Impact of access to irrigation on agricultural productivity: Evidence from community-led lift irrigation schemes in India

Contents

Overview

This project contains all Stata code and cleaned data that reproduces the analysis conducted for the paper titled "Impact of access to irrigation on agricultural productivity: Evidence from community-led lift irrigation schemes in India" by Wyatt Pracht, Jisang Yu, and Felipe Dizon.

Statement about Rights

- I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the LICENSE.txt file.

Data Availability

The data used in this analysis requires access to confidential project administrative data which includes GPS coordinates for irrigation pump sources and their associated outlets. We employed the shapefile of these coordinates to extract data from remotely sensed data products available on Google Earth Engine. These coordinates are not publicly available and cannot be disseminated. For this reason, we removed the coordinates and any identifiable information from the raw data. The final anonymized and cleaned datasets are forthcoming on the Development Data Hub as license data. The code used to clean the raw data and construct the final datasets is included in the reproducibility package. The file is called "Data_Prep.do" in the codes subfolder.

- The subfolder "codes" includes all codes used to reproduce the figures and tables in the main text of the manuscript.
    - The STATA code file "codes\master_do_file.do" runs all the do files in order after the user sets the working directory. This .do file takes 2.5 hours to run.
- The below cleaned analysis datasets forthcoming on the Development Data Hub should be included in the subfolder "data":
    - The file "data\Replication_Whole_Panel.dta" is used to produce the results across all seasons for the overall sample as well as the pump source and outlet subsamples. These include results in Tables 1, 2, and A.1. and all figures.

    - The file “data\Replication_Rabi_Panel.dta” is used to reproduce the Rabi subsample results. These include results in Tables 3 and A.2.
    - The file “data\Replication_Zaid_Panel.dta” is used to reproduce the Zaid subsample results. These include results in Tables 4 and A.3.
    - The file “data\Replication_Kharif_Panel.dta” is used to reproduce the Kharif subsample results. These include results in Tables 5 and A.4.
- All anonymized raw data files forthcoming on the Development Data Hub should be included in “Raw_Data”.
- The subfolder “output\Tables” contains the tables used in the paper.
- The subfolder “output\Figures” contains the figures used in the paper.

Data Sources

Project Administrative Data

To construct our treatment variable and to extract the remotely sensed variables needed for our analysis, we use confidential administrative data from the JOHAR project. The data included information on the GPS coordinates of 491 pumps, their corresponding irrigation outlets, their dates of installation completion, the type of pump, and the command area’s size. The anonymized version of this dataset should be included in the “Raw_Data/Misc Data” subfolder under the name “export_Final_JOHARIrrigation.dta”.

Citation: Author: JOHAR Project Management Unit; Dataset: Jharkhand JOHAR Irrigation GPS Coordinates; Date Accessed: October 2023.

Vegetation Indices

Our outcome variables, the enhanced vegetation index (EVI) and the normalized difference vegetation index (NDVI), were calculated using 30m resolution satellite imagery from the Landsat 8 Level 2, Collection 2, Tier 1 data product. The raw data were extracted using the Google Earth Engine Code Editor. The raw data were exported as CSV files. Anonymized versions of these CSV files should be included the subfolder “Raw_Data\Landsat_Vegetation_Data”. Two intermediary datasets titled “Appended_Landsat.dta” and “Merge_Ready_VI” should be included in the same folder.

- **Source:** Earth Resources Observation and Science (EROS) Center. (2020). Landsat 8-9 Operational Land Imager / Thermal Infrared Sensor Level-2, Collection 2 [dataset]. U.S. Geological Survey. https://doi.org/10.5066/P9OGBGM6.
- **URL:** https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2
- **Access year:** November 2025
- **Access instructions:** We employed the shapefile of confidential GPS coordinates (from the JOHAR project) for irrigation pump sources and their associated outlets coordinates to extract data from remotely sensed data products available on Google Earth Engine.

Rainfall

Our rainfall data come from the Climate Hazards Group InfraRed Precipitation Pentad data product, which includes cumulative rainfall estimates for the six pentads in each calendar month at a 0.05° resolution (roughly 5 km x 5 km). The raw data were extracted using the Google Earth Engine Code Editor. The raw data were exported as CSV files. Anonymized versions of these CSV files should be included in the subfolder "Raw_Data\Rainfall_Data". Two intermediary datasets titled "Appended_Rainfall.dta" and "Merge_Ready_Rainfall" should also be included in the same folder.

Source: Funk, Chris, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell & Joel Michaelsen. (2015). "The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes". Scientific Data 2, 150066. doi:10.1038/sdata.2015.66.

- **URL:** https://developers.google.com/earth-engine/datasets/catalog/UCSB-CHG_CHIRPS_PENTAD
- **Access year:** November 2025
- **Access instructions:** We employed the shapefile of confidential GPS coordinates (from the JOHAR project) for irrigation pump sources, and their associated outlets coordinates to extract data from remotely sensed data products available on Google Earth Engine.

Temperature

Our temperature data come from the ERA5-Land Daily Aggregated data product, a global climate reanalysis dataset from the European Center for Medium-Range Weather Forecasts. Thus data product provides daily minimum and maximum temperatures measured at 2m height. The raw data were extracted using the Google Earth Engine Code Editor. The raw data were exported as CSV files. Anonymised versions of these CSV files should be included in the subfolder "Raw_Data\Temp_data". Two intermediary datasets titled "Appended_Temp.dta" and "Merge_Ready_Temp" should be included in the same folder.

- **Source:** Muñoz Sabater, J., (2019): ERA5-Land monthly averaged data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). doi:10.24381/cds.68d2bb30.
- **URL**: https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_DAILY_AGGR
- **Access year:** November 2025
- **Access instructions:** We employed the shapefile of confidential GPS coordinates (from the JOHAR project) for irrigation pump sources and their associated outlets coordinates to extract data from remotely sensed data products available on Google Earth Engine.

Distance to River

Our distance to the nearest free-flowing river variable comes from the HydroSHEDS Free Flowing Rivers Network data product developed by the World Wildlife Fund. The raw data were extracted using the Google Earth Engine Code Editor. The raw data were exported as a CSV file. Anonymised version of this CSV file should be included in the subfolder "Raw_Data\Misc data" titled "DistanceToRivers.csv". An intermediary dataset titled "DistanceToRivers.dta" should be included in the same folder.

- **Sources:**
  - Lehner, B., Verdin, K., Jarvis, A. (2008): New global hydrography derived from spaceborne elevation data. Eos, Transactions, AGU, 89(10): 93-94.
  - Grill, G., Lehner, B., Thieme, M., Geenen, B., Tickner, D., Antonelli, F., Babu, S., Borrelli, P., Cheng, L., Crochetiere, H. and Macedo, H.E., (2019). Mapping the world's free-flowing rivers. Nature, 569(7755), p.215.
- **URL:** https://developers.google.com/earth-engine/datasets/catalog/WWF_HydroSHEDS_v1_FreeFlowingRivers
- **Access year:** November 2025
- **Access instructions:** We employed the shapefile of confidential GPS coordinates (from the JOHAR project) for irrigation pump sources and their associated outlets coordinates to extract data from remotely sensed data products available on Google Earth Engine.

Elevation, Slope, and Terrain Ruggedness

Our elevation, slope, and the terrain ruggedness index variables are derived from the SRTM Digital Elevation V4 data product. The raw data were extracted using the Google Earth Engine Code Editor. The raw data were exported as a CSV file. An anonymized version of this CSV file should be included in the subfolder "Raw_Data\Misc data" titled "SRTM_Elevation_Slope_TRI_Export.csv". An intermediary dataset titled "Elevation.dta". should be included in the same folder.
Source: Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara. (2008). Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database: https://srtm.csi.cgiar.org.

- **URL:** https://developers.google.com/earth-engine/datasets/catalog/CGIAR_SRTM90_V4
- **Access year:** November 2025
- **Access instructions:** We employed the shapefile of confidential GPS coordinates (from the JOHAR project) for irrigation pump sources and their associated outlets coordinates to extract data from remotely sensed data products available on Google Earth Engine.

Computing Environment

Operating system: Windows 11 Enterprise, version 24H2

Software used: STATA SE 19.5 for Windows.

Packages used:

1. Estout: ssc install estout, replace. Used to export Table 1.
2. Drdid: ssc install drdid, replace. Includes the underlying estimator that the csdid package relies on.
3. Csdid: ssc install csdid, replace. Used to estimate Callaway and Sant'Anna difference-in-differences (CSDID) regressions.
4. Reghdfe: ssc install reghdfe, replace. Used to estimate high-dimensional fixed effects regressions.
5. Ftools: ssc install ftools, replace. Required dependency for reghdfe.
6. Require: ssc install require, replace. Required dependency for reghdfe.

These packages are installed by running the code in "codes\required_packages.do".

Index

This section is an index of the data, code, and figures included in the replication package. Each .do file is paired with the figure or table it produces for straightforward cross-referencing.

1. Data_Prep.do documents how the final datasets were generated from the source data.
2. master_do_file.do details the replication, installs the required STATA packages, and runs each do file to replicate the tables and figures.
3. Required_packages.do installs and each STATA package need for the replication.
4. all_figures.do generates all figures used in the paper.
5. Table_1.do generates Table 1 which displays the descriptive statistics of the outcomes and covariates.
6. Table_2.do generates Table 2 which displays the EVI Treatment Effects across all seasons.
7. Table_3.do generates Table 3 which displays the EVI Treatment Effects for the Rabi (Winter) seasons.
8. Table_4.do generates Table 4 which displays the EVI Treatment Effects for the Zaid (Summer) seasons.
9. Table_5.do generates Table 5 which displays the EVI Treatment Effects for the Kharif (Monsoon) seasons.
10. Table_A1.do generates Table A.1 which displays the NDVI Treatment Effects across all seasons.
11. Table_A2.do generates Table A.2 which displays the NDVI Treatment Effects for the Rabi (Winter) seasons.
12. Table_A3.do generates Table A.3 which displays the NDVI Treatment Effects for the Zaid (Summer) seasons.
13. Table_A4.do generates Table A.4 which displays the NDVI Treatment Effects for the Kharif (Monsoon) seasons.

Below is a list of all figures presented in the paper and their file name in the "Figures" subfolder. All figures were replicated using the "All_Figures.do" file.

1. Figure 1's filename is "Figure1.png". This figure is a combined graph of the following files: "FIGURE1_OVERALL.png", "FIGURE1_PUMP.png", "FIGURE1_Outlet1.png", "FIGURE1_Outlet2.png", "FIGURE1_Outlet3.png", and "FIGURE1_Outlet4.png".
2. Figure 2's filename is "Figure2.png". This figure is a combined graph of the following files: "FIGURE2_Zaid2020.png", "FIGURE2_Kharif2020.png", "FIGURE2_Rabi2020_21.png", "FIGURE2_Zaid2021.png", "FIGURE2_Kharif2021.png", "FIGURE2_Rabi2021_22.png", "FIGURE2_Zaid2022.png".
3. Figure A.1's filename is "FigureA1.png". This figure is a combined graph of the following files: "FIGUREA1_OVERALL.png", "FIGUREA1_PUMP.png", "FIGUREA1_Outlet1.png", "FIGUREA1_Outlet2.png", "FIGUREA1_Outlet3.png", and "FIGUREA1_Outlet4.png".
4. Figure A.2's filename is "FigureA2.png". This figure is a combined graph of the following files: "FIGUREA2_Zaid2020.png", "FIGUREA2_Kharif2020.png", "FIGUREA2_Rabi2020_21.png", "FIGUREA2_Zaid2021.png", "FIGUREA2_Kharif2021.png", "FIGUREA2_Rabi2021_22.png", "FIGUREA2_Zaid2022.png".

Below is a list of the file names for each table located in the "Tables" subfolder, and the associated .do file located in the "Codes" subfolder. All tables are exported from STATA as individual .rtf files. Each .rtf file corresponds to its matching table number in the paper, though minor formatting differences from the published version may exist.

1. Table 1 – saved as "Table1.rtf". Generated by Table_1.do.
2. Table 2 – saved as "Table2.rtf". Generated by Table_2.do.
3. Table 3 – saved as "Table3.rtf". Generated by Table_3.do.
4. Table 4 – saved as "Table4.rtf". Generated by Table_4.do.
5. Table 5 – saved as "Table5.rtf". Generated by Table_5.do.
6. Table A.1 – saved as "TableA1.rtf". Generated by Table_A1.do.
7. Table A.2 – saved as "TableA2.rtf". Generated by Table_A2.do.
8. Table A.3 – saved as "TableA3.rtf". Generated by Table_A3.do.
9. Table A.4 – saved as "TableA4.rtf". Generated by Table_A4.do.

For a list of datasets and the associated tables that are produced from this data, see the data availability section.