AlphaEarth Satellite Embeddings for Modelling Climate Sensitive Diseases Towards Global Health Resilience

Usman Nazir, Sara Khalid

Planetary Health Informatics (PHI) Lab, University of Oxford
{usman.nazir, sara.khalid}@ndorms.ox.ac.uk

Abstract

Introduction: Malaria, childhood acute respiratory infection, and child undernutrition[1] are leading causes of preventable mortality in children under five, concentrated in low and middle-income countries where climate variability directly modulates transmission, exposure, and nutritional outcomes [1–3].. Routine health surveillance in these settings remains sparse, and the utility of satellite-derived representations of the Earth's surface as predictors of population health outcomes is poorly characterised.
Methods: We evaluate AlphaEarth Foundations 64-dimensional satellite embeddings as predictors of population health outcomes using LSTM and Transformer ensembles across three pathologies: malaria, acute respiratory infection, and stunting.
Results: Embeddings provide meaningful predictive value when merged at sufficient spatial granularity: (i) in malaria case prediction, they raise test R^2 from 0.623 to 0.777 in Nigeria and from 0.867 to 0.881 in India; (ii) in childhood ARI prediction across 11 DHS countries, pooled R^2 rises from 0.098 to 0.164, with XGBoost reaching $R^2 = 0.210$; (iii) in child weight-for-height z-score prediction across 35 DHS countries, gains are within seed variance, consistent with country fixed effects already absorbing the between-country variance the static embedding can express.
Interpretation and conclusion: Together, these data show that AlphaEarth satellite embeddings consistently add a predictive value when applied at a location or cluster level across distinct population health outcomes. We close with a request: direct access to the Google Earth AI foundation-model suite, including Population Dynamics embeddings that incorporate health indicators, would substantially accelerate this line of work across all three outcomes.

1 Introduction

Globally, infectious diseases remain the leading cause of mortality in children under the age of five [1], with one in three preventable deaths being attributable to malaria or acute respiratory infection in children aged under five [3]. This is further exacerbated by undernutrition, another significant cause of preventable morbidity and mortality in this age group [2, 3]. The burden of these diseases disproportionately affects low and middle income countries (LMICs), predominantly in sub-Saharan Africa and South Asia [3].

Climate variability is a well established driver of all three of these pathologies. There is a wealth of biological evidence on the impact of climate on disease transmission and vector biology

[1]Following common usage in the child-nutrition community, we use "stunting" as the umbrella term for chronic child undernutrition. The specific outcome modelled in Case 3 is the Weight-for-Height Z-score (WHZ), which is the continuous variable from which the binary *wasting* indicator (WHZ < -2) is derived. Stunting (low height-for-age) and wasting (low weight-for-height) are related but distinct dimensions of undernutrition; our pipeline is agnostic to which of the two is selected as the target.

in malaria [4–9], and malaria incidence has been increasing in recent years [10], and is forecast to continue to do so as the climate changes [11]. Changes in air temperature and air quality are also thought to have a substantial and complex impact on infectious and non-infectious respiratory pathology [12–16]. Finally, climate variability threatening food security is a key factor influencing maternal, fetal and child nutritional status [17–21]. As climate change intensifies these exposures, the need for scalable, near-real-time environmental monitoring tools for health is increasingly urgent.

Satellite and climatic data are an invaluable resource for modelling and prediction, as unfortunately these diseases disproportionately affect resource poor settings with primary survey-based data being sparse and otherwise prone to error. These data have been increasing in popularity in recent years to guide machine learning [9, 11, 22–24]. Here, we go further to see whether embeddings from the geospatial foundation model AlphaEarth [25] further improve forecasting accuracy. We evaluate AlphaEarth Foundations V1 (64-dim, 10-m annual unit-norm embeddings via Google Earth Engine) on three distinct LMIC outcomes spanning three health domains:

- **Malaria** (vector-borne infectious disease): Nigeria and India, 2000–2024.
- **Childhood ARI** (respiratory infectious disease): 11 DHS countries, 2017–2022.
- **Child undernutrition (WHZ)** (chronic / stunting): 35 DHS countries, 2015+.

Each study uses an appropriate baseline (climate-only, pollution-only, or regression-based tabular respectively), and tests whether adding AlphaEarth embeddings improves test-set prediction. The three case studies are presented sequentially below.

2 Case 1 — Malaria Prediction in Nigeria and India

2.1 Covariates

Each training example is a 24-month input sequence $\mathbf{x}_{i,t} = (\mathbf{x}_{i,t-23}, \ldots, \mathbf{x}_{i,t})$ for $5 \times 5\,\text{km}$ MAP pixel i ending at month t of the prediction year. At every month, the feature vector concatenates three blocks:

- **Dynamic environmental (monthly).** Total rainfall from CHIRPS [26] ($\sim$5 km, aggregated from daily totals), NDVI from MODIS [27] (rescaled by 10^{-4}), and mean 2 m air temperature from ERA5-Land [28] ($\sim$11 km, converted from Kelvin to degrees Celsius), all sampled at the pixel centroid. These three channels carry the seasonal signal and drive the baseline configuration.
- **Static context.** Distance to the nearest waterway [29] (100 m), under-18 population count [30] ($\sim$100 m), the harmonised DMSP-VIIRS nighttime-lights composite [31] ($\sim$1 km), and the admin-1 geopolitical zone from GADM [32] (Nigeria: six DHS zones; India: five Zonal Council zones, with Maharashtra and Dadra & Nagar Haveli centroids assigned to South India). All raster sources are sampled at the MAP pixel centroid; the WorldPop products are released only to 2023 and the DMSP-VIIRS composite is annual, so the entire static block is treated as time-invariant and broadcast across the 24 time steps. For India, raster-to-point sampling with 500 m downsampling replaced polygonization to keep extraction tractable at scale.

- **Static AlphaEarth fingerprint (+*AE only*).** The 64-dimensional AlphaEarth Foundations embedding sampled at the pixel centroid for the prediction year on Google Earth Engine [33]. The fingerprint is unit-norm by construction ($\|\cdot\|_2 = 1$ per pixel-year) and is broadcast across all 24 time steps. The baseline configuration omits this block; +AE appends it to the environmental and context channels.

For India only, two additional autoregressive channels are appended to the feature vector: lag-1 and lag-2 annual case counts at the same pixel, log-transformed and per-location z-scored on the training window. Without these, the per-location target normalisation cannot be inverted at test time on low-incidence zones.

The target $y_{i,t}$ is the annual malaria case count from the Malaria Atlas Project [34]. It is transformed to $z_{i,t} = (\log(1+y_{i,t}) - \mu_z)/\sigma_z$, with μ_z, σ_z computed on the training window. For India the log-target is in addition z-scored per pixel, so the model predicts a within-location deviation from the location's own historical mean rather than an absolute count. Predictions are inverted to the count scale with a non-negativity clamp as in Eq. (2).

2.2 Training and testing setup

Models are trained on annual data for 2000–2023 and evaluated on the held-out 2024 year. The one-year forward gap is deliberate: it stresses the model's ability to extrapolate beyond its training window, rather than interpolate inside it, and prevents any test-time leakage from autoregressive lag features. Within the training window, 15% of pixel-years are held out as an internal validation slice for early stopping (patience 5 epochs, maximum 25 epochs).

The same train/validation/test partition is shared exactly between *baseline* (climate + context) and $+AE$ (climate + context + AlphaEarth) configurations, so ΔR^2 isolates the contribution of the embedding rather than split variance. The pipeline is then re-run with $S = 5$ random seeds controlling weight initialisation, dropout masks, and the order of mini-batches; reported metrics are the seed-mean of the five runs, and per-zone seed standard deviations are reported in the choropleth captions.

Numeric channels are mean-imputed and z-scored using statistics computed on the training window only; categorical channels (geopolitical zone) are pinned to the levels observed in training. Optimisation uses AdamW (learning rate 2×10^{-3}, weight decay 10^{-4}), batch size 256, and gradient clipping at $\|\mathbf{g}\|_2 \leq 5$. The ensemble averages the standardised predictions of a single-layer LSTM and a two-layer Transformer encoder, both with $d_{\text{model}} = 64$ followed by an MLP head ($64 \to 64 \to 1$), as detailed in Eq. (1).

2.3 Methodology

We train two encoders on the concatenated input sequence $\mathbf{x}_{i,t} = (x_{i,t-23}, \ldots, x_{i,t})$, where each time step concatenates monthly ERA5-Land climate covariates (temperature, rainfall, NDVI), the contextual covariates listed in Section 2.2, and, where applicable, the 64-dimensional AlphaEarth fingerprint broadcast across the 24 time steps.The target $y_{i,t}$ is the annual malaria case count, transformed by $z = (\log(1 + y) - \mu_z)/\sigma_z$ with μ_z and σ_z computed on the training set. For India, the log-target is additionally normalised per location, and lag-1 and lag-2 case counts are appended to the input vector.

The two encoders are a single-layer LSTM with hidden size 64, and a two-layer Transformer encoder with $d_{\text{model}} = 64$. Each encoder is followed by an MLP head ($64 \to 64 \to 1$) that produces a scalar prediction $\hat{z}_{i,t}$. The final standardized prediction averages the two models:

$$\hat{z}_{i,t} = \tfrac{1}{2}\big(\hat{z}_{i,t}^{\text{LSTM}} + \hat{z}_{i,t}^{\text{Trf}}\big). \tag{1}$$

Predictions are mapped back to the count scale by inverting the z-score and log transforms, with a non-negativity clamp:

$$\hat{y}_{i,t} = \max\left(0,\ e^{\hat{z}_{i,t}\sigma_z + \mu_z} - 1\right) \in \mathbb{Z}_{\geq 0}. \tag{2}$$

Models are trained for 25 epochs and the pipeline is run with five random seeds, with reported metrics taken as the seed mean. Training uses 2000–2023; 2024 is held out for evaluation. Folds are shared across the baseline and +AlphaEarth configurations, so ΔR^2 isolates the contribution of the embedding rather than seed or split variance.

2.4 Geographic distribution of zonal performance

The country-level R^2 and NRMSE numbers and the per-zone radial / radar charts already report what each zone looks like in isolation, but neither view makes the spatial distribution of performance immediately legible. We therefore project the per-zone metrics onto the country map: each zone polygon is built by dissolving Natural Earth admin-1 state polygons via a state → zone lookup, and is filled with the five-seed mean of the metric. India's zone polygons follow the same lat/lon banding scheme used at evaluation time, so the choropleth zone boundaries match the zones the model is actually scored on. Fig. 1 shows Nigeria; Fig. 2 shows India.

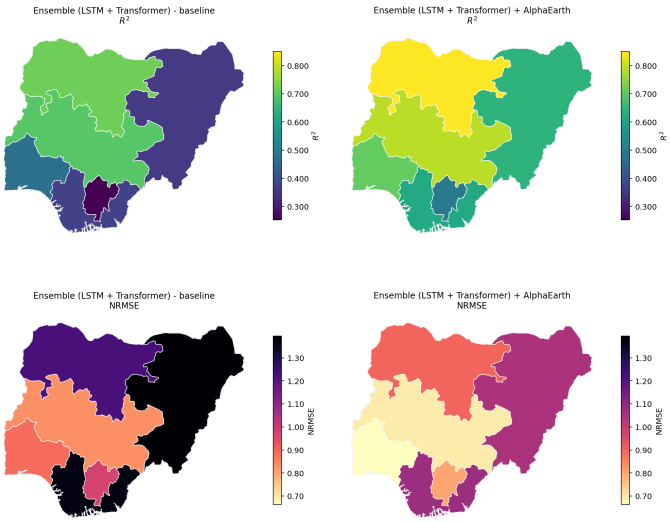


Figure 1: Nigeria ensemble on the held-out 2024 test year, dissolved onto the six DHS geopolitical zones. **Top row:** test R^2. **Bottom row:** test NRMSE = RMSE/$\bar{y}_{\text{test}}$ (lower is better). **Left:** baseline. **Right:** + AlphaEarth. Cell colour is the five-seed mean, on a shared scale per row. Every zone improves under + AlphaEarth on both metrics, with the largest R^2 gains in the North East and South East and the largest NRMSE reductions in the North East and South South.

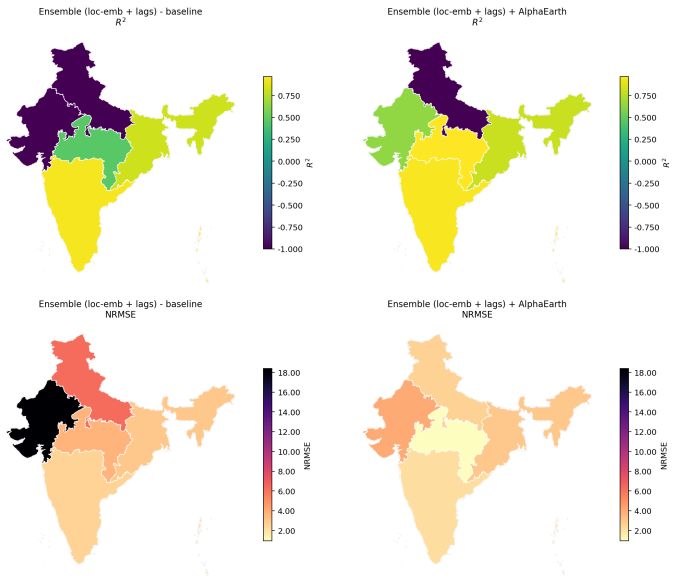


Figure 2: India ensemble (per-location embedding + lag-1/lag-2 + per-location target normalisation) on the held-out 2024 test year, dissolved onto the five Zonal Council zones (Maharashtra and Dadra & Nagar Haveli centroids resolve to South India). **Top row:** test R^2, clipped to $[-1, 1]$ for legibility. **Bottom row:** test NRMSE. **Left:** baseline. **Right:** + AlphaEarth. The gain is concentrated in the low-incidence zones where the panel-only baseline fails: North and West India swing from strongly negative R^2 to positive under + AlphaEarth, Central India follows the same direction at smaller magnitude, and South and East India — already at $R^2 \approx 0.8$ at baseline — are essentially unchanged.

2.5 Result.

On the 2024 test year, adding AlphaEarth embeddings raises test R^2 from 0.623 to 0.777 for Nigeria and 0.867 to 0.881 for India. All Nigerian and Indian regions show positive ΔR^2 in Figure 1 and Figure 2 respectively.

Both choropleths (see Figure 1 and Figure 2) display the five-seed *mean* of the metric. On Nigeria the seed standard deviations are small (within-zone $\leq \sim 0.07$ for R^2, $\leq \sim 0.19$ for NRMSE) and the choropleth mean is a faithful summary; on India's North and West zones the seed standard deviation is large, and the means in Fig. 2 are dominated by which clusters happen to overshoot under each random initialisation.

3 Case 2 —Childhood ARI, 11 DHS Countries

3.1 Covariates

Each DHS cluster contributes a single cross-sectional feature vector $\mathbf{x}_i$. Three feature sets are compared (Table 1): a tabular *baseline*; an embedding-only configuration (*emb_only*) that drops the tabular channels and retains only the AlphaEarth fingerprint; and the combined *emb_plus* configuration. The *emb_only* set is included as a sanity check: it shows what the embedding can predict on its own, without help from pollutants or controls.

Table 1: Feature sets used in the Case 2 ensemble experiments. The denominator of the rate target, `population_in_buffer`, is *excluded* from all three feature sets so the network never sees the target denominator at training time.

Feature set	Variables	# features
`baseline` (without AE)	`CO_mean`, `NO2_mean`, `SO2_mean`, `ALT_DEM`, mean 2 m temperature (100 m buffer), `URBAN_BIN` (derived from `URBAN_RURA == "U"`)	6
`emb_only`	AlphaEarth embedding bands `A00` … `A63`, sampled at the cluster centroid for the survey year over a 2–5 km buffer matching the DHS privacy displacement radius	64
`emb_plus` (with AE)	`baseline` ∪ `emb_only`	70

The target is the per-cluster ARI *rate*

$$y_i = \frac{\texttt{NO\_ARI\_CASES}_i}{\texttt{population\_in\_buffer}_i},$$

rescaled and log-transformed for training as $\tilde{y}_i = \log(1 + 1000\, y_i)$ to give the network a usable gradient range; predictions are inverted via $\hat{y}_i = \max(0, (\exp(\hat{\tilde{y}}_i) - 1)/1000)$ before computing RMSE and R^2 on the original rate scale. Using a rate rather than a raw count removes the trivial signal of cluster size, so the model must explain *prevalence* of ARI from the environmental and satellite channels.

Each scalar feature is tokenised by a learnable $\texttt{Linear}(1 \to 32)$ projection plus a learnable per-feature positional embedding before being passed to the sequence encoders; the length of the resulting token sequence equals the number of features in the active feature set (6, 64, or 70).

3.2 Training and testing setup

Evaluation uses a stratified 5-fold cross-validation (KFold(n_splits=5, shuffle=True, random_state=42)), so each fold trains on 4/5 of the rows and tests on the held-out 1/5. Folds are constructed once and re-used identically across the three feature sets, so ΔR^2 between *baseline*, *emb_only* and *emb_plus* reflects the contribution of the feature set rather than fold variance. Within each training fold, a further $\texttt{VAL\_FRAC} = 0.15$ slice is held out as an internal validation set for early stopping (patience 8 epochs, maximum 60 epochs). Numeric channels are mean-imputed and z-scored using statistics computed on the training fold only.

Two evaluation regimes are reported:

- **Per-country.** The 5-fold CV is run independently on each country's clusters, so countries cannot share clusters between train and test. For India (n = 7,042) this produces folds of roughly 5,634 train and 1,408 test clusters; for the five smallest countries ($n < 139$) the fold size is small enough that the 64-dimensional embedding approaches the per-fold training size, which the headline analysis treats as a known overfitting regime (see Section 3.3 and the small-sample penalty in Fig. 3).
- **Global (pooled).** The same 5-fold CV is applied to the full pooled cohort of 9,271 clusters from all 11 countries, without country stratification. Folds can mix countries, which we accept as a limitation: the i.i.d. split is *not* spatially blocked and does *not* hold out whole countries

or regions, so within-country and within-region spatial autocorrelation is not controlled for. A leave-one-country-out or spatial-block CV would give a more conservative estimate of generalisation, but the consistency of the headline gain across the three estimators in Appendix 7 indicates that the pooled signal is not an artefact of a single fold split.

The ensemble averages the predictions, in log-rate space, of a **TokenLSTM** (per-feature tokenisation $\rightarrow$ single-layer bidirectional LSTM with hidden size 32 $\rightarrow$ mean pool $\rightarrow$ MLP head $64 \rightarrow 32 \rightarrow 1$ with ReLU and dropout 0.1) and a **TokenTransformer** (per-feature tokenisation $\rightarrow$ 2-layer Transformer encoder with $d_{\text{model}} = 32$, 4 heads, FFN 128, GELU, dropout 0.1 $\rightarrow$ mean pool $\rightarrow$ MLP head $32 \rightarrow 32 \rightarrow 1$). Optimisation uses AdamW (lr 2×10^{-3}, weight decay 10^{-4}), batch size 256, and gradient clipping at $\|\mathbf{g}\|_2 \leq 5$. Reported metrics are mean $\pm$ standard deviation across the five outer folds.

3.3 Methodology

Each DHS cluster i contributes the cross-sectional feature vector $\mathbf{x}_i$ defined in Section 3.1, presented as a length-one sequence to both encoders of the ensemble specified in Section 3.2. Writing $\hat{\hat{y}}_i^{\text{LSTM}}$ and $\hat{\hat{y}}_i^{\text{Trf}}$ for the two encoders' predictions in log-rate space, the ensemble averages them as

$$\hat{\hat{y}}_i \;=\; \tfrac{1}{2}\Big(\hat{\hat{y}}_i^{\text{LSTM}} \;+\; \hat{\hat{y}}_i^{\text{Trf}}\Big), \tag{3}$$

and predictions are inverted to the rate scale via $\hat{y}_i = \max\big(0,\, (\exp(\hat{\hat{y}}_i) - 1)/1000\big)$ before computing RMSE and R^2. Reusing the same ensemble across all three case studies isolates spatial granularity and embedding utility, rather than architectural choice, as the variable under test.

To check that the headline ΔR^2 is not an artefact of the inductive bias of a single estimator family, we fit three tree-based regressors — Random Forest, HistGradientBoosting, and XGBoost — to the same three feature sets and the same 5-fold splits used by the ensemble. Results are reported in Appendix C (Fig. 7). All three estimators recover the same ordering, *baseline* < *emb_only* < *emb_plus*, with XGBoost reaching the highest pooled R^2 of 0.210 and Random Forest the lowest pooled RMSE. The cross-model agreement on the direction and magnitude of ΔR^2 indicates that the signal lies in the AlphaEarth features rather than in the architecture of any single estimator.

3.4 Geographic distribution of model performance

Figure 3 maps the per-country metrics onto the 11 study countries, making the geographic structure of the gain—and the small-sample penalty—visible at a glance. The top row shows that absolute R^2 is positive in only a handful of countries under either feature set, but turns greener (improves) in several countries once embeddings are added. The bottom-left panel isolates this shift as ΔR^2, with the largest green polygons over sub-Saharan Africa and South Asia (Mozambique, Nigeria, Nepal, Pakistan, India), and red polygons concentrated in the smaller samples (BD, LB, TJ, PH, KH). The bottom-right ΔNRMSE panel tells a consistent story on a scale-free error metric: the same large-n countries see green (lower error with AlphaEarth), and the same small-n countries see red.

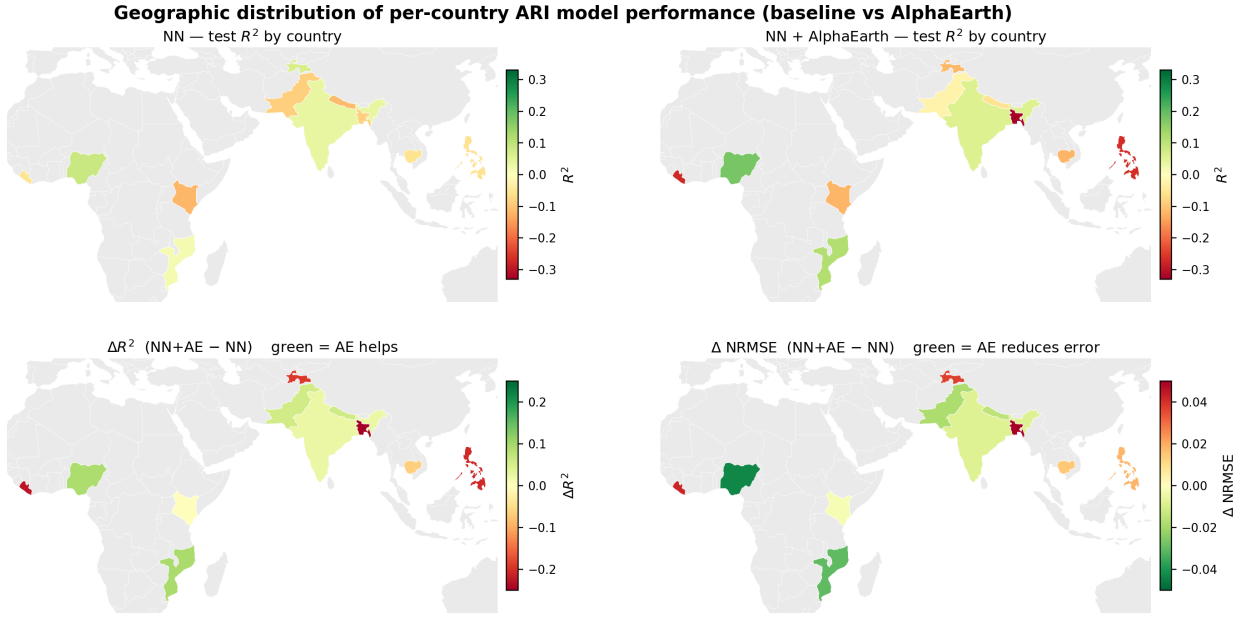


Figure 3: **Per-country ARI prediction performance shown as a 2×2 choropleth over the 11 study countries. Top-left:** test R^2 for the baseline model (gaseous pollutants + controls). **Top-right:** test R^2 for the AlphaEarth model (64-dim embeddings + gaseous + controls). The two top panels share a diverging colormap centered at $R^2 = 0$, so green indicates positive predictive skill and red indicates a model that performs worse than the per-country mean. **Bottom-left:** change in R^2 when AlphaEarth embeddings are added to the baseline ($\Delta R^2 = R^2_{\text{Emb+gas}} - R^2_{\text{baseline}}$); green countries are those where AlphaEarth helps. **Bottom-right:** change in normalized RMSE (ΔNRMSE, where NRMSE is RMSE divided by the country's mean ARI case count, for cross-country comparability); green countries are those where AlphaEarth reduces error. Both bottom panels use diverging scales centered at zero with their own inline colorbars. Country labels follow DHS two-letter codes. The geographic pattern matches the sample-size split discussed in Section 3.4: large-sample countries (IA, NG, PK, NP, MZ) are green in both delta panels, while the five small-sample countries (BD, LB, TJ, PH, KH) are red, consistent with overfitting when the 64-dimensional embedding approaches the per-fold training size.

3.5 Results

Adding AlphaEarth embeddings to the gaseous-pollutant baseline raises pooled test R^2 from 0.098 to 0.164 ($\Delta R^2 = +0.065$) across the 9,271-cluster sample under the LSTM + Transformer ensemble (Table 4 (synthesis)). The effect is geographically structured (Fig. 3): large-sample countries (India, Nigeria, Pakistan, Nepal, Mozambique) are green in both the ΔR^2 and ΔNRMSE panels, indicating that AlphaEarth both improves predictive skill and reduces error in these settings. The five small-sample countries (Bangladesh, Lebanon, Tajikistan, the Philippines, Cambodia) are red in both panels, consistent with overfitting when the 64-dimensional embedding approaches the per-fold training size.

4 Case 3 — Stunting (WHZ), 35 DHS Countries

4.1 Covariates

The analysis cohort comprises 363,353 DHS children under five nested within 47,154 geocoded primary sampling units (`isocluster`) across 35 LMICs surveyed from 2015 onwards, obtained after dropping rows with any missing value in the target, the modelling covariates, the cluster identifier, or geocoordinates, and restricting to the DHS plausibility range $|\text{WHZ}| \leq 5$. Each child carries the original DHS sample weight, which is propagated into all model training and evaluation as a per-row weight.

Two feature blocks are fed to every model (Table 2). The numeric block is standardised to zero mean and unit variance using statistics computed on the training fold only; the categorical block is pinned to the levels observed in training to prevent leakage of test-fold categories. Country enters every model as a learned 8-dimensional embedding (and as a native categorical in the tree-based robustness check).

Table 2: Case 3 covariates. The 64 AlphaEarth bands are present only in the $+AE$ variants. Cleaning drops any row with a missing value in any listed column.

Block	Variable	Description
Numeric (standardised)	`tmax61`	61-day maximum 2 m temperature at the cluster (°C).
	`agemo`	Child age in months.
	`alt`	Cluster elevation above sea level (m).
	`income`	Household income indicator (DHS-derived).
	`cci`	Composite Coverage Indicator ($[0, 1]$).
	`sanitation`	Household sanitation indicator (DHS-derived).
	`sex`	Child sex (encoded 1/2).
	`wiq`	Wealth index quintile (1 poorest – 5 richest).
	`breastfeeding`	Breastfeeding status indicator.
Categorical	`area`	Urban / rural classification.
	`classification`	Climate zone $\in$ {tropical, arid, temperate}.
Geographic	`country`	35 LMIC ISO codes; learned 8-d embedding.
Satellite ($+AE$ *only*)	`A00`–`A63`	64-band AlphaEarth fingerprint sampled at the cluster centroid over a 2–5 km buffer matching the DHS privacy displacement radius, year-matched to the survey (clamped ≥ 2017). L_2-normalised per row in the neural models.

The target is the child's weight-for-height z-score, country-demeaned as $y_{i,c} = \text{WHZ}_{i,c} - \overline{\text{WHZ}}_c$ where $\overline{\text{WHZ}}_c$ is the *training-fold* mean for country c (Eq. (4)). Country fixed effects dominate raw WHZ variance, so the embedding is asked to explain within-country variation rather than between-country differences in nutritional status; this is the analytical choice that makes the Case 3 null result a clean test of *spatial granularity* rather than of the embedding itself.

4.2 Training and testing setup

We use a single *cluster-stratified* 5-fold partition built with `sklearn.model_selection.GroupKFold`, with `isocluster` as the grouping variable, so that no DHS cluster appears in more than one of the train, validation and test folds. This rules out the dominant source of leakage in DHS data —

co-located children sharing the same primary sampling unit — that an i.i.d. row-level split would otherwise admit. Fold 0 is held out as the test fold; fold 1 serves as the validation fold for early stopping and residual-stack blend weights; the remaining three folds form the training set (Table 3).

Table 3: Cluster-stratified train/validation/test partition for Case 3, shared by every model. Cluster counts are disjoint across the three folds; child counts sum to the cleaned cohort of 363,353.

Fold	Children (n)	DHS clusters	Share of cohort
Train	218,011	$\approx$ 28,910	60.0%
Validation	72,671	$\approx$ 9,122	20.0%
Test	72,671	9,122	20.0%
Total	363,353	47,154	100.0%

The test fold spans all 35 cohort countries, ranging from South Africa ($n_{\text{test}} = 151$) to India ($n_{\text{test}} = 34{,}336$), and is never seen during training, validation, hyperparameter selection, or blend tuning. The feature scaler, the categorical level set, the training-fold country means used for demeaning in Eq. (4), and the AlphaEarth–cluster join are all fitted once on the training fold and applied identically to validation and test. Survey weights are normalised within each fold so that the mean weight equals one, preserving relative weighting while keeping the optimiser's effective step size comparable across folds. The same fold assignment is used for the *baseline* and $+AE$ configurations, so ΔR^2 in Table 1 and Fig. 4 isolates the contribution of the embedding rather than fold variance, and all splits are seeded with `seed` $= 42$ for reproducibility.

The ensemble architecture matches Cases 1 and 2: each child's cluster-level feature vector is presented as a length-one sequence to both a single-layer LSTM and a two-layer Transformer encoder with $d_{\text{model}} = 64$, each followed by an MLP head ($64 \to 64 \to 1$), and the two predictions are averaged on the country-demeaned WHZ scale. Reusing the same ensemble across all three case studies isolates spatial granularity and embedding utility, rather than architectural choice, as the variable under test.

4.3 Methodology

Each of the 363,353 children contributes one observation, linked to its DHS cluster's covariates: ambient temperature, child age, altitude, household income, Composite Coverage Indicator, sanitation access, and sex. The target $y_{i,c}$ for child i in country c is the country-demeaned Weight-for-Height Z-score:

$$y_{i,c} = \text{WHZ}_{i,c} - \overline{\text{WHZ}}_c, \tag{4}$$

so the embedding is asked to explain residual within-country variation rather than between-country differences in nutritional status. For each of the 47,154 unique DHS clusters, a 64-dimensional AlphaEarth fingerprint is sampled at the survey year over a 2–5 km buffer matching the DHS privacy displacement radius, and broadcast to every child in that cluster.

4.4 Geographic distribution of model performance

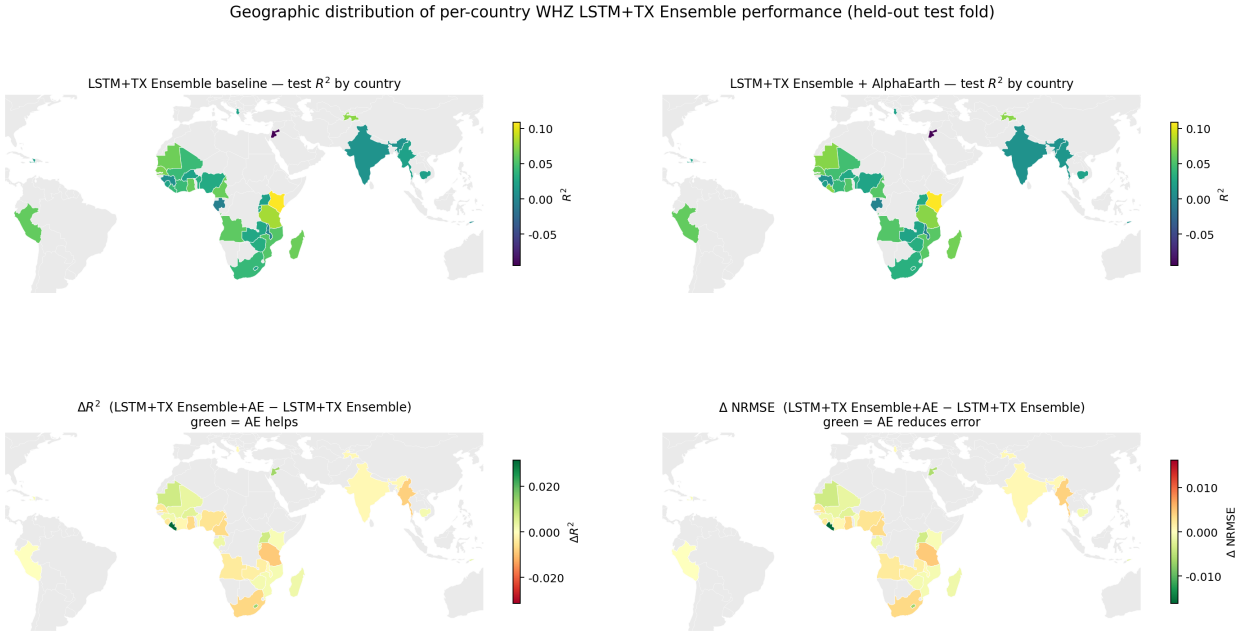


Figure 4: **Geographic distribution of per-country weight-for-height z-score (WHZ) model performance on the held-out test fold. Top row:** per-country test R^2 for the neural-network baseline without AlphaEarth features (left, "NN (no AE)") and the same architecture augmented with AlphaEarth satellite embeddings (right, "NN + AE"), plotted on a shared *viridis* scale so countries can be compared directly across models. **Bottom row:** per-country differences between the two models, $\Delta R^2 = R^2_{\text{NN+AE}} - R^2_{\text{NN}}$ (left) and $\Delta\,\text{NRMSE} = \text{NRMSE}_{\text{NN+AE}} - \text{NRMSE}_{\text{NN}}$ (right), shown on diverging *red–yellow–green* scales centred at zero. In both bottom panels, green indicates that adding AlphaEarth improves the model (higher R^2 or lower NRMSE), while red indicates degradation. Countries included in the study are outlined and coloured; surrounding countries are shown in light grey for geographic context. Hatched fill denotes a country present in our cohort for which a metric was not available.

4.5 Results

Under the LSTM + Transformer ensemble at the cluster level (n = 47,154), test R^2 moves from 0.0408 (baseline) to 0.0418 (+ AlphaEarth), a ΔR^2 of +0.00100 that lies within seed variance (Table 4 (synthesis), Fig. 4). The per-country choropleth (Fig. 4, bottom row) shows the differences scattered symmetrically around zero with no coherent geographic structure — the green and red countries do not cluster by region, sample size, or baseline performance, in contrast to the spatially structured gains seen in Cases 1 and 2. This null result is the expected behaviour when AlphaEarth features cannot express within-country variation beyond what country fixed effects already capture, and it isolates spatial granularity — rather than the embedding itself — as the binding constraint on this outcome.

5 Synthesis: Where AlphaEarth Helps

Across these three outcomes a consistent pattern emerges, summarised in Table 4.

Table 4: AlphaEarth contributions across the three studies. The spatial grain of the merge determines whether the embedding can contribute: country-level broadcast is collinear with country identity, while pixel- or cluster-level merges are informative. Cases 1a and 1b use the MAP 5×5 km pixel as the unit of observation; Cases 2 and 3 use the DHS cluster.

Outcome	Method	Grain	Baseline R^2	+ AE R^2	ΔR^2
Malaria (Nigeria)	Ensemble (LSTM + Transformer)	Pixel (5×5 km, $n = 4{,}233$)	0.623	0.777	+0.154
Malaria (India)	Ensemble (LSTM + Transformer)	Pixel (5×5 km, $n = 28{,}264$)	0.867	0.881	+0.014
ARI (11 LMICs, pooled)	Ensemble (LSTM + Transformer)	Cluster (n=9,271)	0.098	0.164	+0.065
WHZ, 35 LMICs	Ensemble (LSTM + Transformer)	Cluster ($n = 47{,}154$)	0.0408	0.0418	+0.00100

Cases 1, 2, and 3 all operate at pixel or cluster granularity and all show the expected gains, with the magnitude of ΔR^2 scaling with the headroom left by the baseline: largest in malaria (Nigeria) and ARI, where baseline R^2 is modest, and smallest in WHZ.

6 Conclusion

Across three physiologically distinct health outcomes in LMICs, AlphaEarth Foundations satellite embeddings consistently add predictive value when applied at cluster or location level. The gains in malaria prediction (Δ R^2=+0.154, Nigeria; Δ R^2=+0.014, India) and childhood ARI prediction (Δ R^2=+0.065) are geographically uniform and robust. For childhood stunting, the improvement at cluster granularity is modest ($\Delta R^2 = +0.001$) and lies within seed variance. We interpret this as the expected behaviour when country fixed effects already absorb the between-country variance the static AlphaEarth fingerprint can express, leaving within-country variation in WHZ — driven by household-level structural determinants such as income, sanitation, and breastfeeding — as the residual the embedding is asked to explain. This isolates spatial granularity, rather than the embedding itself, as the binding constraint on this outcome, and motivates the Population Dynamics extension proposed in Appendix Section D.

Taken together, these results establish AlphaEarth embeddings as a promising and generalisable feature class for population health modelling in low-resource settings, with clear implications for disease surveillance, early warning systems, and the targeting of public health interventions.

7 Limitations

Several limitations should be noted. First, DHS cluster coordinates are displaced by up to 5 km (urban) or 10 km (rural) for privacy protection [35]; this introduces measurement error in the spatial merge that is likely to attenuate, rather than inflate, the observed embedding effects. Second, AlphaEarth embeddings are static annual composites and cannot capture intra-annual environmental dynamics; monthly or seasonal embeddings would be needed to model outbreak-scale variation.

References

[1] Villavicencio F, Perin J, Eilerts-Spinelli H, Yeung D, Prieto-Merino D, Hug L, et al. Global, regional, and national causes of death in children and adolescents younger than 20 years: an open data portal with estimates for 2000–21. The Lancet Global Health. 2024 1;12:e16-7.

[2] for Health Metrics I, (IHME) E. Global Burden of Disease 2023: Findings from the GBD 2023 Study. Institute for Health Metrics and Evaluation (IHME); 2025.

[3] UNICEF. Levels and Trends in Child Mortality Report 2025. United Nations Inter-agency Group for Child Mortality Estimation (UN IGME); 2026.

[4] Guerra CA, Reiner RC, Perkins TA, Lindsay SW, Midega JT, Brady OJ, et al. A global assembly of adult female mosquito mark-release-recapture data to inform the control of mosquito-borne pathogens. Parasites & Vectors. 2014 12;7:276.

[5] Suh E, Grossman MK, Waite JL, Dennington NL, Sherrard-Smith E, Churcher TS, et al. The influence of feeding behaviour and temperature on the capacity of mosquitoes to transmit malaria. Nature Ecology & Evolution. 2020 5;4:940-51.

[6] Agyekum TP, Botwe PK, Arko-Mensah J, Issah I, Acquah AA, Hogarh JN, et al. A Systematic Review of the Effects of Temperature on Anopheles Mosquito Development and Survival: Implications for Malaria Control in a Future Warmer Climate. International Journal of Environmental Research and Public Health. 2021 7;18:7255.

[7] Stopard IJ, Churcher TS, Lambert B. Estimating the extrinsic incubation period of malaria using a mechanistic model of sporogony. PLOS Computational Biology. 2021 2;17:e1008658.

[8] Mordecai EA, et al. Thermal biology of mosquito-borne disease. Ecology Letters. 2019;22(10):1690-708.

[9] Nazir U, Quddoos MT, Uppal M, Khalid S. Predicting malaria outbreaks using earth observation measurements and spatiotemporal deep learning modelling: a South Asian case study from 2000 to 2017. The Lancet Planetary Health. 2024 4;8:S17.

[10] World Health Organization. World Malaria Report 2025. Geneva: World Health Organization; 2025.

[11] Symons TL, Moran A, Balzarolo A, Vargas C, Robertson M, Lubinda J, et al. Projected impacts of climate change on malaria in Africa. Nature. 2026 3;651:390-6.

[12] Mirsaeidi M, Motahari H, Khamesi MT, Sharifi A, Campos M, Schraufnagel DE. Climate Change and Respiratory Infections. Annals of the American Thoracic Society. 2016 8;13:1223-30.

[13] Cicco MED, Ferrante G, Amato D, Capizzi A, Pieri CD, Ferraro VA, et al. Climate Change and Childhood Respiratory Health: A Call to Action for Paediatricians. International Journal of Environmental Research and Public Health. 2020 7;17:5344.

[14] Chang JH, Lee YL, Chang LT, Chang TY, Hsiao TC, Chung KF, et al. Climate change, air quality, and respiratory health: a focus on particle deposition in the lungs. Annals of Medicine. 2023 12;55.

[15] He Y, Liu WJ, Jia N, Richardson S, Huang C. Viral respiratory infections in a rapidly changing climate: the need to prepare for the next pandemic. eBioMedicine. 2023 7;93:104593.

[16] Charnley GEC, Kelman I. Perspectives on climate change and infectious disease outbreaks: is the evidence there? npj Climate Action. 2024 7;3:61.

[17] Lloyd S, Bangalore M, Chalabi Z, Kovats RS, Hallegatte S, Ronberg J, et al. Potential impacts of climate change on child stunting via income and food price in 2030: a global-level model. The Lancet Planetary Health. 2019 9;3:S1.

[18] Agostoni C, Baglioni M, Vecchia AL, Molari G, Berti C. Interlinkages between Climate Change and Food Systems: The Impact on Child Malnutrition—Narrative Review. Nutrients. 2023 1;15:416.

[19] Helldén D, Andersson C, Nilsson M, Ebi KL, Friberg P, Alfvén T. Climate change and child health: a scoping review and an expanded conceptual framework. The Lancet Planetary Health. 2021 3;5:e164-75.

[20] Tusting LS, Bradley J, Bhatt S, Gibson HS, Weiss DJ, Shenton FC, et al. Environmental temperature and growth faltering in African children: a cross-sectional study. The Lancet Planetary Health. 2020 3;4:e116-23.

[21] Phalkey RK, Aranda-Jan C, Marx S, Höfle B, Sauerborn R. Systematic review of current efforts to quantify the impacts of climate change on undernutrition. Proceedings of the National Academy of Sciences of the United States of America. 2015 8;112:E4522-9.

[22] Nazir U, Khalid S. Foundation Models for Mapping Emission Sources and Acute Respiratory Infection (ARI) Hotspots [Poster Presentation]. NeurIPS; 2025.

[23] Shi S, Lin H, Jiang L, Zeng Z, Lin C, Li P, et al. Development of a respiratory virus risk model with environmental data based on interpretable machine learning methods. npj Climate and Atmospheric Science. 2025 2;8:39.

[24] Bachwenkizi J, He C, Zhu Y, Mugisha A, Tlou B, Moshiro C, et al. Predicting the effects of temperature variability on nutritional status of children under five in Sub-Saharan Africa using machine learning. Scientific Reports. 2026 2;16:8055.

[25] Brown CF, Kazmierski MR, Pasquarella VJ, Rucklidge WJ, Samsikova M, Zhang C, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv preprint arXiv:250722291. 2025.

[26] Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Scientific Data. 2015;2:150066.

[27] Didan K. MODIS/Terra Vegetation Indices Monthly L3 Global 1km SIN Grid V061. NASA EOSDIS Land Processes DAAC; 2021. Distributed by NASA EOSDIS Land Processes DAAC.

[28] Muñoz-Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, et al. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. Earth System Science Data. 2021;13(9):4349-83.

[29] Woods D, Cunningham A, Utazi CE, Bondarenko M, Shuaib F, Garcia AJ, et al.. Distance to waterways layers, WorldPop. WorldPop, University of Southampton; 2024. https://www.worldpop.org/.

[30] Bondarenko M, Nieves JJ, Stevens FR, Gaughan AE, Tatem AJ, Sorichetta A. WorldPop gridded population estimates: age– and sex–structured layers. WorldPop, University of Southampton; 2025. https://www.worldpop.org/.

[31] Li X, Zhou Y, Zhao M, Zhao X. A harmonized global nighttime light dataset 1992–2018. Scientific Data. 2020;7:168.

[32] GADM. GADM database of Global Administrative Areas, version 4.1; 2022. https://gadm.org/.

[33] Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment. 2017;202:18-27.

[34] Malaria Atlas Project. Global *Plasmodium falciparum* Incidence Rate, 2000–2024; 2024. Geospatial estimates at ∼5 km resolution; accessed via the MAP data portal. https://malariaatlas.org/.

[35] Burgert CR, Brady J, Colston J, et al. Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys. Calverton, Maryland: ICF International; 2013. 7.

A Methodological choices across the three case studies

The three case studies share an ensemble architecture (LSTM + Transformer), a feature-merge protocol (AlphaEarth fingerprint appended to a per-case baseline), and a fairness convention (identical folds and seeds between the baseline and the +AE configuration, so ΔR^2 isolates the contribution of the embedding rather than split variance). Several other design choices differ across cases. The differences are not arbitrary: each is dictated by the structure of the underlying data, and choosing a single common setting across all three would have introduced a worse problem than the inconsistency it removed. Table 5 summarises the differences; the rest of this section justifies each row.

Table 5: Methodological choices that differ across the three case studies, and the data-structural reason for each difference. All other modelling choices — ensemble architecture, AlphaEarth merge protocol, identical folds/seeds across baseline and +AE — are held fixed.

	Case 1 (Malaria)	**Case 2 (ARI)**	**Case 3 (WHZ)**
Unit of observation	Pixel-year (5×5 km)	DHS cluster (cross-sectional)	Child (nested in cluster)
Temporal structure	24-month sequence	Length-1 sequence	Length-1 sequence
Target transform	log + z-score; per-location for IN	$\log(1 + 1000y)$ on rate	Country-demeaned WHZ
Split	Forward holdout (test = 2024)	Random 5-fold KFold	GroupKFold by cluster
Repetition	5 seeds, one split	5 outer folds, one seed	Single fold-0 holdout, one seed
Internal validation	15% slice, patience 5	15% slice, patience 8	Fold 1 as validation

Unit of observation. The unit is fixed by the source data, not by a modelling choice. The Malaria Atlas Project distributes annual case counts on a 5×5 km raster, so pixel-year is the natural row. DHS surveys ask the ARI symptom question at the household level and aggregate cluster-level counts as a survey product, so the cluster is the natural row. Anthropometric measurements in DHS are taken on individual children, and using the child as the row preserves the within-cluster variance in WHZ that the embedding is asked to explain. Forcing a common unit across cases — for instance, aggregating WHZ to the cluster level to match Case 2 — would have discarded most of the variance in the only outcome where the AlphaEarth signal is expected to be weakest, and would have made the Case 3 null result less interpretable rather than more.

Temporal structure. Only malaria has a usable monthly history. ERA5-Land climate is monthly and malaria transmission is strongly seasonal, so a 24-month sequence ending at the prediction year is the standard formulation in the malaria forecasting literature and is what carries the seasonal signal the LSTM and Transformer encoders are designed to capture. The DHS covariates used in Cases 2 and 3 are recorded once per survey, and the 2–5 km AlphaEarth buffer is a static annual composite at the survey year, so there is no monthly axis for either encoder to act on. Presenting the cross-sectional feature vector as a length-1 sequence to the same encoders keeps the architecture fixed across cases while respecting what the data can actually provide.

Target transform. Each target is transformed to bring it into a range the network can optimise stably. Annual malaria counts span several orders of magnitude across pixels, so the log + z-score transform is standard; the per-location refinement for India is necessary because the panel-only baseline fails on low-incidence Indian zones whose clusters cannot be distinguished by climate or demographics alone, as shown in Fig. 2. ARI is modelled as a rate (`cases/population_in_buffer`) so that the model must explain prevalence rather than cluster size; the $\log(1 + 1000y)$ rescaling spreads the rate over a useful gradient range. WHZ is already a z-score by construction, so no log transform is needed; country demeaning is applied because country fixed effects otherwise dominate raw WHZ variance and the embedding would be scored mostly on between-country differences in nutritional status it cannot plausibly explain. Each transform is chosen so the model is asked to explain a variance component the AlphaEarth fingerprint can plausibly reach, and inverted before reporting metrics so all R^2 values in the paper are on the natural scale of the outcome.

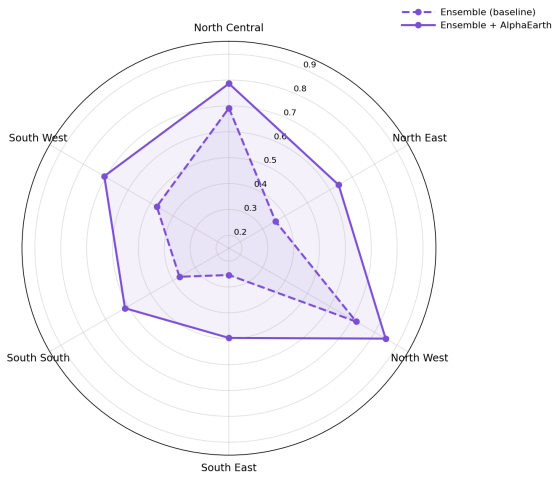
Split. The split is the most case-specific choice, because the dominant leakage risk differs between cases. Malaria is a forecasting problem: the test-time use case is predicting the next annual case count from a trained model, and the autoregressive lag-1/lag-2 features used for India would leak future information into the test set under any random split. The 2024 forward holdout enforces that the model must extrapolate beyond its training window, which is the realistic deployment regime. ARI in Case 2 is cross-sectional with a single survey year per cluster, so temporal holdout is not meaningful; i.i.d. 5-fold KFold is what the rate prediction is being evaluated on, and the limitation that this does not control for spatial autocorrelation is stated explicitly in Section 3.2. WHZ has many children per cluster (389,035 children in 47,154 clusters), so the dominant leakage risk is co-located children sharing a primary sampling unit appearing in both train and test under an i.i.d. row split; GroupKFold with `isocluster` as the grouping variable is the only split that rules this out, and the stricter standard is appropriate because Case 3 is also the weakest signal — a generous split would have made a null result indistinguishable from a small positive one.

Repetition and reported uncertainty. Repetition strategy follows from the split. Case 1 has a single forward-holdout test set, so uncertainty is quantified by running the pipeline with 5 random seeds and reporting the seed-mean and seed standard deviation in the choropleths. Case 2 has 5 outer folds, so uncertainty is quantified across folds. Case 3 uses fold 0 as a held-out test fold under GroupKFold; the residual-stack blend weight is tuned on fold 1 and the per-country differences in Fig. 4 carry the geographic interpretation of the result. These different quantities — seed standard deviation in Case 1, fold standard deviation in Case 2, single-test-fold point estimate in Case 3 — are not interchangeable, and the paper reports each case on its own terms rather than imposing a common "±" that would conflate them. The synthesis in Section 5 and Table 4 reports point estimates of ΔR^2 rather than uncertainty intervals for exactly this reason.

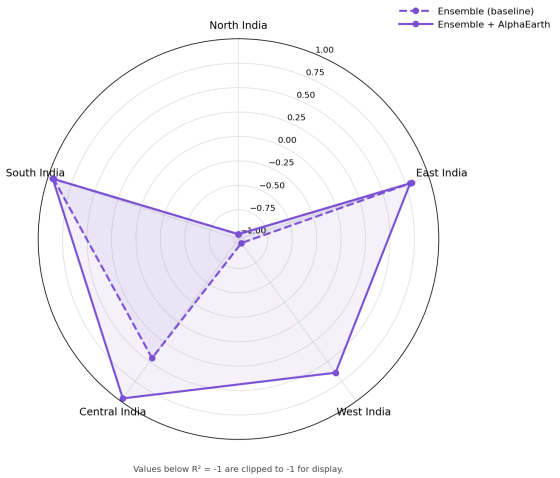
Internal validation and early stopping. Patience and maximum-epoch settings (patience 5, max 25 for malaria; patience 8, max 60 for ARI; the residual-stack blend weight in Case 3) were chosen per case from the validation-loss curves of the baseline configuration on a small grid, before any test-set evaluation. The +AE configuration inherits the same settings without re-tuning. This asymmetry is deliberate: re-tuning the early stopping schedule on the +AE configuration would give that configuration two sources of advantage — better features and a better schedule — which would make ΔR^2 no longer a clean test of the embedding.

What is held fixed across cases. The ensemble architecture (LSTM + Transformer averaged in the transformed target space), the AlphaEarth merge protocol (cluster or pixel centroid, year-matched, 2–5 km buffer where applicable), and the fairness convention (identical folds and seeds between baseline and +AE) are common to all three cases. The cross-estimator robustness check in Appendix C additionally shows that the Case 2 result holds across three tree-based estimators, ruling out a single-architecture artefact. The remaining differences in the table above are exactly those forced by the data; the common elements are those that can be held fixed without distorting any case.

B Case 1 — Malaria Incidence Count Prediction in Nigeria and India



(a) **Per-region 2024 test R^2 across Nigerian states.** Each spoke is one state; the inner polygon is the climate-only baseline and the outer polygon is the same model with the 64-dim AlphaEarth fingerprint appended. The outer polygon strictly dominates on every spoke, indicating that the lift is geographically uniform rather than driven by a few high-burden states.



(b) **Per-region 2024 test R^2 across Indian states** The very negative baseline R^2 on North India is not a real failure across the zone – the model gets most clusters about right, but on one or two clusters it predicts a huge number (e.g. 1,900) when the truth is at most 204, and because North India has very small malaria counts overall, that single mistake is enough to drag R^2 deep below zero.

Figure 5: **Case 1 — Malaria case prediction in Nigeria (NMEP, 2000–2024; train 2000–2023, test 2024).** AlphaEarth embeddings provide a geographically uniform R^2 gain (left) and emerge as the dominant feature group in the importance decomposition (right), supporting the interpretation that static landscape structure carries malaria-transmission signal not captured by monthly climate covariates.

C Case 2 — Childhood ARI, 11 DHS Countries

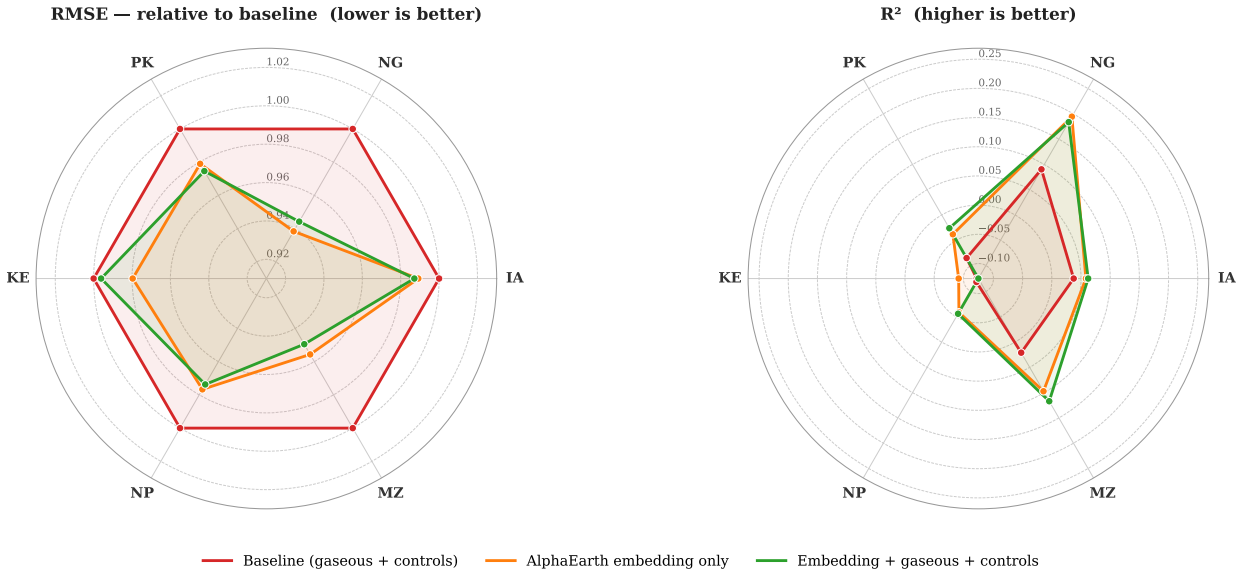


Figure 6: **Per-country performance for the six large sample DHS countries with $n \geq 139$ clusters** (India, Pakistan, Nigeria, Kenya, Nepal, Mozambique; 8,696 of 9,271 clusters total). Left panel: RMSE relative to baseline (lower is better). Right panel: test R^2 (higher is better). The three polygons compare the gas+controls baseline (red), AlphaEarth alone (orange), and the combined model (green). The combined model strictly dominates baseline on every spoke; the largest absolute gains are in Mozambique ($\Delta R^2 = +0.096$) and Nigeria (+0.093).

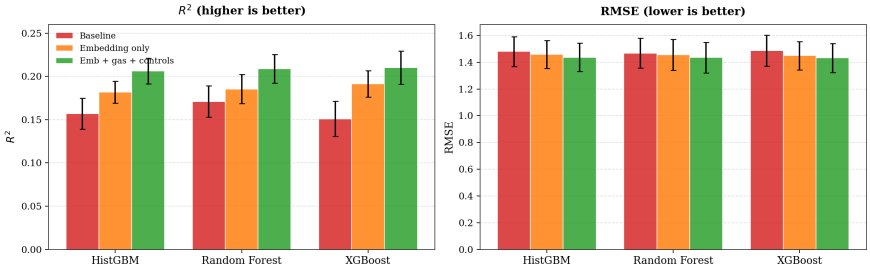


Figure 7: Pooled-model ablation. R^2 (left) and RMSE (right) for each estimator–feature-set combination. Error bars denote cross-fold standard deviation across 5 folds. All three models agree that adding embeddings to the gaseous baseline yields a consistent improvement, with XGBoost reaching the highest $R^2 = 0.210$ and Random Forest the lowest RMSE.

D Future work

We see four concrete ways the Google Earth AI model could accelerate this line of work:

1. **Population Dynamics Foundation embeddings.** The Population Dynamics model includes health and socio-demographic indicators that would complement the landscape-focused

AlphaEarth representation. A combined feature set would be particularly valuable for stunting (where nutritional status depends on both environmental and structural determinants) and for ARI (where health-system covariates partially compete with pollution features in the baseline). Direct access to pre-extracted Population Dynamics embeddings matched to DHS cluster coordinates would enable a three-way ablation (tabular only / + AlphaEarth / + AlphaEarth + Population Dynamics) analogous to what Case 1 demonstrates for single-modality embeddings.

2. **Multi-temporal embeddings.** Moving from static annual composites to quarterly or monthly would allow models to capture how seasonality, through weather, migration, land-use changes, etc., correlate with outcomes of interest.

3. **Cluster-level AlphaEarth at scale.** Our cluster-level extraction across $\sim$ 45,775 DHS clusters with 2–5 km buffers is computationally heavy; A batched extraction pipeline or direct access to pre-computed cluster-level fingerprints for the DHS program would substantially shorten the time to a definitive result.

4. **Methodological feedback.** We would value the AE team's perspective on two specific choices: (a) whether to weight within-buffer pixels uniformly or by population density during cluster-level averaging, and (b) whether multi-year embedding composites (e.g. 3-year means) or survey-year point estimates are preferable for linking to DHS surveys with staggered fieldwork.