farmers might be pumping water from canals uphill. This would be much more feasible in areas that are close to the threshold of zero relative elevation (for example, one meter above a canal) than for areas further uphill (for example, ten meters above a canal). Including areas close to the threshold would bias the RD estimates towards zero. Therefore, I exclude a donut hole of grid cells that lie less than 3 meters below or less than 3 meters above the nearest canal[15].

[15] By doing so, I follow the same bandwidth restriction on the relative elevation as Asher et al. (2022)

4.3 Identifying Assumptions

In order to interpret the RD estimates as causal effects, three key assumptions must be satisfied. First, treatment must be at least partially determined by the running variable. In other words, land below a canal must have a higher probability of being irrigated, while the land above a canal must have a lower probability of being irrigated. Figure 1 shows the probability of land being irrigated as a function of the running variable for agricultural and full samples. There is an apparent discontinuity of the irrigation probability around the threshold but it is not very large in magnitude, particularly for the agricultural sample. This is due to the fact that although it is costly, some commercial farmers are able to pump irrigation water from the canals uphill. This implies that my RD specification follows a fuzzy design.

Second, all the observable and unobservable variables that also determine the outcome must be continuous at the threshold. In other words, there should be no sudden discontinuity in any characteristics (weather variables, topography, etc.) between places that are just above and just below a canal. The only source of discontinuity must be access to irrigation through the canals. Note that the purpose is to estimate the *long-run* effect of irrigation canals. In the long-run, farmers might adjust their input use to take into account the increase in water supply due to irrigation (Hagerty, 2021). Therefore, a discontinuity at the threshold in, for instance, the use of fertilizer would not be a threat to identification, but rather one of the components of the long-run effect. In order to better articulate this concept, suppose that crop yields $Y(W, I)$ are a function of water supply W and other inputs I (fertilizer, pesticides, labor, etc.). We assume that in the long-run, inputs are adjusted to the level of water supply. The long-run effect of water supply on crop yields is then given by the total derivative of $Y(W, I)$ with respect to W which consists of a direct effect and an indirect effect:

$$\frac{dY(W,I)}{dW} = \underbrace{\frac{\partial Y(W,I)}{\partial W}}_{\text{direct effect}} + \underbrace{\frac{\partial Y}{\partial I}\frac{dI(W)}{dW}}_{\text{indirect effect}} \tag{2}$$

In this paper, I estimate the total effect which includes both the direct and indirect effects.

I verify the continuity assumption for five observed geophysical characteristics: average maximum monthly temperature, average total annual precipitation, terrain ruggedness index, distance to the nearest canal, and distance to the nearest river. Figure 2 shows a binned scatter plot for the four variables with fitted linear regression lines at each side of the cutoff. There seem to be points of inflection around the threshold for several characteristics but there is no apparent discontinuity. Table 2 reports statistical tests of the presence of discontinuity at the threshold. The estimates are obtained by estimating equation 1 on the sample of grid cells within 10 kilometers of distance and within 50 meters of relative elevation to the nearest canal, excluding a 3 meters donut hole. I conduct the tests on both the agricultural and the full samples. For the agricultural sample, none of the coefficients are statistically significant nor large in magnitude when compared to the means in the control group. For the full sample which includes both agricultural and non-agricultural land, I find that the treated areas are further away from the nearest canal and from the nearest river but the magnitude of these discontinuities is not very large. I control for all the geophysical covariates in the main analysis.

5 Results

5.1 Agricultural seasons in South Africa

South Africa has two distinct agricultural seasons during which winter and summer crops are grown. Winter crops are typically planted during the autumn months of April and May, and then harvested during the winter and early spring months of September to November. Wheat is the most important winter crop with a production of 2,263,000 tonnes during the 2021/2022 season (SAGIS, Monthly Producer Deliveries). Most of the production is centered in Western Cape. Summer crops are typically planted in the early summer months of October to December, and are harvested during the late summer and early autumn months of February to April. The most important summer crop is maize with a production of 15,810,000 tonnes during the 2021/2022 season (SAGIS, Monthly Producer Deliveries). Maize is produced in almost all parts of South Africa, but the majority of production is concentrated in the Free State, Mpumalanga, North West, and KwaZulu-Natal provinces. The planting, growing, and harvesting months are summarized in Figure A–3 in the Appendix. I use the temporal ranges of growing and harvesting seasons of wheat and maize (east) to construct the EVI measures of agricultural productivity for the wheat and maize seasons.

5.2 Treatment effects on the intensive margin

I first examine the effects of irrigation canals on the intensive margin of agricultural production. The results of estimating equation 1 are shown in Table 3. I restrict the analysis to only agricultural land to ensure that my results are not picking up any non-crop vegetation. I find that irrigation canals increase agricultural productivity by 4% both the wheat season and the maize season (columns (3) and (4)), however, the effect is precisely estimated only for the maize season. I find qualitatively the same results without performing the log-transformation of EVI (columns (1) and (2)). Note that these are reduced-form estimates that are not scaled by the change in probability of irrigated land at either side of the threshold.

5.3 Treatment effects on the extensive margin

Next, I consider the question of how irrigation canals affect the extensive margin of agricultural production. In other words, do irrigation canals increase cultivated area? To answer this question, I estimate equation 1 on the full RD sample (*not* restricting to agricultural land only). The results are reported in Table 4. Areas below canals have a 30% higher share of agricultural land (increase of 6.3 pp). This increase is driven by an increase in land currently cultivated with annual crops (increase of 41% or 6.8 pp). Areas under the canals are 12% (0.6 pp) less likely to consist of fallow land.

5.4 Treatment effects on the structure of the agricultural sector

In the previous two subsections I established that irrigation canals lead to a long-term positive agricultural productivity shock because they increase crop yields as measured by EVI. I also established that irrigation is a land-augmenting technical change.[16] Irrigation makes land more productive than it would have been under rainfed conditions, which leads farmers to expand area under production and decrease the area of fallowed land.[17] Now I examine the broader implications of these two findings. How does a persistent, positive

[16] I borrow this term from Bustos, Caprettini and Ponticelli (2016) who use it to describe the introduction of a second harvesting season for maize in Brazil. This constitutes a land-augmenting technical change because it makes one unit of land on average more productive. They contrast it to labor-augmenting technical change (introduction of a genetically engineered soy variety) which makes one unit of labor on average more productive.

[17] Fallow land is land that is deliberately left uncultivated for a certain period. The purpose of this practice is to restore soil nutrients that are depleted through crop cultivation. Fallowing land thus increases future land productivity.

productivity shock affect the structure of the agricultural sector? In particular, are both the commercial and subsistence farmers able to benefit from the shock?

To answer this question I study how better access to irrigation affects commercial and subsistence farmers both at the intensive and the extensive margins. First, I estimate the RD effects of canals on maize season and wheat season EVI separately for land classified in SANLC 2018 as commercial annuals pivot irrigated, commercial annuals non-pivot irrigated, commercial annuals rainfed/non-irrigated, and subsistence annuals.[18] The results are reported in Table 5.

Panels A and B show the effect of being below a canal on land classified as commercial irrigated (pivot or non-pivot). As expected, the estimated effect on agricultural productivity is zero since all the grid cells in these two samples are by definition treated. In other words, areas on both sides of the zero relative elevation threshold have access to irrigation.

Panel C of Table 5 shows the results for land classified as commercial rainfed, which is the most common category in my agricultural sample. By similar logic, we would expect a null effect of being below a canal since the grid cells are classified as non-irrigated by SANLC 2018. However, I find positive and statistically significant effects of 3.9% for wheat yields and 4.6% for maize yields. Therefore, commercial farmers in areas below a canal seem to have better quality land than commercial farmers in areas above a canal. One possible explanation is that commercial farmers are acquiring high-quality land in potentially irrigable areas with the expectation of making an investment into an irrigation system in the future. Note that the possibility that the observed effects are due to a misclassification of irrigated land into non-irrigated land category in the SANLC data is very unlikely.[19]

[18] A center pivot irrigation system uses rotating sprinklers mounted on wheeled towers to irrigate a circular area of farmland. The circular shape of the fields is what allows an automated mapping algorithm to distinguish between pivot and non-pivot irrigated crops. Non-pivot irrigation systems in South Africa include flood irrigation, sprinkler irrigation, micro-drip irrigation, and micro-spray irrigation.

[19] SANLC 2018 used 6,570 reference points for accuracy assessment. For the commercial non-irrigated category, the user's accuracy was 92.44% and the producer's accuracy was 96.95%. This means that 92.44%

Panel D of Table 5 shows the canal effects on land farmed by subsistence farmers. I find a negative and statistically significant effect of 4% on maize EVI, and a negative insignificant effect of 4.2% on wheat EVI. Therefore, subsistence farmers in (potentially irrigable) areas below a canal perform worse than subsistence farmers in areas above a canal. A possible explanation is that subsistence farmers below canals adopt water-intensive varieties as a result of having access to irrigation, but due to lack of maintenance of the irrigation infrastructure or low priority within the water distribution network, they face uncertain levels of water supplies, which negatively impacts their yields.

Second, I consider the question of whether access to irrigation leads to an expansion of the agricultural sector, and if so, whether commercial and subsistence farmers are affected differently. Table 6 shows the RD results for changes in land use between 1990 and 2020 derived from the SANLC 1990/2020 dataset. The effects are estimated for three outcomes: expansion of commercial pivot irrigated land, expansion of commercial non-pivot irrigated or non-irrigated land[20], and expansion of subsistence land. I find positive and statistically significant effects on the expansion of commercial land but virtually a zero effect on the expansion of land farmed by subsistence farmers. In the control group (areas above a canal), 2.9% of grid cells became commercial pivot-irrigated between 1990 and 2020. In treated areas, the conversion toward commercial pivot-irrigated farming was 66% higher (1.9 pp). Similarly, 2.1% of grid cells in the control group became commercial non-pivot between 1990 and 2020, but this fraction was 29% (0.6 pp) higher in the treatment group. Irrigation canals, therefore, contribute to the expansion of the commercial agriculture in South Africa.

of reference sites that were classified as commercial non-irrigated were in fact commercial non-irrigated and 96.95% of commercial non-irrigated reference sites were classified correctly.

[20] These two categories of commercial land are lumped together because SANLC 1990 data cannot distinguish between irrigated and non-irrigated land without the circular shape of the fields. As explained earlier, center pivot irrigation results in easily distinguishable circular patterns.

5.5 Heterogeneous treatment effects: Former homelands and drought

A possible concern for my analysis of agricultural productivity is “temporal" external validity since I only use data for one year. Maybe the negative effects of irrigation canals on the subsistence farmers’ yields would not be observed in other periods. Indeed, 2018 was a drought year that affected the whole country. It could be that the subsistence farmers situated below canals are negatively impacted only during droughts.[21] To address this issue, I extend the analysis to multiple years.

Instead of SANLC 2018, which covers only one year, I use the Global Cropland datasets produced by the University of Maryland Global Land Analysis & Discovery (UMD GLAD) for the period of 2013–2019 (Potapov et al., 2022) to generate a new crop mask. In particular, I use two of the UMD GLAD datasets. First, UMD GLAD Global Cropland 2015 classifies a grid cell as cropland if active crop was detected any time between 2012 and 2015. Second, UMD GLAD Global Cropland 2015 classifies a grid cell as cropland if active crop was detected any time between 2016 and 2019. When I compare the crop masks generated according to UMD GLAD 2019 and SANLC 2018, I find them mostly consistent with each other. For non-homelands, 87% of grid cells classified as cropland by UMD GLAD 2019 are also classified as agricultural in my previous analysis using SANLC 2018. For homelands, only 37% of grid cells classified as cropland by UMD GLAD 2019 are also classified as agricultural in my previous analysis. This is mostly due to the presence of sugarcane, which is correctly detected as cropland by UMD GLAD, but which I do not include in my previous analysis where I focus only on annual crops. For a more detailed picture of how SANLC and UMD GLAD datasets compare, Table A–2 in the Appendix shows the breakdown of NLC categories for both homeland and non-homeland detected cropland.

[21] In the period of droughts, there is less water available in the irrigation system than usual. Commercial farmers could be prioritized over subsistence farmers in getting access to the limited water supplies during droughts, which could lead to lower yields for subsistence farmers below canals relative to those above canals who usually do not have access to any water from the canal networks.

UMD GLAD datasets only detect cropland and do not allow me to distinguish between commercial and subsistence farming. However, I find that 95% of subsistence land in my sample is concentrated in the former homelands,[22] which allows me to consider former homelands as a proxy for subsistence farmland.

In addition to heterogeneity by the former homelands status, I also exploit heterogeneity by the incidence of drought to evaluate whether the canals alleviate the negative impact of droughts. I derive a normalized measure of drought based on a coarse grid-level Palmer Drought Severity Index (PDSI) derived from the University of Idaho's TerraClimate dataset.[23] The PDSI values in my sample range from -5.9 to 4.2 with positive values representing wet conditions and negative values representing dry conditions. Following Hornbeck and Keskin (2014), I set the index to zero for wet years and normalize it to have a mean of zero and a standard deviation of one. I also derive four different measures of drought. I calculate the average PDSI over the planting and growing months for wheat and maize separately. Drought that occurs at the planting stage might have different consequences for the crops than drought with onset during the growing stage. Table A–1 in the Appendix shows that in the control areas (above canals), drought occurring at the growing stage depresses crop yields more than drought during the planting stage. Moreover, wheat appears to be more drought-sensitive than maize.

The new empirical specification uses variation in the former homelands status and in the normalized PDSI calculated for the planting and growing stages to estimate the differential impact of irrigation canals during droughts on crop yields separately in the former homelands and non-homelands. The maize season and wheat season log crop yields Y in grid cell i, district d, and year t are regressed on the treatment indicator,

[22] Moreover, the agricultural land in the former homelands in my sample consists of 68% subsistence farmland, 31% fallow land, and only 1% commercial farmland.

[23] PDSI is available at a resolution of 4638.3 meters, which is significantly coarser than the resolution of my outcome variable (30 m).

the running variable (Rel_Elev_{id}), the normalized PDSI at planting or growing stages ($Drought_{idt}$), and the relevant interaction terms between treatment and the running variable ($Treat_{id} \times Rel\_Elev_{id}$), drought and treatment ($Drought_{idt} \times Treat_{id}$), drought and the running variable ($Drought_{idt} \times Rel\_Elev_{id}$), and drought, treatment and running variable ($Drought_{idt} \times Treat_{id} \times Rel\_Elev_{id}$). The regression is run separately for homelands and non-homelands and separately for drought at the planting and the growing stages. Geophysical covariates (X_{idt}), district fixed effects (μ_d), and year fixed effects (γ_t) are included as well. Standard errors are clustered at the canal level to account for potential spatial correlation. The empirical specification is:

$$\begin{aligned} Y_{idt} =& \alpha_0 + \alpha_1 Treat_{id} + \alpha_2 Rel\_Elev_{id} + \alpha_3 Treat_{id} \times Rel\_Elev_{id} + \alpha_4 Drought_{idt} \\ & + \alpha_5 Drought_{idt} \times Treat_{id} + \alpha_6 Drought_{id} \times Rel\_Elev_{id} + \\ & \alpha_7 Drought_{idt} \times Treat_{id} \times Rel\_Elev_{id} + \alpha_8 X_{idt} + \mu_d + \gamma_t + \epsilon_{idt} \end{aligned} \tag{3}$$

The effect of irrigation canals on crop yields in non-drought conditions is captured by α_1 and reflects variation spanning multiple years (2013–2019). The impact of an increase of one standard deviation in the drought index on the canal treatment effect is captured by α_5. The results of estimating the equation 3 are reported in Table 7. In non-homelands, I find that irrigation canals increase yields in wet conditions. Moreover, treatment effects are even higher in the dry conditions, although most of the coefficients are imprecisely estimated. In non-drought conditions canals increase wheat yields by 2.8% and maize yields by 1.5%. An increase of one standard deviation in the drought index further increases the effect of canals. The treatment effect becomes 4.7% for wheat and 2.2% for maize.

More importantly, I still find a negative treatment effect of irrigation canals in the former homelands, where subsistence farmers are concentrated, even in non-drought conditions. Wheat yields in areas below canals decline by 6.1 to 6.6% and maize yields decline by 3.6 to 5.9%. Drought conditions further worsen the negative impact of droughts for wheat, although this is imprecisely estimated. In case of maize, however, I find that canals alleviate the adverse impact of droughts in the former homelands.

6 Discussion and concluding remarks

I find that irrigation canals cause a positive agricultural productivity shock, which contributes to the expansion of commercial farming in South Africa. However, subsistence farmers seem largely unaffected and unable to benefit from potential productivity gains. This is not necessarily a negative outcome in the light of structural transformation which is often considered a key component in the process of economic development.

Structural transformation involves a shift from an agriculture-based economy toward an economy based on manufacturing and services and commercialization of agriculture can help spur such change (Suri and Udry, 2022). For instance, a transition from subsistence farming, where smallholder farmers grow crops for own consumption, towards commercial farming, where large-scale farmers grow crops for sale, can provide enough and sufficiently cheap food surpluses to support the growing labor force in the manufacturing sector. Furthermore, smallholder farms are often plagued by inefficiencies and low productivity. As Collier and Dercon (2014) remark, the organization of the agricultural sector in Africa has to change significantly to enable economic development, and such re-organization might be at odds with the primary focus on increasing productivity of smallholder farmers as a means for poverty reduction. Commercialization, and the associated increases in farm size could therefore be a step in the right direction. The typical argument favoring the focus by donors and practitioners on smallholder farmers over large-scale producers involves the “inverse farm size/productivity" relationship (van Zyl, Binswanger and Thirtle, 1995) but more recent works question these conclusions (Foster and Rosenzweig, 2022). In Sub-Saharan Africa, 70–80% of farms are smaller than 2 ha, although a recent rise in farm sizes has been documented (Jayne et al., 2016). However, little is known of what factors play a role in this changing structure.

One limitation of this paper is the inability to make any welfare statements. Crop yields do not represent farmer profits nor wages of agricultural workers. Also, it is unclear whether commercialization of agriculture is welfare-improving. One way in which the rise of commercial agriculture could improve living standards is through inducing rural-urban migration and/or industrialization. Subsistence farmers could thus move towards more productive sectors and earn higher wages. I plan to examine the effects of commercialization of agriculture on employment patterns, and other economic outcomes in my future research.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?*." *The Quarterly Journal of Economics*, 138(1): 1–35.

Altchenko, Y., and K. G. Villholth. 2015. "Mapping irrigation potential from renewable groundwater in Africa – a quantitative hydrological approach." *Hydrology and Earth System Sciences*, 19(2): 1055–1067.

Asher, Sam, Alison Campion, Douglas Gollin, and Paul Novosad. 2022. "The Long-Run Development Impacts of Agricultural Productivity Gains: Evidence from Irrigation Canals in India." *STEG WP004*.

Asher, Sam, and Paul Novosad. 2020. "Rural Roads and Local Economic Development." *American Economic Review*, 110(3): 797–823.

Bablin, Elisabeth. 2021. "South African Water History." *Proceedings of GREAT Day*, 2020(4).

Blakeslee, David, Aaditya Dar, Ram Fishman, Samreen Malik, Heitor S. Pellegrina, and Karan Singh Bagavathinathan. 2023. "Irrigation and the spatial pattern of local economic development in India." *Journal of Development Economics*, 161: 102997.

Blanc, Elodie, and Eric Strobl. 2014. "Is Small Better? A Comparison of the Effect of Large and Small Dams on Cropland Productivity in South Africa." *The World Bank Economic Review*, 28(3): 545–576.

Burke, Marshall, and David B. Lobell. 2017. "Satellite-based assessment of yield variation and its determinants in smallholder African systems." *Proceedings of the National Academy of Sciences*, 114(9): 2189–2194.

Bustos, Paula, Bruno Caprettini, and Jacopo Ponticelli. 2016. "Agricultural Productivity and Structural Transformation: Evidence from Brazil." *American Economic Review*, 106(6): 1320–65.

Chipfupa, U, and E Wale. 2019. "Smallholder willingness to pay and preferences in the way irrigation water should be managed: a choice experiment application in KwaZulu-Natal, South Africa." *Water SA*, 45: 383 – 392.

Collier, Paul, and Stefan Dercon. 2014. "African Agriculture in 50Years: Smallholders in a Rapidly Changing World?" *World Development*, 63: 92–101. Economic Transformation in Africa.

Dennis, H. J., and W. T. Nell. 2002. “Precision irrigation in South Africa.” *Paper prepared for presentation at the 13th International Farm Management Congress, Wageningen, The Netherlands, July 7-12, 2002*.

Dillon, Andrew, and Ram Fishman. 2019. “Dams: Effects of Hydrological Infrastructure on Development.” *Annual Review of Resource Economics*, 11(1): 125–148.

Donaldson, Dave, and Adam Storeygard. 2016. “The View from Above: Applications of Satellite Data in Economics.” *Journal of Economic Perspectives,* 30(4): 171–98.

Duflo, Esther, and Rohini Pande. 2007. “Dams.” *The Quarterly Journal of Economics*, 122(2): 601–646.

Foster, Andrew D., and Mark R. Rosenzweig. 2004. “Agricultural Productivity Growth, Rural Economic Diversity, and Economic Reforms: India, 1970–2000.” *Economic Development and Cultural Change*, 52(3): 509–542.

Foster, Andrew D., and Mark R. Rosenzweig. 2022. “Are There Too Many Farms in the World? Labor Market Transaction Costs, Machine Capacities, and Optimal Farm Size.” *Journal of Political Economy*, 130(3): 636–680.

Funk, Chris, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, and Andrew Hoell and Joel Michaelsen. 2015. “The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes.” *Scientific Data 2, 150066*.

Gelman, Andrew, and Guido Imbens. 2019. “Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs.” *Journal of Business & Economic Statistics*, 37(3): 447–456.

Gollin, Douglas, Casper Worm Hansen, and Asger Mose Wingender. 2021. “Two Blades of Grass: The Impact of the Green Revolution.” *Journal of Political Econonmy*, 129(8): 2344–2384.

Grill, G., B. Lehner, M. Thieme, B. Geenen, D. Tickner, F. Antonelli, S. Babu, P. Borrelli, L. Cheng, H. Crochetiere, and H.E. Macedo. 2019. “Mapping the world’s free-flowing rivers.” *Nature*, 569(7755): 215.

Hagerty, Nick. 2021. “Adaptation to Surface Water Scarcity in Irrigated Agriculture.” *Working Paper*.

Hornbeck, Richard, and Pinar Keskin. 2014. “The Historically Evolving Impact of the Ogallala Aquifer: Agricultural Adaptation to Groundwater and Drought.” *American Economic Journal: Applied Economics*, 6(1): 190–219.

Huete, A, K Didan, T Miura, E.P Rodriguez, X Gao, and L.G Ferreira. 2002. “Overview of the radiometric and biophysical performance of the MODIS vegetation indices.” *Remote Sensing of Environment*, 83(1): 195–213. The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring.

Imbens, Guido W., and Thomas Lemieux. 2008. “Regression discontinuity designs: A guide to practice.” *Journal of Econometrics*, 142(2): 615–635. The regression discontinuity design: Theory and applications.

Jayne, T.S., Jordan Chamberlin, Lulama Traub, Nicholas Sitko, Milu Muyanga, Felix K. Yeboah, Ward Anseeuw, Antony Chapoto, Ayala Wineman, Chewe Nkonde, and Richard Kachule. 2016. “Africa’s changing farm size distribution patterns: the rise of medium-scale farms.” *Agricultural Economics*, 47(S1): 197–214.

Jones, Maria, Florence Kondylis, John Loeser, and Jeremy Magruder. 2022. “Factor Market Failures and the Adoption of Irrigation in Rwanda.” *American Economic Review*, 112(7): 2316–52.

Kuznets, Simon. 1957. “Quantitative aspects of the Economic Growth of Nations: II: Industrial Distribution of National Product and Labor Force.” *Economic Development and Cultural Change*, 5(4): 1–111.

Lewis, W. A. 1954. “Economic Development with Unlimited Supplies of Labour.” *The Manchester School*.

Lichtenberg, Erik. 1989. “Land Quality, Irrigation Development, and Cropping Patterns in the Northern High Plains.” *American Journal of Agricultural Economics*, 71(1): 187–194.

Lobell, David B, George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic, and Siobhan Murray. 2020. “Eyes in the Sky, Boots on the Ground: Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis.” *American Journal of Agricultural Economics*, 102(1): 202–219.

Lowder, Sarah K., Jakob Skoet, and Terri Raney. 2016. “The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide.” *World Development*, 87: 16–29.

Mettetal, Elizabeth. 2019. “Irrigation dams, water and infant mortality: Evidence from South Africa.” *Journal of Development Economics*, 138: 17–40.

Nunn, Nathan, and Diego Puga. 2012. “Ruggedness: The Blessing of Bad Geography in Africa.” *Review of Economics and Statistics*, 94(1): 20–36.

Olmstead, Sheila M., and Hilary Sigman. 2015. “Droughts, dams, and economic activity.” *Working Paper*.

Potapov, Peter, Svetlana Turubanova, Matthew C. Hansen, Alexandra Tyukavina, Viviana Zalles, Ahmad Khan, Xiao-Peng Song, Amy Pickens, Quan Shen, and Jocelyn Cortez. 2022. "lobal maps of cropland extent and change show accelerated cropland expansion in the twenty-first century." *Nature Food*, 3: 19–28.

Ranis, Gustav, and John C. H. Fei. 1961. "A Theory of Economic Development." *The American Economic Review*, 51(4): 533–565.

Sinyolo, Sikhulumile, Maxwell Mudhara, and Edilegnaw Wale. 2014. "The impact of smallholder irrigation on household welfare: The case of Tugela Ferry irrigation scheme in KwaZulu-Natal, South Africa." *Water SA*, 40(1): 145–55.

Strobl, Eric, and Robert O. Strobl. 2011. "The distributional impact of large dams: Evidence from cropland productivity in Africa." *Journal of Development Economics*, 96(2): 432–450.

Suri, Tavneet, and Christopher Udry. 2022. "Agricultural Technology in Africa." *Journal of Economic Perspectives*, 36(1): 33–56.

Tadono, T., H. Ishida, F. Oda, S. Naito, K. Minakawa, and H. Iwamoto. 2014. "Precise Global DEM Generation By ALOS PRISM." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-4: 71–76.

Tatlhego, Mokganedi, Davide Danilo Chiarelli, Maria Cristina Rulli, and Paolo D'Odorico. 2022. "The value generated by irrigation in the command areas of new agricultural dams in Africa." *Agricultural Water Management*, 264: 107517.

van Zyl, Johan, Hans Binswanger, and Colin Thirtle. 1995. "The Relationship Between Farm Size and Efficiency in South African Agriculture." *World Bak Policy Research Working Paper 1548*.

Wan, Z., S. Hook, and G. Hulley. 2021. "MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061." Distributed by NASA EOSDIS Land Processes Distributed Active Archive Center.

Wardlow, Brian D., and Stephen L. Egbert. 2010. "A comparison of MODIS 250-m EVI and NDVI data for crop mapping: a case study for southwest Kansas." *International Journal of Remote Sensing*, 31(3): 805–830.

Zaveri, Esha, Jason Russ, and Richard Damania. 2020. "Rainfall anomalies are a significant driver of cropland expansion." *Proceedings of the National Academy of Sciences*, 117(19): 10225–10233.

Tables

Table 1: Summary statistics

		RD sample			
	Full sample mean	Control mean	Treatment vs. control diff.	Non-homeland mean	Homeland vs. non-homeland diff
	(1)	(2)	(3)	(4)	(5)
Ag. productivity					
Wheat season EVI	0.319	0.288	0.063***	0.308	0.000**
			(0.000)		(0.000)
Maize season EVI	0.496	0.476	0.072***	0.496	0.017***
			(0.000)		(0.000)
Wheat season EVI (log)	-1.241	-1.337	0.167***	-1.289	0.030***
			(0.000)		(0.000)
Maize season EVI (log)	-0.766	-0.806	0.140***	-0.768	0.053***
			(0.000)		(0.000)
As share of total area:					
Agri land	0.167	0.213	0.064***	0.232	0.012***
			(0.000)		(0.000)
As share of agri land:					
Commercial pivot irrig.	0.158	0.146	0.099***	0.201	-0.200***
			(0.000)		(0.001)
Commercial non-pivot irrig.	0.032	0.027	0.031***	0.042	-0.042***
			(0.000)		(0.000)
Commercial rainfed	0.51	0.530	-0.047***	0.566	-0.557***
			(0.000)		(0.001)
Subsistence	0.089	0.077	-0.021***	0.004	0.680***
			(0.000)		(0.000)
Fallow land	0.212	0.221	-0.062***	0.187	0.119***
			(0.000)		(0.001)
Geo. controls:					
Terrain Ruggedness Index	2.06	1.21	0.04***	1.19	0.32***
			(0.000)		(0.001)
Max monthly temperature	35.2	36.2	-0.822***	35.9	0.612***
			(0.001)		(0.002)
Mean annual precipitation	565	523	31.5***	528	50.8***
			(0.065)		(0.104)
Distance to nearest canal	5833	5283	-197***	5129	1025***
			(1.062)		(1.690)
Distance to nearest river	1329	1120	-280***	1035	-6.79***
			(0.338)		(0.548)
Elevation	1077	1101	-52.7***	1109	-263***
			(0.162)		(0.255)
Number of obs.	56,465,817	18,959,806	27,340,027	24,844,695	27,340,027

Note: This table shows summary statistics for the main outcomes and the control variables for different samples of the data. Column (1) includes all the grid cells within 10km distance of the canals. Columns (2) to (5) include the RD sample of grid cells that are $\leq$ 50m and $\geq$ 3m of relative elevation to the nearest canal. Column (2) shows the mean of the control group (grid cells above the canals) and column (3) shows the result of the t-test of a difference between treatment and control means (treatment minus control) with standard errors in the parentheses. Column (4) shows the mean of the non-homeland areas and column (5) shows the result of the t-test of a difference between non-homeland and homeland means (homeland minus non-homeland) with standard errors in the parentheses. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 2: Balance in geophysical characteristics

		Max monthly temp.	Annual precip.	Distance	Distance
	TRI	(in °C)	(in mm)	to canal	to river
	(1)	(2)	(3)	(4)	(6)
Panel A. Agricultural grid cells					
Below canal	-0.021	0.040	-4.33	267	75.4
	(0.038)	(0.102)	(3.70)	(185)	(52.8)
Control mean	0.901	36.0	549	5,401	1,175
R2	0.370	0.539	0.781	0.223	0.189
N	6,363,358	6,363,358	6,363,358	6,363,358	6,363,358
Panel B. All grid cells					
Below canal	0.001	-0.017	-2.67	351***	56.5*
	(0.032)	(0.105)	(2.84)	(107)	(32.8)
Control mean	1.21	36.2	523	5,283	1,120
R2	0.348	0.604	0.806	0.162	0.101
N	27,331,978	27,331,978	27,331,978	27,331,978	27,331,978

Note. This table reports regression discontinuity estimates of the coefficients on the treatment indicator obtained by estimating equation 1 and omitting the outcome variable from the list of controls. Standard errors are clustered at the level of 144 canals. The sample is restricted to grid cells within 10 km of distance and $\leq$ 50m and $\geq$ 3m of relative elevation to the nearest canal. The Terrain Ruggedness Index (TRI) is a topographic measure that captures variability of elevation of a given area and is derived from Nunn and Puga (2012). Maximum monthly temperature is calculated as an average over the maximum temperatures of each month in the period of 2014–2018 and is derived from MODIS Terra Land Surface Temperature dataset. Annual precipitation is calculated as an average of total annual rainfall over the period of 2014–2018 and is derived from the CHIRPS dataset (Funk et al. 2015). * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 3: Regression discontinuity results for agricultural outcomes: intensive margin

	Wheat season EVI (1)	Maize season EVI (2)	Wheat season EVI (log) (3)	Maize season EVI (log) (4)
Below canal	0.014	0.023*	0.040	0.041**
	(0.011)	(0.012)	(0.024)	(0.019)
Control mean	0.327	0.587	-1.28	-0.596
R2	0.537	0.362	0.573	0.354
Clusters	144	144	144	144
N	6,363,358	6,363,358	6,363,349	6,363,358

Note. This table reports regression discontinuity estimates of the coefficients on the treatment indicator obtained by estimating equation 1. Standard errors are clustered at the canal level. The sample is restricted to grid cells within 10 km of distance and $\leq$ 50m and $\geq$ 3m of relative elevation to the nearest canal. The sample includes only grid cells classified as agricultural in the SA NLC 2018 data (categories 38,39,40,41,43,44,45 which comprise commercial and subsistence annual crops and corresponding fallow land). Enhanced Vegetation Index (EVI) is a remote sensing measure that is generated from the Near-IR, Red and Blue bands of each satellite image, and ranges in value from -1 to 1. It is derived from the Landsat 8 Collection 1 Tier 1 8-Day EVI Composite. I extract the maximum EVI over the growing and harvesting seasons of wheat and maize respectively. Since a crop mask is applied to the data, I consider the measure as a proxy for agricultural productivity that is not contaminated by non-crop vegetation. Columns (1) and (2) report the raw measure, whereas columns (3) and (4) report a log-transformation of EVI. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 4: Regression discontinuity results for agricultural outcomes: extensive margin

	Any agri land (1)	Annuals (2)	Fallow land (3)
Below canal	0.063***	0.068***	-0.006
	(0.0172)	(0.0167)	(0.00422)
Control mean	0.213	0.166	0.049
r2	0.086	0.086	0.032
Clusters	144	144	144
N	27,331,978	27,331,978	27,331,978

Note. This table reports regression discontinuity estimates of the coefficients on the treatment indicator obtained by estimating equation 1. Standard errors are clustered at the canal level. The sample is restricted to grid cells within 10 km of distance and $\leq$ 50m and $\geq$ 3m of relative elevation to the nearest canal. The sample includes both agricultural and non-agricultural grid cells. “Any agricultural land" is equal to 1 if the grid cell is classified as agricultural in the SA NLC 2018 data (categories 38,39,40,41,43,44,45 which comprise commercial and subsistence annual crops and corresponding fallow land), and 0 otherwise. “Annuals" is equal to 1 if the grid cell is classified as growing commercial or subsistence annual crops in the SA NLC 2018 data (categories 38,39,40,41) and 0 otherwise. “Fallow land" is equal to 1 if the grid cell is classified as fallow land in the SA NLC 2018 data (categories 43,44,45) and 0 otherwise. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 5: Regression discontinuity results for agricultural outcomes disaggregated by ownership and irrigation status

	Wheat season	Maize season	Wheat season	Maize season
	EVI	EVI	EVI (log)	EVI (log)
	(1)	(2)	(3)	(4)
Panel A. Commercial pivot irrigated				
Below canal	-0.004	0.007	-0.001	0.011
	(0.021)	(0.011)	(0.034)	(0.015)
Control mean	0.709	0.845	-0.410	-0.196
R2	0.191	0.093	0.160	0.117
Clusters	105	105	105	105
N	241,396	241,396	241,396	241,396
Panel B. Commercial non-pivot irrigated				
Below canal	-0.007	-0.010	-0.020	-0.014
	(0.020)	(0.010)	(0.040)	(0.014)
Control mean	0.685	0.863	-0.524	-0.174
R2	0.332	0.142	0.342	0.127
Clusters	99	99	99	99
N	1,157,777	1,157,777	1,157,777	1,157,777
Panel C. Commercial rainfed				
Below canal	0.011**	0.024**	0.038***	0.045***
	(0.005)	(0.010)	(0.013)	(0.016)
Control mean	0.254	0.570	-1.44	-0.613
R2	0.487	0.227	0.504	0.235
Clusters	135	135	135	135
N	3,264,046	3,264,046	3,264,043	3,264,046
Panel D. Subsistence				
Below canal	-0.012	-0.019**	-0.041	-0.039***
	(0.009)	(0.008)	(0.029)	(0.014)
Control mean	0.266	0.462	-1.40	-0.810
R2	0.761	0.611	0.789	0.655
Clusters	40	40	40	40
N	440,017	440,017	440,014	440,017

Note. This table reports regression discontinuity estimates of the coefficients on the treatment indicator obtained by estimating equation 1. Standard errors are clustered at the canal level. The sample is restricted to grid cells within 10 km of distance and $\leq$ 50m and $\geq$ 3m of relative elevation to the nearest canal. Panel A restricts the sample to grid cells classified as commercial pivot-irrigated (cat. 38). Panel B restricts the sample to grid cells classified as commercial non-pivot-irrigated (cat. 39). Panel C restricts the sample to grid cells classified as commercial rainfed (cat. 40). Panel D restricts the sample to grid cells classified as subsistence (cat. 41). * significant at 10%, ** significant at 5%, *** significant at 1% The EVI measure is described in the note of Table 3.

Table 6: Regression discontinuity results for land use outcomes

	Expansion commercial pivot	Expansion commercial non-pivot	Expansion subsistence
	(1)	(2)	(3)
Below canal	0.019**	0.006***	-0.000
	(0.008)	(0.002)	(0.001)
Control mean	0.029	0.021	0.003
R2	0.074	0.012	0.022
Clusters	144	144	144
N	27,331,978	27,331,978	27,331,978

Note. This table reports regression discontinuity estimates of the coefficients on the treatment indicator obtained by estimating equation 1. Standard errors are clustered at the canal level. The sample is restricted to grid cells within 10 km of distance and $\leq$ 50m and $\geq$ 3m of relative elevation to the nearest canal. The sample includes both agricultural and non-agricultural grid cells. “Expansion commercial pivot" is equal to 1 if the grid cell is classified as commercial pivot-irrigated in the SA NLC 2018 data but not in the SA NLC 1990 data, and 0 otherwise. “Expansion commercial non-pivot" is equal to 1 if the grid cell is classified as commercial non-pivot irrigated or commercial rainfed in the SA NLC 2018 data but not in the SA NLC 1990 data and 0 otherwise. “Expansion subsistence" is equal to 1 if the grid cell is classified as subsistence land in the SA NLC 2018 data but not in the SA NLC 1990 data and 0 otherwise. * significant at 10%, ** significant at 5%, *** significant at 1%.

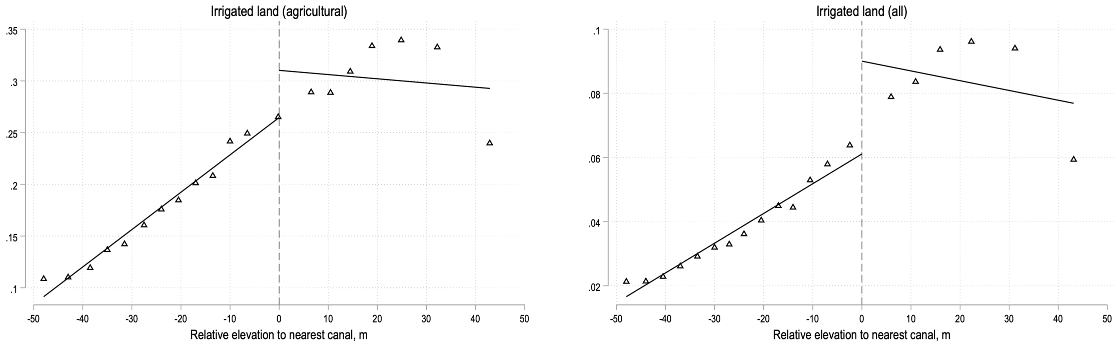
Table 7: Differential RD effects of canals by drought and the former homelands status

	Non-homelands		Homelands	
Drought at:	Growing stage (1)	Planting stage (2)	Growing stage (3)	Planting stage (4)
Panel A. Log wheat EVI				
Below canal	0.027	0.028	-0.064*	-0.059*
	(0.017)	(0.018)	(0.031)	(0.031)
Below canal × drought	0.009	0.018*	-0.055	-0.043
	(0.009)	(0.010)	(0.034)	(0.028)
Panel B. Log maize EVI				
Below canal	0.014*	0.015*	-0.035	-0.057**
	(0.008)	(0.008)	(0.027)	(0.023)
Below canal × drought	0.007	0.007	0.041***	0.036**
	(0.007)	(0.005)	(0.015)	(0.015)
Clusters	136	136	26	26
N	24,120,228	24,120,228	675,443	675,443

Note. This table reports regression discontinuity estimates of the coefficients on the treatment indicator obtained by estimating equation 3. Standard errors are clustered at the canal level. The sample is restricted to grid cells that lie within 10 km of distance and $\leq$ 50m and $\geq$ 3m of relative elevation to the nearest canal. In columns (1) and (2), the sample is further restricted to grid cells that lie *outside* the territory of former homelands, whereas in columns (3) and (4), it is restricted to grid cells that lie *within* the territory of former homelands. The sample includes only grid cells classified as agricultural in the UMD GLAD Global Cropland data. Enhanced Vegetation Index (EVI) is a remote sensing measure that is generated from the Near-IR, Red and Blue bands of each satellite image, and ranges in value from -1 to 1. It is derived from the Landsat 8 Collection 1 Tier 1 8-Day EVI Composite. I extract the maximum EVI over the growing and harvesting seasons of wheat and maize respectively. Since a crop mask is applied to the data, I consider the measure as a proxy for agricultural productivity that is not contaminated by non-crop vegetation. * significant at 10%, ** significant at 5%, *** significant at 1%.

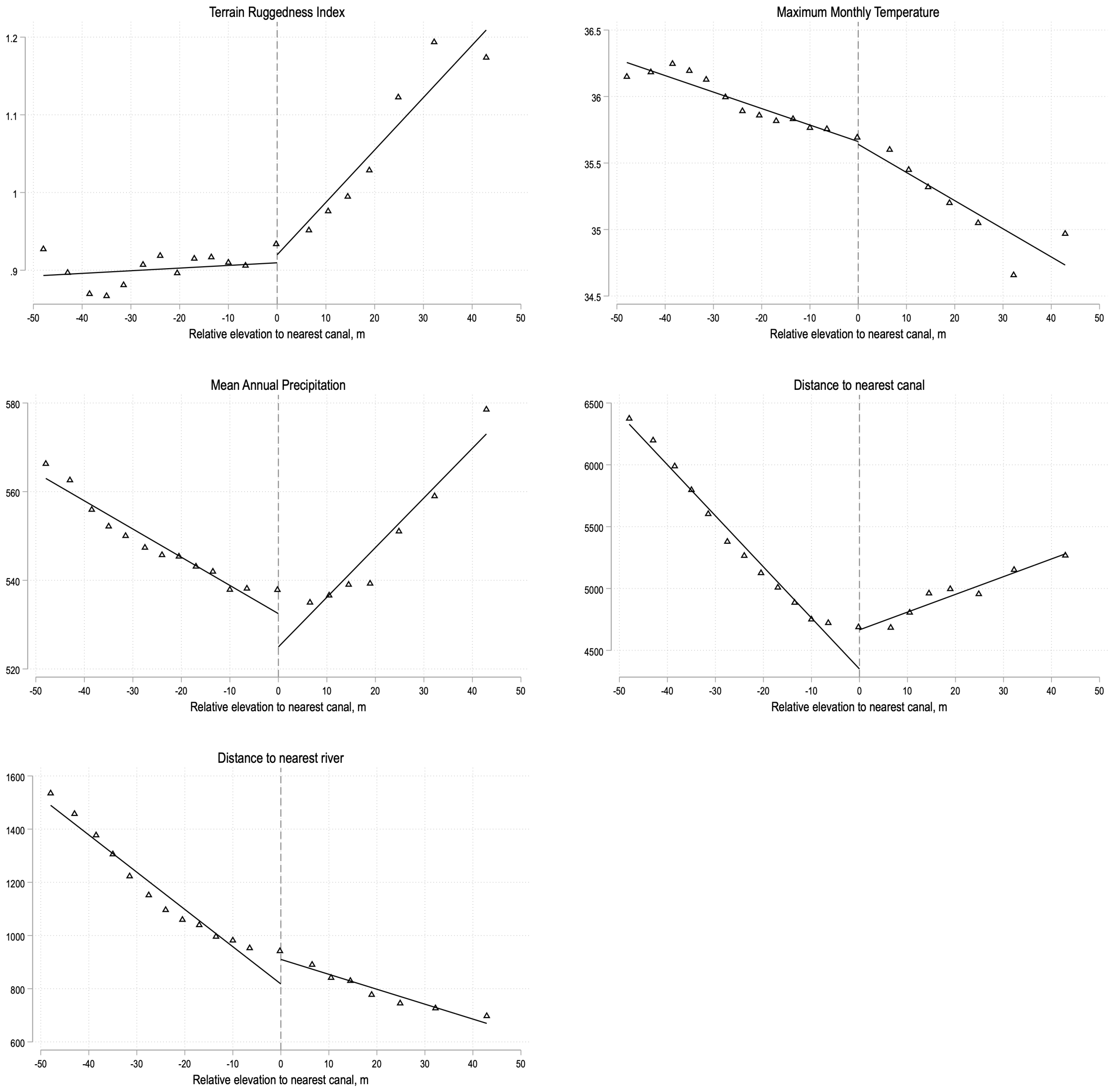
Figures

Figure 1: Discontinuity in probability of irrigated land



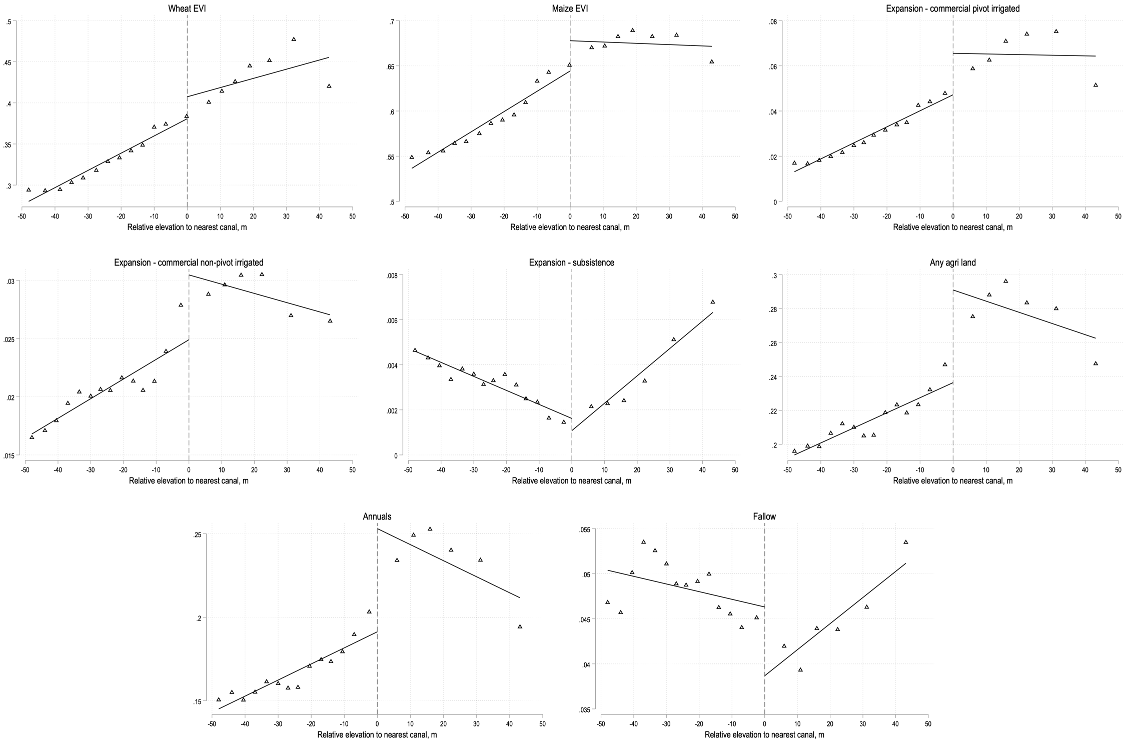
Note. Each panel plots the average probability of being irrigated within each quantile bin of relative elevation to the nearest canal. The negative values to the left of 0 represent the control units (above a canal) and the positive values to the right of 0 represent the treatment units (below a canal). Fitted linear regression lines of the underlying data are plotted separately for each side of the threshold. The sample is restricted to grid cells within 10 km of distance and $\leq$ 50m and $\geq$ 3m of relative elevation from the nearest canal. The left figure is generated for the agricultural sample (where the crop mask is applied) and the right figure is generated for the full sample (both agricultural and non-agricultural land).

Figure 2: Continuity through threshold of geophysical characteristics. Agricultural land



Note. Each panel plots the average outcome within each quantile bin of relative elevation to the nearest canal. The negative values to the left of 0 represent the control units (above a canal) and the positive values to the right of 0 represent the treatment units (below a canal). Fitted linear regression lines of the underlying data are plotted separately for each side of the threshold. The sample is restricted to grid cells within 10 km of distance and $\leq$ 50m and $\geq$ 3m of relative elevation from the nearest canal. Crop mask is applied.

Figure 3: Regression discontinuity binned scatterplots for main outcomes



Wheat EVI

Maize EVI

Expansion - commercial pivot irrigated

Expansion - commercial non-pivot irrigated

Expansion - subsistence

Any agri land

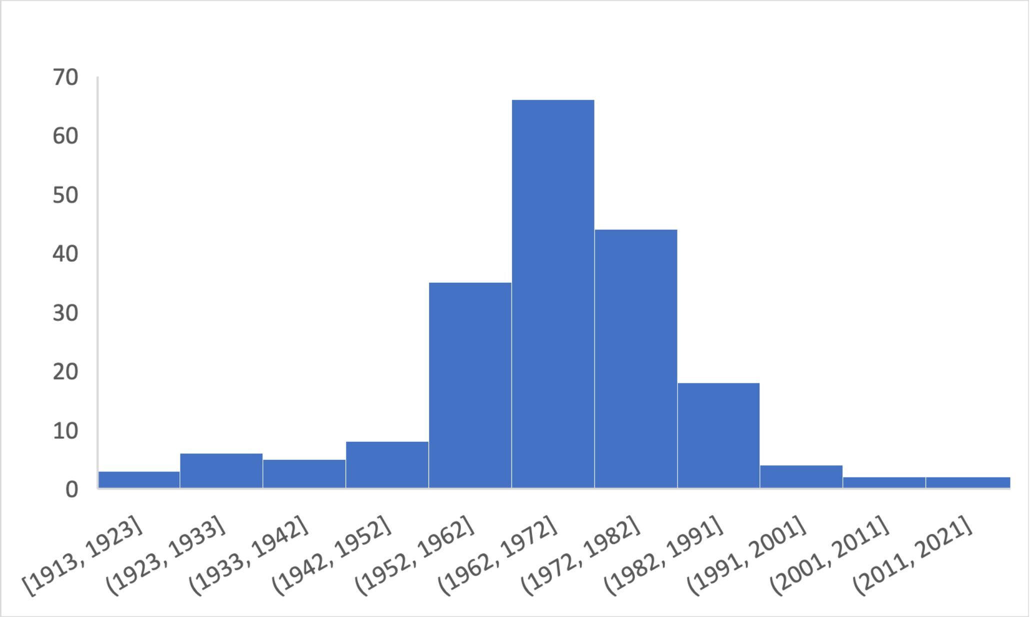
Annuals

Fallow

Note. Each panel plots the average outcome within each quantile bin of relative elevation to the nearest canal. The negative values to the left of 0 represent the control units (above a canal) and the positive values to the right of 0 represent the treatment units (below a canal). Fitted linear regression lines of the underlying data are plotted separately for each side of the threshold. The sample is restricted to grid cells within 10 km of distance and $\leq$ 50m and $\geq$ 3m of relative elevation from the nearest canal.

Appendix

Figure A–1: Histogram of years of construction of irrigation dams



The data comes from AQUASTAT, the geo-referenced database on dams in Africa. Only dams with the listed purpose of irrigation are included. There are 200 irrigation dams and 550 dams in total.

Figure A–2: Schematic representation of elevation-based RDD

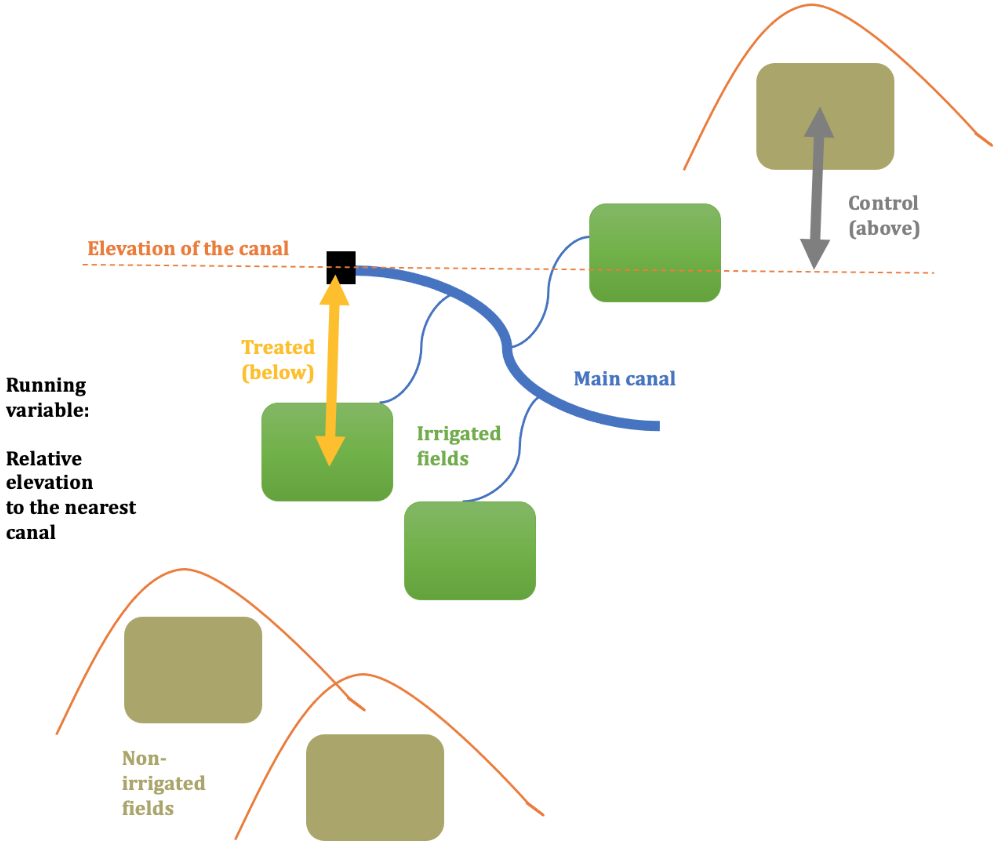
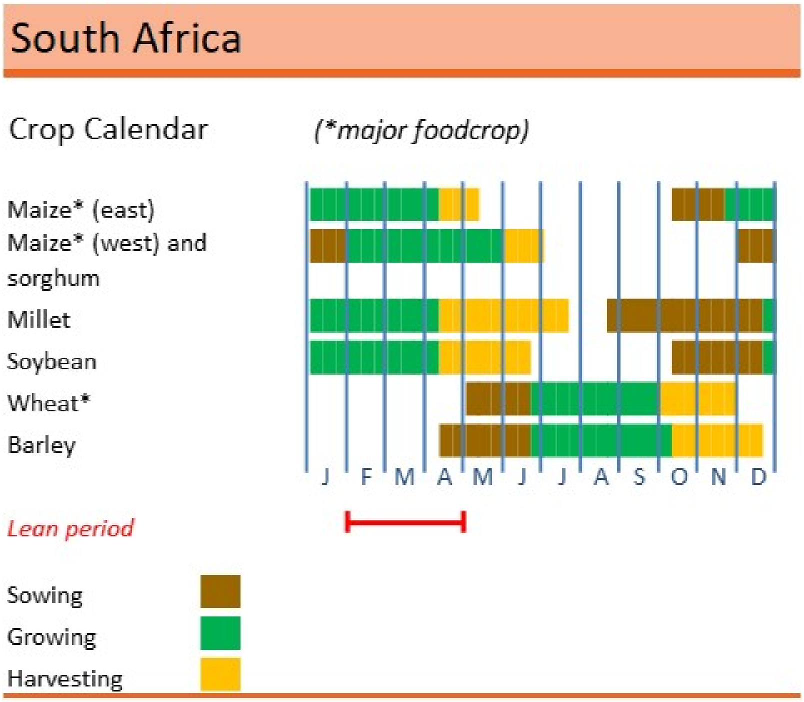


Figure A–3: FAO Crop Calendar



Source: `https://www.fao.org/giews/countrybrief/country.jsp?code=ZAF`.

Table A–1: Drought effects on crop yields (control group)

	Log wheat season EVI	Log maize season EVI
	(1)	(2)
Drought at:		
Planting stage	-0.043**	-0.015**
	(0.019)	(0.007)
Growing stage	-0.029*	0.007
	(0.016)	(0.005)
R2	0.561	0.254
Clusters	142	142
N	14,566,770	14,568,564

Note. This table reports the results of estimating the equation $y_{idt} = \beta_0 + \beta_1 plant\_pdsi_{idt} + \beta_2 grow\_pdsi_{idt} + \beta_3 X_{idt} + \mu_d + \gamma_t + \epsilon_{idt}$ by OLS on the control group sample (above canals), where y_{idt} proxies the crop yields by log wheat season EVI or maize season EVI in grid cell i, district d, and year t, $plant\_pdsi_{idt}$ and $grow\_pdsi_{idt}$ are the normalized values of the average Palmer Drought Severity Index of a given crop during its planting stage and growing stage respectively, X_{idt} are geophysical controls variables (temperature, precipitation, TRI, distance to nearest canal), μ_d are district fixed effects, and γ_t are year fixed effects. Standard errors in parentheses are clustered at the canal level to account for potential spatial correlation. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table A–2: Comparison of UMD GLAD 2019 and SANLC 2018

	Proportion of UMD GLAD cropland grid cells in a given NLC cat.
Panel A. Homelands	
Subsistence / Small-Scale Annual Crops	33%
Cultivated Emerging Farmer Sugarcane Non-Pivot	25%
Open Woodland	12%
Cultivated Commercial Sugarcane Non-Pivot	4%
Residential Formal (Bush)	4%
Panel B. Non-homelands	
Commercial Annuals Crops Rain-Fed / Dryland / Non-Irrigated	49%
Commercial Annuals Pivot Irrigated	32%
Commercial Annuals Non-Pivot Irrigated	6%
Natural Grassland	4%
Cultivated Commercial Sugarcane Non-Pivot	2%

Note. Only five most frequent categories for both homelands and non-homelands are included.