Linking Hourly Climate Data to Weekly Health Outcomes: A Sri Lanka Guide and Methods Note

Author: Jordan Clark, PhD (Duke Global Health Institute | Klimo Insights)

Who this is for**.** Health researchers who want to add climate context to their research, such as diseases surveillance (e.g., leptospirosis, dengue) without assuming prior expertise in meteorology or GIS.

What you will learn**.** How to turn hourly gridded climate data into epidemiology-ready weekly features at an aggregate level, why each step is necessary, where errors creep in, and how to avoid common pitfalls when modeling delayed and seasonal climate–health relationships.

Notes on references and appendices. In-text citations point to sources for deeper methodological and data background (e.g., WMO climate normals, ERA5 documentation, DLNMs, measurement-error and spatial-misalignment texts). Appendix A provides a condensed, end-to-end walkthrough; it is most useful after reading the background sections and can feel terse if climate data are new. Appendix B summarizes climate variables most commonly used in more detail.

# 1) Purpose and Scope

What this guide delivers. A practical, end-to-end pathway for converting hourly, gridded climate fields into district-week exposure features aligned to public-health surveillance. The method addresses three recurring challenges—spatial assignment, unit semantics, and time alignment—and then builds epidemiology-ready features, validates them with quick QA checks, and outlines modeling strategies suitable for delayed and seasonal climate–health relationships (Carroll et al., 2006; Gryparis et al., 2009; Fotheringham & Wong, 1991). The outcome is a documented pipeline that produces tidy, merge-ready tables for analysis and reporting, specifically for an example looking at disease prevalence in Sri Lanka alongside climate data.

# 2) Scale-mismatches and the three questions to answer

Public-health surveillance is reported in epidemiological weeks or months, while climate data arrive at mismatched scales: station observations, satellite overpasses, or hourly reanalyses on fixed grids. Three ingredients determine whether the resulting exposure makes epidemiologic sense: where the climate signal is mapped, what quantity is summarized, and when it is counted.

Where (space). Surveillance units are administrative areas, not grid cells. Nearest-cell assignment can bias exposures in coastal or elongated districts. A more defensible approach constructs district-level exposures using area-weighted overlays in an equal-area projection, so each cell contributes in proportion to the land area it overlaps—standard practice to reduce spatial misalignment and hedge against the MAUP (Fotheringham & Wong, 1991; Gryparis et al., 2009).

What (quantity). Variables behave differently across time. For temperature and humidity proxies, daily mean/min/max retain biologically meaningful variation. For precipitation, unit semantics matter: reanalyses often provide an accumulated depth (e.g., total precipitation) and a rate (e.g., mean total precipitation rate). Weekly totals should be built from daily sums of depth, with explicit unit checks, while also retaining event descriptors such as the largest three-day total or longest wet spell, which are frequently more predictive of water- and vector-borne risks (Simmons et al., 2018; Tran et al., 2019).

When (time). After daily district values are in hand, align them to the jurisdiction’s epidemiological week and, as a safeguard, convert timestamps to local civil time before any day/week roll-ups to avoid “day-splitting” of night-time observations. This preserves the temporal integrity of extremes without letting time-zone mechanics drive the analysis.

Even with careful alignment, residual exposure error—especially for tropical precipitation—can attenuate associations. Small effects warrant caution and should be stress-tested across alternative exposure definitions and aggregation choices (Carroll et al., 2006; Gryparis et al., 2009). The Sri Lanka example that follows simply instantiates these steps (district overlays, ERA5 inputs, Asia/Colombo time) but the logic extends to other diseases, geographies, and data sources.

Note: From this point forward, each methodological section includes a concrete Sri Lanka anchor so readers can map concepts to practice. Code and data pointers are available at: [GitHub Link](https://github.com/DGHI-CHI/southeast-asia/tree/main/sri-lanka-disease-surveillance).

# 3. Data Inventory and Conventions (Climate-data primer)

## What “climate data” usually means in practice

1. Surface stations measure weather at specific points (thermometers, rain gauges). Advantages: direct observation. Limitations: sparse coverage, varying quality control, siting issues (urban heat islands, obstructions).
2. Satellites infer surface and atmospheric properties from radiances. Advantages: broad coverage, consistent methods. Limitations: indirect retrievals, algorithm changes, orbital sampling (overpass times).
3. Reanalyses (e.g., ERA5) blend physical models with all available observations to produce a complete, hourly, global picture on a regular grid (Hersbach et al., 2020). Advantages: spatial/temporal completeness, many variables. Limitations: model and data-assimilation biases; some variables (e.g., precipitation) are less constrained by observations than temperature or pressure.

## Grids, levels, and variables

Grid: Most reanalyses store data on latitude–longitude grids (e.g., ~0.25° for ERA5, ~28 km near the equator). Grid cells are not equal in area—cells shrink toward the poles—hence the recommendation to compute statistics in an equal-area projection when aggregating to districts.

Levels: Some variables are at the surface (2-m air temperature, 10-m wind); others are at pressure levels aloft (e.g., 850 hPa temperature). For health applications focused on human exposure, surface or near-surface variables are typical.

## **ERA5 variables used in this guide**

Much of climate–health variability can be captured with a compact set of meteorological indicators. We use ERA5 reanalysis variables spanning temperature, humidity, precipitation, shortwave radiation, and 10-m winds, delivered as hourly gridded fields. These are aggregated to daily and weekly periods to match surveillance windows and support feature engineering. Each variable carries a standard physical interpretation and established mechanistic pathways to health outcomes (e.g., thermal stress, vector life cycles, environmental persistence).

(See Appendix A for a table and more thorough breakdown of each variable).

2-m Air Temperature (TA): Hourly near-surface air temperature, representing the conditions people and vectors actually experience just above ground. Daily means, minima, and maxima are aggregated to weekly summaries. Why: Governs human thermal stress, affects agricultural growth, and sets biological rates for pathogens and vectors.

2-m Dewpoint Temperature (TD): The temperature at which air becomes saturated, a direct measure of atmospheric moisture. Combined with T2M, it yields relative humidity and vapor pressure deficit. Why: Moisture strongly shapes mosquito survival, biting activity, and the persistence of pathogens in the environment.

Total Precipitation (TP): Hourly accumulation of liquid-equivalent precipitation, summed to daily and then weekly totals. Why: Drives runoff, flooding, and standing water that form exposure pathways for water-borne diseases (e.g., leptospirosis) and create or flush breeding habitat for mosquitoes.

Mean Precipitation Rate (MTPR): Average intensity of precipitation within an hour, integrated to daily and weekly totals. Why: Helps distinguish intense short-duration events from light but prolonged rain, which can have different effects on flooding, vector ecology, and human exposure.

Surface Shortwave Radiation Downward (SSRD): Solar energy reaching the surface, expressed as daily or weekly means of hourly values. Why: Controls surface heating, drying of soils and containers, and evaporation rates — all of which modify habitat suitability and interact with temperature and moisture stress on humans.

10-m Winds (U10, V10)**:** Horizontal wind components at 10 m height, combined into wind speed and direction. Daily means (and maxima) can be summarized weekly. Why: Winds affect dispersion of vectors, transport of spores and pathogens, and indicate synoptic weather systems that organize rainfall or drive drying conditions.

See Appendix B for more information about wind speed and direction.

Relative humidity (RH), vapor pressure deficit (VPD), wet bulb globe temperature (WBGT), wet-day counts, longest wet spell, max 3-day precipitation, anomalies, percent-of-normal, lags, rolling windows, EWAP, and coverage are not ERA5 variables. These appear later as derived features (Section 6) and within the workflow (Section 5) to avoid duplication.

## Sri Lanka anchoring details:

Climate. ERA5 reanalysis at hourly resolution on a ~0.25° grid. We use air temperature (ta), dew point (td), wind, downward shortwave radiation (ssrd), and precipitation. ERA5 provides two precipitation series and they are not the same:

* tp (“total precipitation”): an accumulated depth; once converted from meters, daily sums express mm/day and weekly sums express mm/week.
* mtpr (“mean total precipitation rate”): an average rate; summing the rate over a day yields a comparable depth via the rate pathway.
* We compute weekly sums for both and cross-check magnitudes, clipping small negative artifacts to zero. Because “climatological normal” concepts are central to interpreting departures, we follow WMO guidance (WMO, 2017).

Health. Weekly case counts by district from Sri Lanka’s [*Weekly Epidemiological Reports (WER)*](https://www.epid.gov.lk/weekly-epidemiological-report). Each record carries a district name and a week-ending date. We harmonize district names (e.g., standardizing “Nuwara-Eliya”) so merges are lossless.

Spatial boundaries. Administrative districts (ADM2) from [Global Administrative Areas](https://gadm.org/) (GADM). We validate geometry, standardize names, and compute overlaps in an equal-area projection, so weights are in true square meters (a best practice for areal interpolation; Mennis, 2003).

Population (recommended). Mid-year population by district to compute rates (per 100k) and to furnish an offset in count models.

Time zone. ERA5 timestamps are in UTC; reporting is local. We shift to Asia/Colombo before aggregations so days and weeks reflect local experience.

# 4. Spatial setup for district-level aggregation (projections & overlays)

## Coordinate reference systems (CRS)

Latitude–longitude describes angles, not areas. For area-weighted statistics (e.g., district rainfall totals), use an equal-area projection so one square meter near the coast counts the same as one inland. We project both the grid-cell polygons and district boundaries to an equal-area CRS (e.g., EPSG:6933) before computing overlaps and weights. Results are reported back in geographic coordinates for mapping.

Why this matters. In degrees, coastal grid cells can straddle land and ocean. If we average values without accounting for land fraction, coastal districts inherit ocean bias (e.g., cool, dry), distorting comparisons.

## From cells to districts (polygons and overlaps)

Steps:

1. Build polygons for all cells intersecting the study region.
2. Intersect cell polygons with district polygons in an equal-area CRS.
3. Use overlap area as weights to compute district-level means/sums (Mennis, 2003).  
   This guards against coastal cells with large degree-extent but little land dominating estimates (Fotheringham & Wong, 1991).

Plain-language example. If a grid cell overlaps Colombo by 40% of its land area and Gampaha by 60%, then 40% of that cell’s daily rainfall contributes to Colombo’s total and 60% to Gampaha’s.

## Resampling vs overlay rationale

Resampling (nearest/bilinear) selects or blends values at centers to create a smooth surface. It can be acceptable for intensive variables (e.g., temperature) when only a representative average is needed. For extensive variables (e.g., precipitation), resampling can distort totals because it ignores how much of each cell lies inside the district—coastal districts, for example, may be over-influenced by ocean pixels.

Area-weighted overlay respects geometry and conservation: each cell contributes in proportion to its land area inside the district. For precipitation, this preserves sums; for state variables it yields a defensible mean. Overlay is preferred for district-level summaries, especially for rainfall.

## Practical Implementation notes

CRS: Project both districts and grid polygons to EPSG:6933 (or another equal-area CRS covering Sri Lanka).

Coastal cells: Keep only the land fraction inside each district when forming weights.

Masking: If available, apply a land mask before polygonization to skip pure ocean cells.

Performance: Precompute and cache the cell×district weight table once; reuse across years/variables.

QA checks:

* District-area-weighted precipitation summed over all districts ≈ country total (within a few %).
* No district receives contributions from cells entirely over ocean.
* Visual spot-check a coastal district (map overlaps) to confirm land fractions look reasonable.

# 5) Workflow: hourly grid → district-day → district-week (step-by-step)

Step 1 — Convert to local time, then aggregate to daily.

Shift ERA5 timestamps to Asia/Colombo, then compute cell-level daily statistics. Aggregating in UTC can split nocturnal events across days and push storms into the wrong week.

Step 2 — Build area weights once.

Construct cell polygons; intersect with district polygons in an equal-area CRS; compute overlap area (m²). Store cell→district weights for reuse across variables and dates (Mennis, 2003).

Step 3 — Make daily district values.

For each district–day, compute the area-weighted average of cell-day means (e.g., temperature) and the area-weighted sum for precipitation totals. Clip small negative precipitation to zero.

Step 4 — Create weekly features aligned to surveillance.

Using the WER week-ending rule, gather the days in that local-time week and compute weekly features (totals; wet-day counts; longest wet spell; max 3-day precipitation; weekly mean temperature; mean of daily maxima; 95th percentile of daily maxima). Record n\_days\_week. Apply minimum-coverage rules during analysis rather than in data creation.

Step 5 — Add short-term memory.

Create 1–6 week lags for precipitation, temperature, humidity, and VPD. Build 2- and 4-week rolling means/sums with tolerant windows (e.g., allow 1 of 2 weeks) to avoid NA propagation. Compute EWAP for rainfall memory with defensible α and K settings (Bhaskaran et al., 2013; Gasparrini, 2010).

Step 6 — Derive climatologies and anomalies.

For each district and week-of-year, compute a baseline using the available historical period (or WMO “normal,” e.g., 1991–2020). Subtract to get anomalies and divide to get percent-of-normal (WMO, 2017).

Step 7 — Join to health and population.

Harmonize district names across sources; merge weekly climate features to WER case counts by district/week; add population to compute rates and to provide a Poisson/NB offset (Bhaskaran et al., 2013; Peng et al., 2006).

Sri Lanka note. Cell-day statistics are computed in Asia/Colombo and area-weighted to district-day series using stored weights. Precipitation totals are clipped at zero after conversion. District-day features are aggregated to district-week features aligned to WER weeks. Merge keys: district, date\_start, date\_end, year, week\_of\_year, plus n\_days\_week.

# 6) Derived features: humidity metrics, precipitation events, anomalies, and short-term memory

Derived features transform ERA5’s native variables into epidemiology-ready signals (ordered to avoid leakage). *Leakage* occurs when features for week *t* use information from after week *t* or from the outcome itself. Prevent it by (1) left-aligning windows so features use only weeks *t, t−1, …, t−K*; (2) computing intra-week metrics strictly within week boundaries; and (3) fitting any transforms (scaling, imputation, selection) inside the training fold only.

## Humidity & heat-stress metrics (RH, VPD, optional WBGT)

Relative Humidity (RH) (%) from t2m/d2m; Vaper pressure Deficit (VPD) (kPa) from saturation minus actual vapor pressure. Daily values averaged from hours; weekly values are means of daily values. Avoid co-modeling RH and VPD without strong rationale due to algebraic linkage (Dormann et al., 2013; Vatcheva et al., 2016; Novick et al., 2016). WBGT may be estimated when inputs permit (Liljegren et al., 2008).

Sri Lanka. Weekly rh\_mean\_week and vpd\_mean\_week are produced from localized t2m/d2m. WBGT is optional.

## Precipitation event descriptors (beyond totals)

### What these are and how they’re built. From district-day totals:

* Wet-day count: number of days ≥ threshold (default 10 mm; also compute 5 and 20 mm for sensitivity).  
  Names: wet\_days\_ge10\_tp, wet\_days\_ge10\_mtpr.
* Longest wet spell: longest run of consecutive wet days within the week.  
  Name: wet\_spell\_maxlen\_tp.
* Maximum 3-day total: largest sliding 3-day sum within the week.  
  Names: max3d\_tp, max3d\_mtpr.

Why these matter. The pattern of rain within a week (frequency, clustering, run length) can change risk even when weekly totals are identical—affecting flood/overflow, standing water persistence, and larval habitat dynamics (Zhang et al., 2011; Karl et al., 1999).

Sri Lanka. Computed in local time; thresholds at 5/10/20 mm retained for sensitivity, with 10 mm used for default summaries.

## Climatology and anomalies (levels vs departures)

Baseline (“normals”). For each district and week-of-year, compute a baseline using 1991–2020 (or your chosen period). This yields x\_normal(district, week\_of\_year).

### Derived metrics. Anomalies and percent-of-normal

* Anomaly: x\_anom = x\_obs – x\_normal.
* Percent-of-normal: x\_pct\_normal = x\_obs / max(x\_normal, ε), with a small ε (e.g., 1 mm for precipitation) to avoid divide-by-zero in normally dry weeks.
* Targets: precipitation totals, temperature summaries, and moisture demand (e.g., VPD). RH anomalies are often less informative.

Why use departures. Levels indicate absolute exposure; departures indicate unusualness for the season, which often aligns with biological and behavioral responses (e.g., early wet spell during a typically dry season).

### Implementation guardrails.

* Compute normals by district and week-of-year (ISO week).
* Require a minimum history per week (e.g., ≥15 years with data) to form a normal; otherwise mark as missing.
* Apply a clamp flag when x\_normal < ε and document ε in methods.

Sri Lanka. Normals use 1991–2020; precipitation percent-normal uses ε = 1 mm with a clamp flag recorded.

## Short-term memory: lags, rolling windows, EWAP (concise)

Why memory features exist: Outcomes often reflect recent history—vector life cycles, environmental persistence, and reporting delays mean conditions from prior weeks still matter (Bhaskaran et al., 2013; Gasparrini, 2010).

Leakage occurs when features for week *t* use information from after week *t* or from the outcome itself. Prevent it by (1) left-aligning windows so features use only weeks *t, t−1, …, t−K*; (2) computing intra-week metrics strictly within week boundaries; and (3) fitting any transforms (scaling, imputation, selection) inside the training fold only.

### Lags (discrete delays).

* What: separate predictors for previous weeks, \*\_lag1 … \*\_lag6.
* How: for week *t*, lagk = value at t−k; use past weeks only.
* Use: identify plausible delays (e.g., 0–2 weeks for leptospirosis–rain; 2–6 weeks for dengue–heat/moisture); supports DLNM to map exposure–lag responses.

### Rolling windows (short aggregates).

* What: summarize the last N weeks including the current week (e.g., \*\_roll2w\_mean, \*\_roll4w\_sum).
* How: tolerant windows to limit NAs (≥50% for 2-week, ≥75% for 4-week).
* Use: capture background conditions (sustained wetness/heat) instead of a single spike.

### EWAP (Exponentially Weighted Antecedent Precipitation).

* Idea: a single index that weights recent rain more than older rain—like a sponge that dries over time.

A black background with a black square

AI-generated content may be incorrect.

* Tuning: α controls decay (0.7–0.9 typical: higher = longer memory); K sets horizon (3–6 weeks).
* Use vs others: EWAP compresses multi-week history into one predictor (handy with small samples); unlike rolling windows, it down-weights older weeks.

No-leakage rule (applies to all memory features). Build features for week *t* using only weeks ≤ *t*; never include data from *t+1* onward.

Sri Lanka settings. Create lags 1–6 for precip\_tp\_sum\_week, precip\_mtpr\_sum\_week, tmean\_mean, and vpd\_mean\_week; rolling features as above; and EWAP on tp and mtpr with α ∈ {0.7, 0.8, 0.9} and K ∈ {3, 4, 5, 6} (e.g., ewap\_tp\_a0.8\_K4). Use 0–2 weeks (leptospirosis–precipitation) and 2–6 weeks (dengue–heat/moisture) as starting lag windows, then refine with blocked cross-validation.

## 6.5 Guardrails for multicollinearity and feature proliferation

Avoid pairing RH and VPD; treat tp vs integrated mtpr primarily as cross-checks; with large, correlated feature sets consider penalization (LASSO/elastic net) or stability selection (Tibshirani, 1996; Zou & Hastie, 2005; Meinshausen & Bühlmann, 2010). Center/scale predictors; report back-transformed effects.

## 6.6 Order-of-operations (to avoid leakage)

1. Build weekly base features (Section 5) and record n\_days\_week.
2. Compute weekly RH/VPD and event descriptors from daily inputs.
3. Construct climatologies; then anomalies/percent-normal from weekly bases.
4. Create lags/rolls/EWAP from levels (and from anomalies, only if required).
5. Preserve exact naming per the dictionary (Section 10).

# 7) Quality assurance & diagnostics

Range/sign checks. Negative precipitation after clipping should be zero; dew point ≤ air temperature; near-saturated RH should align with very low VPD.

One-week hand check. For a chosen district/week, recompute the weekly precipitation sum from daily values to confirm builder logic.

Cross-source check. Compare ERA5 weekly precipitation totals against station-derived weekly totals in a few districts. Perfect agreement is not expected; order-of-magnitude mismatches usually indicate unit or integration errors (Simmons et al., 2018).  
Reproducibility pins. Record CRS codes, time-zone handling, random seeds, and package versions.

Sri Lanka. QA records include EPSG:6933 overlays, Asia/Colombo day formation, the GADM release, and any station sources used for spot checks. A short QA vignette for a recent month demonstrates all three checks.

# 8) Modeling weekly outcomes (templates and evaluation)

Count models. While modeling choices should be unique to the data and the outcome, robust starting points include negative binomial or quasi-Poisson with a population offset:

A black background with a black square

AI-generated content may be incorrect.

Handle seasonality with smoothers and evaluate with blocked/forward-chaining cross-validation to respect time order (Peng et al., 2006; Bergmeir & Benítez, 2012; Bhaskaran et al., 2013).

Delayed/nonlinear effects. Use Distributed Lag Non-Linear Models (DLNM) to estimate exposure–lag–response surfaces (Gasparrini, 2010). For transient risks with strong within-district control of time-invariant confounding, consider case-crossover or the case time-series design (Maclure, 1991; Janes et al., 2005; Gasparrini, 2021, 2022).

Measurement error & the modifiable areal unit problem (MAUP). Assigning gridded exposures to districts induces spatial misalignment; classical error tends to attenuate coefficients; Berkson error inflates variance. Area-weighted overlays reduce, but do not eliminate, this issue; results may depend on aggregation units (Carroll et al., 2006; Gryparis et al., 2009; Fotheringham & Wong, 1991).

Interpretable ML (optional). For early-warning tasks, machine-learning models can help; prefer ALE for global structure and apply SHAP cautiously with correlated predictors (Apley & Zhu, 2020; Lundberg & Lee, 2017).

Sri Lanka. The template above is used with Sri Lanka’s feature set. Sensitivity analyses vary wet-day thresholds (5/10/20 mm), EWAP parameters (α, K), and lag windows (e.g., 0–2 or 2–6 weeks) and report robustness.

# 9) Worked example — Colombo, Sri Lanka (step-by-step)

## Epi week ending Friday, 2020-05-08 (Colombo district).

1. Localize time: convert ERA5 hours UTC→Asia/Colombo.
2. Define the week window: Saturday–Friday in local time.
3. Cell-day statistics: daily means for t2m, RH, VPD, ssrd; daily min/max for temperature; sum tp (m→mm) and integrate mtpr to daily depth; clip tiny negatives.
4. Area-weight to district-day: apply EPSG:6933 weights; means for state variables, sums for precipitation.
5. Weekly features: total precipitation (precip\_tp\_sum\_week, precip\_mtpr\_sum\_week); wet days (≥10 mm); longest wet spell; max 3-day total; weekly mean temperature; mean of daily maxima; 95th percentile of daily maxima; record n\_days\_week.
6. Short-term memory: add lag-1 and lag-2 precipitation and temperature.
7. Contextualize: compute anomalies and percent-normal for week-19 using 1991–2020 climatology.
8. Join to health: merge the climate feature row to the WER case count for Colombo for that week (WMO, 2017; Zhang et al., 2011).

# Appendix A: Implementation Example - Sri Lanka Case Study

## 1) The Sri Lanka Datasets in Detail

Health Data Requirements

This section demonstrates the general principles through a concrete implementation linking climate data to Sri Lanka's Weekly Epidemiological Reports (WER).

Health Data: Weekly Epidemiological Reports

Weekly case counts by district from Sri Lanka's *Weekly Epidemiological Reports (WER)*. Each record carries a district name and a week-ending date. District names are harmonized (e.g., standardizing "Nuwara-Eliya") to ensure lossless merges across data sources.

Climate Data: ERA5 Reanalysis

ERA5 reanalysis at hourly resolution on a ~0.25° grid. Variables include air temperature (ta), dew point (td), wind, downward shortwave radiation (ssrd), and precipitation. ERA5 provides **two precipitation series** that are not identical:

* **tp** ("total precipitation"): an accumulated depth; once converted from meters to mm, daily sums express **mm/day** and weekly sums express **mm/week**.
* **mtpr** ("mean total precipitation rate"): an average rate; summing the rate over a day yields a comparable depth via the rate pathway.

Weekly sums are computed for **both** series and cross-checked for magnitude consistency, with small negative artifacts clipped to zero. Climatological normal concepts follow WMO guidance (WMO, 2017).

Spatial Boundaries: GADM Districts

Administrative districts (ADM2) from GADM. Geometry is validated, names are standardized, and overlaps are computed in an **equal-area projection** so weights are in true square meters (following best practices for areal interpolation; Mennis, 2003).

Population Data

Mid-year population by district to compute rates (per 100,000) and to provide an **offset** in count models.

Time Zone: Asia/Colombo

ERA5 timestamps are in UTC; Sri Lankan reporting is local. All timestamps are shifted to **Asia/Colombo** time zone before aggregations so days and weeks reflect local experience.

## 2) Sri Lanka Implementation Workflow

Step 1: Convert to Local Time, Then Aggregate to Daily

Shift ERA5 timestamps to Asia/Colombo time zone, then compute cell-level daily statistics. Aggregating in UTC can shift storms across local day boundaries and into the wrong epidemiological week; conversion first avoids this temporal misalignment.

Step 2: Build Area Weights Once

Construct polygons for grid cells from center coordinates; intersect with district polygons in an equal-area coordinate reference system (e.g., EPSG:6933) and compute overlap area (m²). Store those cell→district weights for reuse across all days and variables (Mennis, 2003).

Step 3: Create Daily District Values

For each district–day, compute the area-weighted average of cell-level daily means (e.g., temperature) and the area-weighted sum for precipitation totals. Set small negative precipitation artifacts to zero.

Step 4: Create Weekly Features Aligned to WER

Using the WER's week-ending date, gather the days in that local-time week and compute weekly features. Record the number of valid days per week. Rather than discarding weeks with incomplete coverage, filter based on coverage thresholds during analysis.

Step 5: Add Memory, Lags, and Rolling Windows

Create 1–6 week lags for precipitation, temperature, and humidity variables. Build 2- and 4-week rolling means/sums with tolerant windows (e.g., allow 1 of 2 weeks) to avoid excessive missing value propagation. Compute EWAP for rainfall memory.

Step 6: Derive Climatologies and Anomalies

For each district and week-of-year, compute a baseline using the available historical period (or a WMO "normal," e.g., 1991–2020). Subtract to get anomalies and divide to get percent-of-normal values (WMO, 2017).

Step 7: Join to Health and Population Data

Harmonize district names across data sources, merge weekly climate features to WER case counts by district and week, and add population data to compute rates and provide Poisson/negative binomial offsets (Bhaskaran et al., 2013; Peng et al., 2006).

## 3) Sri Lanka Variable Dictionary

* Daily District Inputs (After Area Weighting)
* ta\_mean, ta\_min, ta\_max: Temperature variables (°C)
* td\_mean: Dew point temperature (°C)
* rh\_mean: Relative humidity (%)
* vpd\_mean: Vapor pressure deficit (kPa)
* wbgt\_mean: Wet bulb globe temperature (°C, when available)
* ssrd\_MJ\_mean: Solar radiation (MJ m⁻² day⁻¹)
* tp\_sum, mtpr\_sum: Precipitation totals (mm day⁻¹)
* date: Local date
* district: District identifier
* Weekly Features (Illustrative Selection)
* Temperature Features:
* tmean\_mean: Weekly mean temperature
* tmax\_mean: Mean of daily maximum temperatures
* tmax\_p90, tmax\_p95: 90th and 95th percentiles of daily maximum temperatures
* tmax\_range: Temperature range (maximum minus minimum)
* Humidity/Dryness Features:
* rh\_mean\_week: Weekly mean relative humidity
* vpd\_mean\_week: Weekly mean vapor pressure deficit
* Radiation Features:
* ssrd\_MJ\_mean\_week: Weekly mean solar radiation
* Precipitation Features:
* precip\_tp\_sum\_week, precip\_mtpr\_sum\_week: Weekly precipitation totals
* wet\_days\_ge10\_tp, wet\_days\_ge10\_mtpr: Number of wet days (≥10 mm)
* max3d\_tp, max3d\_mtpr: Largest 3-day precipitation total within the week
* wet\_spell\_maxlen\_tp: Longest consecutive wet-day sequence
* Coverage and Quality:
* n\_days\_week: Number of valid days in the week

Anomalies**:**

* \*\_anom: Anomaly (observed minus climatology)
* \*\_pct\_normal: Percent of normal (observed divided by climatology)
* Memory and Lags:
* \*\_lag1 through \*\_lag6: 1-6 week lags
* \*\_roll2w\_mean, \*\_roll4w\_sum: 2- and 4-week rolling statistics
* ewap\_tp, ewap\_mtpr: Exponentially weighted antecedent precipitation indices
* Merge Keys:
* district: District identifier
* date\_start, date\_end: Epidemiological week boundaries
* year, week\_of\_year: Temporal identifiers

## 4) Conceptual Walk-Through: Colombo Example

Consider Colombo district's epidemiological week ending Friday, 2020-05-08. The workflow proceeds as follows:

1. Convert hourly ERA5 timestamps from UTC to Asia/Colombo local time
2. Identify the six-day window (Saturday through Friday) corresponding to that epidemiological week
3. Compute cell-level daily statistics for each day in that window
4. Apply area weights to obtain district–day series for Colombo
5. Aggregate to weekly features:
   * Weekly precipitation sum: arithmetic sum of the six daily precipitation totals
   * Longest wet spell: longest consecutive run of days ≥10 mm within the week
   * Maximum 3-day total: largest sliding 3-day precipitation sum within the week
   * Weekly mean temperature: arithmetic mean of daily mean temperatures
   * Mean of daily maxima: arithmetic mean of the six daily maximum temperatures
   * 95th percentile of daily maxima: 95th percentile of the six daily maximum temperatures
6. Add lagged variables (e.g., lag-1 and lag-2 precipitation and temperature from previous weeks)
7. Compute anomalies relative to week-19's long-term climatology
8. Join to the WER case count for Colombo district in that specific week

This process is repeated for all districts and all weeks in the study period (WMO, 2017; Zhang et al., 2011).

## 5) Sri Lanka-Specific Quality Assurance

For the Sri Lanka implementation, quality assurance involves:

1. Fixed coordinate reference system: Use EPSG:6933 (World Cylindrical Equal Area) for area-weighted overlays
2. Seeded cross-validation splits: Ensure reproducible model evaluation
3. Coverage documentation: Maintain records of data availability by district and week
4. Validation against local data: When possible, compare ERA5 weekly precipitation totals against meteorological department station data from selected districts
5. QA vignette: Include a documented example showing quality checks on a recent month of data

The Sri Lanka implementation demonstrates all general principles while providing concrete variable names, time zones, and administrative boundaries that researchers can directly adapt or use as a template for similar surveillance systems.

# Appendix B: Implementation Example - Sri Lanka Case Study

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Units (native) | Daily aggregation | Weekly aggregation | Why it matters (epidemiology & climate) |
| 2-m Air Temperature (T2M) | Kelvin (K); convert to °C | Mean, min, max of hourly | Mean of daily means; max; high quantiles (90th/95th) | Governs human thermal stress; affects crop growth; sets rates for vector and pathogen development |
| 2-m Dewpoint Temperature (D2M) | Kelvin (K); convert to °C | Mean of hourly | Mean of daily means | Captures atmospheric moisture; combined with T2M to derive RH and VPD; influences mosquito survival, biting, and pathogen persistence |
| Total Precipitation (TP) | m per hour (accumulated); convert to mm | Sum hourly accumulations → mm/day | Sum daily totals → mm/week | Drives flooding/runoff, leptospirosis exposure, and mosquito habitat creation/flush events |
| Mean Total Precipitation Rate (MTPR) | m s⁻¹; convert to mm/hr | Integrate hourly rate → mm/day | Sum daily totals → mm/week | Distinguishes intense bursts vs. prolonged rain; complements TP for intensity/exposure relationships |
| Surface Shortwave Radiation Downward (SSRD) | J m⁻² per hour; convert to MJ m⁻² | Sum hourly → MJ m⁻²/day | Mean (or sum) of daily values | Controls surface heating, evaporation, and drying; shapes water habitat persistence and heat stress |
| 10-m Wind Components (U10, V10) | m s⁻¹ (u = east–west, v = north–south) | Mean (and max) wind speed/direction | Mean of daily means (and max) | Drives dispersion of vectors and airborne pathogens; reflects synoptic-scale circulation affecting rainfall and drying |

## 2-m Air Temperature (TA)

* **Definition:** The modeled “dry-bulb” air temperature 2 meters above the surface, consistent with the height of standard meteorological thermometers. The 2-m level is used because it reflects the near-surface environment experienced by humans and crops, while avoiding direct ground heating biases.
* **Units & sampling:** Kelvin (K), provided as instantaneous hourly values in ERA5. Convert to Celsius with °C = K − 273.15. While ERA5 outputs hourly fields, these are derived from 1-hourly model snapshots rather than direct observations.
* **Daily/weekly aggregation:**  
  • *Daily:* Compute mean, minimum, and maximum from 24 hourly values.  
  • *Weekly:* Average daily means and maxima; also calculate percentiles of daily maxima (e.g., 90th or 95th) to capture heat extremes.
* **Why it matters:**  
  • Human health: Key determinant of heat stress, thermoregulation, and heat illness risk.  
  • Agriculture: Influences plant growth, crop yield timing, and stress thresholds.  
  • Vectors/pathogens: Many biological processes (mosquito larval development, parasite replication rates) are temperature-sensitive, with nonlinear acceleration above specific thresholds.
* **Notes & caveats:**  
  • ERA5 “2-m” is a diagnostic from model levels, bias-corrected using observations, but may not fully capture microclimates (urban heat islands, shaded rural areas).  
  • Even modest biases (0.5–1 °C) can shift apparent thresholds for epidemiological outcomes.  
  • High temperatures can act synergistically with humidity to produce dangerous apparent heat, so T2M is best used in combination with moisture measures.

## 2-m Dewpoint Temperature (TD)

* **Definition:** The temperature at which air becomes saturated with moisture, i.e. the point where relative humidity would reach 100%. Dewpoint provides a direct measure of atmospheric water vapor content.
* **Units & sampling:** Kelvin (K), hourly instantaneous; convert to °C as above. Often paired with T2M to calculate relative humidity (RH) and vapor pressure deficit (VPD).
* **Daily/weekly aggregation:**  
  • *Daily:* Mean of hourly dewpoint values.  
  • *Weekly:* Mean of daily means, with optional percentiles to capture unusual moist/dry spells.
* **Why it matters:**  
  • Vector ecology: Mosquito longevity and biting activity increase under humid conditions; dry conditions shorten lifespan and suppress activity.  
  • Pathogen persistence: Moist air favors survival of certain airborne or surface pathogens.  
  • Human comfort: High dewpoint exacerbates heat stress by limiting sweat evaporation.
* **Notes & caveats:**  
  • Dewpoint is more stable than relative humidity (less diurnal swing), making it robust for weekly summaries.  
  • Negative dewpoints (possible at high elevations or during cool/dry conditions) are valid and signal very dry air.  
  • ERA5 dewpoint is model-derived; local biases can occur in areas with sparse humidity observations.

## Total Precipitation (TP)

* **Definition:** Accumulated depth of liquid water equivalent during the hour, including rainfall and melted frozen precipitation. This is a flux variable: ERA5 reports the total depth that fell in each hourly interval.
* **Units & sampling:** Meters (m) per hour, accumulated. Commonly converted to millimeters (mm) by multiplying by 1000. Occasionally small negative values appear due to reanalysis numerical noise; these should be clamped to zero.
* **Daily/weekly aggregation:**  
  • *Daily:* Sum all hourly accumulations within the local day → mm/day.  
  • *Weekly:* Sum daily totals → mm/week.
* **Why it matters:**  
  • Water-borne exposure: Flooding and runoff increase leptospirosis exposure and transport pathogens through contaminated water.  
  • Vector ecology: Standing water provides mosquito breeding sites; heavy rainfall can both create habitat and flush larvae.  
  • Infrastructure: Influences drainage, sanitation, and housing integrity, all of which mediate disease risk.
* **Notes & caveats:**  
  • Spatial scale: ERA5’s 0.25° grid (~28 km) smooths localized convective rainfall, potentially underestimating extremes.  
  • Cumulative precipitation should be compared to intensity (MTPR) to separate frequent light rains from rare intense bursts.  
  • Data lag: reanalysis may slightly underestimate very high local rainfall without sufficient station calibration.

## Mean Total Precipitation Rate (MTPR)

* **Definition:** The time-averaged precipitation intensity during the reported hour. It complements TP by describing the rate of precipitation rather than just the depth.
* **Units & sampling:** Reported as meters per second (m s⁻¹); multiply by 3600 to convert to meters per hour, then ×1000 for mm/hr.
* **Daily/weekly aggregation:**  
  • *Daily:* Sum hourly intensities (converted to mm) → mm/day.  
  • *Weekly:* Sum daily totals → mm/week.
* **Why it matters:**  
  • Differentiates intensity patterns: Short intense bursts (high MTPR) can cause flash flooding, while prolonged low rates may sustain breeding habitats.  
  • Data quality check: Provides an independent estimate of precipitation depth; comparing TP vs. MTPR can reveal unit inconsistencies or integration errors.  
  • Epidemiological relevance: Intense rain can increase exposure risks rapidly (e.g., flooding of contaminated water), while light persistent rain sustains larval habitats.
* **Notes & caveats:**  
  • Some confusion exists between TP and MTPR; always confirm units when extracting.  
  • Numerical artifacts may occur during transitions between model timesteps — always validate with observed station or satellite rainfall.

## Surface Shortwave Radiation Downward (SSRD)

* **Definition:** The total incoming solar shortwave radiation reaching the Earth’s surface, integrated over the hour. Represents available energy for heating, photosynthesis, and evaporation.
* **Units & sampling:** Joules per square meter (J m⁻²), accumulated per hour. Convert to megajoules (MJ m⁻²) by dividing by 10⁶.
* **Daily/weekly aggregation:**  
  • *Daily:* Sum hourly values to obtain MJ m⁻² per day.  
  • *Weekly:* Mean of daily totals (or weekly sum, depending on application).
* **Why it matters:**  
  • Land–atmosphere interactions: Drives surface heating, evapotranspiration, and soil drying.  
  • Vector ecology: Influences persistence of shallow water habitats; high radiation accelerates drying.  
  • Human health: Contributes to outdoor heat stress and radiant load on individuals.
* **Notes & caveats:**  
  • Cloud cover strongly modulates SSRD; hence, it can indirectly indicate rainfall/cloudiness conditions.  
  • ERA5 radiation estimates are model-based, not direct solar observations; regional biases exist (e.g., underestimating cloud cover).  
  • Units sometimes misunderstood: SSRD is an accumulated flux, not instantaneous irradiance.

## 10-m Wind Components (U10, V10)

* **Definition:** East–west (U10) and north–south (V10) components of horizontal wind at 10 m height, consistent with the World Meteorological Organization’s standard for wind measurements. Together, they define wind speed and direction.
* **Units & sampling:** Meters per second (m s⁻¹), instantaneous hourly. Wind speed is √(u10² + v10²). Wind direction = arctangent(v10/u10).
* **Daily/weekly aggregation:**  
  • *Daily:* Mean (and optionally maximum) wind speed and direction.  
  • *Weekly:* Mean of daily means; optionally mean of daily maxima.
* **Why it matters:**  
  • Vector dispersal: Mosquitoes and other vectors can be transported by wind, influencing spread.  
  • Pathogen transport: Airborne spores or pollutants are carried downwind, affecting exposure zones.  
  • Weather context: Winds signal synoptic-scale circulation (e.g., monsoon flow, tropical disturbances) that shape rainfall and temperature variability.
* **Notes & caveats:**  
  • ERA5 wind fields are generally reliable, but local terrain and urban roughness are not resolved at 0.25°.  
  • Extreme gusts are not captured; these are mean hourly values.  
  • For epidemiological modeling, weekly averages are more relevant than instantaneous values, but context (e.g., storm events) may require finer resolution.

# References

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82*(4), 1059–1086. <https://doi.org/10.1111/rssb.12377>

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Computational Statistics & Data Analysis, 70*, 324–335. <https://doi.org/10.1016/j.csda.2012.10.020>

Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L., & Armstrong, B. (2013). Time series regression studies in environmental epidemiology. *International Journal of Epidemiology, 42*(4), 1187–1195. <https://doi.org/10.1093/ije/dyt092>

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models* (2nd ed.). Chapman & Hall/CRC. <https://doi.org/10.1201/9781420010138>

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>

Fotheringham, A. S., & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A, 23*(7), 1025–1044. <https://doi.org/10.1068/a231025>

Gasparrini, A. (2010). Distributed lag non-linear models. *Statistics in Medicine, 29*(21), 2224–2234. <https://doi.org/10.1002/sim.3940>

Gasparrini, A. (2021). The case time series design. *Epidemiology, 32*(6), 829–837. <https://doi.org/10.1097/EDE.0000000000001390>

Gasparrini, A. (2022). A tutorial on the case time series design for small-area analysis. *International Journal of Health Geographics, 21*, 13. <https://doi.org/10.1186/s12942-022-00300-2>

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J., & Coull, B. A. (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics, 10*(2), 258–274. <https://doi.org/10.1093/biostatistics/kxn033>

Janes, H., Sheppard, L., & Lumley, T. (2005). Case-crossover analyses of air pollution exposure data: Referent selection strategies and their implications for bias. *Epidemiology, 16*(6), 717–726. <https://doi.org/10.1097/01.ede.0000181315.18836.9d>

Karl, T. R., Nicholls, N., & Ghazi, A. (1999). CLIVAR/GCOS/WMO workshop on indices and indicators for climate extremes: Workshop summary. *Climatic Change, 42*(1), 3–7. <https://doi.org/10.1023/A:1005491526870>

Liljegren, J. C., Carhart, R. A., Lawday, P., Tschopp, S., & Sharp, R. (2008). Modeling the wet bulb globe temperature using standard meteorological measurements. *Journal of Occupational and Environmental Hygiene, 5*(10), 645–655. <https://doi.org/10.1080/15459620802310770>

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*. <https://doi.org/10.48550/arXiv.1705.07874>

Maclure, M. (1991). The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology, 133*(2), 144–153. <https://doi.org/10.1093/oxfordjournals.aje.a115853>

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(4), 417–473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer, 55*(1), 31–42. <https://doi.org/10.1111/0033-0124.10042>

Novick, K. A., Ficklin, D. L., Stoy, P. C., Williams, C. A., Bohrer, G., Oishi, A. C., Papuga, S. A., Blanken, P. D., Noormets, A., Sulman, B. N., Scott, R. L., Wang, L., & Phillips, R. P. (2016). The increasing importance of atmospheric demand for ecosystem water and carbon fluxes. *Nature Climate Change, 6*(11), 1023–1027. <https://doi.org/10.1038/nclimate3114>

Peng, R. D., Dominici, F., & Louis, T. A. (2006). Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(2), 179–203. <https://doi.org/10.1111/j.1467-985X.2006.00410.x>

Simmons, A. J., Poli, P., Dee, D. P., Berrisford, P., Hersbach, H., Kobayashi, S., & Peubey, C. (2018). A reassessment of temperature variations and trends from global reanalyses. *Quarterly Journal of the Royal Meteorological Society, 144*(713), 29–49. <https://doi.org/10.1002/qj.3311>

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>

Tran, H., Nguyen-Le, D., Le, D. H., Truong, H. Q., Ngo-Duc, T., Ho, T. C., & Pham, T. T. (2019). A multi-model assessment of precipitation products over Vietnam. *Atmospheric Research, 227*, 190–202. <https://doi.org/10.1016/j.atmosres.2019.05.017>

Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in regression analyses in medical research. *North American Journal of Medical Sciences, 8*(7), 306–311. <https://doi.org/10.4103/1947-2714.187146>

Vicedo-Cabrera, A. M., Sera, F., Guo, Y., Tong, S., Coelho, M., Saldiva, P., Lavigne, E., Correa, P. M., Ortega, N. V., Kan, H., Osorio, S., Kyselý, J., Urban, A., Orru, H., Indermitte, E., Jaakkola, J. J. K., Ryti, N. R. I., Pascal, M., Goodman, P. G., … Gasparrini, A. (2019). Temperature-related mortality impacts under and beyond Paris Agreement climate change scenarios. *Climatic Change, 153*(3–4), 381–395. <https://doi.org/10.1007/s10584-019-02483-2>

World Meteorological Organization. (2017). *Guidelines on the calculation of climate normals* (WMO-No. 1203). <https://doi.org/10.25607/OBP-38>

Zhang, X., Alexander, L., Hegerl, G. C., Jones, P., Klein Tank, A., Peterson, T. C., Trewin, B., & Zwiers, F. W. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. *WIREs Climate Change, 2*(6), 851–870. <https://doi.org/10.1002/wcc.147>

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>