Learned Geospatial Priors for Malaria Forecasting

Usman Nazir[1], **Sina Rezaei**[2], **Arundhati Wuppalapati**[3], **Lais Soares**[4], **Sara Khalid**[1]
[1]Planetary Health Informatics Lab, NDORMS, University of Oxford
[2]Technical University of Munich
[3]Department of Paediatrics, University of Oxford
[4] Universidade Estadual de Campinas
[1]`{usman.nazir, sara.khalid}@ndorms.ox.ac.uk`
[2] `sina.rezaei@tum.de`
[3]`arundhati.wuppalapati@paediatrics.ox.ac.uk`
[4]`l173455@dac.unicamp.br`

Abstract

Malaria remains one of the leading communicable causes of death globally, with approximately half the world's population at risk, predominantly across African and South Asian countries. Accurate, region-specific outbreak prediction is hindered by strong spatio-temporal heterogeneity in environmental, climatological, and sociodemographic risk factors, as well as by the over-dispersed, count-valued nature of incidence data that violates assumptions of standard regression and many off-the-shelf machine learning approaches. We address these challenges through a unified framework combining count-aware statistical modelling, sequence learning, and geospatial foundation model representations, evaluated across India and Nigeria using monthly data from 2000 to 2024 integrating rainfall, temperature, vegetation index, nighttime lights, population counts, distance to water bodies, and historical malaria case counts. First, we model malaria incidence using lag-structured negative binomial regression, explicitly accounting for over-dispersion while capturing biologically grounded temporal effects; this yields interpretable estimates of how environmental and sociodemographic covariates drive transmission. Second, we develop an ensemble of LSTM and Transformer models that captures complementary temporal dependencies in incidence dynamics. Third, we augment both approaches with embeddings from the AlphaEarth foundation model, which encode high-dimensional satellite-derived environmental context. Incorporating these embeddings consistently improves predictive performance, with the ensemble model increasing R^2 from 0.63 to 0.78 in Nigeria and from 0.87 to 0.88 in India, while remaining stable across random seeds. Together, these results suggest that foundation model embeddings convert previously unexplained variance into structured environmental signal, reducing effective over-dispersion and providing a robust and transferable framework for forecasting climate-sensitive infectious diseases.

1 Introduction

Malaria represents a global public health challenge. It remains one of the leading communicable causes of death Gelband et al. (2020); Cowman et al. (2016), with approximately half the world's population considered at risk, predominantly in Africa and South Asia Rosenthal et al. (2019). Malaria incidence has been rising since 2015, despite initial progress at the start of the century WHO (2025). Though this is multifactorial - being caused in part by increasing resistance to anti-malarial

Preprint.

drugs and increasing mosquito resistance to insecticides - a notable driver of this increase in malaria incidence is climate change WHO (2025).

There is a wealth of biological evidence demonstrating the substantial and complex effects of climate change on vector biology and disease transmission Guerra et al. (2014); Suh et al. (2020); Agyekum et al. (2021); Stopard et al. (2021); Mordecai et al. (2019). Additionally, as our understanding of how the natural history of malaria is influenced by the climate has developed, so have attempts to extrapolate how malaria incidence may change with the changing climate, using statistical modelling initially, and more recently machine learning classification techniques Tjaden et al. (2018). Statistical modelling studies very early on showed an association between temperature and rainfall and malaria transmission, though these were and largely continue to be location and season specific Craig et al. (1999); Martens et al. (1995); Rogers and Randolph (2000); Hoshen and Morse (2004); Beloconi et al. (2023).

As machine learning has grown in popularity, techniques such as Random Forests, KNN, Naive Bayes and Extreme Gradient Boost have also shown some promise when linking climatic data and malaria incidence Breiman (2001); Kalipe et al. (2018); Tjaden et al. (2018). Some of these models have provided evidence that temperature effects alone, including sea temperature, can be valuable predictors of malaria incidence Ryan et al. (2020); Mordecai et al. (2020); Martineau et al. (2022), while we and others take a more comprehensive approach and include rainfall, humidity and other climactic indicators to improve predictions Pillay et al. (2023); Nazir et al. (2024); Symons et al. (2026). Additionally, more recently, approaches such as conv-LSTM Shi et al. (2015) and multidimensional-LSTM Nazir et al. (2023); Sebastianelli et al. (2024) appear to lend themselves especially well to climate-based infectious disease predictions.

The key limitations of these studies, as they stand, are two fold. First, most of these only focus on one country, or region - likely due to data availability, which can be challenging in resource poor settings. However, it is becoming increasingly important to go beyond this, with the possibility of climate change expanding the region in which malaria is transmitted. Second, many of the existing studies only use climactic indicators, when in practice, many socioeconomic and demographic factors play a major role in malaria transmissionBaeza et al. (2017); Kibret et al. (2017); Shah et al. (2022); Beloconi et al. (2023); Batheja et al. (2026). Thus, for more accurate prediction of malaria outbreaks, which can inform early warning systems, there is a need to integrate these socio-economic and demographic variables with climactic variables in a model.

Here, we develop an ensemble framework combining LSTM and Transformer architectures with geospatial, climatic, and demographic covariates. The proposed approach further integrates AlphaEarth embeddings Brown et al. (2025), derived from a geospatial foundation model built upon Space Time Precision (STP) learning, enabling the capture of both local and global spatiotemporal relationships through annual embedding field layers (2017–2024).

To further assess the contribution of AlphaEarth embeddings, we employ Negative Binomial Regression analysis to evaluate their effect on over-dispersion reduction in malaria incidence modelling. Finally, AlphaEarth embeddings are incorporated into the ensemble framework to enrich spatial feature representations and improve malaria outbreak forecasting.

Contributions.

1. A reproducible pipeline integrating geospatial foundation models with malaria panels.
2. Negative binomial regression demonstrating reduced over-dispersion with AlphaEarth.
3. Sequence model ensemble with controlled evaluation of AlphaEarth embeddings.

2 Data and Covariates

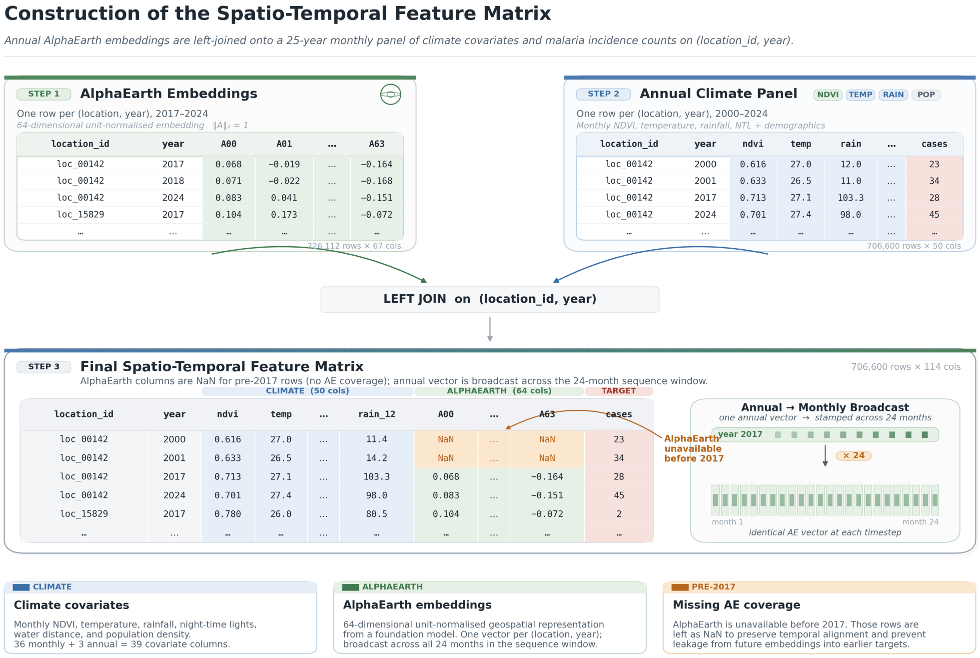
2.1 Malaria incidence counts

Malaria incidence data were obtained from the Malaria Atlas Project (MAP) Malaria Atlas Project (2025), consisting of global estimates of *Plasmodium falciparum* (Pf) incidence counts at an approximate spatial resolution of 5 × 5 km. These estimates are generated using geostatistical models that integrate parasite rate (PR) survey data, routine surveillance records, and environmental covariates.

Table 1: Summary of datasets and covariates used for malaria forecasting. Variables include dynamic spatio-temporal features (malaria incidence, climate, vegetation, population, and nighttime lights) and a static geographic feature (distance to water bodies). All data are aligned to a common spatial grid and monthly or annual temporal resolution as specified.

Category	**Variable**	**Temporal Coverage**	**Spatial Resolution**	**Source**
Dynamic	Malaria Counts	2000–2024	∼5 km	MAP
Dynamic	Night-time Lights	2000–2024	∼1 km	DMSP/VIIRS
Dynamic	NDVI	2000–2024	∼1 km	MODIS
Dynamic	Rainfall	2000–2024	∼5 km	CHIRPS
Dynamic	Temperature	2000–2024	∼11 km	ERA5-Land
Static	Distance to Water Bodies	2023	∼100 m	WorldPop
Static	Population Counts	2023	∼100 m	WorldPop

Raster data were processed and converted into vector representations via polygonization to enable spatial alignment and extraction at the target analysis units.


Construction of the Spatio-Temporal Feature Matrix

Annual AlphaEarth embeddings are left-joined onto a 25-year monthly panel of climate covariates and malaria incidence counts on (location_id, year).

AlphaEarth Embeddings

One row per (location, year), 2017–2024

64-dimensional unit-normalised embedding $\|A\|_2 = 1$

location_id	year	A00	A01	…	A63
loc_00142	2017	0.068	−0.019	…	−0.164
loc_00142	2018	0.071	−0.022	…	−0.168
loc_00142	2024	0.083	0.041	…	−0.151
loc_15829	2017	0.104	0.173	…	−0.072
…	…	…	…	…	…

226,112 rows × 67 cols

Annual Climate Panel

One row per (location, year), 2000–2024

Monthly NDVI, temperature, rainfall, NTL + demographics

location_id	year	ndvi	temp	rain	…	cases
loc_00142	2000	0.616	27.0	12.0	…	23
loc_00142	2001	0.633	26.5	11.0	…	34
loc_00142	2017	0.713	27.1	103.3	…	28
loc_00142	2024	0.701	27.4	98.0	…	45
…	…	…	…	…	…	…

706,600 rows × 50 cols

Final Spatio-Temporal Feature Matrix

706,600 rows × 114 cols

AlphaEarth columns are NaN for pre-2017 rows (no AE coverage); annual vector is broadcast across the 24-month sequence window.

		CLIMATE (50 cols)				ALPHAEARTH (64 cols)			TARGET
location_id	year	ndvi	temp	…	rain_12	A00	…	A63	cases
loc_00142	2000	0.616	27.0	…	11.4	NaN	…	NaN	23
loc_00142	2001	0.633	26.5	…	14.2	NaN	…	NaN	34
loc_00142	2017	0.713	27.1	…	103.3	0.068	…	−0.164	28
loc_00142	2024	0.701	27.4	…	98.0	0.083	…	−0.151	45
loc_15829	2017	0.780	26.0	…	80.5	0.104	…	−0.072	2
…	…	…	…	…	…	…	…	…	…

Climate covariates

Monthly NDVI, temperature, rainfall, night-time lights, water distance, and population density. 36 monthly + 3 annual = 39 covariate columns.

AlphaEarth embeddings

64-dimensional unit-normalised geospatial representation from a foundation model. One vector per (location, year); broadcast across all 24 months in the sequence window.

Missing AE coverage

AlphaEarth is unavailable before 2017. Those rows are left as NaN to preserve temporal alignment and prevent leakage from future embeddings into earlier targets.

Figure 1: Construction of the merged feature matrix via a left join of the annual climate panel onto AlphaEarth (AE) embeddings on (`location_id`, `year`).

2.2 Environmental covariates

Environmental covariates were derived from satellite observations and climate reanalysis datasets spanning the period 2000–2024, using Google Earth Engine (GEE) Gorelick et al. (2017). Monthly rainfall estimates were obtained from CHIRPS Funk et al. (2015) at an approximate spatial resolution of 5 km, with daily observations aggregated to monthly totals. Vegetation dynamics were characterised using the Normalized Difference Vegetation Index (NDVI) from MODIS Didan (2021), rescaled by a factor of 0.0001. Surface temperature was extracted from the ERA5-Land dataset Muñoz-Sabater et al. (2021) at a spatial resolution of approximately 11 km and converted from Kelvin to degrees Celsius.

Using a raster-to-vector extraction pipeline based on polygonization, distance to waterways from WorldPop Woods et al. (2024) was obtained at a spatial resolution of 100,m. Owing to data availability being restricted to 2023, the covariate was treated as temporally static throughout the study period.

2.3 Population Counts

Population estimates for individuals under 18 years of age were obtained from WorldPopBondarenko et al. (2025) for 2023 at a spatial resolution of 3 arc-seconds (100 m), corresponding to the aggregated sum of male and female counts per grid cell. Raster datasets were transformed into tabular format via polygonization, with centroid-based coordinate extraction enabling spatial alignment with point-level data.

In the case of India, due to the large spatial extent and data volume, polygonization was replaced by direct raster-to-point sampling, with spatial downsampling to 500 m to ensure computational tractability.

2.4 Nighttime lights

Nighttime light (NTL) data were extracted from the global harmonized DMSP-VIIRS dataset (2000–2024)Li et al. (2020), which standardizes spatial coordinates and brightness intensity values into a consistent longitudinal format. To perform the extraction at the national level, country-specific boundaries were obtained from the GADM (Database of Global Administrative Areas)GADM (2022) and utilized as spatial masks to clip the NTL data.

2.5 Data summary

Table 1 summarizes the datasets and their characteristics.

3 Methodology

Let i index spatial units (locations) and t index time in months. We model malaria counts $y_{i,t}$ using a combination of climate panel covariates, lagged outcomes, learned location embeddings, and AlphaEarth satellite embeddings $e_{i,y}$ (annual, indexed by year y). We compare a classical Negative Binomial baseline against deep sequence models that consume a 24-month window of features.

3.1 Data Construction

The feature matrix is built by left-joining an annual climate panel (one row per location-year, 2000–2024) with AlphaEarth embeddings (available 2017–2024) on the key (`location_id`, `year`), as illustrated in Figure 1. Climate covariates (NDVI, temperature, rainfall, NTL, water distance, population) vary monthly within each year, while AlphaEarth embeddings are unit-normalized 64-dimensional vectors that are constant within a year and broadcast across the 24 months of each input window. Pre-2017 rows have missing AlphaEarth values, which we handle by zero-imputation with a binary availability indicator.

3.2 Negative Binomial Baseline

To account for overdispersion in count data, we fit a Negative Binomial (NB) regression:

$$y_{i,t} \sim \mathrm{NB}(\mu_{i,t}, \alpha), \tag{1}$$

where $\mu_{i,t}$ is the conditional mean and α the dispersion parameter. The mean is modeled on the log scale as

$$\log \mu_{i,t} = \log(\mathrm{Pop}_{i,t}) + u_i + \gamma_t + \sum_{k=0}^{K} \beta^{(k)\top} x_{i,t-k}, \tag{2}$$

with a population offset, spatial fixed effects u_i, temporal effects γ_t, and distributed lag coefficients on covariates $x_{i,t-k}$. To assess the contribution of satellite features, we fit an extended specification:

$$\log \mu_{i,t} = \cdots + \delta^\top e_{i,y(t)}, \tag{3}$$

where $e_{i,y(t)}$ is the AlphaEarth embedding for location i in the year containing month t.

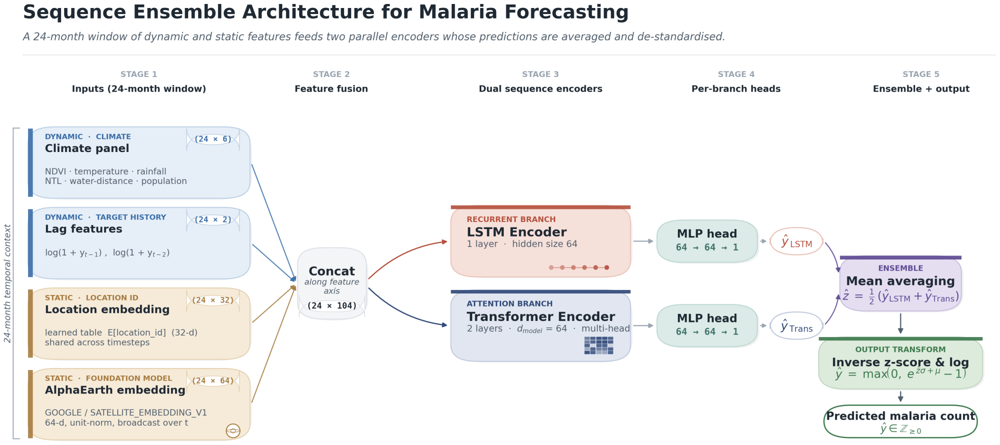


Figure 2: Overview of the sequence-model architecture. For each location, a 24-month window of covariates (NDVI, temperature, rainfall, nighttime lights, distance to water, and population) is combined with lagged malaria features and a learned location embedding. Optional AlphaEarth embeddings provide additional satellite-derived geospatial context. The resulting feature sequence is processed by two parallel encoders: an LSTM and a Transformer. Each encoder produces a prediction via a small MLP head, and the final output is obtained by averaging the two predictions on the count scale. Models are trained on a transformed target and inverted at inference.

3.3 Sequence Models

To capture nonlinear temporal dynamics, we reformulate the task as sequence-to-one regression over a 24-month window. As shown in Figure 2, the input at each timestep concatenates four feature blocks: (i) a climate panel of shape 24×6 (NDVI, T, RF, NTL, water distance, population); (ii) lag features 24×2 corresponding to $\log(1 + y_{t-1})$ and $\log(1 + y_{t-2})$; (iii) a learned location embedding $E[\texttt{location\_id}] \in \mathbb{R}^{32}$ shared across the window; and (iv) the AlphaEarth embedding 24×64 broadcast across timesteps. The target is standardized via a z-score on the log-transformed counts:

$$z_{i,t} = \frac{\log(1 + y_{i,t}) - \mu_z}{\sigma_z}, \tag{4}$$

where μ_z, σ_z are computed on the training set.

We train two encoders on the concatenated input sequence $\mathbf{x}_{i,t} = (x_{i,t-23}, \ldots, x_{i,t})$: 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 $\hat{z}_{i,t}$.

3.4 Ensemble and Inverse Transform

The final standardized prediction averages the two models:

$$\hat{z}_{i,t} = \tfrac{1}{2}\left(\hat{z}_{i,t}^{\text{LSTM}} + \hat{z}_{i,t}^{\text{Trf}}\right). \tag{5}$$

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

$$\hat{y}_{i,t} = \max\left(0,\ e^{\hat{z}_{i,t}\sigma_z + \mu_z} - 1\right) \in \mathbb{Z}_{\geq 0}. \tag{6}$$

3.5 Training

The sequence models are trained end-to-end by minimizing the MSE between predicted and observed standardized log-counts $z_{i,t}$, optimized with Adam. We use a temporal hold-out split for validation and early stopping. The NB baseline is fit by maximum likelihood on the same training period.

4 Results and Evaluation

4.1 Experimental Setup

All experiments were conducted on our institution's Advanced Research Computing (ARC) cluster (anonymised for review), a high-performance computing environment supporting large-scale parallel workloads, GPU acceleration, and high-memory nodes.

ARC provides distributed and shared memory systems with SLURM-based scheduling, enabling efficient training of both statistical and deep learning models.

We used GPU-enabled nodes for sequence model training and CPU nodes for regression experiments.

4.2 Evaluation Metrics

Performance is evaluated on the held-out 2024 test set using:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{7}$$

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\bar{y}}, \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{8}$$

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}. \tag{9}$$

We evaluate four model families—Negative Binomial, LSTM, Transformer, and their ensemble—each in baseline and +AlphaEarth variants. All models are trained and evaluated on identical (i, t) pairs to isolate the effect of AlphaEarth embeddings.

4.3 Cross-Country Evaluation

We begin by analysing the best-performing configuration to understand cross-country behaviour, before presenting a full comparison across all model families.

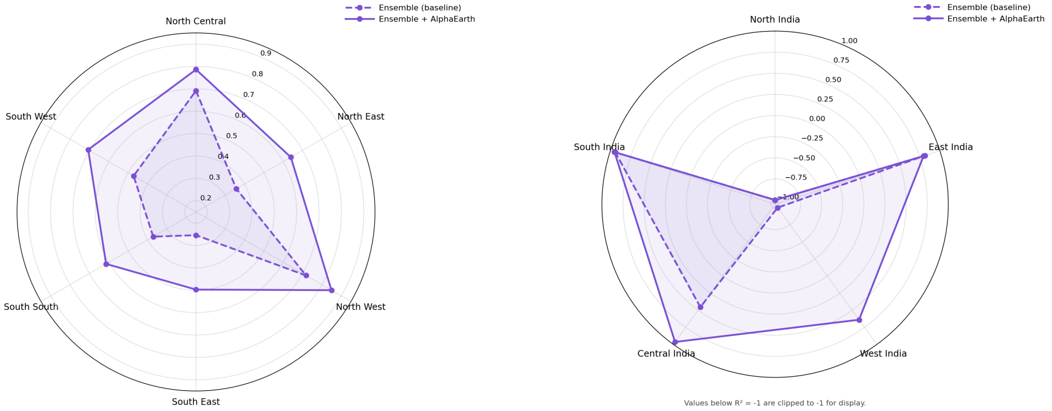
For Nigeria and India, we evaluated a multi-seed LSTM+Transformer ensemble with per-location embeddings and lagged malaria features. Models are trained on a z-score-normalised target, and results are reported for baseline and +AlphaEarth variants, averaged over five seeds. Table 2 reports the overall held-out 2024 metrics. On Nigeria, the AlphaEarth-augmented ensemble lifts R^2 from 0.623 to 0.777 (a $+0.154$ improvement, well beyond the seed std of ≈ 0.06). On India, the architecture alone already brings the ensemble to $R^2 = 0.867$, and AlphaEarth nudges it further to $R^2 = 0.881$ ($+0.014$, within the range of seed variability). The same direction of effect holds for NRMSE on both countries, though the magnitude of AlphaEarth's contribution is substantially larger on Nigeria than on India.

Per-zone view. The per-zone radar (R^2, Fig. 3) and radial-bar (NRMSE, Fig. 4) charts visualise where AlphaEarth helps most in each country. On Nigeria, AlphaEarth expands the R^2 polygon outward in all six DHS zones, with the largest gains in *North East* ($0.36 \rightarrow 0.64$), *South South* ($0.37 \rightarrow 0.61$), and *South West* ($0.47 \rightarrow 0.71$). On India, the per-zone story is dominated by two zones whose baseline R^2 is severely negative because of extreme-count clusters – *West India* ($-4.79 \rightarrow +0.64$) and *North India* ($-12.36 \rightarrow -1.29$); AlphaEarth roughly halves NRMSE in both zones, turning *West India* from the worst-fit zone into a usable one (Fig. 4). *Central India* improves from $R^2 = 0.46$ to 0.96, indicating near-saturation of predictive performance with the addition of AlphaEarth.

Discussion Three observations emerge. First, the AlphaEarth gain is robust to multi-seed averaging: the seed standard deviation is well below the observed ΔR^2 on Nigeria, and the random-embedding

Table 2: Best-of-paper ensemble results on the 2024 held-out test year, mean ± standard deviation over five seeds. NRMSE $= \mathrm{RMSE}/\bar{y}_{\text{test}}$. *Nigeria* uses the multi-seed ensemble (z-score target, 25 epochs); *India* uses the architecture with the sweep-selected configuration (per-location embeddings + per-location target normalisation + lag-1/lag-2 annual malaria features; hidden 128, layers 2, embed_dim 32, dropout 0.2). Best variant per country is in bold and in particular, red (1st) and violet (2nd).

Country	**Variant**	R^2 (mean ± std)	NRMSE (mean ± std)
Nigeria	Ensemble (baseline)	0.623 ± 0.061	1.134 ± 0.091
Nigeria	Ensemble + AlphaEarth	$\mathbf{0.777 \pm 0.021}$	$\mathbf{0.874 \pm 0.040}$
India	Ensemble (baseline)	0.867 ± 0.025	3.867 ± 0.388
India	Ensemble + AlphaEarth	$\mathbf{0.881 \pm 0.086}$	$\mathbf{3.439 \pm 1.441}$



(a) Nigeria. Six DHS geopolitical zones.

(b) India. Five lat/lon-derived zones.

Figure 3: Per-zone R^2 on the 2024 held-out test set, mean across five seeds. Dashed: ensemble baseline; solid: ensemble + AlphaEarth. AlphaEarth pushes the polygon outward in every Nigeria zone and in every extreme-count India zone.

ablation rules out a capacity-driven explanation. Second, India's performance ceiling is primarily representational rather than capacity-limited: incorporating per-location embeddings and lagged malaria features raises the baseline ensemble to $R^2 = 0.867$, comparable to Nigeria. Third, AlphaEarth provides a consistent additional improvement, particularly in zones with rare extreme-count clusters where existing features leave systematic variance unexplained.

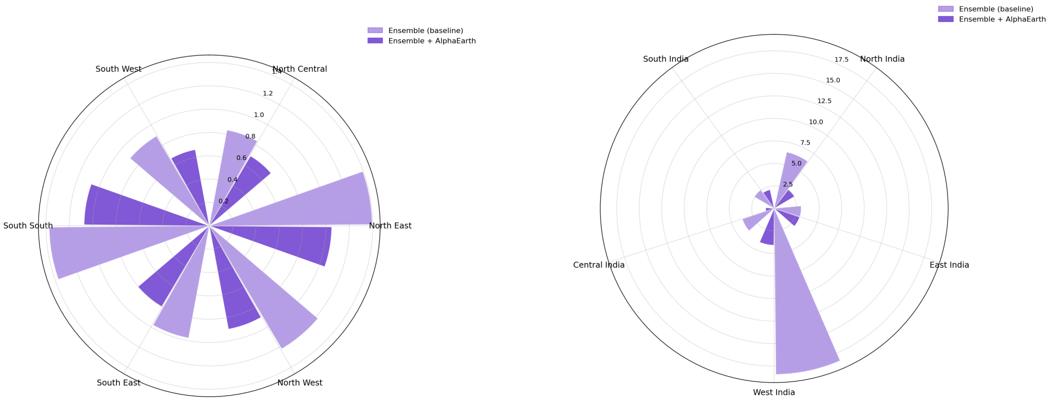
4.4 Results Across Models

We compare four model families on the held-out 2024 test year: Negative Binomial regression, LSTM, Transformer, and an LSTM+Transformer ensemble, each evaluated with and without AlphaEarth embeddings.

For each country, we report the strongest configuration of each model family. Sequence-model results are mean ± standard deviation over five seeds, while the negative binomial regression is deterministic. The sequence models incorporate per-location embeddings, lagged malaria features, and a z-score-normalised target. The NB regression uses per-cluster training-mean malaria and lagged counts without a population offset. NRMSE is defined as $\mathrm{RMSE}/\bar{y}_{\text{test}}$, with $\bar{y}_{\text{test}} = 9{,}821.4$ (Nigeria) and 21.4 (India).

Discussion. Five observations emerge from Tab. 3.

1. **AlphaEarth improves performance in most settings.** It outperforms the baseline in 6 of 8 (model, country) cells. The two exceptions (India NB and Transformer) occur when models already encode strong cluster-level signals (via per-cluster means or learned embeddings),



(a) Nigeria.

(b) India. The dominant bar is *West India* under the baseline (NRMSE ≈ 18); AlphaEarth reduces it to ≈ 4.

Figure 4: Per-zone NRMSE on the 2024 held-out test set as a radial grouped-bar chart, mean across five seeds. Lighter bars are the ensemble baseline; darker bars are AlphaEarth-augmented. Lower is better. AlphaEarth shrinks the bars in every zone of both countries, with the largest absolute reductions on India's extreme-count zones.

Table 3: Headline results across all four model families on the 2024 held-out test year. Sequence-model rows are mean ± std over five seeds; the negative binomial regression is single-seed. Best variant per (model, country) cell is in bold and in particular, red (1st) and violet (2nd).

		Nigeria		India	
Model	**Variant**	R^2	NRMSE	R^2	NRMSE
Negative binomial	baseline	**0.986**	**0.217**	**0.810**	**4.642**
Negative binomial	+ AlphaEarth	**0.986**	**0.219**	0.796	4.810
LSTM	baseline	0.620 ± 0.032	1.140 ± 0.049	0.737 ± 0.076	5.418 ± 0.784
LSTM	+ AlphaEarth	$\mathbf{0.756 \pm 0.027}$	$\mathbf{0.913 \pm 0.049}$	$\mathbf{0.844 \pm 0.102}$	$\mathbf{3.961 \pm 1.561}$
Transformer	baseline	0.490 ± 0.213	1.303 ± 0.253	$\mathbf{0.875 \pm 0.040}$	$\mathbf{3.723 \pm 0.627}$
Transformer	+ AlphaEarth	$\mathbf{0.698 \pm 0.033}$	$\mathbf{1.016 \pm 0.055}$	0.569 ± 0.468	6.082 ± 3.859
Ensemble	baseline	0.623 ± 0.061	1.134 ± 0.091	0.867 ± 0.025	3.867 ± 0.388
Ensemble	+ AlphaEarth	$\mathbf{0.777 \pm 0.021}$	$\mathbf{0.874 \pm 0.040}$	$\mathbf{0.881 \pm 0.086}$	$\mathbf{3.439 \pm 1.441}$

making AlphaEarth redundant or prone to overfitting. In contrast, LSTM and ensemble models still gain $+0.014$ to $+0.107$ in R^2.

2. **NB is the strongest single-seed predictor on Nigeria.** With per-cluster means and lagged malaria features, NB achieves $R^2 = 0.986$, substantially above the best sequence model ($R^2 = 0.777$). This reflects a feature advantage: NB directly leverages strong temporal autocorrelation, whereas the sequence models here do not include lag features.

3. **The gap narrows on India.** NB reaches $R^2 = 0.810$, while the ensemble + AlphaEarth attains $R^2 = 0.881$. The relative gain of deep models reflects India's more challenging distribution, with more clusters, heavier tails, and rare extreme values.

4. **Transformer + AlphaEarth is unstable on India.** The high variance ($\sigma = 0.468$ vs 0.040 baseline) indicates overfitting when combining deep architectures with high-dimensional embeddings. The ensemble mitigates this by averaging across models.

5. **NRMSE trends mirror R^2.** Improvements in R^2 consistently correspond to reductions in NRMSE. The larger NRMSE values on India reflect data distribution differences rather than model performance.

Limitations. The proposed framework remains sensitive to heavy-tailed count distributions and rare extreme outbreaks, particularly in high-variance regions. Additionally, while AlphaEarth embeddings improve predictive performance, their environmental representations remain only partially interpretable.

5 Conclusion

We presented a framework for malaria forecasting that combines count-aware statistical modelling, sequence-based deep learning, and geospatial foundation model embeddings. Across India and Nigeria, AlphaEarth embeddings provide consistent and multi-seed-stable improvements for models that lack explicit mechanisms for encoding location-specific context. When such information is already captured through per-cluster statistics or learned embeddings, the marginal benefit of AlphaEarth diminishes, indicating that its primary contribution lies in encoding latent environmental structure not present in standard covariates.

These results highlight the potential of geospatial foundation models as transferable representations for epidemiological forecasting, particularly in data-sparse or weakly structured settings. Future work will explore integrating such embeddings more tightly into model architectures and extending this approach to other climate-sensitive diseases.

References

Agyekum, T. P., Botwe, P. K., Arko-Mensah, J., Issah, I., Acquah, A. A., Hogarh, J. N., Dwomoh, D., Robins, T. G., and Fobil, J. N. (2021). 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*, 18:7255.

Baeza, A., Santos-Vega, M., Dobson, A. P., and Pascual, M. (2017). The rise and fall of malaria under land-use change in frontier regions. *Nature Ecology & Evolution*, 1:0108.

Batheja, D., Samria, D., Wimberly, M. C., Pascual, M., Baharia, R. K., Mohanty, A. K., Desai, V., Vaishnav, K., Sharma, R., Kohli, V., Sharma, S., Anvikar, A., Murdock, C., and Nandi, A. (2026). Socioeconomic and household water management determinants of malaria and other vector-borne disease prevention in urban gujarat, india. *Malaria Journal*, 25:141.

Beloconi, A., Nyawanda, B. O., Bigogo, G., Khagayi, S., Obor, D., Danquah, I., Kariuki, S., Munga, S., and Vounatsou, P. (2023). Malaria, climate variability, and interventions: modelling transmission dynamics. *Scientific Reports*, 13:7367.

Bondarenko, M., Priyatikanto, R., Tejedor-Garavito, N., Zhang, W., McKeen, T., Cunningham, A., Woods, T., Hilton, J., Cihan, D., Nosatiuk, B., Brinkhoff, T., Tatem, A., and Sorichetta, A. (2025). Estimates of total number of people under the age of 18 years old and broken down by male and female for each year 2015-2030 at a resolution of 3 arc (approximately 100m at the equator) r2025a version v1. Global Demographic Data Project - Funded by The Bill and Melinda Gates Foundation (INV-045237).

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., et al. (2025). Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*.

Cowman, A. F., Healer, J., Marapana, D., and Marsh, K. (2016). Malaria: biology and disease. *Cell*, 167(3):610–624.

Craig, M., Snow, R., and le Sueur, D. (1999). A climate-based distribution model of malaria transmission in sub-saharan africa. *Parasitology Today*, 15:105–111.

Didan, K. (2021). Modis/terra vegetation indices monthly l3 global 1km sin grid v061.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., et al. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2(1):1–21.

GADM (2022). Global administrative areas, version 4.1. University of California, Berkeley.

Gelband, H. et al. (2020). Is malaria an important cause of death among adults? *The American Journal of Tropical Medicine and Hygiene*, 103(1):41.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

Guerra, C. A., Reiner, R. C., Perkins, T. A., Lindsay, S. W., Midega, J. T., Brady, O. J., Barker, C. M., Reisen, W. K., Harrington, L. C., Takken, W., Kitron, U., Lloyd, A. L., Hay, S. I., Scott, T. W., and Smith, D. L. (2014). A global assembly of adult female mosquito mark-release-recapture data to inform the control of mosquito-borne pathogens. *Parasites & Vectors*, 7:276.

Hoshen, M. B. and Morse, A. P. (2004). A weather-driven model of malaria transmission. *Malaria Journal*, 3:32.

Kalipe, G., Gautham, V., and Behera, R. K. (2018). Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis. In *2018 International Conference on Information Technology (ICIT)*, pages 33–38. IEEE.

Kibret, S., Wilson, G. G., Ryder, D., Tekie, H., and Petros, B. (2017). The influence of dams on malaria transmission in sub-saharan africa. *EcoHealth*, 14:408–419.

Li, X., Zhou, Y., Zhao, M., and Zhao, X. (2020). Harmonization of dmsp and viirs nighttime light data from 1992-2018 at the global scale. *Scientific Data*, 7(1):1–15. Dataset available at https://doi.org/10.6084/m9.figshare.9828827.

Malaria Atlas Project (2025). Number of newly diagnosed plasmodium falciparum cases, 2000–2024. Malaria Atlas Project (MAP). Version 202508.

Martens, W. J., Niessen, L. W., Rotmans, J., Jetten, T. H., and McMichael, A. J. (1995). Potential impact of global climate change on malaria risk. *Environmental Health Perspectives*, 103:458–464.

Martineau, P., Behera, S. K., Nonaka, M., Jayanthi, R., Ikeda, T., Minakawa, N., Kruger, P., and Mabunda, Q. E. (2022). Predicting malaria outbreaks from sea surface temperature variability up to 9 months ahead in limpopo, south africa, using machine learning. *Frontiers in Public Health*, 10:962377.

Mordecai, E. A., Caldwell, J. M., Grossman, M. K., Lippi, C. A., Johnson, L. R., Neira, M., Rohr, J. R., Ryan, S. J., Savage, V., Shocket, M. S., Sippy, R., Ibarra, A. M. S., Thomas, M. B., and Villena, O. (2019). Thermal biology of mosquito-borne disease. *Ecology Letters*, 22:1690–1708.

Mordecai, E. A., Ryan, S. J., Caldwell, J. M., Shah, M. M., and LaBeaud, A. D. (2020). Climate change could shift disease burden from malaria to arboviruses in africa. *The Lancet Planetary Health*, 4:e416–e423.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., et al. (2021). Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383.

Nazir, U., Ejaz, A., Quddoos, M. T., Uppal, M., and Khalid, S. (2023). Towards a spatio-temporal deep learning approach to predict malaria outbreaks using earth observation measurements in south asia. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*.

Nazir, U., Quddoos, M. T., Uppal, M., and Khalid, S. (2024). 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*, 8:S17.

Pillay, M. T., Minakawa, N., Kim, Y., Kgalane, N., Ratnam, J. V., Behera, S. K., Hashizume, M., and Sweijd, N. (2023). Utilizing a novel high-resolution malaria dataset for climate-informed predictions with a deep learning transformer model. *Scientific Reports*, 13:23091.

Rogers, D. J. and Randolph, S. E. (2000). The global spread of malaria in a future, warmer world. *Science*, 289:1763–1766.

Rosenthal, P. J., John, C. C., and Rabinovich, N. R. (2019). Malaria: how are we doing and how can we do better? *The American Journal of Tropical Medicine and Hygiene*, 100(2):239.

Ryan, S. J., Lippi, C. A., and Zermoglio, F. (2020). Shifting transmission risk for malaria in africa with climate change: a framework for planning and intervention. *Malaria Journal*, 19:170.

Sebastianelli, A. et al. (2024). A reproducible ensemble machine learning approach to forecast dengue outbreaks. *Scientific Reports*, 14:3807.

Shah, H. A., Carrasco, L. R., Hamlet, A., and Murray, K. A. (2022). Exploring agricultural land-use and childhood malaria associations in sub-saharan africa. *Scientific Reports*, 12:4124.

Shi, X. et al. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, volume 28.

Stopard, I. J., Churcher, T. S., and Lambert, B. (2021). Estimating the extrinsic incubation period of malaria using a mechanistic model of sporogony. *PLOS Computational Biology*, 17:e1008658.

Suh, E., Grossman, M. K., Waite, J. L., Dennington, N. L., Sherrard-Smith, E., Churcher, T. S., and Thomas, M. B. (2020). The influence of feeding behaviour and temperature on the capacity of mosquitoes to transmit malaria. *Nature Ecology & Evolution*, 4:940–951.

Symons, T. L., Moran, A., Balzarolo, A., Vargas, C., Robertson, M., Lubinda, J., Saddler, A., McPhail, M., Harris, J., Rozier, J., Browne, A., Amratia, P., Bertozzi-Villa, A., Bhatt, S., Cameron, E., Golding, N., Smith, D. L., Noor, A. M., Rumisha, S. F., Palmer, M. D., Weiss, D. J., Desai, N., Potere, D., Sukitsch, N., Woods, W., and Gething, P. W. (2026). Projected impacts of climate change on malaria in africa. *Nature*, 651:390–396.

Tjaden, N. B., Caminade, C., Beierkuhnlein, C., and Thomas, S. M. (2018). Mosquito-borne diseases: Advances in modelling climate-change impacts. *Trends in Parasitology*, 34:227–245.

WHO (2025). World malaria report 2025: Addressing the threat of antimalarial drug resistance. Technical report, World Health Organisation.

Woods, D., McKeen, T., Cunningham, A., Priyatikanto, R., Sorichetta, A., Tatem, A. J., and Bondarenko, M. (2024). Worldpop high resolution, harmonised annual global geospatial covariates. version 1.0. Global Demographic Data Project - Funded by The Bill and Melinda Gates Foundation (INV-045237).

A Why two India zones look catastrophic on the headline metrics

The per-zone radar / radial chart for India (Fig. 3b and Fig. 4b) reports two cells that look much worse than the country-level numbers would suggest: the baseline $R^2 = -12.36 \pm 13.39$ on **North India**, and the baseline $\mathrm{NRMSE} = 18.41 \pm 5.31$ on **West India**. Both numbers are five-seed means of the V2 ensemble (per-location embedding, lag-1/lag-2, per-location target normalisation) on the held-out 2024 test year. This appendix dissects them at the per-seed, per-cluster level. Both are the same structural story – a small per-zone denominator multiplied by occasional single-cluster prediction overshoots produced when the model's z-scale prediction is inverted through $\exp(\cdot) - 1$ on the count scale – but they manifest in different metrics, so we explain each in turn.

A.1 West India: tiny $\bar{y}$ + heavy-tailed truth ⇒ inflated NRMSE

Recall $\mathrm{NRMSE} = \mathrm{RMSE}/\bar{y}_{\mathrm{test}}$, the coefficient-of-variation of the RMSE. West India is the lowest-mean zone in the country panel: $\bar{y}_{\mathrm{test}} = 4.26$ malaria counts per cluster-month-year, against 40.5 for East India and 30.1 for South India. Any RMSE is therefore divided by a small denominator. Compounding this, the truth distribution in the zone is extremely heavy-tailed – 95th percentile $= 8.95$, 99th percentile $= 52.8$, but maximum $= 1{,}593$ – so a single freak cluster in Rajasthan or Gujarat sits more than 30× above the 99th percentile of its zone. When the V2 ensemble inverts its z-scale prediction back to the count scale via $\exp(\cdot) - 1$, a modest over-prediction in z becomes a massive over-prediction in counts on exactly that cluster.

Table 4: V2 ensemble on West India, per seed ($n = 3{,}142$, $\bar{y}_{\mathrm{test}} = 4.26$, true max $= 1{,}593$). *pred max* is the largest ensemble prediction on the zone; *resid max* is the largest absolute residual. The inflation of NRMSE is driven by one to three single-cluster blow-ups per seed in the baseline column: a residual of 2,500–5,500 on a cluster whose true value is $\leq 1{,}593$ produces a per-row squared error of 10^7 and dominates the average. AlphaEarth shrinks every pred-max and resid-max column by 3–10× and brings NRMSE down by roughly the same factor.

	Baseline				**+ AlphaEarth**			
Seed	pred max	resid max	RMSE	NRMSE	pred max	resid max	RMSE	NRMSE
0	3086	2540	52.7	12.36	1777	539	12.9	3.03
1	3330	2784	59.7	14.01	1545	437	9.1	2.13
2	6116	5570	108.9	25.53	2330	1784	37.5	8.79
3	5032	4486	83.2	19.51	1583	519	16.3	3.83
4	4984	4438	88.0	20.63	1513	479	11.0	2.59
mean	4510	3964	78.5	18.41	1750	751	17.4	4.07

Three things therefore stack to inflate the West India NRMSE in Fig. 4b: a tiny denominator ($\bar{y} = 4.26$, the smallest of the five zones), a long-tailed truth (one cluster at 30× the 99th percentile), and exponential blow-ups when the z-scale prediction is exponentiated on those tail clusters. The seed-to-seed standard deviation of ± 5.31 in baseline NRMSE confirms the metric is dominated by which clusters happen to blow up under each random initialisation. AlphaEarth absorbs enough landscape-context information to clip those overshoots – pred-max drops from a mean of 4,510 to 1,750, resid-max from 3,964 to 751, and NRMSE from 18.4 to 4.07, a 4.5× improvement.

A.2 North India: tiny σ_y ⇒ a single cluster's residual exceeds the entire R^2 baseline

The same diagnosis with the metric swapped: $R^2 = 1 - \mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$, where $\mathrm{SS}_{\mathrm{tot}} = \sum_i (y_i - \bar{y})^2 \approx n \cdot \mathrm{Var}(y)$. North India has the lowest standard deviation in the country: $\sigma_y = 6.22$ counts. The baseline $\mathrm{SS}_{\mathrm{tot}} = 5{,}249 \times 6.22^2 \approx 0.20\,\mathrm{M}$. Any blow-up that contributes more than $0.20\,\mathrm{M}$ on its own already drives R^2 below zero before the other 5,248 rows contribute anything.

Seed 4 is a single-seed worst case: a prediction of 1,899 on a cluster whose truth is at most 204 produces a residual of 1,797, contributing $1{,}797^2/\mathrm{SS}_{\mathrm{tot}} \approx 16$ to $\mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$ on its own – i.e. that one cluster alone forces $R^2 \leq -15$ before the other 5,248 rows are summed. The remaining squared errors take the seed to $R^2 = -35.81$, and that single seed drags the five-seed mean from ~ -6.5 to -12.37. The seed standard deviation of ± 13.39 in the radar plot is a visual diagnostic of exactly this instability: in the regime where one cluster's overshoot can swamp the entire $\mathrm{SS}_{\mathrm{tot}}$, R^2 is dominated by which clusters happen to blow up. AlphaEarth removes the overshoot in three of the five seeds (pred-max under 470, $R^2 > 0$); on the other two it does not, but its mean of -1.29 is an order-of-magnitude improvement over -12.37.

A.3 Common cause and metric sensitivity

Both pathologies are the same phenomenon – a few single-cluster overshoots when the z-scale prediction is exponentiated – viewed through two metrics with small per-zone denominators ($\bar{y}_{\mathrm{West}} = 4.26$ and $\sigma^2_{y,\mathrm{North}} \cdot n \approx 0.20\,\mathrm{M}$). Two methodological notes follow.

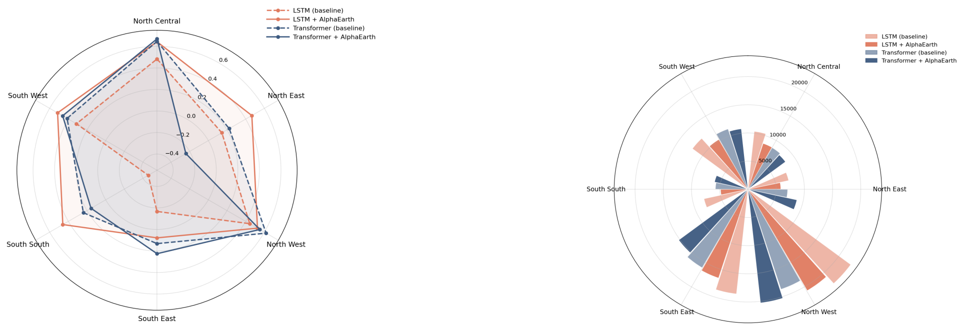
Table 5: V2 ensemble on North India, per seed (n = 5,249, $\bar{y}$ = 3.27, σ_y = 6.22, true max = 204, $\mathrm{SS}_{\mathrm{tot}} \approx 0.20\,\mathrm{M}$). *pred max* is the largest ensemble prediction; *resid max* the largest absolute residual; ΔR^2 *from worst row* is the contribution of that single residual, $(\mathrm{resid\,max})^2/\mathrm{SS}_{\mathrm{tot}}$ – it is the *minimum* amount by which R^2 is already pushed below zero before any other cluster is counted. Several baseline seeds overshoot to predictions in the 500–1,900 range when truth caps at 204, and a single such residual already exceeds the entire $\mathrm{SS}_{\mathrm{tot}}$. AlphaEarth fixes three of five seeds outright (pred-max under 470, R^2 positive); the remaining two seeds still overshoot and pull the mean back down.

	Baseline					**+ AlphaEarth**				
Seed	pred max	resid max	RMSE	$\Delta R^2_{\min}$	R^2	pred max	resid max	RMSE	$\Delta R^2_{\min}$	R^2
0	847	745	21.2	2.7	−10.60	238	138	4.6	0.1	0.44
1	724	719	13.6	2.5	−3.80	826	724	13.3	2.6	−3.55
2	824	819	18.1	3.3	−7.44	384	228	5.6	0.3	0.19
3	525	520	14.2	1.3	−4.18	1017	915	13.4	4.1	−3.65
4	1899	1797	37.7	15.9	−35.81	462	258	5.9	0.3	0.11
mean	964	920	21.0	5.1	−12.37	585	453	8.6	1.5	−1.29

First, both metrics are doing what they should: NRMSE *should* penalise large RMSE more on a zone with small $\bar{y}$, and R^2 *should* go negative when the model is worse than predicting the constant mean. We retain both as headline metrics rather than swap to a more outlier-robust alternative (e.g. medianRMSE / Spearman R), because aggregating across the country is dominated by zones with larger $\bar{y}$ and σ_y (East and South India, $\bar{y} = 40.5$ and 30.1), where the metrics behave benignly and the AlphaEarth ordering is preserved. The country-level numbers in Tab. **??** are not affected by these zone-level pathologies in a meaningful way.

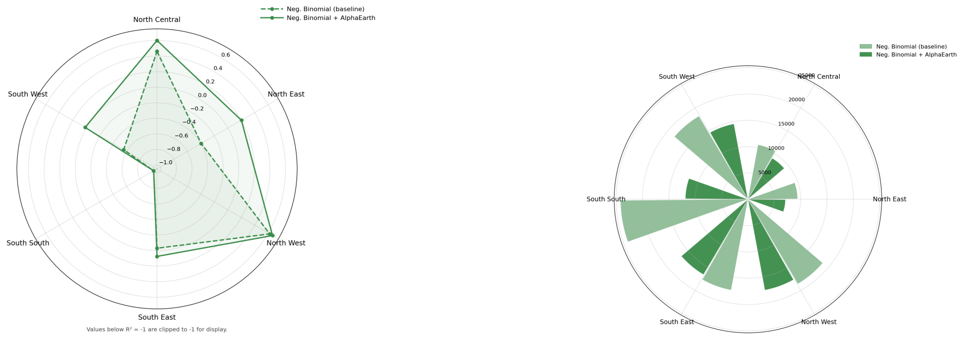
Second, what the appendix is genuinely flagging is a model-side issue: the per-location-embedding V2 still occasionally latches onto a high-magnitude lag value and overshoots on exponentiation. The two straightforward fixes – already on the open follow-up list – are (i) tighter dropout (e.g. 0.4) on the embedding projection, and (ii) early stopping on validation MSE rather than fixed 25 epochs. Either should compress the right tail of single-cluster predictions without disturbing the headline numbers, which would in turn collapse the seed standard deviations on these two cells specifically.

B Additional Per-Zone Results for Nigeria



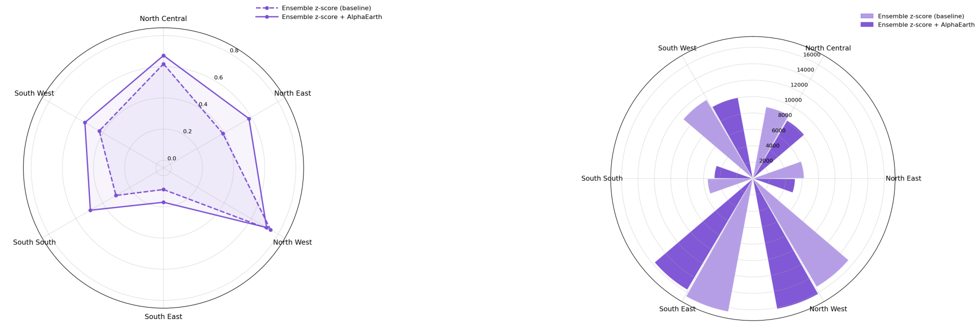
(a) Per-zone R^2 – LSTM and Transformer (baseline vs. +AlphaEarth). Higher is better; outer ring marks $R^2 = 1$.

(b) Per-zone RMSE – same models. Within each zone the four bars are LSTM, LSTM + AE, Transformer, Transformer + AE; lower is better.



(c) Per-zone R^2 – negative binomial panel regression. Values below $R^2 = -1$ are clipped to -1 in the plot for readability; unclipped values are in Tab. **??** (notably *South South*: $R^2 = -12.4 \rightarrow -2.2$).

(d) Per-zone RMSE – negative binomial panel regression. Lighter bar: baseline. Darker bar: +AlphaEarth.



(e) Per-zone R^2 – LSTM + Transformer ensemble trained with a z-score-normalised target. Dashed: baseline ensemble; solid: +AlphaEarth ensemble.

(f) Per-zone RMSE – the same z-score ensemble. Lighter bar: baseline ensemble. Darker bar: +AlphaEarth ensemble.

Figure 5: All per-zone Nigeria 2024 test-set results in a single grid. Rows correspond to model families (top: LSTM and Transformer sequence models; middle: negative binomial panel regression; bottom: LSTM + Transformer ensemble trained with a z-score-normalised target). Columns correspond to the metric (R^2 on the left, RMSE on the right) shown across Nigeria's six DHS geopolitical zones (*North Central*, *North East*, *North West*, *South East*, *South South*, *South West*). AlphaEarth expands the R^2 polygon outward and shrinks the RMSE bars in essentially every zone-by-model cell, with the largest gains on the LSTM and the negative binomial regression. The overall best configuration – the AlphaEarth-augmented z-score ensemble in the bottom row – attains $R^2 = 0.625$ and RMSE $= 11{,}144$ on the held-out test set.