README for Reproducibility Package of the Policy Research Working Paper

Mapping the risk posed to groundwater-dependent ecosystems by the uncontrolled access to photovoltaic water pumping in sub-Saharan Africa

Guillaume Zuffinetti[1,2], Simon Meunier[1,2,*]

[1] *Université Paris-Saclay, CentraleSupélec, CNRS, GeePs, Gif-sur-Yvette 91192, France*

[2] *Sorbonne Université, CNRS, GeePs, Paris 75252, France*

[*] *Corresponding author email address: simon.meunier@centralesupelec.fr*

This README is designed to provide guidance for replicating the results of the Policy Research Working Paper. The paper is included as a PDF file in the package or can be downloaded at the following link:

https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099419309302454145/idu17970fd3c1b6c4148521940c1908ba81aeae1

Content

1. Overview

This reproducibility package includes all the necessary datasets, model, projects, and codes to replicate the results presented in the Policy Research Working Paper.

The references for each dataset used in this study are listed in Section **Error! Reference source not found.**. To replicate the results, please refer to the instructions in Section 3.

Section 4 documents the tables and figures from the Policy Research Working Paper that can be reproduced. Technical requirements for running the codes and handling the datasets are outlined in Section 0. Lastly, detailed code descriptions are provided in Section 6.

2. Data Availability Statement

This section outlines where and how the datasets supporting the findings of the study can be accessed and used.

Note that all data are publicly available.

- **Filename 1:** Global Horizontal Irradiance (GHI)
  - **Source:** CAMS Radiation Service
  - **URL:** https://www.tsv.soda-pro.com/web-services/radiation/cams-radiation-service
  - **Access year:** 2021
  - **Path to the dataset in the package:** Input_data\GHI\ghi_2020.h5
- **Filename 2:** Groundwater productivity map
  - **Source:** British Geological Survey. Bonsor, H. C., & MacDonald, A. M. (2011). An initial estimate of depth to groundwater across Africa.
  - **URL:** https://www2.bgs.ac.uk/groundwater/international/africanGroundwater/mapsDownload.html
  - **Access year:** 2022
  - **Note:** A readme file is also provided with the dataset. The groundwater productivity map is used to obtain the transmissivity map following the rules from Table 2 from Bonsor and MacDonald (2011).
  - **Path to the dataset in the package:** Input_data\Raw_data\xyzASCII_gwprod_v1
- **Filename 3:** Groundwater storage map
  - **Source:** British Geological Survey. MacDonald, A. M., Bonsor, H. C., Dochartaigh, B. É. Ó., & Taylor, R. G. (2012). Quantitative maps of groundwater resources in Africa. Environmental Research Letters, 7(2), 024009.
  - **URL:** https://www2.bgs.ac.uk/groundwater/international/africanGroundwater/mapsDownload.html
  - **Access year:** 2022

  - **Note:** A readme file is also provided with the dataset.
  - **Path to the dataset in the package:** Input_data\Raw_data\xyzASCII_gwprod_v1

- **Filename 4:** Africa Water Table Depth
  - **Source:** Earth2Observe. Fan, Y., Li, H., & Miguez-Macho, G. (2013). Global patterns of groundwater table depth. Science, 339(6122), 940-943.
  - **URL:** https://wci.earth2observe.eu/thredds/catalog/usc/water-table-depth/catalog.html
  - **Access year:** 2022
  - **Note:** The URL no longer seems to work. However, this link was obtained by contacting the first author (i.e. Y. Fan) directly by e-mail in 2022.
  - **Path to the datasets in the package:** Input_data\Raw_data\WTD.tif

- **Filename 5:** Reliance on unimproved water sources – Number - 2017
  - **Source:** Institute for Health Metrics and Evaluation
  - **URL:** https://cloud.ihme.washington.edu/s/bkH2X2tFQMejMxy?path=%2F
  - **Access year:** 2022
  - **Path to the dataset in the package:** Input_data\Raw_data\IHME_LMIC_WASH_2000_2017_W_UNIMP_NUMBER_UPPER_2017_Y2020M06D02

- **Filename 6:** Aquifer typology
  - **Source:** World Bank
  - **URL:** The download link for the data will be provided shortly.
  - **Access year:** 2022
  - **Path to the dataset in the package:** Input_data\Raw_data\Aquifer_type
  - **Variable name:**
    - Alluvial
    - Complexe
    - Karstic
    - Shallow

- **Filename 7:** Groundwater resource
  - **Source:** World Bank
  - **URL:** The download link for the data will be provided shortly.
  - **Access year:** 2022
  - **Path to the dataset in the package:** Input_data\Raw_data\ aqtyp_gwresource_grid05deg.gpkg
  - **Variable name:** resource
- **Filename 8:** GDEs
  - **Source:** World Bank
  - **URL:** The download link for the data will be provided shortly.
  - **Access year:** 2022
  - **Note:** Several shape files are provided. The used ones in the codes are "SSA_GDEs_Polygons", "SSA_GDEs_Points MAJ", "SSA_GDEs_Lines", and "Studied_countries".
  - **Path to the dataset in the package:** Input_data\Raw_data\GDEs

Note: for filename 6,7,8 The data is associated with the World Bank report, "The Hidden Wealth of Nations: Groundwater in Times of Climate Change". The report can be downloaded from the following link: https://www.worldbank.org/en/topic/water/publication/the-hidden-wealth-of-nations-groundwater-in-times-of-climate-change. These data were directly provided by the World Bank team to the authors of this paper. At the time of sharing, the authors were informed that data is being prepared for publication. In the meantime, the data can be included with this reproducibility package.

Note that the folder "Input_data" is organized into three subfolders:

- "GHI": Contains the raw GHI dataset in .h5 format.
- "Raw_data": Stores every raw input dataset.
- "Rasterized_data": Includes all the input datasets that have been rasterized thanks to "getGHI.mat" MatLab function or thanks to the "Map_construction" QGIS project and the "model.model3" model in QGIS (see Sections 3 and 6).

Additionally, the folder "Input_data" contains "contour_Africa.mat", a MatLab matrix only containing 1 and NaN values to clearly delineate the contour of each African gridded dataset used in this study.

3. Instructions for Replicators

Note that all the model and codes have been pre-run, and all the parameters have been pre-entered. Therefore, one can open every MatLab code, QGIS project or QGIS model and run it independently from the others and visualize the results without the need to go through each of the following steps.

Note: During the replication process, the authors noted that the replicators were not able to obtain the same results when running the code `Plot_results`. However, all other results and dataframes were correctly reproduced. To address this, we are including the intermediate data created by the authors, which allows for the reproduction of Fall figures directly, without needing to run the `Plot_results` code.

To run the package successfully from scratch, the following steps should be followed:

- Open the Reproducibility package folder.
- Open "getGHI.mat" with MatLab and run the code by clicking on the green run button. To open the code, it is recommended to right click on it and open it with an appropriate version of MatLab (see Section 5.2). **This code requires powerful computation resources to run (see Section 5.1). However, even though the code cannot be run on your computer, the result of the code is already available as a .tif file in the folder "Input_data\Raw_data", and is named "GHI.tif". Therefore, if the code cannot be run on your computer, simply skip this point.**
- When the "getGHI.mat" code has been run (or skipped, see above), go back to the Reproducibility package folder and open the project "Map_construction" with QGIS. To open the project, it is recommended to right click on it and open it with an appropriate version of QGIS (see Section 5.2). Then, the following steps should be followed:
  - Go to Processing → Graphical Modeler…
  - Open the Graphical Modeler by double-clicking on it
  - In Model, double click on "Open model…"
  - Open the model "model.model3" in the Reproducibility package
  - Run it by clicking on the green run button: a window opens with parameters that can be filled in. The parameters should be as follow:
    - Alluvial_label → Alluvial [EPSG:4326]
    - Complex_label → Complexe [EPSG:4326]
    - GDE_label → GDE
    - GDE_Lines → SSA_GDE_Lines [EPSG:4326]

  - GDE_Points → SSA_GDEs_Points MAJ [EPSG :4326]
  - GDE_Polygon → SSA_GDEs_Polygons [EPSG: 4326]
  - GWProductivity → xyzASCII_gwprod_v1 [EPSG:4326]
  - GWresources_initial_resolution → 0.5
  - GWressource_&_AQtype → aqtyp_gwresource_grid05deg [EPSG :4326]
  - GWStorage → xyzASCII_gwstor_v1 [EPSG:4326]
  - Karstic_label → Karstic [EPSG :4326]
  - Output_extent → GHI [EPSG:4326]
  - Population_density_initial →gpw_v4_population_density_rev11_2020_2pt5_min [EPSG: 4326]
  - Resolution → 0.2
  - Resource_label → Resource
  - Shallow_label → Shallow [EPSG:4326]
  - Storage_label → Storage
  - Surface_label → Surface
  - Transmissivity_label → Transmissivity
  - type_class_label → type_class
  - WTD_initial → WTD [EPSG:4326]

- *Optional:* Check the checkboxes if you want to plot the raster maps.
- Save the raster files in the folder "Input_data\Rasterized_data" as:
  - GWresource_0.2_Africa_km3 → GWresources.tif
  - Population_density → Popdensity.tif
  - GDE_Lines_raster → GDELines.tif
  - AlluvialTIFF → AlluvialTIFF.tif
  - ComplexeTIFF → ComplexeTIFF.tif
  - KarsticTIFF → KarsticTIFF.tif
  - GDE_Polygons_raster → GDEPolygons.tif

    - GDE_Points_raster → GDEPoints.tif
    - Transmissivity → Transmissivity.tif
    - Storage → Storage.tif
    - ShallowTIFF → ShallowTIFF.tif
    - Hbs → Hbs.tif
  - Click on "Run".

- When the model has been run, go back to the Reproducibility package folder and open "Main_AHP.mat" with MatLab and run the code by clicking on the green run button. To open the code, it is recommended to right click on it and open it with an appropriate version of MatLab (see Section 5.2).

- When the "Main_AHP.mat" code has been run, go back to the Reproducibility package folder and open the project "Plot_results" with QGIS. To open the project, it is recommended to right click on the project and open it with an appropriate version of QGIS (see Section 5.2).

- Check or uncheck the layers you want to plot by checking or unchecking the checkboxes in the "Layers" window.
  - *To plot the results of Figure 2a*, check the layers "PointRisk_AHP", "LinesRisk_AHP", "PolygonRisk_AHP", and "Studied_countries".
  - *To plot the results of Figure 2b*, check the layers "PointRisk_sameweight", "LinesRisk_ sameweight", "PolygonRisk_ sameweight", and "Studied_countries".
  - *To plot the results of Figure 3a*, check the layers "PointRisk_AHP", "LinesRisk_AHP", "PolygonRisk_AHP", and "IHME_LMIC_WASH_2000_2017_W_UNIMP_NUMBER_UPPER_2017_Y2020M06D02".
  - *To plot the results of Figure 3b*, check the layers "PointRisk_sameweight", "LinesRisk_ sameweight", "PolygonRisk_ sameweight", and "IHME_LMIC_WASH_2000_2017_W_UNIMP_NUMBER_UPPER_2017_Y2020M06D02".
  - *To plot the results of Figure 4a*, check the layers "PointRisk_AHP", "LinesRisk_AHP", "PolygonRisk_AHP", "Alluvial", "Shallow", "Karstic", and "Complexe", and uncheck "Studied_countries".
  - *To plot the results of Figure 4c*, check the layers "PointRisk_sameweight", "LinesRisk_ sameweight", "PolygonRisk_ sameweight", "Alluvial", "Shallow", "Karstic", and "Complexe", and uncheck "Studied_countries".

4. List of Exhibits

The provided projects, model, and codes reproduce all numbers, tables and figures provided in the paper. Detailed information on where to find them is provided below:

Exhibit name	Code/Project	Note
Table 2	Main_AHP.mat	Found in the “Command window” once the code has been run
Table 3	Main_AHP.mat	Found in the “Command window” once the code has been run
Table 4	Main_AHP.mat	Found in the “Command window” once the code has been run
Figure 1	Map Construction	Point and click approach detailed below
Figure 2a and b	Plot_results	Check the desired layers in the “Layer” window from QGIS (see Section 3)
Figure 3a and b	Plot_results	Check the desired layers in the “Layer” window from QGIS (see Section 3)
Figure 4a and c	Plot_results	Check the desired layers in the “Layer” window from QGIS (see Section 3)
Figure 4b and d	Main_AHP.mat and “Results” folder	The bar charts are plotted automatically when the code is run and are saved in the folder “Results”

To reproduce **Figure 1**, Open the “Map Construction” project in QGIS. Navigate to “Processing” → Double click “Graphical Modeler”. In the “Model” menu, click “Open Model” and select the file "model.model3" from the package. To run the model, left-click the green arrow and ensure that every box labeled "Open output file after running algorithm" is checked. After the model has finished running, close both the “Model” window and the “Model Designer” window, then return to the “Map Construction” project. Finally, select the box corresponding to the desired data frame:
Figure 1a. → GHI (done
Figure 1b. → WTD d
Figure 1c. → Transmissivity
Figure 1d. → Storage
Figure 1e. → Population_density
Figure 1f. → GWresource_0.2_Africa_km3 (GWresources.tif)
Figure 1g. → SSA_GDEs_Polygons, SSA_GDEs_Points_MAJ, and SSA_GDEs_lines, Studied_countries d

To modify the colorbars of individual data frames, double-click on the box corresponding to the desired data frame. This will open the "Layer Properties" window. Navigate to the "Symbology" tab. In the "Band Rendering" section, choose the "Render Type" option as "Singleband Pseudocolor". You can then customize the colorbar according to your preferences.

5. Requirements

5.1. Computational Requirements

The whole package is ~2 Go and can therefore be downloaded and opened on conventional laptops and computers. The QGIS model and “Main_AHP.mat” code have been run on a classical Dell laptop (Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz 2.11 GHz, 16 Go, 8 cores) and are therefore assumed to be runnable on conventional laptops or computers.

Regarding the “geGHI.mat” code, the code has been run with the following computer server: Intel Xeon, CPU E5-2643 0, 3.40GHz, 16 cores. If you run the code on a computer that is not powerful enough, you might get the following error message in the “Command window” from MatLab:

“Error using h5readc

The HDF5 library encountered an error and produced the following stack trace information:

H5FL__malloc   memory allocation failed for chunk”

Nevertheless, even though your computer does not manage to run the code, the rasterized GHI dataset is already available in the folder “Input_data\Rasterized_data”.

5.2. Software Requirements

The codes and model to obtain the results have been run using 2 software:

- **MatLab R2022a**
- **QGIS Desktop 3.26.3**

6. Code Description

- “Date2jd.mat”: This MatLab function calculates the Julian day number for a specified date in the Gregorian calendar. It is used by “getGHI.mat” to process GHI time-series data.
- “getGHI.mat”: This MatLab code converts GHI data from the initial .h5 format to a raster format, making it compatible with QGIS for visualization and analysis. It also calculates the annual GHI average from GHI time series with a 15-minute time step. It finally saves the annual GHI average in a .tif format in the folder “Input_data\Rasterized_data”.
- “Map_construction” and “model.model3”: Together, this QGIS project and model enable the visualization of all input datasets and rasterize each dataset to a 0.2-degree resolution, using the coordinate system EPSG:4326 and the following extent: -20.1,-37.1 : 54.1,40.1. This ensures compatibility with MatLab for further analysis. It finally saves all the rasterized datasets in the folder “Input_data\Rasterized_data”.
- “Main_AHP.mat”: This MatLab code normalizes the datasets and classifies them by calculating their associated weights, either using the Analytic Hierarchy Process (AHP) or by assuming equal weights, as outlined in the Policy Research Working Paper. It also computes the risk of over-exploitation for each GDE. The code generates Tables 2, 3, and 4, as well as Figures 4b and 4d from the Policy Research Working Paper. Finally, it rasterizes the results on the risk of over-exploitation, making them compatible with QGIS for visualization of the results, and store them in the folder “Results”. The results are named “XXXRisk_YYY”. XXX can be “Polygon”, “Lines” or “Point”, depending on how the GDEs were initially stored in the raw shape files. YYY can be either “AHP” or “sameweight” depending on the way the weights have been

estimated. “AHP” means that the weights have been estimated using the AHP method, while “sameweight” means that the all the datasets have been equally weighted.

- “figure_PP.mat” saves the barcharts produced by “Main_AHP.mat” in .png in the folder “Results”.
- “Plot_results”: This QGIS project enables the visualization of the final results, i.e., Figures 2, 3 and 4a and c from the Policy Research Working Paper.

7. Disclaimer

The findings presented in this Policy Research Working Paper are based on big data analysis conducted using the open-source software QGIS. This reproducibility package was requested and its required format was specified two years after the study's results were initially shared. To build this package and comply to the specified format, we thus changed the architecture of the code to make it one-click runnable. This update introduced minor variations in the rasterization processes (i.e., the data cleaning steps required to transform raw datasets into usable formats for our analysis). These slight differences are shown in Figure 1, which illustrates the rasterization of GDEPolygons and GWresources. While the precise cause of these changes is unclear, they appear to be random, minor and to occur only along polygon/pixel borders.

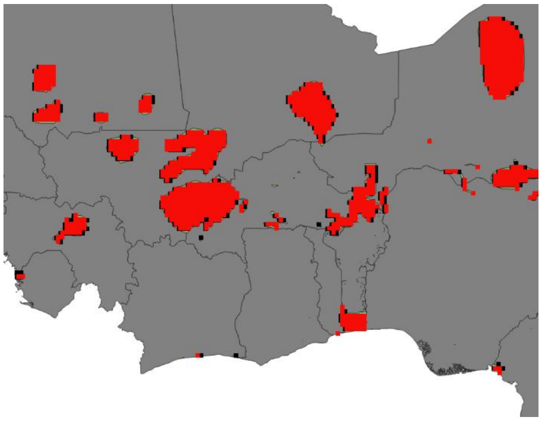

Figure 1 - Variations due to the rasterization process of “SSA_GDEs_Polygons.shp”: in red, original rasterized data, and in black, the newly introduced rasterized data with the reproducibility package. If the rasterization was identical, the black pixels should not be visible.

These minor discrepancies in the rasterization lead to slight differences in some of the numbers reported in the original Policy Research Working Paper (variations of up to +/- 2% difference). However, since these variations are very small, the overall results, trends, and analyses presented in the article remain unaffected, as illustrated for instance in Figure 2 (which is the main Figure of our paper).

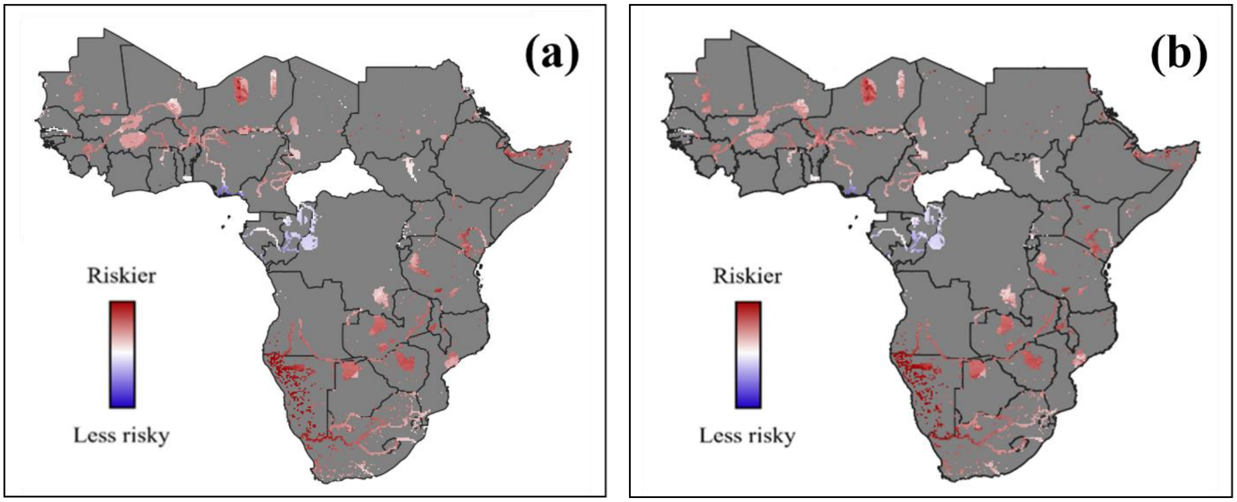

Figure 2 – Comparison between (a) the reproduction of Figure 2a of the Policy Research Working Paper with the reproducibility package, and (b) Figure 2a from the Policy Research Working Paper.

The only notable difference between the numbers reported in the Policy Research Working Paper and those from the reproducibility package is observed in Figure 4b of the paper. This discrepancy is highlighted in Figure 3 below. Specifically, the percentage of GDEs with a risk of overexploitation between 4 and 5 relying on Complex aquifers shifts from 11% to 19%, while the percentage of GDEs with the same risk level relying on Major Alluvial aquifers shifts from 10% to 0%. However, GDEs with a risk of overexploitation between 4 and 5 represent only ~57 out of the 4133 cells (~1% of the total). Therefore, the ~10% difference between the results of the reproducibility package and the paper in this 4-5 column corresponds to just 6 cells (10% of 57) out of 4133. Also, as this column concerned so few pixels, it was anyway not analyzed in the Policy Research Working Paper. Thus, this difference observed here also does not influence the analysis and discussion of the paper. However, the paper was updated to match the most updated results ran by the authors on December 23th.

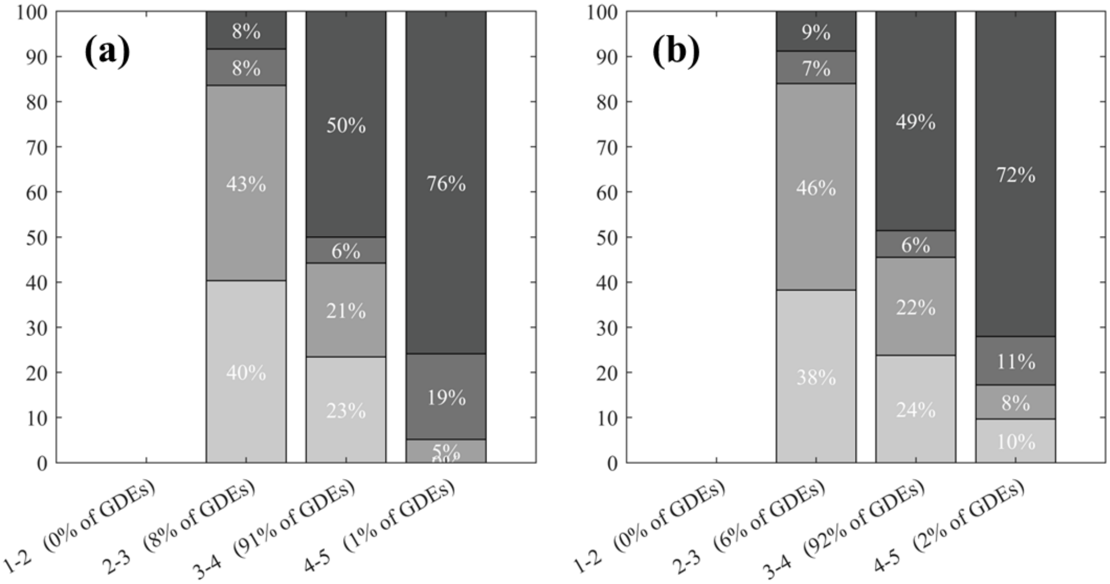

Figure 3 – Comparison between (a) the reproduction of Figure 4b of the Policy Research Working Paper with the reproducibility package, and (b) Figure 4b from the Policy Research Working Paper.