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

*Author: Jordan Clark, PhD (DGHI / 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 aggregare level, why each step is necessary, where errors creep in, and how to avoid common pitfalls when modeling delayed and seasonal climate–health relationships.

Ok based on that summary of what I had changed, please continue that theme of changes and provide a fully revised version of the entire document. I have attached the latest version to this message from which you can work. Avoid the use of words like “we” or or “I”

# 1) Why this is tricky—and how we bridge the clocks

Public-health surveillance is often 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 being summarized, and **when** it is assigned to the health period.

We begin with **space**. Surveillance units are administrative areas, not grid cells. Assigning the nearest grid cell 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—a standard remedy for spatial misalignment and a practical hedge against the modifiable areal unit problem (MAUP) (Fotheringham & Wong, 1991; Gryparis et al., 2009).

Next is **what** we aggregate. Different variables behave differently across time. For temperature and humidity proxies, daily **means/min/max** retain biologically meaningful variation; for precipitation, the unit semantics matter: reanalyses often provide both 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).

Only then do we resolve **when** the exposure is counted. After daily district values are in hand, we align them to the **jurisdiction’s epidemiological week definition** and, as a safeguard, convert timestamps to **local civil time** before any day/week roll-ups to avoid “day-splitting” of night-time events. 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, so small effects should be interpreted cautiously and stress-tested across alternative exposure definitions and aggregation choices (Carroll et al., 2006; Gryparis et al., 2009). The Sri Lanka WER example that follows simply instantiates these steps (district overlays, ERA5 inputs, and Asia/Colombo time), but the same logic extends to other diseases, geographies, and data sources.

# 2) The datasets, in plain language

* **Health.** Weekly case counts by district from Sri Lanka’s *Weekly Epidemiological Reports (WER)*. Each record carries a district name and a week-ending date. We harmonize district names (e.g., standardizing “Nuwara-Eliya”) so merges are lossless.
* **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).
* **Spatial boundaries.** Administrative districts (ADM2) from 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.

# 3) Key ideas in this example workflow

* **Grid cells versus districts.** Climate is gridded; surveillance is by district. We construct polygons for ERA5 grid cells from their **center coordinates**, intersect them with district polygons in an equal-area CRS (e.g., EPSG:6933), and compute the overlap area for each cell–district pair. Each cell contributes in proportion to that area. This prevents, for example, a large coastal cell (large in degrees but small in land area) from dominating an average (Mennis, 2003; Fotheringham & Wong, 1991).
* **Daily first, then weekly.** We compute **cell-level daily** statistics (means for temperature, humidity proxies, radiation; min/max for daily extremes; **sums** for precipitation), convert these to **area-weighted district–day** values, and only then assemble **district–week** features aligned to the epidemiological week.
* **Climatology and anomaly.** For each district and week-of-year we compute a long-run average (“climatology,” typically 1991–2020). Weekly **anomaly** is the departure from that expected value, and **percent-of-normal** is the ratio. Both are retained because levels and departures tell complementary stories (WMO, 2017; Zhang et al., 2011; Karl et al., 1999).

# 4) Units, definitions, and what the columns mean

We keep variable names explicit so they map directly into your analysis tables.

* **Temperature (°C).** Daily district values include ta\_mean, ta\_min, and ta\_max. Weekly features include the week’s mean (tmean\_mean), mean of daily maxima (tmax\_mean), range (max minus min), and upper quantiles (e.g., tmax\_p95). Extreme heat is mechanistically relevant for vector and host behavior (Vicedo-Cabrera et al., 2019).
* **Humidity and atmospheric dryness.** From ta and td we compute **relative humidity** (rh) and **vapor pressure deficit** (vpd, kPa; higher is drier). Weekly features typically use rh\_mean\_week and vpd\_mean\_week. Avoid placing RH and VPD in the same model unless you have a strong reason; they are algebraically linked and can inflate uncertainty (Dormann et al., 2013; Vatcheva et al., 2016). VPD is increasingly recognized as a biologically meaningful dryness metric (Novick et al., 2016).
* **Solar radiation.** ERA5 ssrd is converted to **MJ m⁻² day⁻¹**; the weekly feature is the mean of daily values (ssrd\_MJ\_mean\_week). Radiation modulates temperature and surface moisture.
* **Precipitation.** Daily totals include tp\_sum and mtpr\_sum (both as **mm/day** after conversion). Weekly features are sums over the week (precip\_tp\_sum\_week, precip\_mtpr\_sum\_week). We also derive: the **number of wet days** (e.g., days ≥ 10 mm; wet\_days\_ge10\_tp), the **largest three-day total** within the week (max3d\_tp), and the **longest wet spell** (consecutive wet-day count; wet\_spell\_maxlen\_tp). Ten millimeters is a defensible starting point because many water-borne/vector mechanisms are sensitive near that range; sensitivity at 5 and 20 mm is advisable (Zhang et al., 2011; Karl et al., 1999).
* **Coverage.** n\_days\_week counts days with usable data inside the week. Rather than hard-masking, we keep this column so analysts can set explicit filters (e.g., keep weeks with ≥5 valid days).
* **Anomalies.** For selected variables we compute \*\_anom (observed minus climatology) and \*\_pct\_normal (observed divided by climatology with safeguards near zero). These enable comparisons across districts and seasons (WMO, 2017).
* **Memory, lags, and rolling windows.** For key features we generate \*\_lag1 … \*\_lag6 (weeks) and **2- and 4-week rolling means/sums** that include the current week. Because pathogens and vectors respond to **recent history**, we also compute an **Exponentially Weighted Antecedent Precipitation (EWAP)** index: the current week plus exponentially decayed prior weeks, tuned by **α** (decay) and **K** (horizon). Choices should be biologically plausible and validated empirically (Bhaskaran et al., 2013; Gasparrini, 2010).

# 5) The workflow, 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 shift storms across local day boundaries and into the wrong epi week; conversion first avoids this.

**Step 2 — Build area weights once.** Construct polygons for grid cells from center coordinates; intersect with district polygons in an equal-area CRS and compute overlap area (m²). Store those cell→district weights for reuse across all days and variables (Mennis, 2003).

**Step 3 — Make daily district values.** For each district–day, take 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 surveillance.** Using the WER’s week-ending date, gather the days in that local-time week and compute the features above. Record n\_days\_week. If you want a minimum coverage rule, filter later rather than discarding informative weeks.

**Step 5 — Add memory, lags, and short windows.** Create 1–6 week lags for precipitation, temperature, humidity, 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.

**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 (WMO, 2017).

**Step 7 — Join to health and population.** Harmonize district names across sources, merge weekly climate features to WER case counts by district and week, and add population to compute rates and to provide a Poisson/NB offset (Bhaskaran et al., 2013; Peng et al., 2006).

# 6) Design choices and why they matter

Three choices do most of the quality work. First, **local-time aggregation before daily/weekly roll-ups** ensures exposure belongs to the week people actually experienced. Second, **area-weighted exposure** performs better than nearest-cell lookups, especially along coasts or in elongated districts where the gridded footprint and the administrative footprint differ (Mennis, 2003; Fotheringham & Wong, 1991). Third, we **retain both ERA5 precipitation measures (tp and mtpr)** and cross-validate them to catch unit or conversion problems early (Simmons et al., 2018; Tran et al., 2019).

Two further choices protect inference. To reduce **collinearity**, we avoid including algebraically redundant features (e.g., RH together with VPD) unless needed and consider penalization when feature sets are large (Dormann et al., 2013; Tibshirani, 1996; Zou & Hastie, 2005; Meinshausen & Bühlmann, 2010). To reduce **exposure attenuation** from spatial misalignment, we use area weights and interpret small coefficients with caution, particularly for precipitation (Carroll et al., 2006; Gryparis et al., 2009).

# 7) What can go wrong (and how to catch it quickly)

Before trusting any model output, run three short checks. First, plot district-day rainfall histograms and verify totals are non-negative and plausible; weekly sums should not all be zero. Second, pick a district and a specific week and **hand-calculate** the weekly precipitation sum from its daily values to confirm the weekly builder behaves as expected. Third, compare ERA5 weekly totals against any available station-based weekly totals in a few districts. Exact agreement is not expected, but order-of-magnitude mismatches flag unit or conversion bugs (Bhaskaran et al., 2013; Simmons et al., 2018).

# 8) Modeling weekly outcomes - generally

For counts, a robust starting point is a **negative binomial or quasi-Poisson** regression with a population offset:

cases\_{d,w} ~ offset(log(pop\_{d,w})) + precipitation (levels, anomalies, lags) + temperature/VPD (levels, anomalies, lags) + district effects + smooth seasonality f(week-of-year) + year effects.

Control long- and short-term seasonality with **flexible smoothers** (e.g., penalized splines of time or week-of-year) and check stability to reasonable smoothing choices (Peng et al., 2006; Bhaskaran et al., 2013). Respect time order in evaluation with **blocked or forward-chaining cross-validation**; random folds leak future information (Bergmeir & Benítez, 2012). When you expect **nonlinear delayed effects**, fit **Distributed Lag Non-Linear Models (DLNMs)** to estimate an exposure–lag–response surface (Gasparrini, 2010). For short-term risk 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).

If the goal is **early warning**, machine-learning models can help; prefer **interpretation tools** that behave well under correlated features—**Accumulated Local Effects (ALE)** for global structure and, with care, **SHAP** for local contributions (Apley & Zhu, 2020; Lundberg & Lee, 2017).

# 9) Exposure alignment, measurement error, and why small effects aren’t always “no effect”

Assigning gridded climate to districts introduces **spatial misalignment** and thus measurement error. Area-weighting reduces, but does not eliminate, this error. Classical measurement error tends to **attenuate** (bias toward zero) regression coefficients; Berkson error inflates variance (Carroll et al., 2006; Gryparis et al., 2009). Treat very small effect sizes with humility—particularly for precipitation in the tropics—and prioritize robust patterns over marginal p-values.

Beyond exposure error, the **modifiable areal unit problem (MAUP)** reminds us that results can depend on the aggregation unit. We work at the reporting unit (districts) and use equal-area weights to keep the exposure definition faithful to those units (Fotheringham & Wong, 1991).

# 10) What to tune for your disease and setting

Three knobs matter most. The **wet-day threshold** should be plausible for your hydrology and infrastructure; sensitivity at 5, 10, and 20 mm demonstrates robustness (Zhang et al., 2011; Karl et al., 1999). The **EWAP decay α** and **horizon K** represent process “memory”; ranges like α=0.7–0.9 and K=3–6 weeks cover many contexts (Bhaskaran et al., 2013). The **lag window** reflects plausible delays from exposure to symptom onset and reporting; use a prior (e.g., 0–2 weeks for leptospirosis precipitation effects; 2–6 weeks for dengue heat/moisture) and let blocked cross-validation fine-tune (Gasparrini, 2010; Bergmeir & Benítez, 2012).

Report the settings you chose and show that conclusions are stable to reasonable alternatives.

# 11) A small conceptual walk-through

Suppose Colombo’s epidemiological week ends on Friday, 2020-05-08. Convert hourly ERA5 timestamps to local time, compute six daily values for that Saturday–Friday window, and area-weight those to obtain district–day series. The weekly precipitation sum is the arithmetic sum of those daily totals; the longest wet spell is the longest run of days at or above 10 mm; the maximum three-day total is the largest sliding three-day sum inside that week. Compute the weekly mean temperature, the mean of daily maxima, and the 95th percentile of daily maxima. Add lag-1 and lag-2 precipitation and temperature, compute anomalies relative to week-19’s climatology, and join to the WER case count for that week (WMO, 2017; Zhang et al., 2011).

# 12) Variable dictionary (to keep columns unambiguous)

**Daily district inputs (after area weighting).** ta\_mean, ta\_min, ta\_max (°C); td\_mean (°C); rh\_mean (%); vpd\_mean (kPa); wbgt\_mean (°C, when available); ssrd\_MJ\_mean (MJ m⁻² day⁻¹); tp\_sum, mtpr\_sum (mm day⁻¹); date (local); district.

**Weekly features (illustrative, not exhaustive).**  
Temperature: tmean\_mean, tmax\_mean, tmax\_p90, tmax\_p95, tmax\_range.  
Humidity/dryness: rh\_mean\_week, vpd\_mean\_week.  
Radiation: ssrd\_MJ\_mean\_week.  
Precipitation: precip\_tp\_sum\_week, precip\_mtpr\_sum\_week, wet\_days\_ge10\_tp, wet\_days\_ge10\_mtpr, max3d\_tp, max3d\_mtpr, wet\_spell\_maxlen\_tp.  
Coverage: n\_days\_week.  
Anomalies: \*\_anom, \*\_pct\_normal.  
Memory: \*\_lag1 … \*\_lag6; \*\_roll2w\_mean, \*\_roll4w\_sum; ewap\_tp, ewap\_mtpr.  
Keys for merging: district, date\_start, date\_end (epi week), year, week\_of\_year.

# 13) Practical notes on reproducibility and QA

Use **fixed CRS codes** (e.g., EPSG:6933 for equal-area operations) and **seeded splits** for cross-validation. Keep a **coverage** column and a short **QA vignette** in your repo showing the checks above on a recent month. When possible, compare a subset of ERA5 weekly rainfall to station-derived weekly totals to calibrate expectations (Simmons et al., 2018).

# 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>