**Bias-Correction Pipeline (ANA gauges vs. CHIRPS / IMERG / PERSIANN / BR-DWGD “Xavier”)**

This manual explains how to run the Python scripts in **Testing.zip** to estimate station bias factors from gauge–satellite pairs, interpolate them to a grid, check results with plots, and apply the correction to annual-maximum rasters. It also highlights what to change in each script and the key parameters.

**0) What the pipeline does (in one minute)**

1. **Find extreme days at gauges** (e.g., >P99 with declustering) and **sample the product** (CHIRPS/IMERG/PERSIANN/Xavier) on those dates/locations → yearly CSVs with pr\_g, product value, and ratio = pr\_g / product.
2. **Aggregate pairs per station** to a single bias factor **ζ** (median/mean of ratios, or slope through origin) → zeta\_per\_station.csv.
3. **Interpolate ζ to a grid** (IDW or kriging) → zeta\_map\_\*.tif.
4. **Visual QC**: scatter/density plots of gauge vs. product before/after correction (+ optional bias map panel).
5. **Apply ζ grid to rasters** (annual max daily) → bias-corrected GeoTIFFs.

Equation used for correction (time-invariant):

**1) Prerequisites**

* **Python 3.9+** recommended.
* Install packages (use a clean environment):
* pip install pandas numpy matplotlib rasterio geopandas shapely scipy tqdm

Optional for kriging:

pip install pykrige

* **Google Earth Engine (GEE)** for step 01:
* pip install earthengine-api
* earthengine authenticate
* Input: a **wide ANA CSV** with one row per station and **daily columns** named like 1/1/1994, plus metadata columns:
  + Station ID (e.g., Code), Latitude, Longitude, and daily precipitation columns (mm).

**2) Folder & filename conventions**

The scripts assume per-product folders (you can rename them):

* CHIRPS pairs: G:\My Drive\chirps\_bias\_pairs\
* IMERG pairs: G:\My Drive\imerg\_bias\_pairs\
* PERSIANN: G:\My Drive\persiann\_bias\_pairs\
* Xavier: G:\My Drive\xavier\_bias\_pairs\

Each year’s extraction writes a CSV like:

<product>\_bias\_pairs\_YYYY.csv

The station bias table is:

zeta\_per\_station.csv # columns: station\_id, zeta, (optionally lat, lon, n\_pairs)

The interpolated raster is:

zeta\_map\_<res>.tif # EPSG:4326

**Windows paths**: use raw strings r'C:\...\' or forward slashes to avoid escape issues.

**3) Step-by-step**

**STEP 01 — Extract extreme gauge/product pairs in GEE (one task per year)**

Scripts:

* 01\_extract\_extreme\_pairs\_ANA\_CHIRPS\_GEE.py
* 01\_extract\_extreme\_pairs\_ANA\_IMERG\_GEE.py
* 01\_extract\_extreme\_pairs\_ANA\_PERSIANN\_GEE.py
* 01\_extract\_extreme\_pairs\_ANA\_XAVIER\_GEE.py

**Edit these at the top (“USER INPUTS”)**:

* CSV\_PATH – path to your wide ANA CSV.
* STATION\_ID\_COL, LAT\_COL, LON\_COL – column names in your CSV.
* PCT – extreme percentile (e.g., 0.99 for P99).
* MIN\_GAP\_DAYS – declustering window in days (keeps peaks ≥ this gap apart).
* MIN\_PER\_YEAR – minimum number of extremes to keep a station-year.
* YEAR\_START, YEAR\_END – year range to process (clips to CSV coverage).
* GEE\_PROJECT – your GEE project ID (or None).
* OUT\_GDRIVE\_FOLDER – Drive folder where yearly CSVs will be exported.
* SAMPLE\_SCALE\_M – sampling scale (product native resolution):
  + CHIRPS: 5560 (~0.05°)
  + IMERG/Xavier: 11132 (~0.1°)
  + PERSIANN-CDR: 27830 (~0.25°)
* CHUNK\_SIZE – number of stations per batch (controls memory/requests).

**Run** (one product at a time):

python 01\_extract\_extreme\_pairs\_ANA\_<PRODUCT>\_GEE.py

Then open GEE **Tasks** to monitor the yearly exports. Download or reference the CSVs from Drive.

**STEP 02 — Compute per-station bias factor ζ**

Scripts:

* 02\_compute\_station\_bias\_factors\_zeta\_from\_pairs.py (CHIRPS)
* 02\_compute\_station\_bias\_factors\_zeta\_from\_pairs\_IMERG.py
* 02\_compute\_station\_bias\_factors\_zeta\_from\_pairs\_PERSIANN.py
* 02\_compute\_station\_bias\_factors\_zeta\_from\_pairs\_XAVIER.py

**Edit**:

* IN\_DIR / IN\_GLOB – where the yearly pair CSVs live and how to match them (e.g., chirps\_bias\_pairs\_\*.csv).
* WIDE\_ANA\_CSV, STATION\_ID\_COL, LAT\_COL, LON\_COL – to merge coordinates (optional but recommended).
* OUT\_CSV – output path for zeta\_per\_station.csv.
* Filtering and estimator:
  + MIN\_SAMPLES – minimum pairs per station.
  + RATIO\_CLIP – keep ratios within range (e.g., (0.25, 5.0)).
  + MIN\_RAIN\_MM, MAX\_RAIN\_MM – drop tiny/huge event values.
  + ZETA\_METHOD – 'median' (robust default), 'mean', or 'ols\_slope' (slope through origin).

**Run**:

python 02\_compute\_station\_bias\_factors\_zeta\_from\_pairs\_<PRODUCT>.py

**Output**: zeta\_per\_station.csv with station\_id, zeta, n\_pairs, and (if merged) lat, lon.

**STEP 03 — Interpolate ζ to a grid (IDW or kriging)**

Scripts:

* 03\_interpolate\_zeta\_to\_grid\_IDW\_or\_kriging.py
* Product-specific duplicates set different defaults for resolution/output.

**Edit**:

* ZETA\_CSV – path to zeta\_per\_station.csv.
* BOUNDARY\_PATH – optional Brazil boundary (.shp/.geojson) to bound/mask the map (set "" to skip).
* OUT\_TIF – output GeoTIFF.
* Grid & method:
  + GRID\_RES\_DEG – output resolution (e.g., 0.05 for CHIRPS, 0.10 for IMERG/Xavier, 0.25 for PERSIANN).
  + PAD\_DEG – pad the bbox to avoid edge holes (e.g., 0.5).
  + METHOD – "idw" (fast) or "ok" (Ordinary Kriging).
* **IDW params**:
  + IDW\_POWER (e.g., 2.0), IDW\_K neighbors (e.g., 8), IDW\_CAP\_RANGE final clamp (e.g., (0.25, 10)).
* **Kriging params** (needs pykrige):
  + KRIG\_VARIOGRAM ('spherical', 'gaussian', etc.), KRIG\_NLAGS, KRIG\_ENABLE\_ANISOTROPY.
* Filtering:
  + MIN\_PAIRS – drop stations with fewer pairs (if n\_pairs exists).
  + ZETA\_CLIP – drop extreme ζ before interpolation, e.g., (0.05, 10).

**Run**:

python 03\_interpolate\_zeta\_to\_grid\_IDW\_or\_kriging.py

**Output**: zeta\_map\_\*.tif (EPSG:4326).

Tip: Match GRID\_RES\_DEG to the product you will correct to minimize resampling.

**STEP 04 — Visual QC: scatter/density plots (raw vs. corrected)**

Scripts:

* Single-product:  
  04\_apply\_bias\_correction\_and\_density\_plots\_<PRODUCT>.py
* Multi-product panel:  
  06\_apply\_bias\_correction\_and\_density\_plots\_ALL\_4x2.py  
  (header says 4×3; this script can include a bias-map panel as well.)

**Edit (single-product)**:

* PAIRS\_DIR, PAIRS\_GLOB – where the pairs CSVs live.
* ZETA\_CSV – the station ζ file (or it may join from PAIRS\_DIR).
* OUT\_PNG – output figure.
* Plot tuning (optional):  
  MIN\_MM, RATIO\_CLIP, NBINS, DOT\_SIZE, ALPHA, DPI, PCT\_LIM, etc.

**Edit (multi-product)**:

* CONFIGS – a list of dicts per product:
* CONFIGS = [
* {
* "name": "BR-DWGD",
* "pairs\_dir": r'G:\My Drive\xavier\_bias\_pairs',
* "pairs\_glob": "xavier\_bias\_pairs\_\*.csv",
* "zeta\_csv": r'G:\My Drive\xavier\_bias\_pairs\zeta\_per\_station.csv',
* "zeta\_tif": r'G:\My Drive\xavier\_bias\_pairs\zeta\_map\_010deg.tif',
* },
* {
* "name": "IMERG",
* "pairs\_dir": r'G:\My Drive\imerg\_bias\_pairs',
* "pairs\_glob": "imerg\_bias\_pairs\_\*.csv",
* "zeta\_csv": r'G:\My Drive\imerg\_bias\_pairs\zeta\_per\_station.csv',
* "zeta\_tif": r'G:\My Drive\imerg\_bias\_pairs\zeta\_map\_010deg.tif',
* },
* ...
* ]
* OUT\_PNG – final composite figure.
* Optional: STATES\_SHP, STATES\_EDGE\_COLOR/WIDTH for Brazil overlay; font path/size.

**Run**:

python 04\_apply\_bias\_correction\_and\_density\_plots\_<PRODUCT>.py

# or

python 06\_apply\_bias\_correction\_and\_density\_plots\_ALL\_4x2.py

**Output**: one or more .png figures (raw vs. corrected scatter-density; optional ζ map).

**STEP 05 — Apply ζ map to annual-max rasters**

Scripts:

* 05\_apply\_bias\_to\_AMaxDaily\_CHIRPS.py
* 05\_apply\_bias\_to\_AMaxDaily\_IMERG.py
* 05\_apply\_bias\_to\_AMaxDaily\_PERSIANN.py
* 05\_apply\_bias\_to\_AMaxDaily\_XAVIER.py

**Edit**:

* **Input rasters**:
  + CHIRPS: IN\_DIR, IN\_GLOB (e.g., CHIRPS\_MaxDaily\_0p1deg\_\*\_Brazil.tif)
  + IMERG/PERSIANN: IN\_DIR, IN\_GLOB or regex as included.
  + Xavier: IN\_DIR, IN\_REGEX (e.g., ^BR\_DWGD\_prmax\_(\d{4})...)
* ZETA\_TIF – path to your zeta\_map\_\*.tif.
* OUT\_DIR – output directory for corrected max rasters.
* Resampling & caps:
  + RESAMPLE – how ζ is resampled onto the data grid (e.g., Resampling.bilinear).
  + CAP\_ZETA – clamp ζ (e.g., (0.25, 5.0)) to avoid unrealistic corrections.
  + MAX\_RAIN\_MM – cap corrected values if needed (e.g., 400.0).
  + COMPRESS, ZLEVEL – GeoTIFF compression settings.

**Run**:

python 05\_apply\_bias\_to\_AMaxDaily\_<PRODUCT>.py

**Output**: bias-corrected annual-maximum GeoTIFFs written to OUT\_DIR.

**4) “What do I change?” — Quick cheat-sheet by script**

**01\_extract\_extreme\_pairs\_ANA\_<PRODUCT>\_GEE.py**

* Required: CSV\_PATH, STATION\_ID\_COL, LAT\_COL, LON\_COL, YEAR\_START, YEAR\_END, OUT\_GDRIVE\_FOLDER.
* Optional: PCT (e.g., 0.95–0.995), MIN\_GAP\_DAYS (2–5), SAMPLE\_SCALE\_M (product native), CHUNK\_SIZE.
* For Xavier (BR-DWGD) scripts: keep the product-specific sampling and any scale/offset if present in code comments.

**02\_compute\_station\_bias\_factors\_\*.py**

* Required: IN\_DIR, IN\_GLOB, OUT\_CSV.
* Optional: WIDE\_ANA\_CSV (+ ID/lat/lon columns) to attach coordinates.
* Filters/Estimator: MIN\_SAMPLES, RATIO\_CLIP, MIN\_RAIN\_MM, MAX\_RAIN\_MM, ZETA\_METHOD.

**03\_interpolate\_zeta\_to\_grid\_IDW\_or\_kriging.py**

* Required: ZETA\_CSV, OUT\_TIF.
* Grid: GRID\_RES\_DEG, PAD\_DEG.
* Method & params: METHOD, IDW\_POWER, IDW\_K, KRIG\_\*.
* Filtering: MIN\_PAIRS, ZETA\_CLIP.
* Optional boundary mask: BOUNDARY\_PATH.

**04\_apply\_bias\_correction\_and\_density\_plots\_\*.py**

* Required: PAIRS\_DIR, PAIRS\_GLOB, ZETA\_CSV, OUT\_PNG (or CONFIGS for the 4-panel/4×3 script).
* Styling: MIN\_MM, RATIO\_CLIP, NBINS, DOT\_SIZE, DPI, etc.
* Optional Brazil states overlay & custom fonts in the ALL-products script.

**05\_apply\_bias\_to\_AMaxDaily\_\*.py**

* Required: IN\_DIR, file matcher (IN\_GLOB or IN\_REGEX), ZETA\_TIF, OUT\_DIR.
* Optional caps/quality: CAP\_ZETA, MAX\_RAIN\_MM, RESAMPLE, COMPRESS, ZLEVEL.

**5) Tips & troubleshooting**

* **GEE auth**: run earthengine authenticate once; ensure GEE\_PROJECT (if used) matches your Cloud project. Check the **Tasks** tab for progress/errors.
* **Date columns**: the scripts expect **daily headers like** 1/1/1994. If your CSV uses another format, adapt the parser in step 01 (look for a function that parses date headers).
* **Few extremes?** Lower PCT (e.g., 0.98) or reduce MIN\_GAP\_DAYS. Ensure MIN\_PER\_YEAR is not filtering valid stations away.
* **Wild ratios**: tighten RATIO\_CLIP and set reasonable MIN\_RAIN\_MM to avoid dividing by tiny numbers.
* **Missing coordinates in ζ table**: provide WIDE\_ANA\_CSV + column names in step 02 so step 03 can interpolate.
* **CRS/Alignment**: all ζ rasters are EPSG:4326. The apply scripts resample ζ to the product grid. Keep GRID\_RES\_DEG close to the product’s native resolution to minimize smoothing.
* **Memory**: step 01 processes stations in chunks. Increase/decrease CHUNK\_SIZE as needed.
* **Output sanity**: the single- and multi-product plotting scripts are meant for quick QA. You should see the corrected panel cluster along the 1:1 line and ζ map with realistic values (typically ~0.5–2, but product/region dependent).

**6) Minimal run order (example)**

**CHIRPS**

1. python 01\_extract\_extreme\_pairs\_ANA\_CHIRPS\_GEE.py
2. python 02\_compute\_station\_bias\_factors\_zeta\_from\_pairs.py
3. python 03\_interpolate\_zeta\_to\_grid\_IDW\_or\_kriging.py *(set GRID\_RES\_DEG=0.05)*
4. python 04\_apply\_bias\_correction\_and\_density\_plots\_CHIRPS.py
5. python 05\_apply\_bias\_to\_AMaxDaily\_CHIRPS.py

**IMERG / PERSIANN / Xavier**

* Repeat with the corresponding scripts and set GRID\_RES\_DEG to 0.10 (IMERG/Xavier) or 0.25 (PERSIANN).

**7) Parameter defaults by product (suggested)**

* **CHIRPS**: SAMPLE\_SCALE\_M=5560, GRID\_RES\_DEG=0.05, ζ clamp (0.25, 5).
* **IMERG**: SAMPLE\_SCALE\_M=11132, GRID\_RES\_DEG=0.10, ζ clamp (0.25, 5).
* **PERSIANN**: SAMPLE\_SCALE\_M=27830, GRID\_RES\_DEG=0.25, ζ clamp (0.25, 5).
* **Xavier (BR-DWGD)**: similar to IMERG (0.1°); keep any product-specific scaling noted in the script header.

**8) File outputs you should expect**

* .../<product>\_bias\_pairs\_YYYY.csv (one per year)
* .../zeta\_per\_station.csv
* .../zeta\_map\_<res>.tif
* .../density\_raw\_vs\_corrected\_\*.png (and/or the multi-product composite)
* .../<product>\_MaxDaily\_\*\_Brazil\_BC.tif (bias-corrected annual maxima)