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

**Background**

Environmental conditions shape infectious disease risk through multiple pathways—pathogen survival, vector ecology, human behavior, and health-system performance. Yet climate and health datasets rarely line up cleanly. Reanalysis products such as ERA5 arrive as hourly values on a fixed spatial grid, while surveillance typically reports weekly counts for administrative units. Bridging that gap requires explicit choices about time zones, aggregation, spatial weighting, lag structures, and quality control. This note documents a concrete approach we implemented for Sri Lanka (leptospirosis and dengue), but the principles generalize to other settings.

**What and where**

We use hourly ERA5 fields (temperature, dew point, wind, solar radiation, precipitation, WBGT when available) from 1980–2024 and align them to Sri Lanka’s ADM2 districts. District boundaries are read from GADM and normalized to match the naming used in the Weekly Epidemiological Reports (WERs). Because ERA5 is gridded, not administrative, we convert grid-cell centers to polygons, intersect them with districts in an equal-area CRS, and compute cell-to-district area weights. Daily district values are then area-weighted averages (for means) or area-weighted summaries (for mins/maxes and daily precipitation sums) across all cells that overlap a district. This avoids bias toward oddly shaped or coastal cells and ensures total exposure is proportional to the land area actually in the district.

**When and why time matters**

ERA5 timestamps are in UTC; Sri Lankan reporting is local. We first convert each hourly record to **Asia/Colombo** (UTC+5:30) and only then aggregate to daily values. That step prevents “day splitting” where a storm late at night in Colombo is misfiled into the next UTC day. From daily values, we aggregate into epidemiological weeks defined by the WER week-ending date. The resulting weekly features reflect the conditions people actually experienced during the reporting window.

**How features are constructed**

At the cell level, hourly values are summarized to *daily* metrics: means for temperature, humidity, VPD, WBGT and radiation; mins and maxes for daily temperature; and sums for precipitation. After area-weighting to districts, we assemble *weekly* features that are epidemiologically interpretable. For temperature we keep the weekly mean, range, and upper quantiles; for humidity and VPD the weekly means; for solar radiation the weekly mean; for precipitation the weekly total, the number of “wet days” above a threshold, the maximum 3-day total nested within the week, and the longest wet spell. These capture both sustained background conditions and short bursts that often matter for transmission or exposure (e.g., heavy multi-day rain that floods fields).

Precipitation deserves special care. ERA5 provides **tp** (total precipitation, an accumulated depth) and **mtpr** (mean total precipitation rate). Confusion between the two is a common source of unit errors. We carry both tracks explicitly and compute weekly **sums** for each, cross-checking magnitudes to spot inconsistencies. All precipitation inputs are non-negative, with obvious negative artifacts clipped to zero.

**Lags, memory, and seasonality**

Health outcomes rarely respond to the same week’s climate. We therefore generate lagged features one to six weeks back for key variables and compute rolling means and sums over two- and four-week windows. To represent hydrologic “memory,” we add an exponentially weighted antecedent precipitation index (EWAP) that decays recent rainfall (α≈0.8) over multiple prior weeks. Seasonality is handled by computing, for each district and week-of-year, a long-run climatology; weekly **anomalies** (observed minus climatology) and **percent-of-normal** metrics indicate departures from expected conditions. Within-district z-scores put variables on a comparable scale for modeling without imposing a single global metric.

**Data quality and defensible defaults**

Two guardrails keep the pipeline robust without being brittle. First, we winsorize daily inputs at the tails (e.g., 1st–99th percentiles) to reduce the influence of spurious spikes while preserving real extremes. Second, rather than hard-masking weeks with incomplete daily coverage, we carry a coverage counter (n\_days\_week). Analysts can then filter to their tolerance (for example, require ≥5 of 7 days) without losing visibility into why a week was excluded. All spatial operations use valid geometries; intersections are computed in an equal-area projection to ensure weights are true areas, not degrees.

**Linking to health data**

WER PDFs are indexed and parsed to produce district-week counts. After district name harmonization and week alignment, we left-join the climate features to the health table on district and epidemiological week. Population (mid-year) is merged to compute rates and serve as offsets. The result is a clean panel ready for generalized linear models or machine-learning approaches that can include district fixed effects and smooth seasonal terms.

**What we considered and why we landed here**

Several forks in the road matter. We chose **area-weighting** over nearest-cell assignment to improve spatial representativeness; **local-time aggregation** before daily/weekly roll-ups to prevent temporal misclassification; explicit handling of **tp vs mtpr** to avoid unit errors; and **soft coverage rules** so end-users can set thresholds appropriate to their question. We also retain a station-based daily product for cross-checks, but rely on ERA5 for wall-to-wall coverage and long temporal span. The goal is not to declare a single “correct” choice, but to make choices explicit, reproducible, and easy to adapt.

**Using the outputs**

Two files are central. The daily file (srilanka\_district\_daily\_era5\_areawt.csv) holds district-day summaries for diagnostics and custom aggregations. The weekly file (srilanka\_district\_weekly\_era5\_areawt.csv) aligns to epidemiological weeks and already includes lags, rolling windows, anomalies, and coverage. In modeling, pair these with weekly cases and population. For count outcomes, a Poisson or negative binomial model with offset(log(pop)), district fixed effects, and smooth functions of week-of-year is a sensible baseline; GAMs or tree-based models can capture the non-linearities and interactions that are common in climate–health relationships.

**Limits and next steps**

Surveillance noise and reporting delays inevitably dampen signals; even with careful climate alignment, effects may appear weaker than expected during some periods. Future extensions could incorporate sub-district heterogeneity, exposure metrics tied to land use (e.g., paddy extent within districts already computed in the pipeline), and explicit treatment of reporting delays. The current workflow is designed so those additions are modular: spatial weights, temporal calendars, and feature definitions are all transparent and replaceable.