

README for the Replication Package — Cheap Energy Might Not Be Enough: A Trade Model of AI Compute Services

This package reproduces the model calibration and the core quantitative results of the paper from analysis-ready input data, using a 14-step Stata pipeline orchestrated by a single master script.

1. Quick start

1. Install Stata (version **16 or later** recommended; the code uses only base Stata — no user-written packages are required).
2. Unzip / copy this `Replication` folder anywhere on your machine.
3. In Stata, change to this folder and run the master script:

```
cd "C:/path/to/Replication"  
do run_all.do
```

`run_all.do` auto-detects its location from the working directory. If that is inconvenient, hard-code the path on the `global root` line near the top of the script.

Expected runtime: a few minutes on a typical laptop. A full transcript is written to `output/run_all_log.txt`.

2. What this package reproduces

This is a **Stata-only** package that reproduces **every data table in the paper** (Tables 1, 2, 3, A1–A8). Steps 01–14 build the core model (calibration, regimes, capacity equilibrium, inference sourcing, welfare, sensitivity, reliability, Kyrgyzstan DCF); steps 15–19 add the v33 specifications and export the tables:

- **15** Symmetric-LRMC cost-recovery (Table 3/A2 col 2)
- **16** Host-country WACC channel (Table 3 col 4)

- **17** Bilateral sovereignty premium λ_{ij} (Table 3/A2 col 3)
- **18** Construction-cost OLS (Table A7)
- **19** Writes all eleven tables to `output/table*.csv`

The constant tables in steps 15–17 (carbon prices/intensities, OECD scope, cross-subsidies, income-group WACC bands, geopolitical-bloc matrix, EU/APEC/DEPA membership) are transcribed from the paper's Python generator.

Fidelity vs. the published tables (per the agreed "qualitative match")

Published table	Stata reproduction	Status
Table A1 (calibration params)	full	☑ matches
Table 3/A2 col (1) Raw	observed-price ranking	☑ matches
Table 3/A2 col (2) Symmetric LRMC	top-5 KGZ/ETH/XKX/CAN/TJK; p_T = \$1.598 (paper \$1.604)	☑ matches
Table 3 col (4) CR + WACC	HIC \$1.58 / LIC \$1.87 / gap \$0.29	☑ exact
Table A3 (sensitivity)	step 12 scenarios	⚠ same method, scenario set differs from paper's
Table A5 / A6 (DCF)	NPV \$353M / IRR 17.6% / payback Yr 6	☑ matches
Table A7 (construction OLS)	N=37, R ² ≈0.48, same regressors	☑ structure matches (small-N coeffs insignificant)
Table 3/A2 col (3) Bilateral	λ_{ij} matrix + training-side sourcing; HHI_T≈0.58, welfare≈1.7%	⚠ partial — a training-side proxy; does not reproduce the published bilateral welfare (4.7%) / HHI (0.46), which also include the inference side and full deadweight-loss components
Tables 1, 2, A4, A8	definitional/static content	☑ reproduced as content

Caveat on the bilateral column (col 3): step 17 implements the bilateral premium and per-buyer *training* sourcing only; the full bilateral equilibrium (inference hubs + the two-component welfare decomposition) is not ported, so its headline welfare/HHI differ from the paper. Treat col (3) as indicative.

Notes:

- `data/` holds the **analysis-ready inputs**; upstream construction of a few (temperature from raw ERA5, latency from raw pings) was done in Python and is not part of this pipeline (raw files are still included — see §5).
- The package reproduces the table **values** as CSV; it does not typeset the formatted Word tables or Figure 1 (the model-structure diagram).

3. Repository layout

```

Replication/
├── README.md           this file
├── run_all.do          master script – sets parameters & runs all 19 steps
├── code/               the Stata pipeline (driven entirely by run_all.do)
│   ├── _solver_program.do      equilibrium solver (loaded first)
│   ├── 01_prep_electricity.do
│   ├── 02_prep_temperature.do
│   ├── 03_prep_construction.do
│   ├── 04_calibrate_costs.do
│   ├── 05_prep_latency.do
│   ├── 06_regime_assignment.do
│   ├── 07_demand_shares.do
│   ├── 08_capacity_equilibrium.do
│   ├── 09_cost_recovery.do
│   ├── 10_inference_sourcing.do
│   ├── 11_welfare_sovereignty.do
│   ├── 12_sensitivity.do
│   ├── 13_reliability.do
│   ├── 14_kyrgyzstan_dcf.do
│   ├── 15_symmetric_lrnc.do    (Table 3/A2 col 2)
│   ├── 16_wacc_channel.do      (Table 3 col 4)
│   ├── 17_bilateral_lambda.do  (Table 3/A2 col 3)
│   ├── 18_construction_regression.do (Table A7)
│   └── 19_export_tables.do      (writes all table*.csv)
├── data/               all input data (see §5; provenance in the DAS, §6)
├── intermediate_datasets/ the 4 analysis-ready intermediate datasets that
│   │                   bridge raw sources and the pipeline, with full metadata
│   │                   and data dictionaries (the self-contained reproducibility
│   │                   starting point; byte-identical to the copies in data/)
├── preprocessing/     Python scripts that build 3 processed inputs from raw
│   │                   sources (transparency; not needed to run the pipeline)
├── output/             result files + the 11 exported table*.csv; shipped with
│   │                   REFERENCE results from a validated run, overwritten on re-
├── run
└── temp/               intermediate .dta files (created on each run)

```

All model parameters (GPU price, lifetime, utilization, PUE coefficients, latency rate τ , sovereignty premium λ , demand totals, the 13 cost-recovery prices, and the Kyrgyzstan DCF assumptions) are defined **once** at the top of `run_all.do` — change them there to explore alternatives.

4. Pipeline steps and outputs

Step	Script	Produces	Where it appears in the paper
1	01_prep_electricity.do	cleaned electricity prices (temp/)	input to cost calibration
2	02_prep_temperature.do	peak summer temperatures (temp/)	PUE / cooling term
3	03_prep_construction.do	per-watt construction costs (temp/)	construction term in c_j
4	04_calibrate_costs.do	output/calibration_results_stata.csv + temp/calibration_results.dta	cost rankings — Table 3 col (1), Tables A1/A2
5	05_prep_latency.do	symmetric latency matrix (temp/)	inference latency cost
6	06_regime_assignment.do	trade-regime types (temp/)	regime taxonomy — Table 1
7	07_demand_shares.do	demand shares ω_k (temp/)	demand calibration (§3.3)
8	08_capacity_equilibrium.do	training equilibrium p_T , HHI_T, shadow values (temp/)	trade-flow results (§6.2)
9	09_cost_recovery.do	cost-recovery costs/rankings (temp/)	Table 3 col (2) — 13-country spec
10	10_inference_sourcing.do	inference hubs & shares (temp/)	inference results (§6.2)
11	11_welfare_sovereignty.do	welfare cost of sovereignty (temp/)	welfare results (§6.2 / §7)
12	12_sensitivity.do	output/sensitivity_results.dta	sensitivity — Appendix C / Table A3
13	13_reliability.do	output/reliability_rankings.dta	reliability-adjusted robustness
14	14_kyrgyzstan_dcf.do	output/kyrgyzstan_dcf.dta, output/kyrgyzstan_dcf_sensitivity.dta	Kyrgyzstan DCF — Appendix D
15	15_symmetric_lrmc.do	temp/symmetric_lrmc.dta	Table 3/A2 col (2) symmetric LRM; $p_T \approx \$1.60$

16	16_wacc_channel.do	temp/wacc_channel.dta	Table 3 col (4) CR + host WACC
17	17_bilateral_lambda.do	temp/bilateral_lambda.dta	Table 3/A2 col (3) bilateral λ (training side)
18	18_construction_regression.do	output/tableA7_construction_regression.csv	Table A7
19	19_export_tables.do	output/table{1,2,3,A1–A8}_*.csv	all paper tables (see §2 fidelity table)

(Steps 1–3 and 5–11 write intermediate `.dta` files to `temp/`; persistent `output/` files are bolded. Step 19 writes the eleven `table*.csv` files.)

5. Data inputs

`data/` contains every data file from the project, copied as-is. The Stata pipeline reads **only the following eight analysis-ready files**:

File	Used by	Contents
calibration_results_v3.csv	04	per-country calibration inputs (prices, climate, governance)
country_electricity_prices.csv	01	retail electricity prices (\$/kWh)
country_temperatures.csv	02	peak summer temperatures (°C)
predicted_construction_costs.csv	03	per-watt DC construction costs
country_pair_latency.csv	05	bilateral round-trip latency (ms)
dc_capacity_estimates.csv	07	installed data-center capacity (MW) → demand shares
grid_capacity_estimates.csv	08	national grid capacity (MW) → capacity ceilings
reliability_index.csv	13	grid-reliability index

Large raw files EXCLUDED from this package (not read by the Stata pipeline; they are upstream inputs for the Python preprocessing only): the raw ERA5 climate grid (`*.nc`, ~110 MB), `wondernetwork_pings.csv.gz` (~49 MB), and `ripe_atlas_latency.csv` (~37 MB). They were dropped to keep the package small

(~17 MB vs. ~210 MB); the analysis-ready CSVs derived from them (`country_temperatures.csv`, `country_pair_latency.csv`) ARE included, so the Stata pipeline runs fully. The raw files remain available in the full project repository if upstream reconstruction is needed.

Intermediate datasets bundled. The four analysis-ready intermediate datasets (temperature, latency, construction costs, electricity prices) are small, so they are redistributed directly in `intermediate_datasets/` with full metadata and variable-level data dictionaries. The package is therefore self-contained as a reproducibility starting point: verification can begin from these intermediates without retrieving the large raw sources. (The same four files are also in `data/`, where the pipeline reads them; the copies are byte-identical.)

For complete provenance of every input dataset — source, URL, access date, citation, and license — see the **Data Availability Statement** in §6 below.

6. Data Availability Statement

This section documents the provenance of every dataset used as input to the package, following the Social Science Data Editors' guidance on citing data. Datasets fall into three groups:

- **(A) Externally sourced data used directly** by the pipeline.
- **(B) Processed inputs** derived from raw sources by the scripts in `preprocessing/` (see that folder's README). The analysis-ready CSV is shipped; for two of them the large raw source file is not redistributed.
- **(C) Auxiliary files** shipped in `data/` as-is but **not read** by this Stata pipeline (legacy Python-era calibration inputs), listed for completeness at the end.

On access dates and compiled data. All inputs were assembled by the research team during the project (the underlying data were downloaded or compiled in **February 2026**, except the auxiliary ξ -calibration caches in group C, pulled in April 2026). Access dates below reflect that assembly. The underlying datasets refer to the reference periods noted in each entry. Several inputs are **authors' compilations or estimates** rather than verbatim

downloads of a single dataset; these are labelled explicitly. A small number of genuinely open questions (redistribution permissions for proprietary figures) are listed at the end.

(A) Externally sourced data used directly

A1. Data-centre construction cost index

- *Filename(s)*: `data/dcci_2025_construction_costs.csv` (= `data/raw/construction/dcci_construction_cost_2025.csv`)
- *Source*: Turner & Townsend — *Data Centre Cost Index 2025* (52 markets; the file's `index` column is the published relative index, `usd_per_watt` the corresponding cost level).
- *URL*: <https://www.turnerandt Townsend.com/en/perspectives/data-centre-cost-index/>
- *Access date*: February 2026 (report reference year: 2025)
- *Citation*: Turner & Townsend (2025). *Data Centre Cost Index 2025*.
- *License*: Published industry report; per-market figures transcribed by the authors. Proprietary — see redistribution question (1) at the end of this section.

A2. Data-centre installed capacity (MW)

- *Filename(s)*: `data/dc_capacity_estimates.csv` (= `data/raw/capacity/dc_capacity_mw.csv`)
- *Source*: **Authors' estimates**. For most countries (≈ 41 of 86) capacity is *imputed* as data-centre count $\times$ a regional average facility size; for ≈ 20 countries it is anchored to published figures from Synergy Research Group, CBRE, Cushman & Wakefield, Arizton, Mordor Intelligence, and the IEA. The per-row basis is recorded in the file's `source` column (e.g. "DC count x regional avg", "CBRE EMEA estimate", "Synergy Research/IEA 2024").
- *URL*: Synergy Research Group <https://www.srgresearch.com/> ; CBRE <https://www.cbre.com/> ; Cushman & Wakefield <https://www.cushmanwakefield.com/> ; Arizton <https://www.arizton.com/> ; Mordor Intelligence <https://www.mordorintelligence.com/> ; IEA <https://www.iea.org/>
- *Access date*: February 2026 (underlying industry figures: 2024–2025)
- *Citation*: Authors' estimates (2026), anchored to the industry sources named in the file's `source` column.
- *License*: Authors' compiled estimates, released with the package; underlying proprietary figures used only as anchor points — see redistribution question (1).

A3. National grid / generation capacity

- *Filename(s)*: `data/grid_capacity_estimates.csv` (= `data/raw/capacity/grid_capacity_estimates.csv`)

- **Source:** **Authors' construction** from World Bank World Development Indicators (electricity consumption per capita; population) scaled to installed-capacity and GPU-hour ceilings.
- **URL:** <https://databank.worldbank.org/source/world-development-indicators>
- **Access date:** February 2026 (WDI reference year: 2023)
- **Citation:** Derived by the authors from World Bank, *World Development Indicators* (electric power consumption; population).
- **License:** Underlying World Bank data CC BY 4.0; derived series released with the package.

A4. Worldwide Governance Indicators (rule of law, regulatory quality)

- **Filename(s):** `data/raw/governance/wgi_rule_of_law.csv`; values also embedded in `data/calibration_results_v3.csv` and `data/reliability_index.csv` (governance column). (Re-pulled to `data/wgi_cache.csv`, group C, for the auxiliary ξ work.)
- **Source:** World Bank, Worldwide Governance Indicators (WGI) — Rule of Law and Regulatory Quality percentile ranks.
- **URL:** <https://www.worldbank.org/en/publication/worldwide-governance-indicators>
- **Access date:** February 2026 (WGI most recent vintage at access)
- **Citation:** Kaufmann, D., A. Kraay, and M. Mastruzzi. *The Worldwide Governance Indicators*. World Bank.
- **License:** Creative Commons Attribution 4.0 (CC BY 4.0).

A5. World Development Indicators (GDP per capita PPP, population, urban share)

- **Filename(s):** `data/wb_gdp_per_capita_ppp_2023.csv`, `data/wb_population_2023.csv`, `data/wb_urban_share_2023.csv`
- **Source:** World Bank, World Development Indicators (WDI)
- **URL:** <https://databank.worldbank.org/source/world-development-indicators>
- **Access date:** February 2026 (indicator reference year: 2023)
- **Citation:** World Bank. *World Development Indicators*: GDP per capita, PPP (NY.GDP.PCAP.PP.CD); Population, total (SP.POP.TOTL); Urban population (% of total) (SP.URB.TOTL.IN.ZS), 2023.
- **License:** CC BY 4.0.

A6. World Bank country–region classification

- **Filename(s):** `data/wb_country_regions.csv`
- **Source:** World Bank country and lending-groups classification
- **URL:** <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519>
- **Access date:** February 2026
- **Citation:** World Bank. *Country and Lending Groups* (region classification).
- **License:** CC BY 4.0.

A7. Seismic-hazard zone flag

- *Filename(s)*: data/seismic_zones.csv (binary seismic_high $\in \{0,1\}$ for 81 countries)
- *Source*: **Authors' coding**. A coarse high-/low-seismic-hazard indicator assigned by the authors from widely-known seismic-hazard geography (high = countries on or near major plate boundaries / active seismic belts, e.g. Japan, Indonesia, Iran, Türkiye, Chile, New Zealand). It is **not** a download from a single hazard dataset; it is consistent with the qualitative classifications of programs such as the Global Seismic Hazard Assessment Program (GSHAP) and USGS hazard maps.
- *URL*: Reference classifications: GSHAP <https://www.gfz-potsdam.de/en/section/seismic-hazard-and-risk-dynamics/> ; USGS <https://www.usgs.gov/programs/earthquake-hazards>
- *Access date*: February 2026 (authors' coding)
- *Citation*: Authors' coding (2026), informed by GSHAP/USGS seismic-hazard classifications.
- *License*: Authors' coding, released with the package (public domain).

A8. Grid-reliability index

- *Filename(s)*: data/reliability_index.csv (columns: xi_reliability, governance, grid_quality, sanctions_adj)
- *Source*: **Authors' index**. Combines the WGI governance percentile (A4) with a qualitative grid-quality score assigned by the authors (informed by general infrastructure-quality knowledge and reliability reporting), plus a sanctions adjustment. It is a constructed index, not a verbatim third-party dataset.
- *URL*: Governance component: see A4.
- *Access date*: February 2026 (authors' construction)
- *Citation*: Authors' construction (2026) from World Bank WGI (governance) and an authors-assigned grid-quality score.
- *License*: Authors' index, released with the package; WGI component CC BY 4.0.

(B) Processed inputs derived by preprocessing/ scripts

B1. Country temperatures (generating script: `preprocessing/process\_temperature.py`)

- *Filename(s)*: data/country_temperatures.csv (shipped). Raw source: data/raw/climate/era5_2m_temperature_monthly.nc (**not** redistributed, ~110 MB) + Natural Earth boundaries (downloaded at runtime).
- *Source*: ECMWF Copernicus Climate Change Service (C3S), ERA5 reanalysis (2-m temperature); Natural Earth (country boundaries).
- *URL*: <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-monthly-means> ; <https://www.naturalearthdata.com/downloads/110m-cultural-vectors/>
- *Access date*: February 2026 (ERA5 monthly data covering 2020–2024); Natural Earth fetched at runtime

- *Citation*: Hersbach, H., et al. (2023). *ERA5 monthly averaged data on single levels from 1940 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.f17050d7. Boundaries: *Natural Earth, 1:110m Admin-0 Countries*.
- *License*: ERA5 — Copernicus Licence (free reuse and redistribution with attribution). Natural Earth — public domain.

B2. Country-pair latency (*generating script: `preprocessing/process\_latency.py`*)

- *Filename(s)*: data/country_pair_latency.csv (shipped) + data/wondernetwork_servers.csv (shipped). Raw source: data/raw/latency/wondernetwork_pings.csv.gz (**not** redistributed, ~49 MB).
- *Source*: WonderNetwork, Global Ping Statistics (server-to-server round-trip times). (*Note: only WonderNetwork is used; the latency input is not derived from RIPE Atlas.*)
- *URL*: <https://wondernetwork.com/pings>
- *Access date*: February 2026
- *Citation*: WonderNetwork. *Global Ping Statistics* [data set]. <https://wondernetwork.com/pings>
- *License*: Proprietary (website terms of use). The raw ping table is **not** redistributed for this reason; only the aggregated country-pair averages are shipped — see redistribution question (2).

B3. Predicted construction costs (*generating script: `preprocessing/predict\_construction\_costs.py`*)

- *Filename(s)*: data/predicted_construction_costs.csv (shipped; the script reproduces it byte-for-byte from the inputs below).
- *Source*: Derived — OLS on Turner & Townsend DCCI 2025 (A1) with World Bank WDI covariates (A5) and the seismic flag (A7).
- *URL*: see A1, A5, A7.
- *Access date*: February 2026 (see A1, A5, A7).
- *Citation*: Authors' calculation; see A1, A5, A7 for underlying sources.
- *License*: Derived dataset released with the package; underlying-source terms in A1/A5/A7 apply.

B4. Country electricity prices (*merged; no single generating script — see `preprocessing/README.md`*)

- *Filename(s)*: data/country_electricity_prices.csv (= the master data/raw/electricity/non_european_prices.csv). Component raw inputs: data/eurostat_electricity_prices.csv (= data/raw/electricity/eurostat_nrg_pc_205.csv), data/eia_electricity_prices.csv (= data/raw/electricity/eia_average_retail.csv).
- *Source*: (i) Eurostat (European countries); (ii) U.S. Energy Information Administration (United States); (iii) **authors' manual compilation** for ≈20 non-European/non-EIA

countries from GlobalPetrolPrices, national regulators/utilities, and the IEA. The per-row source is recorded in the file's `source` column (rows tagged `research`).

- *URL*: Eurostat https://ec.europa.eu/eurostat/databrowser/view/nrg_pc_205/ ; EIA <https://www.eia.gov/electricity/data.php> ; GlobalPetrolPrices https://www.globalpetrolprices.com/electricity_prices/
- *Access date*: February 2026 (Eurostat reference: 2024 H2; EIA reference: 2024; GlobalPetrolPrices/regulators: 2024–2025 figures)
- *Citation*: Eurostat, *Electricity prices for non-household consumers — bi-annual data* [nrg_pc_205]. U.S. Energy Information Administration, *Electricity data* (average retail price, industrial sector). Authors' compilation from GlobalPetrolPrices and national regulators (per-row `source` column).
- *License*: Eurostat — free reuse with acknowledgement (Commission Decision 2011/833/EU). EIA — U.S. Government work, public domain. GlobalPetrolPrices — proprietary/terms of use; figures used as reference points (see redistribution question (3)).

B5. Master calibration table (generating script: ``Programs/calibrate_model_v3.py``, full project repo)

- *Filename(s)*: `data/calibration_results_v3.csv`
- *Source*: Derived — combines B1, B2, B4, A4, A5, A7 into per-country cost-model inputs.
- *Access date*: February 2026 (derived).
- *Citation*: Authors' calculation; underlying sources as cross-referenced.
- *License*: Derived dataset released with the package.
- *Note*: This Python calibration output is read by Stata step 04. The Stata package independently re-derives the same quantities from the component inputs; this file is shipped so the pipeline runs without the Python step.

(C) Auxiliary files shipped but NOT read by this pipeline

The following are included in `data/` for archival completeness but are **not** read by any step of this Stata pipeline (they are legacy Python-era ξ -calibration inputs and intermediate snapshots, assembled April 2026): `wgi_cache.csv`, `fdi_cache.csv`, `xi_scenarios.xlsx`, `xi_scenarios_data.csv`, `xi_calibration_test.xlsx`, `form_b_simulations.xlsx`, `country_datacenters.csv`, `country_results_v29.csv`, `tableA3_v29.csv`, `calibration_regimes_v3.csv`, `model_parameters.csv`, `us_state_electricity_prices.csv`, and `data/output/table3_recalculated.xlsx`. They can be omitted without affecting reproduction. (`fdi_cache.csv` was pulled from UNCTAD/World Bank FDI statistics; it does not enter the published results.)

Open redistribution questions (proprietary sources)

The provenance above is complete. The only items not fully settled are redistribution permissions for three proprietary sources, handled conservatively in this package:

4. **Turner & Townsend DCCI 2025 (A1) and the industry capacity figures (A2)** — transcribed/anchor figures from proprietary reports. Retain if the journal's policy permits citing transcribed values; otherwise they can be replaced by citations to the reports.

5. **WonderNetwork (B2)** — the raw ping table is excluded; only aggregated country-pair averages are shipped. Confirm this aggregate may be redistributed under the site's terms.

6. **GlobalPetrolPrices (B4)** — individual reference prices used in the manual non-European compilation; values are cited per-row, not redistributed in bulk.

7. Reproducibility notes

- The pipeline is deterministic — no random number generation — so results are bit-stable across runs and platforms.
- `run_all.do` recreates `temp/` and `output/` on each run and overwrites prior outputs; intermediate `.dta` files in `temp/` can be deleted freely between runs.
- No user-written Stata commands (`ssc install`) are required.
- If a step errors, the log at `output/run_all_log.txt` records the exact step and message.

8. Citation

Lokshin, M. (2026). Cheap Energy Might Not Be Enough: A Trade Model of AI Compute Services. Working paper.