**Linking Hourly Climate Data to Weekly Health Outcomes: Sri Lanka Example / Guide**

Author: Jordan Clark, PhD | Data Scientist, Duke Global Health Institute (DGHI) / Co-founder, Klimo Insights

**Who this is for:**

Epidemiologists and public-health analysts who want to add climate context to weekly surveillance data (e.g., leptospirosis, dengue) without assuming prior experience with meteorology or geospatial analysis.

**What you will learn:**

How to turn hourly gridded climate data into epidemiology-ready weekly features at the district level, why each step is necessary, and how to avoid common pitfalls.

**1. The datasets, in plain language**

Health. Weekly case counts by district from Sri Lanka’s Weekly Epidemiological Reports (WER). Each record has a district name and a week-ending date. We harmonize district names so they match across sources.

Climate. ERA5 reanalysis, hourly, on a fixed latitude–longitude grid (~0.25°). We use air temperature, dew point, wind, solar radiation, and precipitation. Two precipitation series appear in ERA5 products and they are not the same:

* tp (“total precipitation”): accumulated depth; daily sums represent mm/day once converted from meters.
* mtpr (“mean total precipitation rate”): an average rate; summing over the day gives a comparable depth but via a rate pathway. We compute weekly sums for both and cross-check magnitudes.

Spatial boundaries. Administrative districts (ADM2) from GADM. We validate geometry and standardize names (e.g., “Nuwara-Eliya”).

Population (optional but recommended). Mid-year population by district to compute rates and offsets.

Time zone. ERA5 timestamps are UTC. Sri Lanka reporting is local time. We convert before we aggregate so that days and weeks reflect local experience.

**2. Key ideas before you start**

Grid cells vs. districts. Climate is reported on a grid; surveillance is by district. We build polygons for ERA5 grid cells and intersect them with district polygons in an equal-area projection so we can compute true areas in square meters. Each cell contributes to a district in proportion to the land area it overlaps. This prevents a coastal cell from dominating just because it happens to be large in degrees.

Daily first, then weekly. We summarize hourly variables to daily values at the cell level (means for temperature/humidity, min/max for daily extremes, sums for precipitation). Then we compute area-weighted daily district values. Finally, we assemble weekly features aligned to the epidemiological week.

Climatology and anomaly. For each district and week-of-year we compute a long-run average (the “climatology”). A weekly anomaly is the departure from that expected seasonal value. Both absolute levels (e.g., mm of rain) and anomalies are informative.

**3. Units, definitions, and what the columns mean**

* Temperature (°C). ta\_mean, ta\_min, ta\_max are daily district values; weekly features include the week’s mean, range, and upper quantiles (e.g., 95th percentile of daily maxima).
* Humidity (%). From temperature and dew point we compute relative humidity (rh) and vapor pressure deficit (vpd, kPa; higher means drier air). Weekly features use means.
* Solar radiation. ssrd converted to MJ/m² per day; weekly feature is the mean of daily values.
* Precipitation. tp\_sum and mtpr\_sum are daily totals in mm after conversion; weekly features are sums over the week. We also derive (a) number of “wet days” (e.g., days ≥ 10 mm), (b) the largest 3-day total inside the week, and (c) the longest wet spell length (number of consecutive wet days).
* Coverage. n\_days\_week is the count of days with usable data inside the week. Keep this for filtering rather than hard-masking.
* Anomalies. For selected variables, \_anom is observed minus the district’s typical value for that week-of-year; \_pct\_normal is observed divided by that typical value (with safeguards when the climatology is near zero).
* Lags and rolling windows. We generate \_lag1 … \_lag6 (weeks) for key features and 2- and 4-week rolling means/sums that include the current week.

*Why 10 mm for a “wet day”?* Many water-borne and vector-related mechanisms are sensitive to totals in the ~10 mm/day range; it’s a defensible starting point. You should test sensitivity at 5 mm and 20 mm in your context.

**4. The workflow, explained**

Step 1 — Convert hourly UTC to local time. Shift ERA5 timestamps to Asia/Colombo (UTC+5:30). Only then aggregate to daily values, otherwise late-night storms drift into the wrong local day.

Step 2 — Build area weights once. Create polygons for grid cells from their centers; intersect with district polygons in an equal-area CRS (e.g., EPSG:6933). For each cell–district pair, store the intersection area. These weights are reused for all days and variables.

Step 3 — Make daily district values. For each day and district, take the area-weighted average of cell-level daily means (e.g., temperature) and the area-weighted sum for precipitation totals. Check for negative precipitation artifacts and clip at zero.

Step 4 — Create weekly features aligned to surveillance. From the daily district series, summarize the days within each epidemiological week (use the WER’s week-ending date). Compute the features listed above. Record n\_days\_week. If you want a minimum coverage rule, filter later (e.g., keep weeks with ≥5 valid days).

Step 5 — Add memory and lags. Create 1–6 week lags for precipitation, temperature, humidity, and VPD. Build 2- and 4-week rolling means/sums with partial windows (e.g., allow 1 of 2 weeks to be present) to avoid propagating NAs. For antecedent rainfall memory, compute EWAP: the current week plus exponentially decayed prior weeks, where α (e.g., 0.8) controls how quickly the effect fades and K (e.g., 4) sets how many weeks matter. Choose α and K by plausibility and cross-validated fit.

Step 6 — Derive climatologies and anomalies. For each district and week-of-year, compute a baseline using the available historical period (or a designated “normal,” such as 1991–2020). Subtract to obtain anomalies and divide to obtain percent-of-normal. Retain both—the level and the departure often tell complementary stories.

Step 7 — Join to health data and population. Harmonize district names, merge weekly climate features to WER case counts by district and week, and add population to compute rates and offsets.

**5. Design choices and why they matter**

The most important choice is local-time aggregation before daily/weekly roll-ups. This avoids assigning exposure to the wrong epi week. The second is area-weighting rather than nearest-cell assignment, which provides a truer district average, especially along coasts or in elongated districts. We keep both ERA5 precipitation measures (tp and mtpr) to cross-validate units and catch problems early. We prefer soft coverage (carry n\_days\_week) over hard masking, because analysts differ in tolerance for missing days and some diseases are dominated by events that can occur in a two-day window.

**6. Quality assurance you can actually run**

Before trusting the features, do three quick checks. First, plot district-day rainfall histograms and ensure totals are non-negative and within plausible ranges; weekly sums should not be all zeros. Second, pick one district and one specific week, and hand-calculate the weekly precipitation total from daily values to confirm your code. Third, compare ERA5 weekly totals against any available station-based weekly totals in a few districts; you won’t expect exact agreement, but gross discrepancies reveal unit or conversion errors.

**7. Using the features in models**

For count outcomes, start with a negative binomial or Poisson model with a population offset:

cases\_d,w ~ offset(log(pop\_d,w)) + precip (levels + anomalies + lags)

+ temp/VPD (levels + anomalies + lags)

+ district fixed effects

+ smooth seasonality f(week-of-year)

For leptospirosis, examine precipitation in the current and prior one to two weeks, the longest wet spell, and EWAP. For dengue, explore broader lags (two to six weeks) and temperature/humidity metrics that influence vector ecology. Use cross-validation or out-of-sample prediction to decide which features help rather than selecting by p-values alone.

**8. A small worked example (conceptual)**

Suppose Colombo’s epidemiological week ends on Friday 2020-05-08. Convert all hourly ERA5 timestamps to local time, summarize to six daily values for that week (Saturday through Friday). Area-weight those daily values by cell-district overlap to obtain district-day series. The weekly precipitation sum is the arithmetic sum of those six daily totals; the longest wet spell is the longest run of days at or above 10 mm; the max 3-day total is the largest three-day moving sum within that week. Compute the weekly mean temperature and 95th percentile of daily maxima. Then add lag-1 and lag-2 values of precipitation and temperature, compute anomalies relative to “week 19” climatology, and join to the reported case count for that week.

**9. Common pitfalls (and how we avoid them)**

Misinterpreting precipitation units is the most frequent error: never assume; confirm whether your tp is meters or millimeters and how it was accumulated. Aggregating in UTC creates “day splitting” that shifts storms across local day boundaries; always convert first. Using geographic degrees for area weights biases district averages; always measure overlaps in an equal-area CRS. Hard-masking weeks with incomplete data can erase informative events; use a coverage flag and filter intentionally. Finally, climatologies built on too short a history will be noisy; when possible choose a stable baseline period.

**10. What to tune for your disease and setting**

The wet-day threshold, the EWAP decay α and horizon K, and the lag window should be chosen with a mix of biological plausibility and empirical validation. For water-exposure diseases like leptospirosis, short lags and wet-spell metrics often matter; for vector-borne diseases like dengue, temperature and humidity over longer windows can be important. Report the settings you chose and show that your conclusions are robust to reasonable alternatives.

**11. Glossary (brief, in order of appearance)**

ERA5: a global, gridded reconstruction of the atmosphere (a “reanalysis”) at hourly resolution.  
ADM2: second-level administrative areas—districts, in this context.  
Equal-area CRS: a map projection where areas are preserved; needed for valid square-meter weights.  
VPD: vapor pressure deficit (kPa), a measure of atmospheric dryness; higher means drier.  
WBGT: wet-bulb globe temperature, a heat-stress index (used when available).  
Climatology: the long-run average for a specific district and week-of-year.  
Anomaly: observed minus climatology (or observed divided by climatology for “percent-of-normal”).  
EWAP: exponentially weighted antecedent precipitation; a way to represent rainfall “memory.”